Primordial Cosmology

Study Notes based on Patrick Peter & Jean-Philippe Uzan

Department of Physics / Cosmology Notes

Personal Study Notes
chandra.pp@alumni.iitg.ac.in

November 25, 2026

Preface

These are my personal study notes on general relativistic cosmology, early-universe thermodynamics, cosmic nucleosynthesis, cosmological perturbation theory, and CMB anisotropies. The core of the book follows Patrick Peter and Jean-Philippe Uzan's *Primordial Cosmology* (Oxford University Press) as its roadmap, though it draws on many other references along the way.

Throughout, the goal has been to **derive every calculation explicitly**: filling in the intermediate steps, gauge-invariant formulations, and technical nuances that standard treatments often leave to the reader, so that the underlying mathematical and physical narrative stays fully transparent. This is as much for my own mastery of the material as for any other student navigating this demanding field.

The later chapters on primordial non-Gaussianity and cosmological collider physics carry the same philosophy into current research territory, well beyond Peter and Uzan's own scope. There, the primary reference is Arkani-Hamed and Maldacena's *Cosmological Collider Physics*, a paper written for an audience that already has a working knowledge of conformal field theory; rather than simply summarizing its results, I have tried to rebuild the calculation from first principles, assuming little more than what the earlier chapters develop, and to work out in full the algebraic steps that the original paper leaves to an appendix or a citation.

Figures and diagrams are, where possible, redrawn or extracted directly from the original source material to stay faithful to the presentation being followed; a number of hand-drawn additions of my own appear alongside them in the later, more research-adjacent chapters.

I welcome feedback, corrections, or discussions regarding the content of these notes – please feel free to reach out by email. I would like to thank Prathamesh Changde for his help with several TikZ diagrams and with LaTeX in a few chapters, without which this project would have taken far longer.

Contact:
chandra.pp@alumni.iitg.ac.in

Disclaimer: These are pedagogical study notes. While I have strived for accuracy, any errors in the calculations, typos, or physical interpretations remain entirely my own. I am actively revising and correcting the notes; the latest version can always be found on the project page at lunar-photon.github.io.

Contents

Preface i

I Overview of the Theoretical Basis 1

1 General relativity 2
1.1 Space-time and gravity . . . 2
1.1.1 Absolute space and time of Newtonian physics . . . 2
1.1.2 Space-time of special relativity . . . 6
1.1.3 General relativity and curved space-time . . . 10
1.2 Elements of differential geometry . . . 12
1.2.1 Manifolds and tensors . . . 12
1.2.2 Geodesic equations and Christoffel symbols . . . 18
1.2.3 Covariant derivative, parallel transport, and Lie derivative . . . 22
1.2.4 Curvature . . . 29
1.2.5 Covariant approach (1+3 / 3+1 splitting) . . . 33
1.3 Equations of motion . . . 38
1.3.1 Einstein's equations . . . 38
1.3.2 Conservation equations . . . 39
1.3.3 ADM Hamiltonian formulation of general relativity . . . 44
1.4 Geodesics in curved space-time . . . 45
1.4.1 Conserved quantities along geodesics . . . 45
1.4.2 Geodesic deviation and tidal forces . . . 45
1.5 The weak-field regime . . . 47
1.5.1 The Newtonian limit . . . 47
1.5.2 Gravitational waves in vacuum . . . 50
1.6 Experimental tests of general relativity . . . 52
1.6.1 Tests of the Equivalence Principle . . . 52
1.6.2 Solar System tests and the Parameterized Post-Newtonian (PPN) framework 55
1.6.3 Gravitational radiation and binary pulsars . . . 58
1.6.4 Need for tests at astrophysical and cosmological scales . . . 58

II The Modern Standard Cosmological Model 60

2 The cosmological solution of Friedmann–Lemaître 61
2.1 The Friedmann–Lemaître spacetime . . . 61
2.1.1 Constructing a cosmological model . . . 61
2.1.2 The Cosmological Principle and the Copernican Principle . . . 62
2.1.3 Spacetime symmetries and the Robertson–Walker metric . . . 62
2.2 Kinematics of the expanding Universe . . . 65

2.2.1 Fundamental observers and the Hubble expansion law . . . 65
2.2.2 Propagation of light and cosmological redshift . . . 65
2.2.3 Dynamics of massive test particles . . . 66
2.3 Cosmological dynamics of the FLRW spacetime . . . 66
2.3.1 Energy-momentum tensor of the cosmological fluid . . . 66
2.3.2 The Friedmann equations and conservation laws . . . 66
2.3.3 Cosmological equations in conformal time . . . 67
2.3.4 Equations of state and cosmic components . . . 67
2.3.5 Dimensionless density parameters and the cosmic sum rule . . . 68
2.3.6 Exact analytical solutions of the Friedmann equations . . . 69
2.3.7 Dynamical system analysis in phase space . . . 70
2.4 Cosmological timescales and distance measures . . . 72
2.4.1 Age of the Universe and look-back time . . . 73
2.4.2 Comoving radial distance . . . 73
2.4.3 Angular diameter distance and reciprocity theorem . . . 74
2.4.4 Luminosity distance and the distance modulus . . . 75
2.4.5 Cosmic volume elements and galaxy number counts . . . 78
2.5 Kinematic expansions at small redshifts ($z \ll 1$) . . . 78
2.5.1 Taylor expansion of the scale factor and deceleration parameter . . . 79
2.5.2 Small-z expansions of cosmological observables . . . 79
2.5.3 The cosmic redshift drift (Sandage–Loeb effect) . . . 79
2.6 Cosmological horizons and global conformal structure . . . 80
2.6.1 The cosmological event horizon . . . 80
2.6.2 The cosmological particle horizon . . . 82
2.6.3 Global conformal structure and Penrose–Carter diagrams . . . 82
2.7 Beyond the cosmological principle: anisotropic and inhomogeneous spacetimes . . . 83
2.7.1 Classification of spacetime isometries and Lie algebras . . . 84
2.7.2 Classification of cosmological models . . . 84
2.7.3 Spatially inhomogeneous spacetimes . . . 85
2.7.4 Spatially homogeneous anisotropic spacetimes (Bianchi models) . . . 86

3 The Standard Cosmological Model: Observational Tests and Paradigm 88
3.1 The Hubble diagram and cosmic age constraints . . . 88
3.1.1 The Hubble expansion law . . . 88
3.1.2 The extragalactic distance ladder . . . 90
3.1.3 Physical, non-ladder methods for measuring H_0 . . . 91
3.2 Type Ia supernovae and the discovery of cosmic acceleration . . . 92
3.2.1 Standardizing Type Ia supernovae . . . 93
3.2.2 The 1998 discovery of cosmic acceleration . . . 93
3.2.3 Dark energy equation of state . . . 95
3.3 Cosmic age determinations and the age crisis resolution . . . 95
3.3.1 Astrophysical age indicators in the Milky Way . . . 96
3.3.2 Resolution of the cosmological age crisis . . . 97
3.4 Thermodynamics of the expanding cosmic plasma . . . 98
3.4.1 The radiation-dominated early Universe . . . 98
3.4.2 Equilibrium relativistic thermodynamics . . . 100
3.4.3 Effective relativistic degrees of freedom $g_*(T)$ . . . 113
3.4.4 Chemical potentials and matter–antimatter asymmetry . . . 114
3.4.5 Comoving entropy conservation and $g_{*S}(T)$ . . . 115
3.4.6 Thermal history: Neutrino decoupling and e^+e^- annihilation . . . 115
3.4.7 Out-of-equilibrium relativistic Boltzmann equation . . . 118

3.4.8 Frozen interactions and thermal relic freeze-out . . . 134
3.4.9 Asymptotic solutions: Cold and hot relics . . . 136
3.4.10 Precision neutrino decoupling and the Standard Model benchmark $N_{\rm eff} = 3.044$ . . . 140
3.5 Primordial Big-Bang Nucleosynthesis (BBN) . . . 149
3.5.1 Chronology of primordial nucleosynthesis . . . 150
3.5.2 Key physical parameters governing BBN . . . 151
3.5.3 Initial state: Nuclear statistical equilibrium . . . 151
3.5.4 Freeze-out of the weak interaction . . . 153
3.5.5 Synthesis of light elements and the nuclear reaction network . . . 155
3.5.6 Observational status and cosmological constraints . . . 157
3.6 The Cosmic Microwave Background Radiation . . . 159
3.6.1 Historical prediction and discovery . . . 159
3.6.2 Thermodynamic recombination and the Saha equation . . . 159
3.6.3 Photon decoupling and the surface of last scattering . . . 161
3.6.4 Recombination dynamics and the Peebles three-level atom . . . 161
3.6.5 The optical depth and the Last Scattering Surface . . . 163
3.6.6 Observed properties of the CMB spectrum . . . 163
3.6.7 Kinematic dipole and galactic foregrounds . . . 165
3.6.8 Direct spectroscopic confirmation of cosmic cooling $T(z) = T_0(1+z)$ . . . 166
3.6.9 The Dark Age and Baryon Thermal Decoupling . . . 166
3.7 The Standard Cosmological Paradigm: Summary and Open Problems . . . 167
3.7.1 Summary of cosmological parameters and thermal history . . . 168
3.7.2 Classic problems of the standard Big-Bang paradigm . . . 168
3.7.3 Synthesis and transition to the inflationary paradigm . . . 172

III The Inhomogeneous Universe 173

4 The Inhomogeneous Universe: Perturbations and Their Evolution 174
4.1 Newtonian perturbations . . . 174
4.1.1 Case of a static space . . . 174
4.1.2 Case of an expanding space . . . 176
4.1.3 Predictions and observables . . . 180
4.1.4 Towards the non-linear regime . . . 182
4.2 Gauge invariant cosmological perturbation theory . . . 184
4.2.1 Perturbed space-time . . . 184
4.2.2 Description of matter . . . 189
4.2.3 Choosing a gauge . . . 191
4.2.4 Einstein equations: derivation . . . 194
4.2.5 Einstein equations: SVT decomposition . . . 195
4.2.6 Perturbed conservation equation for a fluid . . . 199
4.2.7 Interpretation of the perturbation equations . . . 202
4.3 Evolution . . . 203
4.3.1 Vector and tensor modes . . . 204
4.3.2 Evolution of the gravitational potential . . . 207
4.3.3 Scalar modes in the adiabatic regime . . . 212
4.3.4 Mixture of several fluids . . . 218
4.4 Power spectrum of density fluctuations . . . 226
4.4.1 Two equivalent approaches . . . 227
4.4.2 Different regimes . . . 229
4.4.3 Some refinements . . . 250

4.5 The large-scale structure of the Universe . . . 253
4.5.1 Observing the large-scale structure . . . 254
4.5.2 Structures in the non-linear regime . . . 258

5 Initial Conditions for Cosmological Perturbations 262
5.1 Background Dynamics of a Radiation + Matter Universe . . . 262
5.1.1 The Density Ratio Variable y . . . 262
5.1.2 Exact Conformal Time Integration . . . 263
5.2 Two-Fluid Cosmological Perturbations . . . 264
5.2.1 Total and Relative Perturbation Variables . . . 264
5.2.2 The Coupled Bardeen and Kodama–Sasaki Equations . . . 265
5.2.3 Relation to Metric and Gauge-Invariant Variables . . . 265
5.3 Asymptotic Expansions in Conformal Time . . . 266
5.3.1 Rescaling the System to the Dimensionless Variable x . . . 266
5.3.2 Power Series Recurrence Solution . . . 267
5.4 The Physical Initial Modes . . . 268
5.4.1 The Growing Adiabatic Mode . . . 268
5.4.2 The Growing Isocurvature Mode . . . 269
5.5 Gauge Transformation to Synchronous Gauge . . . 269
5.5.1 Synchronous Gauge Fixing . . . 270
5.5.2 Transformation from Comoving to Synchronous Gauge . . . 270
5.5.3 Explicit Synchronous Initial Series . . . 270
5.6 Multi-Fluid Primordial Dynamics in the Real Universe . . . 271
5.6.1 Multi-Fluid Density and Velocity Perturbations . . . 271
5.6.2 The Neutrino Boltzmann Hierarchy and Anisotropic Stress . . . 272
5.7 Conformal Time Expansion for the Multi-Fluid Universe . . . 273
5.7.1 Primordial Metric Amplitudes . . . 273
5.7.2 The Complete Multi-Fluid Linear Matching System . . . 273
5.7.3 Transformation to Synchronous Gauge . . . 274
5.8 The Five Primordial Initial Condition Modes . . . 274
5.8.1 The Multi-Fluid Adiabatic Mode (ADI) . . . 275
5.8.2 Matter Density Isocurvature Modes (CDI and BDI) . . . 275
5.8.3 Neutrino Density Isocurvature Mode (NDI) . . . 276
5.9 Summary and Comparison of Multi-Fluid Initial Conditions . . . 277

6 The cosmic microwave background 279
6.1 Origin of the cosmic microwave background anisotropies . . . 279
6.1.1 Sachs–Wolfe formula . . . 279
6.1.2 Angular power spectrum . . . 284
6.2 Properties of the angular power spectrum . . . 288
6.2.1 Large angular scales . . . 288
6.2.2 Intermediate scales . . . 290
6.2.3 Small scales . . . 295
6.2.4 Other effects . . . 301
6.3 Kinetic description . . . 305
6.3.1 Perturbed Boltzmann equation . . . 306
6.3.2 Gauge invariant expressions . . . 318
6.3.3 Polarization . . . 323
6.3.4 Numerical integration . . . 368
6.4 Anisotropies of the cosmic microwave background . . . 376
6.4.1 Observations of the cosmic microwave background anisotropies . . . 376
6.5 Effects of the parameters on the angular power spectrum . . . 383

6.5.1 Cosmological parameters . . . 383
6.5.2 Parameters describing the primordial physics . . . 388

7 Cosmological parameter estimation 390
7.1 Introduction . . . 390
7.1.1 The Transfer Function Formalism . . . 390
7.1.2 Parameter Measurement vs. Estimation . . . 391
7.2 The physics of parameter dependence . . . 392
7.2.1 The acoustic peaks . . . 392
7.2.2 Neutrinos . . . 396
7.2.3 Gravitational waves . . . 398
7.3 Reionization . . . 398
7.3.1 The Gunn–Peterson Trough . . . 398
7.3.2 Optical Depth to Reionization . . . 398
7.4 CMB data . . . 399
7.4.1 The Data Modeling Paradigm: Signal, Noise, and Likelihood . . . 399
7.4.2 Temperature Fluctuation Maps and Correlation Functions . . . 401
7.4.3 Example window functions . . . 404
7.4.4 Cosmic variance . . . 407
7.5 Statistical methods . . . 408
7.5.1 Bayes' Theorem and the Likelihood Function . . . 409
7.5.2 Fisher matrix and parameter estimation . . . 412
7.5.3 Model comparison and Bayesian evidence . . . 416
7.5.4 Markov chain Monte Carlo methods (MCMC) . . . 419
7.6 Degeneracies . . . 423
7.6.1 Curvature (The Geometrical Degeneracy) . . . 424
7.6.2 Scalar Spectral Index, Tensor Component, and Optical Depth . . . 425
7.6.3 Initial Conditions: Adiabatic vs. Isocurvature Modes . . . 426
7.7 Complementary observations . . . 427
7.7.1 The Hubble Parameter $H(z)$ and Type Ia Supernovae . . . 427
7.7.2 Worked example: constructing a maximum-likelihood estimator . . . 428
7.7.3 Big Bang Nucleosynthesis (BBN) . . . 429
7.7.4 The Galaxy Distribution and Large-Scale Structure . . . 430
7.7.5 Measuring the Late ISW Effect via CMB–LSS Cross-Correlations . . . 432
7.7.6 The Lyman-α Forest . . . 436
7.7.7 Weak Gravitational Lensing . . . 437
7.7.8 Galaxy Clusters . . . 438
7.8 Sources: Active Seeds and Topological Defects . . . 440
7.8.1 Topological Defects . . . 440
7.8.2 Causal Scaling Seeds . . . 442
7.8.3 Calculating CMB Anisotropies from Sources . . . 443
7.8.4 Decoherence of Acoustic Oscillations . . . 444
7.8.5 Observational Results and Constraints on Defects . . . 445

8 Gravitational Lensing and Dark Matter 447
8.1 Gravitational lensing and its applications . . . 447
8.1.1 Gravitational lensing in the thin-lens regime . . . 448
8.1.2 Lensing by galaxies and galaxy clusters . . . 477
8.1.3 Gravitational distortion by the large-scale structure . . . 486
8.1.4 Cosmic convergence . . . 491
8.1.5 Cosmic shear . . . 495
8.1.6 Lensing of the cosmic microwave background . . . 504

8.2 Evidence for the existence of dark matter . . . 533
8.2.1 Dark matter in galaxies . . . 533
8.2.2 Dark matter in clusters and groups of galaxies . . . 537
8.2.3 Cosmological evidence . . . 540
8.2.4 Summary of dark matter evidence . . . 540
8.2.5 Candidates and constraints . . . 540

9 Inflation **556**
9.1 Genesis of the paradigm . . . 556
9.1.1 Original motivations . . . 557
9.1.2 Resolution of the Big-Bang problems . . . 558
9.1.3 First models of inflation . . . 560
9.1.4 Inflation as a de Sitter phase . . . 561
9.2 Dynamics of single-field inflation . . . 564
9.2.1 Equations of evolution . . . 565
9.2.2 Slow-roll parameters . . . 571
9.2.3 End of inflation . . . 577
9.2.4 Some examples . . . 580
9.2.5 Classification of inflationary models . . . 583
9.3 Quantum fluctuations during inflation . . . 584
9.3.1 Massless test scalar field in a de Sitter space-time . . . 585
9.3.2 Massive test field in de Sitter . . . 591
9.3.3 Massive test field during slow-roll inflation . . . 592
9.4 Quantum fluctuations of the inflaton . . . 594
9.4.1 Overview of the questions to address and expected results . . . 595
9.4.2 Perturbed quantities . . . 595
9.4.3 Perturbation equations . . . 596
9.4.4 Perturbations in the slow-roll regime . . . 634
9.4.5 Relation to observations . . . 641
9.4.6 Reconstruction of the potential . . . 644
9.5 End of inflation and reheating . . . 646
9.5.1 Perturbative reheating . . . 646
9.5.2 Theory of preheating . . . 649
9.6 Eternal inflation . . . 654
9.6.1 Heuristic argument . . . 655
9.6.2 Stochastic approach . . . 656
9.7 Extensions of the inflationary paradigm . . . 659
9.7.1 Multifield inflation . . . 659
9.7.2 Primordial non-Gaussianity . . . 665
9.7.3 The trans-Planckian problem . . . 666
9.8 Status of the inflationary paradigm . . . 669
9.8.1 Generic predictions . . . 669
9.8.2 Current observational constraints . . . 670
9.8.3 Conclusions and open questions . . . 671

10 Non-Gaussianity in the Early Universe **672**
10.1 The Wavefunction of the Universe and the Origin of Gaussianity . . . 672
10.1.1 Canonical argument: the field as a collection of oscillators . . . 672
10.1.2 Path-integral argument . . . 674
10.2 Statistics of Random Fields . . . 675
10.2.1 Gaussian statistics of a single variable . . . 675
10.2.2 Multivariate Gaussian statistics and Wick's theorem . . . 675

10.2.3 Homogeneity, isotropy, and the power spectrum . . . 676
10.2.4 Weakly non-Gaussian statistics . . . 677
10.3 Application: Non-Gaussianity in Gravitational Clustering . . . 679
10.4 Non-Gaussianity as a Test of Inflation . . . 681
10.4.1 Power spectrum and scale invariance . . . 681
10.4.2 The bispectrum and its shapes . . . 683
10.5 Multi-Field Inflation and the δN Formalism . . . 684
10.5.1 Single-field check . . . 685
10.5.2 Local-form non-Gaussianity from δN . . . 686
10.5.3 The squeezed bispectrum . . . 687
10.6 The Single-Field Consistency Theorem . . . 688
10.7 The Non-linearity Parameter $f_{\rm NL}$ . . . 689
10.8 Trispectrum and the Multi-Field Consistency Relation . . . 689
10.8.1 The Suyama–Yamaguchi inequality . . . 690
10.9 Measuring Non-Gaussianity from the CMB . . . 691
10.9.1 Toy model: estimating a skewness . . . 691
10.9.2 PDF of the CMB temperature field . . . 691
10.9.3 Relating $a_{\ell m}$ to $\zeta_{\vec{k}}$ . . . 694

11 Cosmological Collider Physics: Poor man's accelerator 696
11.1 Two-Point Functions in de Sitter Space . . . 699
11.2 Two-Point Functions of Fields with Spin . . . 705
11.2.1 Position-space correlation functions . . . 705
11.2.2 Conformal Ward identities in momentum space . . . 709
11.2.3 Momentum-space two-point function with spin . . . 711
11.2.4 Mass-dimension relation, positivity, and partial masslessness . . . 715
11.3 Three-Point Functions in de Sitter Space . . . 716
11.4 Two Conformally Coupled Fields and a General Scalar . . . 717
11.5 Three-Point Functions of Massless Inflaton Fluctuations . . . 721
11.6 Four-Point Functions . . . 721
11.6.1 Contact four-point function ($\lambda\phi^4$ interaction) . . . 722
11.6.2 Exchange four-point function ($\lambda\phi^3$ interaction) . . . 722
11.6.3 Four-point function with massive intermediate state . . . 727
11.6.4 Four-point function with a spinning particle intermediate state . . . 733
11.7 Three Point in inflation . . . 735
11.7.1 Intermediate massive field with spin . . . 739
11.8 Comments on Loop correction . . . 740
11.9 Conclusion and Cosmological Horizons . . . 742
11.9.1 Summary of the Particle Spectrogram . . . 742
11.9.2 Scale-Dependent vs. Ratio-Dependent Oscillations . . . 743
11.9.3 Observational Prospects: CMB, LSS, and 21-cm Tomography . . . 744
11.9.4 Frontiers: The Cosmological Bootstrap and Quantum Gravity . . . 744

IV Appendices 745

A Numerical values 746
A.1 Physical constants . . . 746
A.2 Astrophysical quantities . . . 746
A.3 Units . . . 746
A.3.1 Natural units . . . 747
A.3.2 Conversion factors . . . 748

A.4 Particle physics . . . 748
A.5 Cosmological quantities . . . 749
A.6 Electromagnetic spectrum . . . 749

B Special Functions **753**
B.1 Euler Functions . . . 753
B.2 Spherical Harmonics . . . 755
B.2.1 Definition . . . 755
B.2.2 Expressions . . . 755
B.2.3 Properties . . . 755
B.2.4 Physical Interpretation: Standing Waves on the Sphere . . . 756
B.2.5 Fourier Transform . . . 757
B.2.6 Useful Integrals . . . 760
B.3 Bessel Functions . . . 760
B.3.1 Definition . . . 760
B.3.2 Asymptotic Properties . . . 761
B.3.3 Special Cases . . . 761
B.3.4 Spherical Bessel Functions . . . 762
B.3.5 Some Useful Integrals . . . 763
B.4 Legendre Polynomials . . . 764
B.4.1 Associated Legendre Polynomials . . . 764
B.4.2 Legendre Polynomials . . . 764
B.5 Fourier Transform and Eigenmodes of the Laplacian . . . 766
B.5.1 Fourier Transform . . . 766
B.5.2 Power Spectrum . . . 766
B.5.3 Cartesian Coordinates . . . 767
B.5.4 Spherical Coordinates . . . 767

C Useful Cosmological Quantities **769**
C.1 Background Space . . . 769
C.1.1 Geometry . . . 769
C.1.2 Matter . . . 770
C.2 Perturbed Quantities . . . 772
C.2.1 Geometry . . . 772
C.2.2 Matter . . . 773

D Monte Carlo Methods in CMB Lensing Reconstruction **777**
D.1 Introduction and Observational Motivation . . . 777
D.2 The Same-Sky Quadratic Estimator and its Trispectrum . . . 778
D.3 Separation of Disconnected and Connected Four-Point Statistics . . . 779
D.4 Realization Independence and the Mixed-Sky Formulation . . . 779
D.4.1 Cross-Paired Estimator Construction with Realization Annotations . . . 780
D.4.2 The Mixed Auto-Spectrum: Direct Disconnected Contraction . . . 781
D.4.3 The Mixed Cross-Spectrum: Cross Disconnected Contraction . . . 781
D.4.4 Complete Exact Recovery of $N_L^{(0)}$ from Independent Skies . . . 781
D.5 Algorithmic Pipeline Implementation: Paired Simulation Execution . . . 782
D.5.1 Partitioning into $N_{\rm sim}/2$ Disjoint Realization Pairs . . . 782
D.5.2 Step-by-Step Paired Accumulation . . . 782
D.5.3 Statistical Convergence and Variance Scaling . . . 783
D.6 Master Decomposition of the Observed Power Spectrum . . . 783
D.7 Shared Lensing Potentials and Isolation of the $N_L^{(1)}$ Bias . . . 784
D.8 Practical Considerations and Methodological Nuances . . . 784

D.8.1 The Fatal Pitfall of Unlensed Simulations ($\phi = 0$) for Polarization 785
D.8.2 Realization-Dependent Mean-Field Subtraction 786
D.8.3 Comparison: Monte Carlo Simulation vs. Analytical Fixed-Point Iteration 786

Part I

Overview of the Theoretical Basis

Chapter 1

General relativity

Gravitation holds a uniquely privileged position in astrophysics and cosmology. While the strong and weak nuclear forces operate only across subatomic distances, and electromagnetic forces cancel out due to the large-scale electrical neutrality of astronomical bodies, gravity is an un-screened, universally attractive interaction that acts across infinite distances. Consequently, the geometry, physical evolution, and large-scale cosmic structure of the Universe are dictated entirely by general relativity.

Where this chapter is going

The path from Newton to modern cosmology turns on a single insight: gravity is geometry. We trace how this develops through five connected physical ideas. First, the conflict between Maxwellian electrodynamics and Galilean relativity forces space and time into a single four-dimensional continuum, where the invariant speed of light replaces absolute simultaneity. Second, the universality of free fall—the exact equality of inertial and gravitational mass—reveals that gravity cannot be an ordinary force pulling objects through space, but is rather the curvature of spacetime itself. Third, to do calculus on a curved surface without a global Cartesian grid, we build the intrinsic machinery of differential geometry: tangent vectors as directional derivatives, dual 1-forms as gradients, the metric as the physical ruler, the connection for parallel transport, and the Riemann tensor to detect genuine tidal curvature. Fourth, Einstein's field equations equate this spacetime curvature to the stress, energy, and momentum of matter. Finally, we explore the physical consequences: geodesic motion, gravitational redshift, the generation and propagation of gravitational waves, and the cosmological framework that governs the primordial Universe.

1.1 Space-time and gravity

Every physical theory requires an arena. In classical mechanics, that arena was an immutable, three-dimensional Euclidean space through which an external time flowed uniformly. Special relativity fused space and time into a four-dimensional Minkowski continuum, but left the background rigid and inert. General relativity takes the decisive step: spacetime becomes a dynamic participant in the physics, curved by the matter within it and in turn dictating how that matter moves.

1.1.1 Absolute space and time of Newtonian physics

In classical mechanics, Newton's first law posits that any body subject to no external forces continues in its state of rest or of uniform rectilinear motion. Yet this foundational statement immediately raises a deep physical conundrum: relative to what frame of reference is this uniform

straight-line motion defined? If an observer stands on a rotating carousel, free particles curve and accelerate wildly without any material agent acting on them. If acceleration is only meaningful relative to a reference frame, yet inertial frames are defined as those in which unaccelerated bodies experience no forces, Newtonian mechanics threatens to collapse into circular reasoning. To break this circle and anchor the laws of motion, Newton postulated the existence of **absolute space** $\mathbb{E}^3$—a rigid, immutable three-dimensional Euclidean arena that exists independently of any matter contained within it.

Running parallel to absolute space is an **absolute time** t. Time is an external, universal parameter that flows uniformly across the cosmos, completely unaffected by the motion, position, or gravitational environment of any observer.

Newtonian causality and simultaneity. Because signals can in principle travel at arbitrarily high speeds in Newtonian physics ($v \to \infty$), the causal relationship between any two events P and Q is absolute. Three mutually exclusive possibilities exist:

1. $t_Q > t_P$: A physical trajectory can travel from P to Q; hence event Q lies strictly in the **future** of P.
2. $t_Q < t_P$: A physical trajectory can travel from Q to P; hence event Q lies strictly in the **past** of P.
3. $t_Q = t_P$: No physical particle can travel between P and Q because travel would require infinite velocity. The two events are **simultaneous**.

The set of all events simultaneous with P forms a three-dimensional Euclidean hypersurface $\mathbb{E}^3$ representing "space at time t" (Figure 1.1).

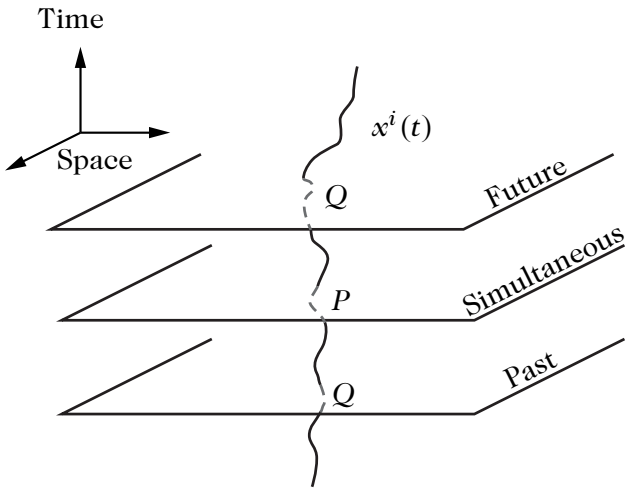


Figure 1.1: The causal structure of Newtonian space-time. If an event Q can be connected to P by a trajectory $x^i(t)$ with finite speed ($\mathrm{d}x^i / \mathrm{d}t < \infty$), Q lies unambiguously in the past or the future of P. If $\mathrm{d}t = 0$, the two events are strictly simultaneous. The set of all simultaneous events forms a three-dimensional Euclidean space.

Spatial distance and the metric. Because Newtonian space is assumed to be Euclidean, Pythagoras' theorem gives the infinitesimal spatial distance $\mathrm{d}\ell$ between two neighboring points in Cartesian coordinates (x^1, x^2, x^3):

$$\mathrm{d}\ell^2 = (\mathrm{d}x^1)^2 + (\mathrm{d}x^2)^2 + (\mathrm{d}x^3)^2. \tag{1.1.1}$$

Adopting the **Einstein summation convention**—where repeated upper and lower indices are automatically summed over—we write the spatial line element as:

$$\mathrm{d}\ell^2 = \sum_{i=1}^{3} (\mathrm{d}x^i)^2 = \sum_{i,j=1}^{3} \delta_{ij}\, \mathrm{d}x^i\, \mathrm{d}x^j \equiv \delta_{ij}\, \mathrm{d}x^i\, \mathrm{d}x^j\,, \tag{1.1.2}$$

where δ_{ij} is the Kronecker delta symbol ($\delta_{ij} = 1$ if $i = j$, and 0 otherwise), and Latin indices $i, j, k \ldots$ run from 1 to 3. The scalar product of two spatial vectors $\mathbf{T}$ and $\mathbf{V}$ is thus $\mathbf{T} \cdot \mathbf{V} = \delta_{ij} T^i V^j = T^1 V^1 + T^2 V^2 + T^3 V^3$.

A physical body traces a path parameterized by absolute time, $x^i(t)$. The elapsed time between point A at t_A and point B at t_B is simply $\Delta t = t_B - t_A$, which is identical for all observers regardless of the spatial path taken.

Coordinates as labels versus the metric as ruler. At the foundation of geometry lies an essential conceptual distinction that is frequently a source of confusion for the beginner: the distinction between the coordinate system used to label points and the intrinsic geometry of the space itself. Coordinates are purely arbitrary, human-assigned book-keeping labels; they carry no intrinsic physical meaning or physical length. The physical geometry resides entirely in the metric tensor.

To make this distinction vivid, suppose we pass from standard Cartesian coordinates (x, y, z) to spherical polar coordinates (r, θ, ϕ) in Euclidean three-space:

$$x = r \sin\theta \cos\phi, \qquad y = r \sin\theta \sin\phi, \qquad z = r \cos\theta. \tag{1.1.3}$$

Under an arbitrary coordinate transformation $x^j(y^i)$, displacement differentials transform via the multivariable chain rule, $\mathrm{d}x^j = (\partial x^j / \partial y^i)\, \mathrm{d}y^i$. Substituting these differentials into Equation (1.1.2) yields

$$\mathrm{d}\ell^2 = g_{ij}(y^k)\, \mathrm{d}y^i\, \mathrm{d}y^j\,, \qquad g_{ij}(y^k) \equiv \frac{\partial x^m}{\partial y^i} \frac{\partial x^n}{\partial y^j} \delta_{mn}, \tag{1.1.4}$$

where g_{ij} are the components of the metric tensor in the spherical coordinate basis. Computing the partial derivatives produces $g_{rr} = 1$, $g_{\theta\theta} = r^2$, and $g_{\phi\phi} = r^2 \sin^2\theta$, giving the line element:

$$\mathrm{d}s^2 = \mathrm{d}r^2 + r^2 \left(\mathrm{d}\theta^2 + \sin^2\theta\ \mathrm{d}\phi^2 \right). \tag{1.1.5}$$

Notice that the metric components now depend explicitly on the spatial coordinate r and the polar angle θ. One might naively jump to the conclusion that because the metric coefficients are non-trivial and position-dependent, the space has become curved. But that conclusion would be entirely mistaken: the space is identical to the flat Euclidean space we started with. The position-dependence of g_{ij} is simply a consequence of choosing coordinates whose coordinate surfaces curve and diverge. Coordinates (y^1, y^2, y^3) provide addresses for points, while the metric g_{ij} serves as the physical ruler that converts infinitesimal coordinate increments into invariant spatial lengths. As we will see, genuine physical curvature is an intrinsic property that cannot be created or eliminated by coordinate transformations.

Galilean group

Which coordinate transformations preserve the form of the Euclidean line element (1.1.1)? The transformations that preserve the mutual distances between all points at any given instant t are rigid rotations and spatial translations:

$$x^i \longrightarrow y^i = \Lambda^i{}_j(t) x^j + T^i(t), \qquad \Lambda^i{}_j(t) \Lambda_k{}^j(t) = \delta^i_k. \tag{1.1.6}$$

Here, the orthogonal rotation matrix $\Lambda^i{}_j(t)$ may in principle depend on time, but it is spatially rigid (independent of position x^j).

To ensure that the laws of Newtonian mechanics take the exact same form in two different reference frames—preserving the definition of an **inertial frame** where free particles experience no acceleration—the relative velocity must be constant. Restricting Equation (1.1.6) to constant rotations and linear translations in time ($T^i(t) = T_0^i - v^i t$) yields the **Galilean group**:

$$x^i \longrightarrow y^i = \Lambda^i{}_j x^j + T_0^i - v^i t, \qquad \Lambda^i{}_j \Lambda_k{}^j = \delta_k^i. \tag{1.1.7}$$

Taxonomy of the Galilean group. The Galilean group is a **10-parameter Lie group** composed of:

- **3 spatial rotations** ($\Lambda^i{}_j$ parameterized by 3 Euler angles): encoding the spatial isotropy of the Universe.
- **3 spatial translations** (T_0^i): encoding the spatial homogeneity of the Universe.
- **3 Galilean boosts** (v^i): describing relative uniform motion between inertial observers.
- **1 time translation** ($t \to t + t_0$): encoding the time-translation invariance of physical laws.

Trajectory of a massive body

In Newtonian mechanics, the concept of "mass" enters in two entirely different physical contexts:

1. **Inertial mass (m_I):** Appears in Newton's second law, $\mathbf{F} = m_\mathrm{I}\mathbf{a}$. It measures an object's resistance to being accelerated by *any* physical force.
2. **Gravitational mass (m_G):** Appears in Newton's universal law of gravitation, $\mathbf{F}_\mathrm{grav} = -m_\mathrm{G}\boldsymbol{\nabla}\phi$. It acts as the "gravitational charge", quantifying how strongly a body couples to a gravitational potential ϕ.

To see the consequence of this distinction, consider the action for a particle moving in a gravitational potential $\phi(x^k)$ formulated in arbitrary spatial coordinates x^i:

$$L = \frac{1}{2} m_\mathrm{I} v^2 + m_\mathrm{G}\phi = \frac{1}{2} m_\mathrm{I} g_{ij}(x^k)\dot{x}^i\dot{x}^j + m_\mathrm{G}\phi(x^k), \tag{1.1.8}$$

where $\dot{x}^i \equiv \mathrm{d}x^i / \mathrm{d}t$. Applying the Euler–Lagrange equations,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}^i}\right) - \frac{\partial L}{\partial x^i} = 0, \tag{1.1.9}$$

we obtain the exact classical equation of motion in arbitrary coordinates:

$$\ddot{x}^i + \Gamma^i_{jk}\dot{x}^j\dot{x}^k = -\frac{m_\mathrm{G}}{m_\mathrm{I}} g^{ij}\partial_j\phi, \tag{1.1.10}$$

where Γ^i_{jk} are the spatial Christoffel connection coefficients:

$$\Gamma^i_{jk} \equiv \frac{1}{2} g^{ip}\left(\partial_j g_{kp} + \partial_k g_{jp} - \partial_p g_{jk}\right). \tag{1.1.11}$$

Inertial forces and the coincidence of mass. Equation (1.1.10) reveals two profound physical lessons:

- In Cartesian coordinates ($g_{ij} = \delta_{ij}$), the connection coefficients vanish ($\Gamma^i_{jk} = 0$). In curvilinear or accelerating coordinates, $\Gamma^i_{jk} \neq 0$, and the term $\Gamma^i_{jk}\dot{x}^j\dot{x}^k$ generates the familiar **fictitious inertial forces** (centrifugal, Coriolis, etc.).
- If and only if **inertial mass equals gravitational mass** ($m_\mathrm{I} = m_\mathrm{G}$), the mass of the body completely cancels out from both sides of Equation (1.1.10)!

When $m_\mathrm{I} = m_\mathrm{G}$, all bodies placed at the same initial position with the same initial velocity trace **identical trajectories in a gravitational field**, regardless of their mass, shape, or chemical composition. This empirical fact is known as the **Weak Equivalence Principle (WEP)** or the **Universality of Free Fall**.

In Newtonian physics, the exact equality $m_\mathrm{I} = m_\mathrm{G}$ is a complete accident—there is no theoretical reason why an object's inertial resistance should match its gravitational charge. Newton himself recognized the profound nature of this coincidence and used pendulums of different materials (gold, silver, lead, glass, wheat) to verify that:

$$\frac{|m_\mathrm{G} - m_\mathrm{I}|}{m_\mathrm{G} + m_\mathrm{I}} < 10^{-3}. \tag{1.1.12}$$

What was a happy coincidence for Newton became the foundational cornerstone of general relativity for Einstein.

1.1.2 Space-time of special relativity

In the late nineteenth century, physics faced a severe contradiction. Maxwell's equations of electromagnetism predict that light is an electromagnetic wave propagating in vacuum at an invariant speed:

$$c = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \approx 3 \times 10^8 \,\mathrm{m/s}. \tag{1.1.13}$$

However, under Galilean relativity (1.1.7), velocities add vectorially: $\mathbf{v}' = \mathbf{v} - \mathbf{u}$. An observer chasing a light beam at speed u should measure the light speed as $c - u$.

This clash between electrodynamics and Newtonian mechanics posed a profound puzzle: does light propagate through a pervasive mechanical medium—the luminiferous aether—or is the speed of light an invariant constant of nature for all inertial observers? In the nineteenth century, physicists naturally assumed that Maxwell's equations held strictly in the rest frame of this hypothetical aether. An observer moving through the aether with velocity $\mathbf{u}$ was expected to measure an altered speed of light $c \pm u$, experiencing an apparent "aether wind" that would reveal Earth's absolute motion through the cosmos.

Lorentz transformations

The Michelson–Morley experiment (1887) delivered a decisive blow to the aether hypothesis: it demonstrated to high precision that the speed of light is completely isotropic and independent of Earth's orbital velocity through space.

In 1905, Albert Einstein resolved the paradox through two foundational postulates:

1. **The Principle of Relativity:** The laws of physics are identical in all inertial reference frames.
2. **Constancy of the Speed of Light:** The speed of light in vacuum is an invariant constant c, identical for all inertial observers regardless of their relative motion or the motion of the source.

Because the speed of light is invariant, the transformation between two inertial frames moving with relative speed v along their mutual x-axis cannot be the Galilean boost. Instead, it must be the **Lorentz boost**:

$$ct' = \gamma\left(ct - \frac{v}{c}x\right), \qquad x' = \gamma\left(x - \frac{v}{c}ct\right), \qquad y' = y, \qquad z' = z, \tag{1.1.14}$$

where γ is the Lorentz factor:

$$\gamma \equiv \frac{1}{\sqrt{1 - v^2/c^2}}. \tag{1.1.15}$$

In the non-relativistic limit $v \ll c$, $\gamma \to 1$ and $vx/c \to 0$, so the Lorentz boost smoothly reduces to the Galilean transformation $x' = x - vt, t' = t$. At relativistic speeds, time and space become intimately entwined.

Minkowski space-time

In special relativity, individual measurements of spatial distance Δx and time intervals Δt are observer-dependent. However, Hermann Minkowski (1908) recognized that a specific combination of space and time is invariant for all inertial observers: the **spacetime interval** $\mathrm{d}s^2$.

Replacing the Euclidean line element (1.1.1), the interval between two infinitesimally separated events is:

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + (\mathrm{d}x^1)^2 + (\mathrm{d}x^2)^2 + (\mathrm{d}x^3)^2 = -(\mathrm{d}x^0)^2 + (\mathrm{d}x^1)^2 + (\mathrm{d}x^2)^2 + (\mathrm{d}x^3)^2 \equiv \eta_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu\,, \tag{1.1.16}$$

where $x^0 \equiv ct$, Greek indices run from 0 to 3, and $\eta_{\mu\nu}$ is the **Minkowski metric tensor**, which in Cartesian coordinates takes the diagonal form:

$$\eta_{\mu\nu} = \mathrm{diag}(-1, +1, +1, +1). \tag{1.1.17}$$

(We adopt the "mostly plus" signature $(-,+,+,+)$ standard in modern cosmology and general relativity.)

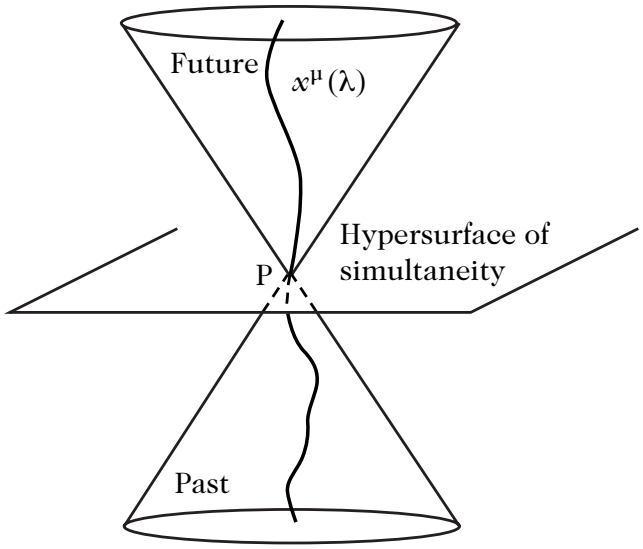


Figure 1.2: The causal structure of Minkowski space-time. The Newtonian concept of universal simultaneity is replaced by the **light cone** defined by $\mathrm{d}s^2 = 0$. Massive particle worldlines $x^\mu(\lambda)$ lie strictly within the interior of the light cone ($\mathrm{d}s^2 < 0$). Events outside the light cone ($\mathrm{d}s^2 > 0$) are spacelike separated and cannot be causally influenced by P.

Taxonomy of spacetime intervals. Unlike the Euclidean distance $\mathrm{d}\ell^2 \geq 0$, the pseudo-Euclidean spacetime interval $\mathrm{d}s^2$ can be positive, negative, or zero. This provides an absolute, observer-independent **causal classification** of spacetime (Figure 1.2):

- **Time-like interval ($\mathrm{d}s^2 < 0$):** The events lie inside each other's light cones. A physical particle with mass can travel between them at speed $v < c$. An inertial frame always exists in which the two events occur at the same spatial position ($\mathrm{d}x^i = 0$), separated purely by time.
- **Null / Light-like interval ($\mathrm{d}s^2 = 0$):** The events lie on the boundary of the light cone. They can be connected only by signals propagating at the speed of light c (photons, gravitons).
- **Space-like interval ($\mathrm{d}s^2 > 0$):** The events lie outside each other's light cones. No causal signal can travel between them without exceeding the speed of light. An inertial frame always exists in which the two events occur simultaneously ($\mathrm{d}t = 0$), separated purely by spatial distance.

Proper time

Coordinate time versus proper time. In classical mechanics, one unthinkingly identifies the coordinate time t that labels points on a trajectory with the elapsed time recorded by a physical clock carried along that trajectory. In special relativity, this naive identification breaks down completely. Coordinate time t is merely an arbitrary coordinate assigned to events by a chosen observer's grid of synchronized clocks. In contrast, the **proper time** τ is the invariant physical time measured by an ideal clock carried directly along the particle's own worldline.

Along a timelike trajectory $x^\mu(\lambda)$, the infinitesimal proper time $\mathrm{d}\tau$ is defined directly in terms of the invariant spacetime interval:

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}\tau^2\,. \tag{1.1.18}$$

Integrating this interval along the worldline from an initial event to a final event, the total elapsed proper time is:

$$\tau = \frac{1}{c}\int \sqrt{-\eta_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda}}\;\mathrm{d}\lambda\,. \tag{1.1.19}$$

When the trajectory is parameterized by proper time τ itself, the particle's tangent vector is its **four-velocity**, $u^\mu \equiv \mathrm{d}x^\mu / \mathrm{d}\tau$. It satisfies the universal normalization condition:

$$u^\mu u_\mu = \eta_{\mu\nu}u^\mu u^\nu = -c^2. \tag{1.1.20}$$

To see the physical consequences of this distinction, consider an observer $\mathcal{O}$ at rest in an inertial frame $\mathcal{S}$ ($\mathrm{d}x^i = 0$), whose proper time coincides with coordinate time, $\mathrm{d}\tau_{\mathcal{O}} = \mathrm{d}t$. Now consider a second observer $\mathcal{O}'$ moving through $\mathcal{S}$ with spatial velocity $v^i = \mathrm{d}x^i / \mathrm{d}t$. The proper time interval experienced by $\mathcal{O}'$ is:

$$\mathrm{d}\tau^2_{\mathcal{O}'} = \mathrm{d}t^2 - \frac{\delta_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j}{c^2} = \left(1 - \frac{v^2}{c^2}\right)\mathrm{d}t^2 \implies \mathrm{d}\tau_{\mathcal{O}'} = \sqrt{1 - \frac{v^2}{c^2}}\;\mathrm{d}\tau_{\mathcal{O}}\,. \tag{1.1.21}$$

Equation (1.1.21) is the familiar relation of **relativistic time dilation**: a moving clock runs slow relative to stationary coordinate clocks. From the geometric perspective of Minkowski spacetime, time dilation is not a dynamical distortion of clock mechanisms, but rather the geometric statement that different timelike curves connecting the same two events have different spacetime arc lengths.

Poincaré group

Just as in Euclidean space, the Minkowski metric can be written in arbitrary curvilinear coordinates $y^\mu(x^\nu)$:

$$\mathrm{d}s^2 = g_{\mu\nu}(y^\lambda)\,\mathrm{d}y^\mu\,\mathrm{d}y^\nu\,, \qquad g_{\mu\nu}(y^\lambda) = \frac{\partial x^\alpha}{\partial y^\mu}\frac{\partial x^\beta}{\partial y^\nu}\eta_{\alpha\beta}. \tag{1.1.22}$$

For example, an observer undergoing constant proper acceleration α along the x-axis naturally uses **Rindler coordinates** (T, R, y', z'), related to Minkowski coordinates by:

$$ct = R\sinh\left(\frac{\alpha T}{c}\right), \qquad x = R\cosh\left(\frac{\alpha T}{c}\right), \qquad y = y', \qquad z = z'. \tag{1.1.23}$$

In these coordinates, the line element takes the non-trivial form:

$$\mathrm{d}s^2 = -\left(\frac{\alpha R}{c}\right)^2 \mathrm{d}T^2 + \mathrm{d}R^2 + \mathrm{d}y'^2 + \mathrm{d}z'^2\,. \tag{1.1.24}$$

Even though the metric components depend on R, the spacetime is completely flat.

The symmetries of flat spacetime. The coordinate transformations that leave the functional form of the Minkowski metric invariant are the transformations of the **Poincaré group** (the inhomogeneous Lorentz group):

$$x^\mu \longrightarrow y^\mu = \Lambda^\mu{}_\nu x^\nu + T^\mu, \qquad \Lambda^\mu{}_\nu \Lambda_\sigma{}^\nu = \delta^\mu_\sigma, \tag{1.1.25}$$

where $\Lambda^\mu{}_\nu$ is a constant Lorentz matrix and T^μ is a constant four-displacement vector.

The Poincaré group has **10 independent parameters**:

- 3 spatial rotations (mixing spatial coordinates among themselves),
- 3 Lorentz boosts (mixing time and space coordinates),
- 4 spacetime translations (3 spatial translations and 1 time translation).

Poincaré algebra

The structure of the Poincaré group is illuminated by its infinitesimal generators. For an infinitesimal translation $T^\mu = \epsilon^\mu$, a scalar field transforms as $f(x+\epsilon) = (1 - i\epsilon^\mu P_\mu)f(x)$, which identifies the **spacetime translation generator**:

$$P_\mu = i\partial_\mu. \tag{1.1.26}$$

For an infinitesimal Lorentz transformation $y^\mu = x^\mu + \omega^\mu{}_\nu x^\nu$ with antisymmetric parameter $\omega_{\mu\nu} = -\omega_{\nu\mu}$, expanding $f(y) = (1 - \frac{i}{2}\omega^{\mu\nu}J_{\mu\nu})f(x)$ yields the **Lorentz generators** (rotations and boosts):

$$J_{\mu\nu} = i\left(x_\mu\partial_\nu - x_\nu\partial_\mu\right). \tag{1.1.27}$$

These generators satisfy the Lie algebra commutation relations of the **Poincaré algebra** $\mathfrak{iso}(3,1)$:

$$[P_\alpha, P_\beta] = 0, \tag{1.1.28}$$

$$[J_{\mu\nu}, P_\alpha] = i\left(\eta_{\mu\alpha}P_\nu - \eta_{\nu\alpha}P_\mu\right), \tag{1.1.29}$$

$$[J_{\mu\nu}, J_{\alpha\beta}] = i\left(\eta_{\mu\alpha}J_{\beta\nu} + \eta_{\mu\beta}J_{\alpha\nu} + \eta_{\nu\alpha}J_{\mu\beta} + \eta_{\nu\beta}J_{\alpha\mu}\right). \tag{1.1.30}$$

Equation (1.1.30) defines the closed subalgebra of the Lorentz group $\mathfrak{so}(3,1)$. Notice that translations commute among themselves ($[P_\alpha, P_\beta] = 0$), reflecting the translational flatness of Minkowski spacetime.

Trajectory of a free body

In special relativity, the trajectory of a free massive particle is obtained by extremizing the relativistic action. In arbitrary coordinates, the Lagrangian is:

$$L = \frac{1}{2} g_{\mu\nu}(x^\lambda) \dot{x}^\mu \dot{x}^\nu, \tag{1.1.31}$$

where $\dot{x}^\mu \equiv \mathrm{d}x^\mu / \mathrm{d}\lambda$ with respect to an affine parameter λ. The Euler–Lagrange equations yield the four-dimensional **geodesic equation**:

$$\ddot{x}^\mu + \Gamma^\mu_{\nu\lambda} \dot{x}^\nu \dot{x}^\lambda = 0, \qquad \Gamma^\mu_{\nu\lambda} \equiv \frac{1}{2} g^{\mu\sigma} \left(\partial_\nu g_{\lambda\sigma} + \partial_\lambda g_{\nu\sigma} - \partial_\sigma g_{\nu\lambda} \right). \tag{1.1.32}$$

In standard Minkowski Cartesian coordinates, $g_{\mu\nu} = \eta_{\mu\nu}$, so $\Gamma^\mu_{\nu\lambda} = 0$, and Equation (1.1.32) reduces to the statement of uniform rectilinear motion:

$$\ddot{x}^\mu = 0 \iff u^\nu \partial_\nu u^\mu = 0. \tag{1.1.33}$$

Why Newtonian gravity cannot fit into special relativity. Can we simply add a Newtonian gravitational potential Φ to Equation (1.1.31)? Three fatal physical obstacles prevent it:

1. **Causality Violation:** Newton's gravitational Poisson equation $\nabla^2 \Phi = 4\pi G_\mathrm{N} \rho$ is instantaneous. If a mass moves, its gravitational potential changes everywhere in the universe at that exact instant, violating the relativistic cosmic speed limit c.
2. **Relativistic Sources of Gravity:** In relativity, mass is not an invariant scalar; it is related to energy by $E = mc^2$. Pressure, momentum, and stress also carry energy and must gravitate. A relativistic theory of gravity cannot have a single scalar density ρ as its source; it must couple to the entire four-dimensional stress-energy tensor $T_{\mu\nu}$.
3. **Light Deflection:** Light has zero rest mass ($m = 0$), but carries energy and momentum ($E = \hbar\omega$). A scalar gravitational theory that couples purely to rest mass would leave light completely undeflected, directly contradicting observation.

1.1.3 General relativity and curved space-time

Equivalence principle

How can we construct a relativistic theory of gravity if gravitational forces cannot act instantaneously at a distance? Einstein's breakthrough came in 1907 with what he called "the happiest thought of my life":

> *"If a person falls freely, he will not feel his own weight. All objects falling with him will accelerate at the exact same rate, remaining at rest relative to him."*

Einstein realized that this thought experiment reveals an exact physical equivalence: can an observer enclosed within a small, freely falling laboratory distinguish their environment from that of an inertial observer floating in deep, gravity-free space? The answer is no. Because all objects fall with the exact same acceleration in a gravitational field ($m_\mathrm{I} = m_\mathrm{G}$), an observer inside the elevator who releases an apple and a lead weight sees both hover motionlessly alongside them. If the observer shines a laser across the elevator, the beam travels in a straight line relative to the elevator's interior walls.

Consequently, **within a sufficiently small, freely falling laboratory, all effects of gravity vanish entirely**. The local laws of physics are indistinguishable from those of special relativity in an unaccelerated inertial frame in deep space.

This complete cancellation is unique to gravitation:

- In electromagnetism, the acceleration of a charged particle in an electric field is $a = (q/m)E$. Because different particles have different charge-to-mass ratios q/m, an accelerated frame can never eliminate the electric field for all particles simultaneously.
- In gravity, the "gravitational charge-to-mass ratio" is $m_{\rm G}/m_{\rm I} \equiv 1$ for every physical object in the universe. Therefore, a single accelerated frame eliminates gravity for *all* bodies at once.

Tidal forces and space-time curvature

Eliminating gravity locally versus globally. Having established that gravity can be eliminated in a freely falling elevator, one might jump to the conclusion that gravity is nothing more than a coordinate artifact—a fictitious force like the centrifugal or Coriolis force, capable of being transformed away globally across the entire universe.

This conclusion would be mistaken, because it overlooks the vital fact that real gravitational fields are non-uniform. A coordinate transformation can eliminate gravity at a single point or within an infinitesimally small laboratory, but it cannot eliminate it across an extended region of space. To see why, consider two test bodies in orbit around Earth: body A at position $\mathbf{r}_A$ and body B at position $\mathbf{r}_B = \mathbf{r}_A + \delta\mathbf{r}$. Taylor expanding the Newtonian gravitational field to first order in the spatial separation $\delta\mathbf{r}$, the relative acceleration of body B with respect to body A is:

$$\delta\mathbf{a}_{AB} \equiv \mathbf{a}_B - \mathbf{a}_A = -\frac{G_{\rm N}M_\oplus}{r_A^3}\left[\delta\mathbf{r} - 3\frac{\mathbf{r}_A\cdot\delta\mathbf{r}}{r_A^2}\mathbf{r}_A\right]. \tag{1.1.34}$$

No coordinate acceleration can eliminate this relative acceleration, because it is non-uniform:

- Along transverse directions perpendicular to the radius ($\mathbf{r}_A\cdot\delta\mathbf{r} = 0$), the relative acceleration is $\delta\mathbf{a}_{AB} = -(G_{\rm N}M_\oplus/r_A^3)\delta\mathbf{r}$. The two particles accelerate toward each other because both are falling along converging radial lines toward Earth's center.
- Along the radial direction parallel to the radius ($\delta\mathbf{r} \parallel \mathbf{r}_A$), the relative acceleration is $\delta\mathbf{a}_{AB} = +2(G_{\rm N}M_\oplus/r_A^3)\delta\mathbf{r}$. The particles accelerate away from each other because the lower particle is closer to Earth and experiences a stronger gravitational attraction.

An initially spherical cloud of freely falling particles is stretched along the radial direction and compressed along transverse directions, forming an ellipsoid of constant volume ($\boldsymbol{\nabla}\cdot\mathbf{g} = 0$ in vacuum).

These differential accelerations are the familiar **tidal forces**. In Newtonian physics, tidal forces are governed by the second spatial derivatives of the gravitational potential, $\partial_i\partial_j\Phi$. In general relativity, they represent the direct physical manifestation of **spacetime curvature**. Gravity is not an external physical force that deflects bodies from straight paths; rather, freely falling bodies simply follow the straightest possible paths (geodesics) through a four-dimensional spacetime that is intrinsically curved.

Gravity as a manifestation of geometry

Taxonomy of the equivalence principles. Precision experimental tests distinguish three hierarchically nested formulations of the equivalence principle:

1. **Weak Equivalence Principle (WEP):** The trajectory of an uncharged, freely falling test body with negligible self-gravitational energy is completely independent of its mass and internal chemical composition (universality of free fall).
2. **Einstein Equivalence Principle (EEP):**

- WEP is valid;
- **Local Lorentz Invariance (LLI):** The outcome of any local non-gravitational test experiment is independent of the velocity of the freely falling laboratory;
- **Local Position Invariance (LPI):** The outcome of any local non-gravitational test experiment is independent of where and when in spacetime it is performed.

3. **Strong Equivalence Principle (SEP):** Extends EEP to include bodies with substantial *gravitational self-energy* (such as stars, neutron stars, and black holes), as well as local experiments that measure gravitational phenomena itself. All bodies, regardless of their own self-gravitational binding energy, fall with identical acceleration in an external gravitational field.

Any theory of gravitation that satisfies the Einstein Equivalence Principle is necessarily a **metric theory of gravity**, defined by three fundamental postulates:

- Spacetime is endowed with a symmetric metric tensor $g_{\mu\nu}$.
- Freely falling test particles follow the geodesics of this metric.
- In a local freely falling frame (Riemann normal coordinates), the non-gravitational laws of physics reduce to the familiar laws of special relativity.

General relativity satisfies not only EEP but also the **Strong Equivalence Principle**. Because gravitational binding energy contributes to gravitational mass in general relativity, "gravity itself gravitates"—making Einstein's field equations intrinsically non-linear.

1.2 Elements of differential geometry

In flat spacetime, geometry is deceptively simple: a single Cartesian coordinate system can cover the entire manifold, and a vector can be visualized as an arrow connecting an origin to an arbitrary point. On a curved manifold such as a sphere, however, this simple picture breaks down completely. No single coordinate system can cover a sphere without coordinate singularities or multi-valuedness, there is no preferred origin, and straight arrows connecting points on the surface would slice through the interior space rather than remaining on the manifold itself. To formulate the laws of physics on curved spacetime without relying on an external embedding space, we must construct the intrinsic machinery of differential geometry.

1.2.1 Manifolds and tensors

Manifolds

Recall that an open ball $B_r(\mathbf{y})$ of radius r centered at $\mathbf{y} \in \mathbb{R}^n$ is the set of points $\mathbf{x} \in \mathbb{R}^n$ satisfying $\|\mathbf{x} - \mathbf{y}\| < r$. An open set in $\mathbb{R}^n$ is any union of open balls.

A **differentiable manifold** $\mathcal{M}$ of dimension n is a topological space that locally resembles flat Euclidean space $\mathbb{R}^n$, smoothly stitched together from local pieces. Formally, $\mathcal{M}$ is endowed with a collection of open sets $\{O_i\}$ satisfying three precise criteria:

1. **Open Cover:** The union of the open sets covers the entire manifold: $\bigcup_i O_i = \mathcal{M}$. Every point $p \in \mathcal{M}$ belongs to at least one set O_i.

2. **Coordinate Charts (Homeomorphisms):** For each open set O_i, there exists an invertible, continuous map (a homeomorphism) $\psi_i : O_i \to U_i \subset \mathbb{R}^n$ mapping O_i onto an open subset U_i of $\mathbb{R}^n$. The map assigns to each point $p \in O_i$ an n-tuple of real numbers $(x^1, x^2, \ldots, x^n)$, called the **coordinates** of p in the chart (O_i, ψ_i).

These maps allow one to deform the points on the manifold such that they fit on $\mathbb{R}^n$. The key point being that geometry of $\mathcal{M}$ and geometry of $\mathbb{R}^n$ are not identical, however the information regarding the geometry of $\mathcal{M}$ is not lost but encoded in the metric tensor components written in that coordinate system.

3. **Smooth Overlap Maps (Differentiability):** If two open coordinate neighborhoods O_i and O_j overlap ($O_i \cap O_j \neq \emptyset$), a point p in the intersection has coordinates in both charts. The transition map

$$\psi_j \circ \psi_i^{-1} : \psi_i(O_i \cap O_j) \subset U_i \longrightarrow \psi_j(O_i \cap O_j) \subset U_j \tag{1.2.1}$$

is an infinitely differentiable ($\mathcal{C}^\infty$) function from $\mathbb{R}^n$ to $\mathbb{R}^n$.

A complete collection of smoothly overlapping charts $\{(O_i, \psi_i)\}$ that covers $\mathcal{M}$ is called an **atlas** (Figure 1.3). Familiar examples include:

- **The Euclidean plane $\mathbb{R}^2$:** Can be covered by a single chart (trivial atlas).
- **The 2-sphere S^2:** Cannot be covered by a single non-singular chart; it requires at least two charts (for example, via stereographic projections from the North and South poles).

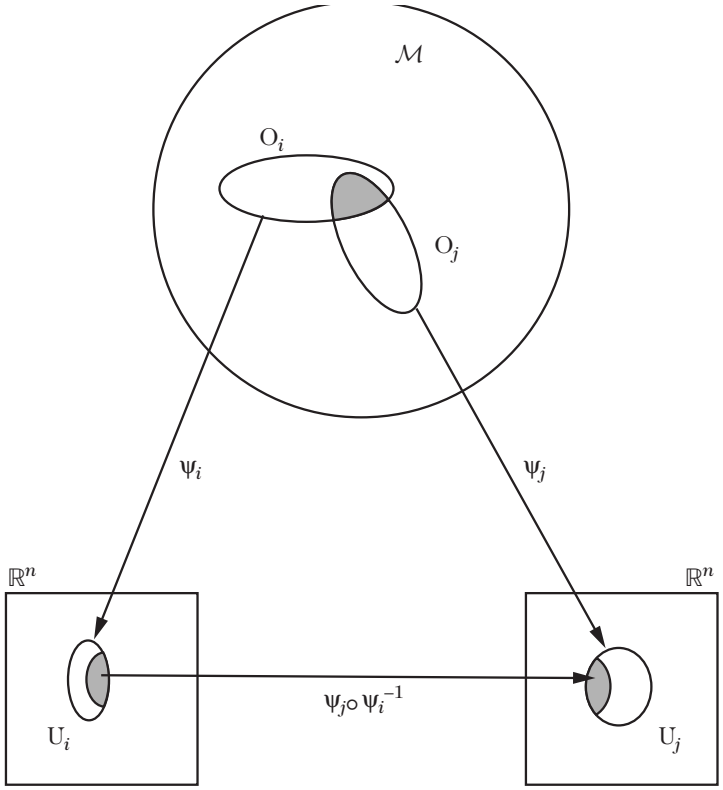


Figure 1.3: Illustration of the relationship between different charts of the same manifold. Locally, every small neighborhood can be mapped homeomorphically to an open subset of $\mathbb{R}^n$. Globally, this is accomplished by an atlas of overlapping charts whose transition functions $\psi_j \circ \psi_i^{-1}$ are smooth.

Vectors

In elementary physics, a vector is typically conceived as a directed line segment pointing from one point to another in Euclidean space. On a curved manifold, this simple definition fails

immediately: a straight line connecting two cities on Earth would cut directly through the planet's mantle, leaving the two-dimensional surface entirely. Furthermore, on an arbitrary manifold, points cannot be added or subtracted as if they were elements of a vector space.

To define tangent vectors purely in terms of intrinsic properties of the manifold, without relying on an ambient embedding space, modern differential geometry defines a vector by its operational action: a tangent vector specifies a direction and an instantaneous rate of change of smooth scalar functions along a curve.

Let $\mathcal{F} \equiv \mathcal{C}^\infty(\mathcal{M})$ denote the set of all infinitely differentiable real-valued scalar fields $f : \mathcal{M} \to \mathbb{R}$. A **tangent vector y** at a point $p \in \mathcal{M}$ is defined as a linear mapping $y : \mathcal{F} \to \mathbb{R}$ that satisfies the **Leibniz product rule**:

1. **Linearity:** $y(af + bg) = a\,y(f) + b\,y(g)$ for all $f, g \in \mathcal{F}$ and all $a, b \in \mathbb{R}$.
2. **Leibniz Rule (Product Rule):** $y(fg) = f(p)\,y(g) + g(p)\,y(f)$.

The set of all tangent vectors at point p forms an n-dimensional real vector space $T_p\mathcal{M}$ (or V_p), called the **tangent space** at p.

Given a coordinate system $\{x^\mu\}$, a natural **coordinate basis** $\{\partial_\mu\}$ for $T_p\mathcal{M}$ is provided by the partial derivative operators along coordinate lines:

$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu}, \qquad \partial_\mu(f) \equiv \left.\frac{\partial f}{\partial x^\mu}\right|_p . \tag{1.2.2}$$

Any tangent vector $\mathbf{y} \in T_p\mathcal{M}$ is uniquely expanded in this basis as:

$$y = y^\mu \partial_\mu = y^\mu \frac{\partial}{\partial x^\mu}, \tag{1.2.3}$$

where the coefficients y^μ are the **contravariant components** of the vector. For example, on a 2-sphere, the tangent space at any point is a two-dimensional plane spanned by the basis vectors ∂_θ and ∂_ϕ (Figure 1.4).

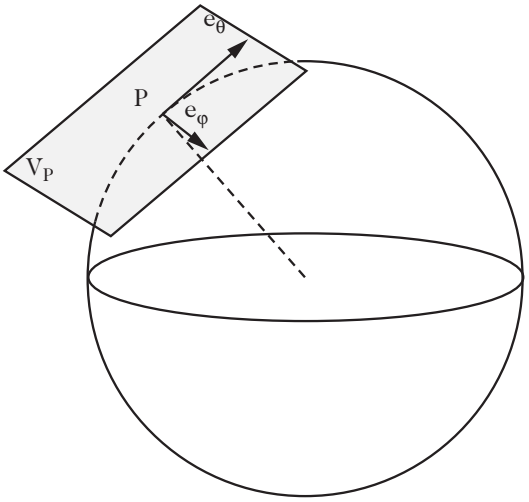
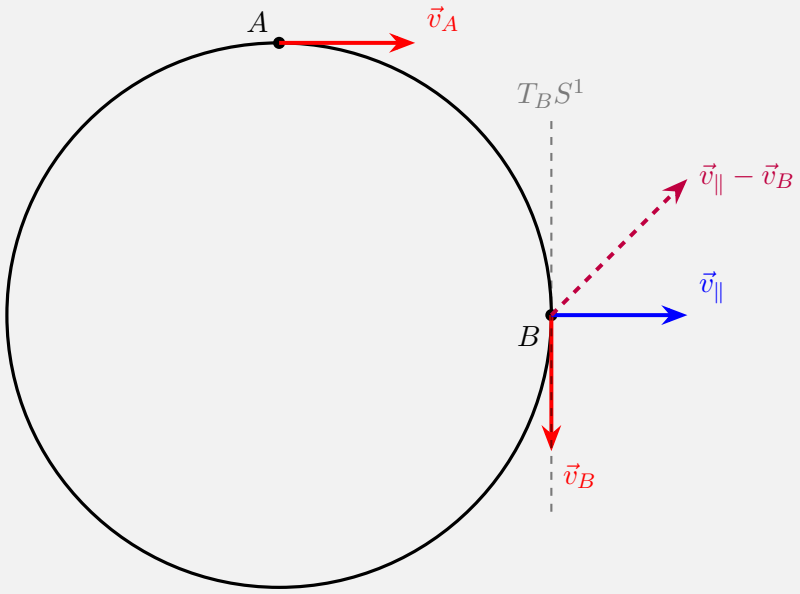


Figure 1.4: The tangent space T_pS^2 to a two-dimensional sphere at a point p is a flat tangent plane whose coordinate basis vectors are $e_\theta = \partial_\theta$ and $e_\phi = \partial_\phi$.

Can We Borrow a Global Vector Space by Embedding?

To add or subtract vectors that reside at different points on a manifold, and thus live in distinct tangent spaces and as such, one requires a map to transport vectors from one space to another. Because diffeomorphism symmetry prevents any unique, canonical identification between different tangent spaces, no such map exists globally by default. However, specifying a path or introducing an additional field (such as a metric or a vector field) provides a prescription for constructing a transport map. This path- and structure-dependent choice is precisely why multiple notions of differentiation (e.g., the covariant derivative, the Lie derivative) exist on curved spaces.

The most naive attempt to bypass this intrinsic structure would be to embed the manifold S^1 into flat Euclidean space $\mathbb{R}^2$ and borrow the global vector space structure of the ambient space. Yet, this approach immediately fails: even though $\vec{v}_A \in T_A S^1$ and $\vec{v}_B \in T_B S^1$ can both be viewed as vectors in $\mathbb{R}^2$, their difference computed in $\mathbb{R}^2$ is not a tangent vector to S^1.



As illustrated above, transporting $\vec{v}_A$ flatly in $\mathbb{R}^2$ yields $\vec{v}_\parallel$, while intrinsic parallel transport along S^1 yields $\vec{v}_B$. The ambient difference $\vec{v}_\parallel - \vec{v}_B$ points outward into $\mathbb{R}^2$, demonstrating that $\vec{v}_\parallel - \vec{v}_B \notin T_B S^1$.

Tangent vector to a curve and transformation law. If a curve on the manifold is parameterized by $x^\mu(\tau)$, its tangent velocity vector $\mathbf{t} \in T_{x(\tau)}\mathcal{M}$ acts on any smooth scalar function $f \in \mathcal{F}$ via the multi-variable chain rule:

$$t(f) = \frac{\mathrm{d}}{\mathrm{d}\tau} f\left[x^\mu(\tau)\right] = \frac{\partial f}{\partial x^\mu}\frac{\mathrm{d}x^\mu(\tau)}{\mathrm{d}\tau} = t^\mu \partial_\mu f, \tag{1.2.4}$$

identifying the vector components as parameter derivatives: $t^\mu = \mathrm{d}x^\mu(\tau)/\,\mathrm{d}\tau$.

Under a smooth change of coordinates $x^\mu \to x'^\mu(x^\nu)$ with non-vanishing Jacobian determinant $\det(\partial x'^\mu/\partial x^\nu) \neq 0$, the chain rule relates the basis partial derivatives:

$$\partial_\mu \longrightarrow \partial'_\mu = \frac{\partial x^\nu}{\partial x'^\mu}\partial_\nu. \tag{1.2.5}$$

Because the physical vector $\mathbf{y} = y^\mu \partial_\mu = y'^\mu \partial'_\mu$ is an invariant geometric object, its components

must transform with the inverse Jacobian matrix:

$$y^\mu \longrightarrow y'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} y^\nu. \tag{1.2.6}$$

This defines the **contravariant transformation law**.

Tensors

Tangent vectors provide directions along a manifold. To measure them—to compute scalar products, evaluate fluxes, and quantify rates of change—we need linear measurement machines that take vectors as inputs and return real numbers.

Dual vectors (1-forms / covariant vectors). Given the tangent space $T_p\mathcal{M}$, we construct its **dual space** $T_p^*\mathcal{M}$ (the **cotangent space**), defined as the vector space of all linear maps from $T_p\mathcal{M}$ to the real numbers $\mathbb{R}$:

$$\omega : T_p\mathcal{M} \longrightarrow \mathbb{R}, \qquad \omega(a\mathbf{u} + b\mathbf{v}) = a\,\omega(\mathbf{u}) + b\,\omega(\mathbf{v}). \tag{1.2.7}$$

The elements of $T_p^*\mathcal{M}$ are called **1-forms** (or covariant vectors, dual vectors).

A natural dual basis for $T_p^*\mathcal{M}$, denoted $\{\mathrm{d}x^\nu\}$, is defined by its action on the coordinate basis vectors $\{\partial_\mu\}$:

$$\mathrm{d}x^\nu\left(\partial_\mu\right) \equiv \delta^\nu_\mu. \tag{1.2.8}$$

The basis 1-form $\mathrm{d}x^\nu$ simply extracts the ν-th component of any vector: $\mathrm{d}x^\nu\left(\mathbf{y}\right) = \mathrm{d}x^\nu\left(y^\mu\partial_\mu\right) = y^\mu\delta^\nu_\mu = y^\nu$.

Vectors as arrows versus 1-forms as surfaces. It is very common in introductory physics to view vectors and 1-forms as mere typographical variations of the same mathematical entity—one written with an upper index and the other with a lower index. Because the metric tensor allows us to lower indices via $V_\mu = g_{\mu\nu}V^\nu$ and raise them via $V^\mu = g^{\mu\nu}V_\nu$, one might wonder why we insist on maintaining such a fundamental distinction between contravariant vectors and covariant 1-forms.

The illusion that vectors and 1-forms are interchangeable arises entirely from our intuition in flat Euclidean space equipped with Cartesian coordinates. In that special setting, the metric components are simply δ_{ij}, meaning that the numerical components of a vector and its dual 1-form are identical ($V_i = V^i$). This numerical coincidence obscures a profound geometric difference. On a general differentiable manifold, vectors and 1-forms represent two completely different geometric concepts with opposite transformation laws under coordinate changes:

- **A contravariant vector (v^μ, upper index)** represents a directed displacement or velocity arrow tangent to a curve (such as the four-velocity $u^\mu = \mathrm{d}x^\mu\,/\,\mathrm{d}\tau$). It transforms with the forward Jacobian matrix $\frac{\partial x'^\mu}{\partial x^\nu}$.
- **A covariant 1-form (ω_μ, lower index)** represents a family of parallel level surfaces or a gradient (such as the exterior differential of a scalar field, $\mathrm{d}f = \partial_\mu f\,\mathrm{d}x^\mu$). Geometrically, evaluating a 1-form on a vector, $\mathrm{d}f\left(\mathbf{v}\right) = v^\mu\partial_\mu f$, counts how many contour surfaces the trajectory pierces per unit parameter time. Under coordinate changes, basis 1-forms and components transform as:

$$\mathrm{d}x^\mu \longrightarrow \mathrm{d}x'^\mu = \frac{\partial x'^\mu}{\partial x^\nu}\,\mathrm{d}x^\nu \implies \omega_\mu \longrightarrow \omega'_\mu = \frac{\partial x^\nu}{\partial x'^\mu}\omega_\nu. \tag{1.2.9}$$

Crucially, on a bare differentiable manifold where no metric has yet been defined, there is no canonical identification between tangent vectors and cotangent 1-forms. The metric tensor provides an isomorphism between $T_p\mathcal{M}$ and $T_p^*\mathcal{M}$, but their geometric roles—directional arrows versus contour slicers—remain fundamentally distinct.

General rank-(r,q) tensors. A **tensor T of rank** (r,q) is a multilinear map that takes r dual 1-forms and q tangent vectors and maps them to a real scalar:

$$T : \underbrace{T_p^*\mathcal{M} \times \cdots \times T_p^*\mathcal{M}}_{r \text{ copies}} \times \underbrace{T_p\mathcal{M} \times \cdots \times T_p\mathcal{M}}_{q \text{ copies}} \longrightarrow \mathbb{R}. \tag{1.2.10}$$

Expanded in the tensor product basis of vectors and 1-forms,

$$T = T^{\mu_1\ldots\mu_r}_{\nu_1\ldots\nu_q}\partial_{\mu_1} \otimes \cdots \otimes \partial_{\mu_r} \otimes \mathrm{d}x^{\nu_1} \otimes \cdots \otimes \mathrm{d}x^{\nu_q}\,, \tag{1.2.11}$$

where T is r times contravariant and q times covariant. Under a coordinate transformation $x \to x'$, its components transform as:

$$T'^{\mu_1\ldots\mu_r}_{\nu_1\ldots\nu_q} = \left(\frac{\partial x'^{\mu_1}}{\partial x^{\alpha_1}} \cdots \frac{\partial x'^{\mu_r}}{\partial x^{\alpha_r}}\right)\left(\frac{\partial x^{\beta_1}}{\partial x'^{\nu_1}} \cdots \frac{\partial x^{\beta_q}}{\partial x'^{\nu_q}}\right) T^{\alpha_1\ldots\alpha_r}_{\beta_1\ldots\beta_q}. \tag{1.2.12}$$

A **tensor field** is an assignment of a rank-(r,q) tensor smoothly to each point of the manifold $\mathcal{M}$.

The metric tensor. To measure physical lengths, angles, and proper times on the manifold, we introduce the **metric tensor** g, which is a symmetric, non-degenerate tensor field of rank $(0,2)$. The spacetime interval between two infinitesimally separated events is:

$$\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu\,. \tag{1.2.13}$$

Because $g_{\mu\nu} = g_{\nu\mu}$, the metric in 4 dimensions possesses $\frac{4\times5}{2} = \mathbf{10}$ independent components. It transforms as a rank-$(0,2)$ tensor:

$$g'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\beta}{\partial x'^\nu} g_{\alpha\beta}. \tag{1.2.14}$$

Taking the determinant of Equation (1.2.14), the metric determinant $g \equiv \det(g_{\mu\nu})$ transforms as a **scalar density of weight** -2:

$$g' = \left[\det\left(\frac{\partial x^\alpha}{\partial x'^\mu}\right)\right]^2 g. \tag{1.2.15}$$

Tensor calculus

Tensor algebra consists of four fundamental operations:

1. **Linear Combination:** Tensors of identical rank (r,q) can be added linearly: $(S+T)^{\mu_1\ldots\mu_r}_{\nu_1\ldots\nu_q} = S^{\mu_1\ldots\mu_r}_{\nu_1\ldots\nu_q} + T^{\mu_1\ldots\mu_r}_{\nu_1\ldots\nu_q}$.

2. **Outer (Tensor) Product:** The tensor product of a rank-(r,s) tensor S with a rank-(p,q) tensor T yields a tensor of rank $(r+p, s+q)$:

$$(S \otimes T)^{\mu_1\ldots\mu_r\alpha_1\ldots\alpha_p}_{\nu_1\ldots\nu_s\beta_1\ldots\beta_q} = S^{\mu_1\ldots\mu_r}_{\nu_1\ldots\nu_s} T^{\alpha_1\ldots\alpha_p}_{\beta_1\ldots\beta_q}. \tag{1.2.16}$$

3. **Contraction:** Setting one contravariant (upper) and one covariant (lower) index equal sums over that index, reducing the tensor rank from (r,q) to $(r-1,q-1)$:

$$R^{\mu_2\ldots\mu_r}_{\nu_2\ldots\nu_q} \equiv S^{\alpha\mu_2\ldots\mu_r}_{\alpha\nu_2\ldots\nu_q}. \tag{1.2.17}$$

   Complete contraction of a rank-$(1,1)$ tensor yields an invariant scalar: $T^\mu{}_\mu = \text{scalar}$.

4. **Raising and Lowering Indices:** The metric $g_{\mu\nu}$ and its inverse $g^{\mu\nu}$ (defined by $g^{\mu\alpha}g_{\alpha\nu} = \delta^\mu_\nu$) act as index-raising and index-lowering operators, establishing the canonical isomorphism between tangent vectors and dual 1-forms:

$$V_\mu = g_{\mu\nu}V^\nu, \qquad V^\mu = g^{\mu\nu}V_\nu. \tag{1.2.18}$$

Symmetries of rank-2 tensors. Any rank-2 tensor $T_{\mu\nu}$ can be decomposed uniquely into its symmetric part $T_{(\mu\nu)}$ and antisymmetric part $T_{[\mu\nu]}$:

$$T_{(\mu\nu)} \equiv \frac{1}{2}\left(T_{\mu\nu} + T_{\nu\mu}\right), \qquad T_{[\mu\nu]} \equiv \frac{1}{2}\left(T_{\mu\nu} - T_{\nu\mu}\right). \tag{1.2.19}$$

In four dimensions, any rank-2 tensor decomposes into an antisymmetric piece, a traceless symmetric piece, and an isotropic trace:

$$T_{\mu\nu} = T_{[\mu\nu]} + \left(T_{(\mu\nu)} - \frac{1}{4}T^{\alpha}{}_{\alpha}g_{\mu\nu}\right) + \frac{1}{4}T^{\alpha}{}_{\alpha}g_{\mu\nu}. \tag{1.2.20}$$

The Levi-Civita pseudotensor. The completely antisymmetric **Levi-Civita tensor** of rank 4 is defined by:

$$\epsilon_{\mu\nu\rho\sigma} \equiv \sqrt{-g}\,[\mu\nu\rho\sigma], \qquad \epsilon^{\mu\nu\rho\sigma} = -\frac{1}{\sqrt{-g}}\,[\mu\nu\rho\sigma], \tag{1.2.21}$$

where $[\mu\nu\rho\sigma]$ is the permutation symbol ($+1$ for even permutations of 0123, -1 for odd, 0 if any indices repeat). The factor of $\sqrt{-g}$ ensures that $\epsilon_{\mu\nu\rho\sigma}$ transforms as a genuine tensor under coordinate transformations.

The contraction identities of the Levi-Civita tensor are indispensable in relativistic field theory:

$$\epsilon^{\alpha\beta\delta\gamma}\epsilon_{\alpha\lambda\nu\mu} = -\delta^{\beta}_{\lambda}\delta^{\delta}_{\nu}\delta^{\gamma}_{\mu} - \delta^{\beta}_{\nu}\delta^{\delta}_{\mu}\delta^{\gamma}_{\lambda} - \delta^{\beta}_{\mu}\delta^{\delta}_{\lambda}\delta^{\gamma}_{\nu} + \delta^{\beta}_{\lambda}\delta^{\delta}_{\mu}\delta^{\gamma}_{\nu} + \delta^{\beta}_{\mu}\delta^{\delta}_{\nu}\delta^{\gamma}_{\lambda} + \delta^{\beta}_{\nu}\delta^{\delta}_{\lambda}\delta^{\gamma}_{\mu}, \tag{1.2.22}$$

$$\epsilon^{\alpha\beta\delta\gamma}\epsilon_{\alpha\beta\nu\mu} = -2\left(\delta^{\delta}_{\nu}\delta^{\gamma}_{\mu} - \delta^{\delta}_{\mu}\delta^{\gamma}_{\nu}\right), \tag{1.2.23}$$

$$\epsilon^{\alpha\beta\delta\gamma}\epsilon_{\alpha\beta\delta\mu} = -6\delta^{\gamma}_{\mu}, \tag{1.2.24}$$

$$\epsilon^{\alpha\beta\delta\gamma}\epsilon_{\alpha\beta\delta\gamma} = -24. \tag{1.2.25}$$

1.2.2 Geodesic equations and Christoffel symbols

In flat Euclidean space, an unaccelerated particle travels in a straight line. Straight lines possess two intimately linked geometric characteristics: they are curves of stationary length between two endpoints, and their tangent velocity vector maintains a constant direction along the path. On a curved spacetime manifold, these concepts generalize to the notion of a **geodesic**.

To determine the equations governing such paths, one naturally seeks to extremize the spacetime interval. For a massive particle, the physical action proportional to the proper time is $S = -mc\int \mathrm{d}\tau = -mc\int \sqrt{-g_{\mu\nu}\dot{x}^{\mu}\dot{x}^{\nu}}\,\mathrm{d}\lambda$. While geometrically transparent, the presence of the square root introduces a practical and conceptual obstacle: for massless particles such as photons, the interval along the path vanishes identically ($\mathrm{d}s = 0$), rendering the Lagrangian singular. To circumvent this difficulty and treat massive and massless trajectories within a unified framework, we instead extremize the **quadratic action**:

$$S = \int L\,\mathrm{d}\lambda\,, \qquad L = \frac{1}{2}g_{\mu\nu}(x^{\alpha})\dot{x}^{\mu}\dot{x}^{\nu}, \tag{1.2.26}$$

where $\dot{x}^{\mu} \equiv \mathrm{d}x^{\mu}\,/\,\mathrm{d}\lambda$. The Euler–Lagrange equations obtained from this quadratic Lagrangian yield trajectories identical to those of the square-root action for timelike curves, remain perfectly well-behaved for null curves, and uniquely pick out an **affine parameter** λ along the trajectory.

We now compute the variation of the action. The Euler–Lagrange equations for the coordinate components $x^{\alpha}(\lambda)$ are:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\partial L}{\partial \dot{x}^{\alpha}}\right) - \frac{\partial L}{\partial x^{\alpha}} = 0. \tag{1.2.27}$$

Differentiating the Lagrangian with respect to the coordinates yields:

$$\frac{\partial L}{\partial x^{\alpha}} = \frac{1}{2}\partial_{\alpha}g_{\mu\nu}\dot{x}^{\mu}\dot{x}^{\nu}, \tag{1.2.28}$$

while differentiating with respect to the velocities gives $\frac{\partial L}{\partial \dot{x}^\alpha} = g_{\mu\alpha}\dot{x}^\mu$. Taking the total derivative with respect to the affine parameter λ, and applying the chain rule to the position-dependent metric, we find:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\partial L}{\partial \dot{x}^\alpha}\right) = g_{\mu\alpha}\ddot{x}^\mu + \partial_\beta g_{\mu\alpha}\dot{x}^\beta\dot{x}^\mu. \tag{1.2.29}$$

Because the contraction $\dot{x}^\beta\dot{x}^\mu$ is symmetric under the exchange of β and μ, we can rewrite the second term symmetrically as $\frac{1}{2}(\partial_\beta g_{\mu\alpha} + \partial_\mu g_{\beta\alpha})\dot{x}^\beta\dot{x}^\mu$. Combining terms in the Euler–Lagrange equation then produces:

$$g_{\mu\alpha}\ddot{x}^\mu + \frac{1}{2}\left(\partial_\beta g_{\mu\alpha} + \partial_\mu g_{\beta\alpha} - \partial_\alpha g_{\mu\beta}\right)\dot{x}^\mu\dot{x}^\beta = 0. \tag{1.2.30}$$

The combination of metric derivatives appearing in parentheses is of paramount importance throughout general relativity. We define it as the **Christoffel symbol of the first kind**:

$$\Gamma_{\lambda\mu\nu} \equiv \frac{1}{2}\left(\partial_\mu g_{\lambda\nu} + \partial_\nu g_{\mu\lambda} - \partial_\lambda g_{\mu\nu}\right). \tag{1.2.31}$$

One can always embed the geometry in higher dimension where the cartesian like coordinate exists. In such a coordinate basis, perform a coordinate transformation such that our embededd manifold becomes a coordinate surface and as such we can express the induced coordinates in terms of the embedding space coordinate which allows us to calculate how the basis changes from point to point. Since

$$\nabla_\mu \mathbf{e}_\nu = \Gamma^\alpha_{\mu\nu}\mathbf{e}_\alpha$$

allows us to infer connection term from the existence of such coordinate system itself.

Example: Consider a two-dimensional sphere of radius R embedded in three-dimensional Euclidean space. Let (θ, ϕ) be the usual spherical coordinates on the sphere, so that the embedding into $\mathbb{R}^3$ is

$$x = R\sin\theta\cos\phi, \qquad y = R\sin\theta\sin\phi, \qquad z = R\cos\theta. \tag{1}$$

Here θ is the polar angle and ϕ is the azimuthal angle.
Let $\{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$ denote the Cartesian basis vectors of the ambient $\mathbb{R}^3$. The coordinate basis vectors on the sphere are obtained by differentiating the embedding:

$$\mathbf{e}_\theta = \frac{\partial}{\partial\theta} = R\cos\theta\cos\phi\,\mathbf{e}_x + R\cos\theta\sin\phi\,\mathbf{e}_y - R\sin\theta\,\mathbf{e}_z, \tag{2}$$

$$\mathbf{e}_\phi = \frac{\partial}{\partial\phi} = -R\sin\theta\sin\phi\,\mathbf{e}_x + R\sin\theta\cos\phi\,\mathbf{e}_y. \tag{3}$$

It is also useful to introduce the outward-pointing unit normal

$$\mathbf{e}_r = \sin\theta\cos\phi\,\mathbf{e}_x + \sin\theta\sin\phi\,\mathbf{e}_y + \cos\theta\,\mathbf{e}_z. \tag{4}$$

Notice that $\mathbf{e}_r$ is normal to the sphere, whereas $\mathbf{e}_\theta$ and $\mathbf{e}_\phi$ are tangent to it.
Since the Cartesian basis is constant,

$$\partial_\mu \mathbf{e}_x = \partial_\mu \mathbf{e}_y = \partial_\mu \mathbf{e}_z = 0. \tag{5}$$

Therefore, when differentiating the coordinate basis vectors, we simply differentiate their components.

First,

$$\partial_\theta \mathbf{e}_\theta = -R\sin\theta\cos\phi\,\mathbf{e}_x - R\sin\theta\sin\phi\,\mathbf{e}_y - R\cos\theta\,\mathbf{e}_z \tag{6}$$
$$= -R\mathbf{e}_r. \tag{7}$$

Similarly,

$$\partial_\theta \mathbf{e}_\phi = -R\cos\theta\sin\phi\,\mathbf{e}_x + R\cos\theta\cos\phi\,\mathbf{e}_y \tag{8}$$
$$= \frac{\cos\theta}{\sin\theta}\mathbf{e}_\phi = \cot\theta\,\mathbf{e}_\phi. \tag{9}$$

For the derivatives with respect to ϕ, we find

$$\partial_\phi \mathbf{e}_\theta = -R\cos\theta\sin\phi\,\mathbf{e}_x + R\cos\theta\cos\phi\,\mathbf{e}_y \tag{10}$$
$$= \cot\theta\,\mathbf{e}_\phi, \tag{11}$$

while

$$\partial_\phi \mathbf{e}_\phi = -R\sin\theta\cos\phi\,\mathbf{e}_x - R\sin\theta\sin\phi\,\mathbf{e}_y. \tag{12}$$

To express this in the spherical basis, we use

$$\partial_\phi \mathbf{e}_\phi = -R\sin^2\theta\,\mathbf{e}_r - \sin\theta\cos\theta\,\mathbf{e}_\theta. \tag{13}$$

Thus the derivatives of the coordinate basis vectors are

$$\partial_\theta \mathbf{e}_\theta = -R\mathbf{e}_r, \qquad \partial_\theta \mathbf{e}_\phi = \cot\theta\,\mathbf{e}_\phi, \tag{14}$$
$$\partial_\phi \mathbf{e}_\theta = \cot\theta\,\mathbf{e}_\phi, \qquad \partial_\phi \mathbf{e}_\phi = -R\sin^2\theta\,\mathbf{e}_r - \sin\theta\cos\theta\,\mathbf{e}_\theta. \tag{15}$$

The terms proportional to $\mathbf{e}_r$ are normal to the sphere and therefore describe the extrinsic curvature of the embedded surface. For the intrinsic connection on the sphere, we retain only the components tangent to the sphere. Thus,

$$\nabla_\theta \mathbf{e}_\theta = 0, \qquad \nabla_\theta \mathbf{e}_\phi = \cot\theta\,\mathbf{e}_\phi, \tag{16}$$
$$\nabla_\phi \mathbf{e}_\theta = \cot\theta\,\mathbf{e}_\phi, \qquad \nabla_\phi \mathbf{e}_\phi = -\sin\theta\cos\theta\,\mathbf{e}_\theta. \tag{17}$$

Comparing these results with

$$\nabla_\mu \mathbf{e}_\nu = \Gamma^\alpha_{\mu\nu}\mathbf{e}_\alpha, \tag{18}$$

we obtain

$$\boxed{\Gamma^\phi_{\theta\phi} = \Gamma^\phi_{\phi\theta} = \cot\theta}, \qquad \boxed{\Gamma^\theta_{\phi\phi} = -\sin\theta\cos\theta}. \tag{19}$$

All other Christoffel symbols vanish.
Indeed, the induced metric on the sphere is

$$\mathrm{d}s^2 = R^2\mathrm{d}\theta^2 + R^2\sin^2\theta\,\mathrm{d}\phi^2, \tag{20}$$

and the Christoffel symbols obtained from this metric are precisely

$$\Gamma^\phi_{\theta\phi} = \Gamma^\phi_{\phi\theta} = \cot\theta, \qquad \Gamma^\theta_{\phi\phi} = -\sin\theta\cos\theta. \tag{21}$$

Contracting Equation (1.2.31) with the inverse metric $g^{\alpha\lambda}$ defines the **Christoffel symbols of the second kind** (the connection coefficients):

$$\Gamma^{\alpha}_{\mu\nu} \equiv g^{\alpha\lambda}\Gamma_{\lambda\mu\nu} = \frac{1}{2}g^{\alpha\lambda}\left(\partial_\mu g_{\lambda\nu} + \partial_\nu g_{\mu\lambda} - \partial_\lambda g_{\mu\nu}\right). \tag{1.2.32}$$

The equation of motion for a freely falling particle then takes the universal form of the **geodesic equation**:

$$\ddot{x}^\mu + \Gamma^{\mu}_{\nu\lambda}\dot{x}^\nu\dot{x}^\lambda = 0. \tag{1.2.33}$$

Because the metric tensor is symmetric ($g_{\mu\nu} = g_{\nu\mu}$), the Christoffel symbols are automatically symmetric in their lower two indices:

$$\Gamma^{\alpha}_{\mu\nu} = \Gamma^{\alpha}_{\nu\mu}, \qquad \Gamma_{\lambda\mu\nu} = \Gamma_{\lambda\nu\mu}. \tag{1.2.34}$$

In four spacetime dimensions, there are $4 \times \frac{4\times5}{2} = \mathbf{40}$ independent Christoffel symbols. Inverting Equation (1.2.32) relates the partial derivatives of the metric directly to the connection:

$$\partial_\alpha g_{\mu\nu} = g_{\lambda\mu}\Gamma^{\lambda}_{\alpha\nu} + g_{\lambda\nu}\Gamma^{\lambda}_{\alpha\mu}. \tag{1.2.35}$$

The connection carries indices, but is not a tensor. A central question that confronts any student of relativity is whether the Christoffel symbols constitute a tensor. With one upper index and two lower indices, the symbol $\Gamma^{\alpha}_{\mu\nu}$ bears the outward appearance of a rank-$(1,2)$ tensor. However, index placement alone does not make an object a tensor; genuine tensorial status requires that the components transform multilinearly under coordinate changes according to the Jacobian matrices.

If we transform coordinates $x^\mu \to x'^\mu(x^\nu)$ and apply the chain rule directly to the definition of $\Gamma^{\alpha}_{\mu\nu}$ in Equation (1.2.32), a direct calculation reveals that the components transform as:

$$\Gamma'^{\alpha}_{\mu\nu} = \frac{\partial x'^\alpha}{\partial x^\beta}\frac{\partial x^\sigma}{\partial x'^\mu}\frac{\partial x^\rho}{\partial x'^\nu}\Gamma^{\beta}_{\sigma\rho} - \frac{\partial^2 x'^\alpha}{\partial x^\sigma \partial x^\rho}\frac{\partial x^\sigma}{\partial x'^\mu}\frac{\partial x^\rho}{\partial x'^\nu}. \tag{1.2.36}$$

The first term on the right-hand side has the expected multilinear structure of a rank-$(1,2)$ tensor, but the second term is inhomogeneous, involving second partial derivatives $\partial^2 x'^\alpha/\partial x^\sigma \partial x^\rho$ of the coordinate transformation. Consequently, the Christoffel symbol is *not* a tensor.

Far from being an unfortunate mathematical defect, the non-tensorial transformation law of the connection is a physical necessity demanded by the Equivalence Principle. Recall that according to the equivalence principle, an observer can always eliminate gravity locally at any chosen spacetime event P by entering a freely falling reference frame (known formally as Riemann normal coordinates). In such a frame, the metric components are locally flat, $g_{\mu\nu}(P) = \eta_{\mu\nu}$, and all first derivatives of the metric vanish, $\partial_\alpha g_{\mu\nu}(P) = 0$. By Equation (1.2.32), this implies that the connection coefficients vanish at that point:

$$\Gamma^{\alpha}_{\mu\nu}(P) = 0. \tag{1.2.37}$$

Now consider what would happen if $\Gamma^{\alpha}_{\mu\nu}$ were a genuine tensor: if all components of a tensor vanish in one coordinate frame, the homogeneous transformation law would force them to vanish in *every* coordinate frame. If the connection were a tensor, it would be impossible to make it vanish by choosing a freely falling frame unless spacetime were globally flat everywhere! The inhomogeneous second-derivative term in Equation (1.2.36) is precisely what allows an observer to transform away the Christoffel symbols by accelerating, reflecting the physical fact that inertial forces can be eliminated by choosing an appropriate reference frame.

Contraction identities of the connection. Using Jacobi's formula for the derivative of a matrix determinant, $\mathrm{d}g = g g^{\mu\nu}\,\mathrm{d}g_{\mu\nu}$, we obtain:

$$\partial_\alpha g = g g^{\mu\nu}\partial_\alpha g_{\mu\nu}. \tag{1.2.38}$$

Contracting the Christoffel symbol on its upper and one lower index yields the identity:

$$\Gamma^\mu_{\mu\alpha} = \frac{1}{2} g^{\mu\lambda}\partial_\alpha g_{\mu\lambda} = \partial_\alpha \ln\sqrt{-g} = \frac{1}{\sqrt{-g}}\partial_\alpha\sqrt{-g}. \tag{1.2.39}$$

Furthermore, differentiating $g_{\mu\nu}g^{\mu\nu} = 4$ gives:

$$g_{\mu\nu}\partial_\alpha g^{\mu\nu} = -g^{\mu\nu}\partial_\alpha g_{\mu\nu}. \tag{1.2.40}$$

1.2.3 Covariant derivative, parallel transport, and Lie derivative

Covariant derivative

In flat Cartesian coordinates, the partial derivative $\partial_\alpha T^\mu$ of a vector field transforms as a genuine rank-$(1,1)$ tensor. In curved spacetime, however, **the partial derivative of a tensor is not a tensor**.

To understand why ordinary differentiation fails on a curved manifold, we differentiate the transformed vector $T'^\mu = \frac{\partial x'^\mu}{\partial x^\nu}T^\nu$ with respect to the new coordinates x'^α:

$$\frac{\partial T'^\mu}{\partial x'^\alpha} = \frac{\partial x'^\mu}{\partial x^\nu}\frac{\partial x^\sigma}{\partial x'^\alpha}\frac{\partial T^\nu}{\partial x^\sigma} + \frac{\partial^2 x'^\mu}{\partial x^\sigma \partial x^\nu}\frac{\partial x^\sigma}{\partial x'^\alpha}T^\nu. \tag{1.2.41}$$

The second term on the right-hand side contains second derivatives of the coordinate map, explicitly violating the tensorial transformation law.

The geometric reason for this failure is profound. The standard difference quotient,

$$\partial_\alpha T^\mu = \lim_{\Delta x^\alpha \to 0}\frac{T^\mu(x+\Delta x) - T^\mu(x)}{\Delta x^\alpha}, \tag{1.2.42}$$

attempts to subtract two vectors that reside in *two completely different vector spaces*—the tangent space $T_x\mathcal{M}$ at x and the tangent space $T_{x+\Delta x}\mathcal{M}$ at $x+\Delta x$. Because the coordinate basis vectors ∂_μ themselves rotate and change lengths from point to point, the difference between their numerical components mixes the physical variation of the vector field with the purely coordinate-dependent variation of the basis vectors.

To construct a genuine tensorial derivative, we must parallel-transport the vector from $x+\Delta x$ back to x before taking the difference. This correction is provided by the **covariant derivative** ∇_α (often denoted by a semicolon $_{;\alpha}$):

$$\nabla_\alpha T^\mu \equiv \partial_\alpha T^\mu + \Gamma^\mu_{\alpha\nu}T^\nu, \tag{1.2.43}$$

$$\nabla_\alpha T_\mu \equiv \partial_\alpha T_\mu - \Gamma^\nu_{\alpha\mu}T_\nu. \tag{1.2.44}$$

The connection term $\Gamma^\mu_{\alpha\nu}T^\nu$ in Equation (1.2.43) transforms with an inhomogeneous piece that *exactly cancels* the non-tensorial second derivative in Equation (1.2.41), ensuring that $\nabla_\alpha T^\mu$ transforms as a true rank-$(1,1)$ tensor.

For a general tensor of rank (p,q), a Christoffel correction term is added for every contravariant index and subtracted for every covariant index:

$$\nabla_\alpha T^{\mu_1\ldots\mu_p}_{\nu_1\ldots\nu_q} = \partial_\alpha T^{\mu_1\ldots\mu_p}_{\nu_1\ldots\nu_q} + \sum_{i=1}^{p}\Gamma^{\mu_i}_{\alpha\lambda}T^{\mu_1\ldots\lambda\ldots\mu_p}_{\nu_1\ldots\nu_q} - \sum_{j=1}^{q}\Gamma^{\lambda}_{\alpha\nu_j}T^{\mu_1\ldots\mu_p}_{\nu_1\ldots\lambda\ldots\nu_q}. \tag{1.2.45}$$

For a scalar field ϕ, $\nabla_\alpha\phi = \partial_\alpha\phi$ is already a genuine 1-form.

Metric compatibility. Computing the covariant derivative of the metric tensor:

$$\nabla_\alpha g_{\mu\nu} = \partial_\alpha g_{\mu\nu} - \Gamma^\sigma_{\alpha\mu} g_{\sigma\nu} - \Gamma^\sigma_{\alpha\nu} g_{\sigma\mu}. \tag{1.2.46}$$

Substituting Equation (1.2.35), we obtain the foundational property of **metric compatibility**:

$$\nabla_\alpha g_{\mu\nu} = 0, \qquad \nabla_\alpha g^{\mu\nu} = 0. \tag{1.2.47}$$

Metric compatibility guarantees that the metric behaves as a constant under covariant differentiation: it can be pulled through covariant derivatives when raising or lowering indices ($\nabla_\alpha(g_{\mu\nu}T^\nu) = g_{\mu\nu}\nabla_\alpha T^\nu$). Physically, it ensures that lengths of vectors and angles between them are strictly preserved under parallel transport.

The **Levi-Civita connection** is the unique affine connection that is simultaneously torsion-free ($\Gamma^\alpha_{\mu\nu} = \Gamma^\alpha_{\nu\mu}$) and metric-compatible ($\nabla_\alpha g_{\mu\nu} = 0$).

Parallel transport

Along a curve $x^\mu(\tau)$ with tangent four-velocity $u^\mu = \mathrm{d}x^\mu / \mathrm{d}\tau$, the rate of change of a vector T^μ evaluated along the trajectory is the **absolute derivative** (directional covariant derivative) $\mathrm{D}/\mathrm{D}\tau \equiv u^\nu \nabla_\nu$:

$$\frac{\mathrm{D}T^\mu}{\mathrm{D}\tau} \equiv u^\nu \nabla_\nu T^\mu = \frac{\mathrm{d}T^\mu}{\mathrm{d}\tau} + \Gamma^\mu_{\nu\alpha} T^\alpha u^\nu. \tag{1.2.48}$$

A vector T^μ is said to be **parallel transported** along the curve if its absolute derivative vanishes: $\frac{\mathrm{D}T^\mu}{\mathrm{D}\tau} = 0$.

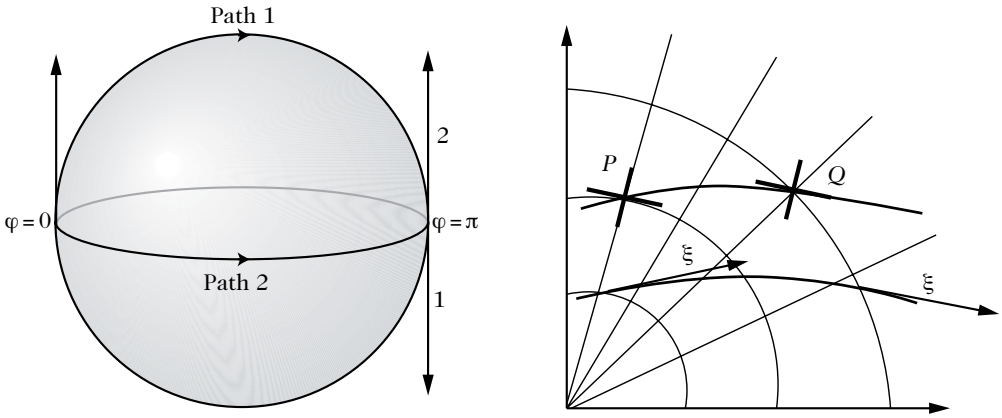


Figure 1.5: (Left): On a two-dimensional sphere, parallel transporting a vector along two different paths (Path 1 along a great circle to the antipodal point vs. Path 2 along the equator) results in two final vectors that point in different directions. Parallel transport in curved space is path-dependent. (Right): The Lie derivative evaluates the distortion of a tensor field by dragging an observer's local frame along the integral flow curves of a vector field ξ^μ.

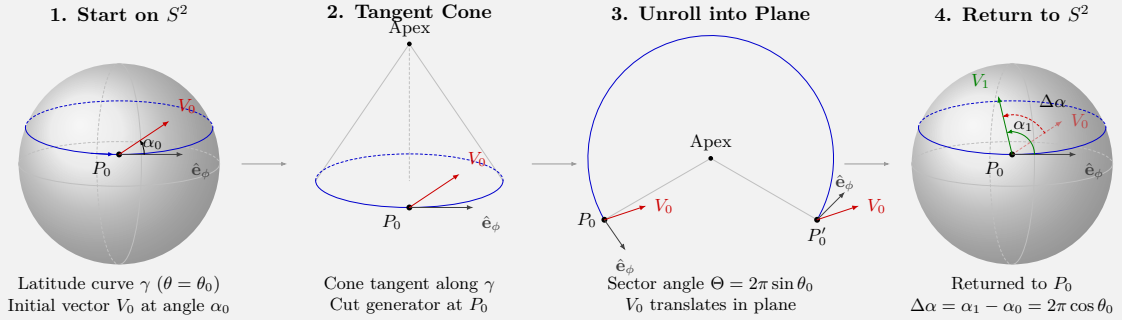
Auto-parallel curves and geodesics. A curve is defined to be **auto-parallel** if it parallel-transports its own tangent vector:

$$\frac{\mathrm{D}u^\mu}{\mathrm{D}\tau} = u^\nu \nabla_\nu u^\mu = 0 \iff \frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{\nu\lambda} \frac{\mathrm{d}x^\nu}{\mathrm{d}\tau} \frac{\mathrm{d}x^\lambda}{\mathrm{d}\tau} = 0. \tag{1.2.49}$$

This is identically the geodesic equation (1.2.33): **geodesics are the auto-parallel paths of spacetime**. In curved spacetime, a freely falling body does not accelerate ($\frac{\mathrm{D}u^\mu}{\mathrm{D}\tau} = 0$); it simply glides along the straightest possible trajectory.

Parallel Transport via embedding and diffeomorphism

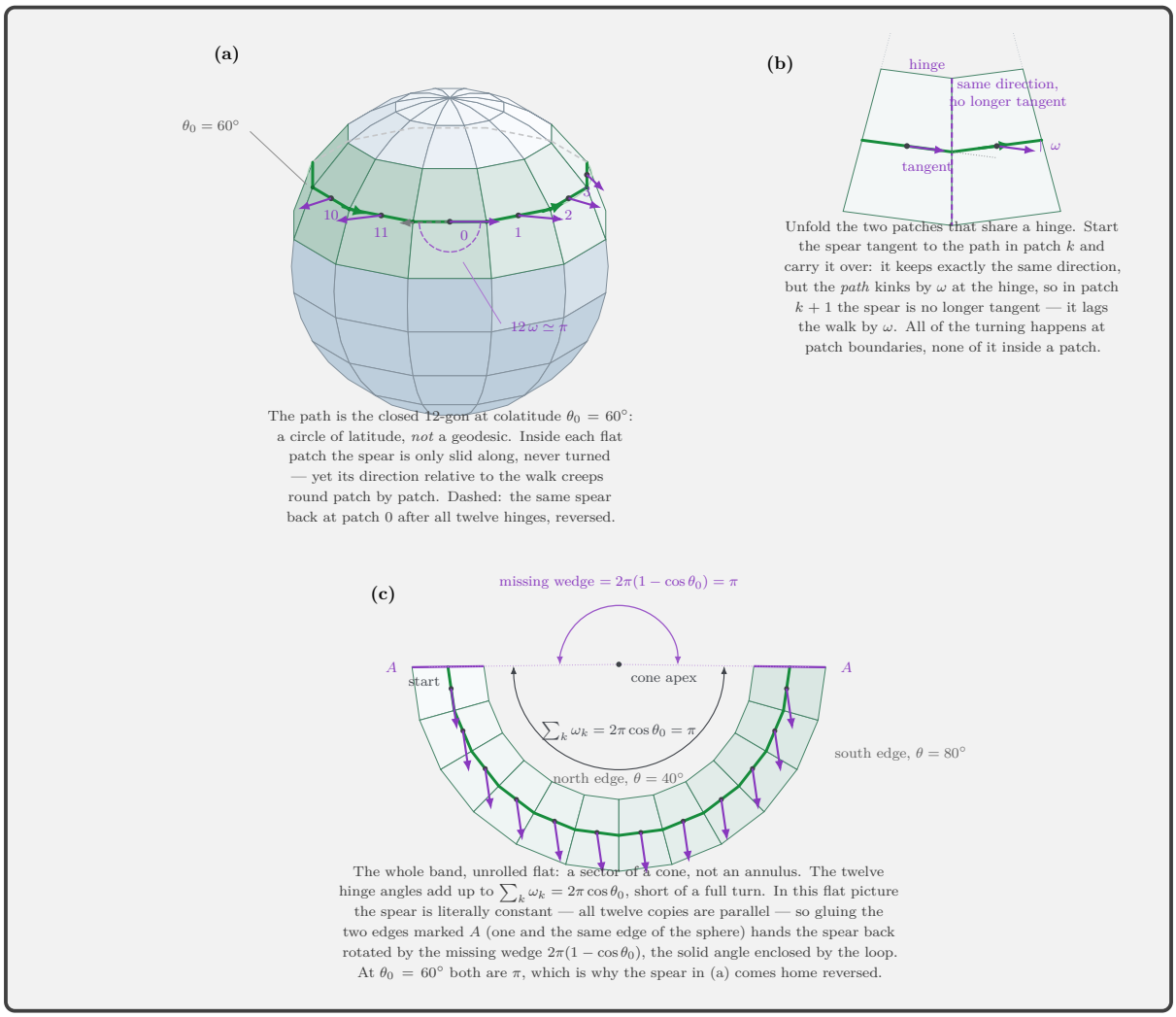
Suppose we have a two-dimensional surface S and a curve γ drawn on it. We want to parallel transport a vector along γ.
Instead of working directly with S, we can sometimes find *another two-dimensional surface* S' that contains the *same curve* γ but is easier to flatten, or *uncurl*.
The important point is that S' should have the same intrinsic geometry along the curve that we care about. Then parallel transport along γ can be understood using this alternative surface.



The reason for doing this is purely practical. *Uncurling a sphere is impossible without distorting its geometry*, whereas surfaces such as a *cylinder or cone can be unrolled into a plane without stretching*. For example, a cylinder can simply be cut along a vertical line and opened into a rectangle, while a cone can be cut along one generator and opened into a circular sector.
We can then perform ordinary parallel transport in the flat, uncurled surface. Since the surface has been flattened without changing its intrinsic distances and angles, parallel transport there corresponds to parallel transport on the original curved surface along the curve.
Finally, we roll the surface back up:

$$\boxed{\text{uncurl} \longrightarrow \text{parallel transport in the plane} \longrightarrow \text{curl back}}. \tag{1}$$

For a *constant-latitude curve on a sphere*, this is why a cone is useful. The latitude is also a curve on a suitable cone, and the cone can be unrolled easily. We therefore use the cone as a convenient geometric representation of the neighborhood relevant to that curve, rather than trying to flatten the entire sphere.
The other way is to descretize the sphere and place a sheet of paper on that, then trace out the curve along with initial vector. Flatten the sheet and perform parallel transport and put it back on as shown in fig. below:



The path is the closed 12-gon at colatitude $\theta_0 = 60°$: a circle of latitude, *not* a geodesic. Inside each flat patch the spear is only slid along, never turned — yet its direction relative to the walk creeps round patch by patch. Dashed: the same spear back at patch 0 after all twelve hinges, reversed.

Unfold the two patches that share a hinge. Start the spear tangent to the path in patch k and carry it over: it keeps exactly the same direction, but the *path* kinks by ω at the hinge, so in patch $k+1$ the spear is no longer tangent — it lags the walk by ω. All of the turning happens at patch boundaries, none of it inside a patch.

The whole band, unrolled flat: a sector of a cone, not an annulus. The twelve hinge angles add up to $\sum_k \omega_k = 2\pi\cos\theta_0$, short of a full turn. In this flat picture the spear is literally constant — all twelve copies are parallel — so gluing the two edges marked A (one and the same edge of the sphere) hands the spear back rotated by the missing wedge $2\pi(1-\cos\theta_0)$, the solid angle enclosed by the loop. At $\theta_0 = 60°$ both are π, which is why the spear in (a) comes home reversed.

Differential operators

In flat space, vector calculus relies on the gradient, divergence, curl, and Laplacian. To generalize these operators to curved spacetime, we must account for the curved metric and the invariant four-volume element $\mathrm{d}V = \sqrt{-g}\,\mathrm{d}^4x$. Using the contracted Christoffel symbol identity $\Gamma^\mu_{\mu\alpha} = \frac{1}{\sqrt{-g}}\partial_\alpha\sqrt{-g}$ from Equation (1.2.39), we obtain covariant formulations that hold in any coordinate system:

1. **Divergence of a Vector Field:** Evaluating the covariant divergence of a vector field T^μ:

$$\nabla_\mu T^\mu = \partial_\mu T^\mu + \Gamma^\mu_{\mu\rho}T^\rho = \partial_\mu T^\mu + T^\rho\partial_\rho \ln\sqrt{-g} = \frac{1}{\sqrt{-g}}\partial_\mu\left(\sqrt{-g}T^\mu\right). \tag{1.2.50}$$

   Notice that the Christoffel symbols disappear, replaced entirely by derivatives of the metric determinant. This identity allows Gauss' divergence theorem to be applied directly: $\int_\mathcal{V}\nabla_\mu T^\mu\sqrt{-g}\,\mathrm{d}^4x = \oint_{\partial\mathcal{V}} T^\mu\,\mathrm{d}\Sigma_\mu$.

2. **Divergence of an Antisymmetric Tensor:** For an antisymmetric rank-2 tensor $F^{\mu\nu} = -F^{\nu\mu}$ (such as the electromagnetic Faraday tensor), the term $\Gamma^\nu_{\mu\rho}F^{\mu\rho}$ vanishes identically because it contracts the symmetric Christoffel symbol with the antisymmetric tensor ($\Gamma^\nu_{(\mu\rho)}F^{[\mu\rho]} = 0$). Hence:

$$\nabla_\mu F^{\mu\nu} = \frac{1}{\sqrt{-g}}\partial_\mu\left(\sqrt{-g}F^{\mu\nu}\right). \tag{1.2.51}$$

3. **The Covariant d'Alembertian Operator:** The d'Alembertian (wave operator) in curved spacetime is $\Box \equiv \nabla_\mu \nabla^\mu = g^{\mu\nu}\nabla_\mu\nabla_\nu$. Acting on a scalar field ϕ (where $\nabla_\nu\phi = \partial_\nu\phi$), Equation (1.2.50) gives:

$$\Box\phi = \nabla_\mu\nabla^\mu\phi = \frac{1}{\sqrt{-g}}\partial_\mu\left(\sqrt{-g}g^{\mu\nu}\partial_\nu\phi\right). \tag{1.2.52}$$

In flat Minkowski spacetime in Cartesian coordinates, $\sqrt{-g} = 1$, and this reduces to $\Box\phi = (-\partial_t^2/c^2 + \nabla^2)\phi$.

4. **Curl of a 1-Form (Exterior Derivative):** Computing the antisymmetrized derivative of a 1-form A_μ:

$$A_{\nu;\mu} - A_{\mu;\nu} = \partial_\mu A_\nu - \Gamma^\alpha_{\mu\nu}A_\alpha - \left(\partial_\nu A_\mu - \Gamma^\alpha_{\nu\mu}A_\alpha\right) = \partial_\mu A_\nu - \partial_\nu A_\mu = A_{\nu,\mu} - A_{\mu,\nu}. \tag{1.2.53}$$

Because the Levi-Civita connection is torsion-free ($\Gamma^\alpha_{\mu\nu} = \Gamma^\alpha_{\nu\mu}$), the connection coefficients cancel out completely. The curl of a 1-form is intrinsically tensorial, requiring no connection at all.

Lie derivative

In differential geometry, the covariant derivative ∇_μ requires an external geometric structure—namely, an affine connection $\Gamma^\alpha_{\mu\nu}$ and, typically, a metric tensor. Yet on a bare differentiable manifold, one often wishes to quantify how a tensor field changes along the direction of a vector field ξ^μ without introducing any metric or connection whatsoever.

The **Lie derivative** $\mathcal{L}_\xi$ achieves this by using the vector field ξ^μ as an intrinsic dynamical conveyor belt. The integral curves of ξ^μ define a one-parameter group of diffeomorphisms (a flow) that drags points, curves, and coordinate meshes smoothly across the manifold (Figure 1.5, right).

Consider an infinitesimal displacement from point $P(x^\alpha)$ to a neighboring point $Q(x^\alpha + \epsilon\xi^\alpha)$. Using the coordinate system at P to label points near Q corresponds to performing a local coordinate transformation:

$$x'^\mu = x^\mu - \epsilon\xi^\mu(x). \tag{1.2.54}$$

To first order in ϵ, the Jacobian matrices of this coordinate shift are:

$$\frac{\partial x'^\mu}{\partial x^\nu} = \delta^\mu_\nu - \epsilon\partial_\nu\xi^\mu, \qquad \frac{\partial x^\nu}{\partial x'^\mu} = \delta^\nu_\mu + \epsilon\partial_\nu\xi^\mu. \tag{1.2.55}$$

Evaluating the vector field T^μ at Q in this transported frame via Taylor expansion:

$$\begin{aligned} T'^\mu(Q) &= \frac{\partial x'^\mu}{\partial x^\nu}T^\nu(x^\alpha + \epsilon\xi^\alpha) \\ &= (\delta^\mu_\nu - \epsilon\partial_\nu\xi^\mu)\left[T^\nu(P) + \epsilon\xi^\alpha\partial_\alpha T^\nu(P)\right] \\ &= T^\mu(P) + \epsilon\left[\xi^\alpha\partial_\alpha T^\mu(P) - T^\nu(P)\partial_\nu\xi^\mu\right] + \mathcal{O}(\epsilon^2). \end{aligned} \tag{1.2.56}$$

The **Lie derivative** $\mathcal{L}_\xi T^\mu$ is defined by comparing this transported vector directly with the original vector $T^\mu(P)$:

$$\mathcal{L}_\xi T^\mu \equiv \lim_{\epsilon\to 0}\frac{T'^\mu(Q) - T^\mu(P)}{\epsilon} = \xi^\nu\partial_\nu T^\mu - T^\nu\partial_\nu\xi^\mu = [\xi, T]^\mu. \tag{1.2.57}$$

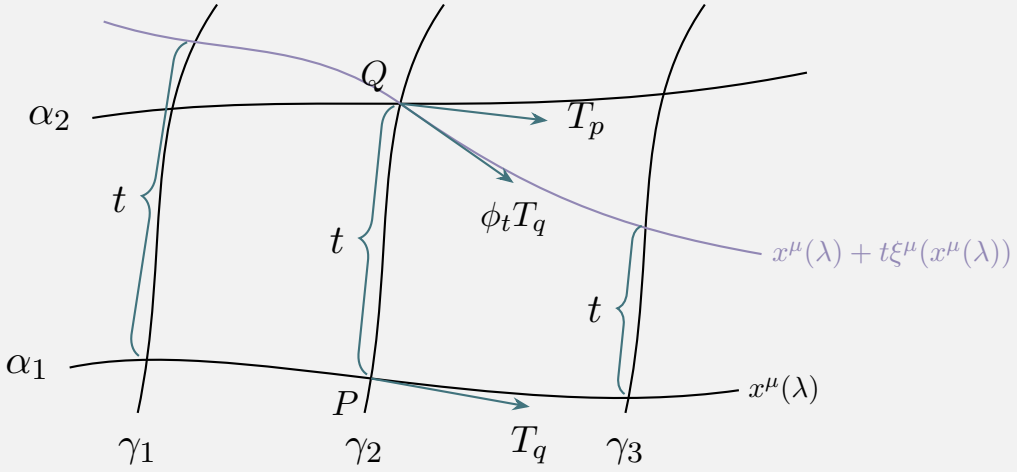
The Lie derivative of a vector field along ξ is precisely the **Lie bracket** $[\xi, T]$.

Generalizing to an arbitrary tensor field of rank (p, q), the Lie derivative is:

$$\mathcal{L}_\xi T^{\mu_1\ldots\mu_p}_{\nu_1\ldots\nu_q} = \xi^\alpha\partial_\alpha T^{\mu_1\ldots\mu_p}_{\nu_1\ldots\nu_q} - \sum_{i=1}^{p} T^{\mu_1\ldots\alpha\ldots\mu_p}_{\nu_1\ldots\nu_q}\partial_\alpha\xi^{\mu_i} + \sum_{j=1}^{q} T^{\mu_1\ldots\mu_p}_{\nu_1\ldots\beta\ldots\nu_q}\partial_{\nu_j}\xi^\beta. \tag{1.2.58}$$

Because the connection is symmetric ($\Gamma^\alpha_{\mu\nu} = \Gamma^\alpha_{\nu\mu}$), all partial derivatives ∂ in Equation (1.2.58) can be replaced by covariant derivatives ∇.

One intuitive way to derive the Lie derivative is with the notion of integral curves.



Let ξ be a vector field and let its integral curves be

$$\frac{dx^\mu(t)}{dt} = \xi^\mu(x).$$

The integral curves generate a one-parameter family of maps

$$\phi_\epsilon : M \to M,$$

called the flow of ξ. For an infinitesimal parameter ϵ,

$$\phi_\epsilon(x)^\mu = x^\mu + \epsilon\xi^\mu(x) + \mathcal{O}(\epsilon^2).$$

Now consider another vector field T. Its integral curves satisfy

$$\frac{dx^\mu(\lambda)}{d\lambda} = T^\mu(x),$$

so that

$$T = T^\mu \partial_\mu.$$

We now ask how T changes under an infinitesimal displacement generated by ξ.
The flow ϕ_ϵ moves the points of the manifold and therefore moves the integral curves of T. In particular, if

$$x^\mu(\lambda)$$

is an integral curve of T, its image under ϕ_ϵ is

$$\begin{aligned}\bar{x}^\mu(\lambda) &= \phi_\epsilon\left(x(\lambda)\right)^\mu \\ &= x^\mu(\lambda) + \epsilon\xi^\mu(x(\lambda)) + \mathcal{O}(\epsilon^2).\end{aligned}$$

The tangent vector to this displaced curve is

$$\begin{aligned}\bar{T} &= \frac{d\bar{x}^\mu}{d\lambda}\bar{\partial}_\mu \\ &= \left[T^\mu + \epsilon T^\nu \partial_\nu \xi^\mu\right]\bar{\partial}_\mu + \mathcal{O}(\epsilon^2).\end{aligned}$$

However, $\bar{T}$ is naturally written in the coordinate basis associated with the flowed coordinates $\bar{x}^\mu = \phi_\epsilon^\mu(x)$. To compare it with the original vector field T at the same point, we express the new basis in terms of the original basis. Since

$$\bar{x}^\mu = \phi_\epsilon^\mu(x) = x^\mu + \epsilon\,\xi^\mu(x) + \mathcal{O}(\epsilon^2),$$

the inverse transformation is

$$x^\mu = \bar{x}^\mu - \epsilon\,\xi^\mu(\bar{x}) + \mathcal{O}(\epsilon^2).$$

Hence,

$$\begin{aligned}\bar{\partial}_\mu \equiv \frac{\partial}{\partial \bar{x}^\mu} &= \frac{\partial x^\nu}{\partial \bar{x}^\mu}\,\partial_\nu \\ &= \left(\delta_\mu^\nu - \epsilon\,\partial_\mu \xi^\nu\right)\partial_\nu + \mathcal{O}(\epsilon^2).\end{aligned}$$

Therefore the pulled-back vector field is

$$\begin{aligned}T_\epsilon^\mu &= [T^\nu + \epsilon\xi^\rho\partial_\rho T^\nu]\,[\delta_\nu^\mu - \epsilon\partial_\nu\xi^\mu] + \mathcal{O}(\epsilon^2) \\ &= T^\mu + \epsilon\,(\xi^\nu\partial_\nu T^\mu - T^\nu\partial_\nu\xi^\mu) + \mathcal{O}(\epsilon^2).\end{aligned}$$

Thus the infinitesimal change of T under the flow generated by ξ is

$$\begin{aligned}\mathcal{L}_\xi T \equiv \lim_{\epsilon\to 0}\frac{T_\epsilon - T}{\epsilon} \\ = (\xi^\nu\partial_\nu T^\mu - T^\nu\partial_\nu\xi^\mu)\,\partial_\mu.\end{aligned}$$

Recognizing the Lie bracket,

$$[\xi, T]^\mu = \xi^\nu\partial_\nu T^\mu - T^\nu\partial_\nu\xi^\mu,$$

we obtain

$$\boxed{\mathcal{L}_\xi T = [\xi, T].} \tag{1}$$

Thus, the Lie derivative measures the infinitesimal change of the vector field T under the flow generated by ξ. The pullback is essential because it brings the flowed vector field back to the original manifold before it is compared with T.

Geometric transport versus dynamical flow. At first glance, both the covariant derivative $\nabla_X Y$ and the Lie derivative $\mathcal{L}_X Y$ appear to accomplish the same basic task: each evaluates the directional rate of change of a vector field Y along another vector field X. Why, then, does differential geometry require two distinct directional derivatives?

The answer lies in the different geometric structures upon which they rely, and the completely different physical questions they address. The covariant derivative requires an affine connection Γ (and in general relativity, a metric tensor). It answers the question: *how does Y change relative to parallel transport along the integral curve of X?* Because the connection evaluates vectors at adjacent points by transporting them along a path, the covariant derivative behaves linearly with respect to smooth functions in its directional argument: $\nabla_{fX}Y = f\nabla_X Y$.

By contrast, the Lie derivative requires *neither a connection nor a metric.* It is an intrinsic differential-topological operator that requires only the smooth manifold structure itself. It answers a dynamical question: *how is the vector field Y distorted when the entire manifold is dragged along the flow generated by X?* Because changing the magnitude of X distorts the

parameterization of the flow itself, the Lie derivative is not linear over functions in its direction slot; instead, it satisfies the Leibniz-like rule $\mathcal{L}_{fX}Y = f\mathcal{L}_X Y - (Yf)X$.

These distinct properties make the Lie derivative the natural tool for discovering spacetime symmetries. If $\mathcal{L}_\xi T = 0$, the tensor field T is invariant under the continuous symmetry generated by ξ^μ. In particular, for the metric tensor $g_{\mu\nu}$, applying Equation (1.2.58) and using metric compatibility yields:

$$\mathcal{L}_\xi g_{\mu\nu} = \nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu. \tag{1.2.59}$$

1.2.4 Curvature

The Riemann curvature tensor

If an observer is given an intimidating metric $g_{\mu\nu}(x)$, how can they determine whether spacetime is truly curved (with real gravitational tidal fields), or whether it is merely flat Minkowski spacetime written in complicated curvilinear coordinates?

As illustrated in Figure 1.5, parallel transport along a closed loop in a curved space does not return a vector to its original orientation. This path-dependence (holonomy) is the defining physical property of curvature.

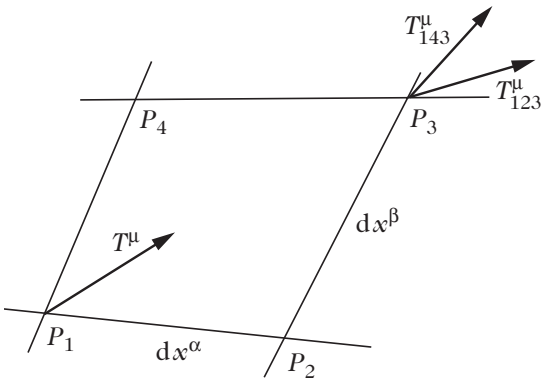


Figure 1.6: Parallel transport of a vector T^μ around an infinitesimal closed quadrilateral $P_1P_2P_3P_4$. Transporting along path $P_1 \to P_2 \to P_3$ yields T^μ_{123}, whereas transporting along $P_1 \to P_4 \to P_3$ yields T^μ_{143}. The failure of these vectors to coincide reveals the presence of the Riemann curvature tensor.

Consider parallel transporting a vector T^μ around an infinitesimal parallelogram with coordinate displacements dx^α and dx^β (Figure 1.6). The net change in the vector upon completing the circuit is determined by the **commutator of covariant derivatives**:

$$(\nabla_\alpha \nabla_\beta - \nabla_\beta \nabla_\alpha) T^\mu = R^\mu{}_{\nu\alpha\beta} T^\nu, \tag{1.2.60}$$

where $R^\mu{}_{\nu\alpha\beta}$ is the **Riemann curvature tensor**:

$$R^\mu{}_{\nu\alpha\beta} \equiv \partial_\alpha \Gamma^\mu_{\nu\beta} - \partial_\beta \Gamma^\mu_{\nu\alpha} + \Gamma^\mu_{\sigma\alpha}\Gamma^\sigma_{\nu\beta} - \Gamma^\mu_{\sigma\beta}\Gamma^\sigma_{\nu\alpha}, \tag{1.2.61}$$

with fully covariant form $R_{\sigma\nu\alpha\beta} \equiv g_{\sigma\mu} R^\mu{}_{\nu\alpha\beta}$.

The unambiguous test of curvature. The Riemann tensor is a genuine rank-$(1,3)$ tensor. Therefore:

- Spacetime is flat if and only if $R^\mu{}_{\nu\alpha\beta} = 0$ everywhere.
- If $R^\mu{}_{\nu\alpha\beta} \neq 0$, no coordinate transformation can ever eliminate it. Genuine gravitational tidal forces are present.

Algebraic symmetries of the Riemann tensor. The fully covariant Riemann tensor $R_{\alpha\mu\nu\sigma}$ possesses remarkable algebraic symmetries that reduce its independent components:

1. **Antisymmetry in the First and Second Index Pairs:**

$$R_{\alpha\mu\nu\sigma} = -R_{\mu\alpha\nu\sigma} = -R_{\alpha\mu\sigma\nu}. \tag{1.2.62}$$

2. **Pair-Exchange Symmetry:**

$$R_{\alpha\mu\nu\sigma} = R_{\nu\sigma\alpha\mu}. \tag{1.2.63}$$

3. **First Algebraic Bianchi Identity (Cyclic Sum):**

$$3R_{\alpha[\mu\nu\sigma]} \equiv R_{\alpha\mu\nu\sigma} + R_{\alpha\nu\sigma\mu} + R_{\alpha\sigma\mu\nu} = 0. \tag{1.2.64}$$

In an n-dimensional space, the pair antisymmetries (1.2.62) and exchange symmetry (1.2.63) reduce the components to a symmetric $\frac{n(n-1)}{2}\times\frac{n(n-1)}{2}$ matrix. The cyclic identity (1.2.64) imposes $\binom{n}{4}$ additional constraints, leaving the total number of algebraically independent components as:

$$N_{\text{Riem}} = \frac{n^2(n^2-1)}{12}. \tag{1.2.65}$$

In four spacetime dimensions ($n = 4$), $N_{\text{Riem}} = \frac{16\times15}{12} = \mathbf{20}$ independent components.

Another way to view the counting is to regard $R_{\mu\nu\rho\sigma}$ as a matrix whose rows are made by dissecting another anti symmetric matrices and same happens for columns. Since

$$R_{\mu\nu\rho\sigma} = -R_{\nu\mu\rho\sigma} = -R_{\mu\nu\sigma\rho}, \qquad R_{\mu\nu\rho\sigma} = R_{\rho\sigma\mu\nu}, \tag{1}$$

we may take each pair $(\mu\nu)$ and $(\rho\sigma)$ to be ordered, leaving

$$\frac{n(n-1)}{2}$$

independent elements in each row and column if additional symmetry constraints were not imposed. Because of this anti symmetry certain rows will repeat themselves and hence we will only have $\frac{n(n-1)}{2}$ independent rows.
The pair-exchange symmetry $R_{\mu\nu\rho\sigma} = R_{\rho\sigma\mu\nu}$ then makes sure that even out of those $\frac{n(n-1)}{2}\times\frac{n(n-1)}{2}$ independent elements some are related. The coloured entries below illustrate this symmetry: entries with the same colour are related by the exchange $(\mu\nu)\leftrightarrow(\rho\sigma)$ and therefore are not independent.

	(11)	(12)	(13)	$\cdots$	$(n-1,n)$	(nn)	number of independent component in each row
(11)	R_{1111}	R_{1112}	R_{1113}	$\cdots$	$R_{11\,n-1\,n}$	R_{11nn}	$\frac{n(n-1)}{2}$
(12)	R_{1211}	R_{1212}	R_{1213}	$\cdots$	$R_{12\,n-1\,n}$	R_{12nn}	$\frac{n(n-1)}{2}-1$
(13)	R_{1311}	R_{1312}	R_{1313}	$\cdots$	$R_{13\,n-1\,n}$	R_{13nn}	$\frac{n(n-1)}{2}-2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$(n-1,n)$	$R_{n-1\,n\,11}$	$R_{n-1\,n\,12}$	$R_{n-1\,n\,13}$	$\cdots$	$R_{n-1\,n\,n-1\,n}$	$R_{n-1\,n\,nn}$	1
(nn)	R_{nn11}	R_{nn12}	R_{nn13}	$\cdots$	$R_{nn\,n-1\,n}$	R_{nnnn}	0

Thus, the first row contains $\frac{n(n-1)}{2}$ independent entries, the second contains $\frac{n(n-1)}{2}-1$,

and so on. Hence,

$$\frac{n(n-1)}{2}+\left(\frac{n(n-1)}{2}-1\right)+\left(\frac{n(n-1)}{2}-2\right)+\cdots+1+0=\frac{1}{2}\frac{n(n-1)}{2}\left(\frac{n(n-1)}{2}+1\right)$$

Then the identity (1.2.64) reduces the number of independent elements by $\binom{n}{4}$ and hence we get

$$N_{\text{Riem}}=\frac{n^2(n^2-1)}{12}. \tag{2}$$

The differential Bianchi identity. In addition to its algebraic symmetries, the covariant derivatives of the Riemann tensor satisfy a fundamental differential identity:

$$3R_{\alpha\mu[\nu\sigma;\beta]}\equiv R_{\alpha\mu\nu\sigma;\beta}+R_{\alpha\mu\sigma\beta;\nu}+R_{\alpha\mu\beta\nu;\sigma}=0. \tag{1.2.66}$$

This is satisfied by any metric regardless of weather it solves the Einstein Field Equation or not. Therefore, it is not a constraint that removes degrees of freedom but rather an identity. In the language of Dirac-Brackets, this constraint appears when we impose $\{\phi_{\text{secondary}},\mathcal{H}_{\text{tot}}\}=0$ and saves us from tertiary or higher order constraints to appear in the theory.

Ricci and Einstein tensors

In Newtonian gravity, Poisson's equation $\nabla^2\Phi=4\pi G_{\text{N}}\rho$ relates the second derivatives of the gravitational potential to the mass density. In general relativity, we seek second-order field equations relating spacetime curvature to the symmetric matter stress-energy tensor $T_{\mu\nu}$ (which has 10 components).

Because $T_{\mu\nu}$ is a symmetric rank-2 tensor, the gravitational side of the field equations must also be a symmetric rank-2 tensor constructed from the Riemann tensor (which has 20 components).

Contracting the Riemann tensor. Due to the antisymmetries of the Riemann tensor, contracting the first two or last two indices yields zero ($g^{\alpha\mu}R_{\alpha\mu\nu\sigma}=0$). There is only one unique non-zero contraction, defining the symmetric **Ricci tensor**:

$$R_{\mu\nu}\equiv R^{\alpha}{}_{\mu\alpha\nu}=-R^{\alpha}{}_{\mu\nu\alpha}. \tag{1.2.67}$$

Because $R_{\mu\nu}=R_{\nu\mu}$, the Ricci tensor in 4 dimensions possesses **10 independent components**.

Taking the scalar trace of the Ricci tensor yields the **Ricci scalar** (scalar curvature):

$$R\equiv g^{\mu\nu}R_{\mu\nu}=R^{\mu}{}_{\mu}. \tag{1.2.68}$$

Why the Ricci tensor cannot serve as the field tensor. A natural first guess for the gravitational field equations, and indeed one that Einstein himself pursued for several years, is to set the Ricci tensor directly proportional to the stress-energy tensor: $R_{\mu\nu}=\kappa T_{\mu\nu}$. Because both $R_{\mu\nu}$ and $T_{\mu\nu}$ are symmetric rank-2 tensors with 10 independent components, this proposal seems the perfect relativistic counterpart to the Newtonian Poisson equation $\nabla^2\Phi=4\pi G_{\text{N}}\rho$.

However, this naive proposal suffers from a fatal physical inconsistency: it violates local energy-momentum conservation. In general relativity, the conservation of energy and momentum requires that the divergence of the stress-energy tensor vanish identically, $\nabla^{\mu}T_{\mu\nu}=0$. If $R_{\mu\nu}=\kappa T_{\mu\nu}$ were true, the Ricci tensor would also have to be divergence-free. But geometry

dictates otherwise. By taking the differential Bianchi identity (1.2.66) and contracting it twice with the metric tensors $g^{\alpha\beta}$ and $g^{\mu\sigma}$, we obtain:

$$\nabla^\alpha R_{\alpha\mu\nu\sigma} - \nabla_\nu R_{\mu\sigma} + \nabla_\sigma R_{\mu\nu} = 0, \tag{1.2.69}$$

$$\nabla^\alpha R_{\alpha\nu} - \frac{1}{2}\nabla_\nu R = 0 \iff \nabla^\mu \left(R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} \right) = 0. \tag{1.2.70}$$

This reveals that the divergence of the Ricci tensor does not vanish; rather, it equals half the gradient of the Ricci scalar: $\nabla^\mu R_{\mu\nu} = \frac{1}{2}\nabla_\nu R$. Postulating $R_{\mu\nu} = \kappa T_{\mu\nu}$ would therefore force $\nabla_\nu R = 0$, implying that the Ricci scalar—and hence the trace of the stress-energy tensor $T^\mu{}_\mu$—must be spatially and temporally constant throughout the universe! Such a constraint would make it impossible to describe a universe containing localized lumps of matter like stars or planets surrounded by vacuum.

This crucial geometric fact points directly to the correct field tensor: the **Einstein tensor**:

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu}, \qquad \nabla^\mu G_{\mu\nu} = 0. \tag{1.2.71}$$

The Einstein tensor is automatically and identically divergence-free for any pseudo-Riemannian geometry, making $G_{\mu\nu} = \kappa T_{\mu\nu}$ mathematically compatible with energy-momentum conservation.

Taxonomy of curvature in 4D spacetime. The 20 independent components of the Riemann tensor split physically into two distinct parts:

- **The Ricci Tensor ($R_{\mu\nu}$, 10 components):** Represents the local trace of curvature. Through Einstein's equations, it is directly determined by the presence of matter and energy at that exact point.
- **The Weyl Tensor ($C^\mu{}_{\nu\alpha\beta}$, 10 components):** The completely traceless part of the Riemann tensor. It represents the tidal gravitational field and gravitational waves that propagate through vacuum ($T_{\mu\nu} = 0$) far from matter sources.

Killing vector fields

A vector field ξ^μ that generates a continuous isometry (a symmetry transformation that leaves the metric tensor invariant) satisfies the **Killing equation**:

$$\mathcal{L}_\xi g_{\mu\nu} = \nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu = 0. \tag{1.2.72}$$

Such a field is called a **Killing vector field**.

Differentiating the Killing equation and commuting covariant derivatives via Equation (1.2.60) yields the integrability condition:

$$\nabla_\alpha \nabla_\beta \xi_\nu = R^\mu{}_{\alpha\beta\nu} \xi_\mu. \tag{1.2.73}$$

Consequently, a Killing vector field everywhere on $\mathcal{M}$ is uniquely fixed once its value $\xi_\nu(p)$ (N parameters) and its antisymmetric derivative $\nabla_{[\mu}\xi_{\nu]}(p)$ ($\frac{N(N-1)}{2}$ parameters) are specified at a single point p. Therefore, an N-dimensional spacetime can admit at most:

$$N_{\text{Killing}} \leq \frac{N(N+1)}{2} \tag{1.2.74}$$

linearly independent Killing vector fields. In four dimensions ($N = 4$), $N_{\text{Killing}} \leq 10$, matching the 10 parameters of the Poincaré group.

Maximally symmetric spacetimes. A spacetime possessing the theoretical maximum of 10 Killing vector fields is called a **maximally symmetric space**. Complete spatial homogeneity and isotropy require the Riemann tensor to be proportional to combinations of the metric:

$$R_{\mu\nu\alpha\beta} = \frac{R}{N(N-1)} \left(g_{\mu\alpha} g_{\nu\beta} - g_{\mu\beta} g_{\nu\alpha} \right), \qquad R_{\mu\nu} = \frac{R}{N} g_{\mu\nu}, \tag{1.2.75}$$

where the Ricci scalar R is constant everywhere. In four dimensions, there exist exactly three maximally symmetric spacetimes:

- **Minkowski space-time** ($R = 0$): Flat empty spacetime.
- **de Sitter space** (dS_4) ($R > 0$): Positive curvature; the fundamental background geometry of cosmological inflation and dark-energy-dominated expansion.
- **Anti-de Sitter space** (AdS_4) ($R < 0$): Negative curvature; central to holographic quantum gravity (AdS/CFT correspondence).

1.2.5 Covariant approach (1+3 / 3+1 splitting)

General relativity treats space and time on an equal four-dimensional footing. In physical cosmology, however, we observe the universe from a specific vantage point: the rest frame of the cosmic fluid, as defined by the rest frame of the cosmic microwave background radiation. To connect four-dimensional general relativity to cosmological observations, we seek a fully covariant, coordinate-independent decomposition of spacetime into "time" and "space".

The 4-velocity congruence. We describe a family of fundamental cosmic observers (or fluid elements) by a congruence of timelike worldlines. Each observer possesses a unit timelike 4-velocity $u^\mu = \mathrm{d}x^\mu / \mathrm{d}\tau$ normalized as (setting $c = 1$):

$$u^\mu u_\mu = -1. \tag{1.2.76}$$

The vector u^μ defines the local time direction for that observer.

3+1 decomposition

The 4-velocity u^μ decomposes the metric tensor into two mutually orthogonal projection operators:

$$\delta^\mu{}_\nu = U^\mu{}_\nu + \gamma^\mu{}_\nu \implies U^\mu{}_\nu \equiv -u^\mu u_\nu, \qquad \gamma_{\mu\nu} \equiv g_{\mu\nu} + u_\mu u_\nu, \tag{1.2.77}$$

where $U^\mu{}_\nu$ projects parallel to u^μ (along the observer's worldline), and $\gamma_{\mu\nu}$ projects orthogonally into the observer's instantaneous spatial rest space. They satisfy the projection algebra:

$$U^\mu{}_\nu U^\nu{}_\alpha = U^\mu{}_\alpha, \qquad U^\mu{}_\mu = 1, \qquad U^\mu{}_\nu u^\nu = u^\mu, \tag{1.2.78}$$

$$\gamma^\mu{}_\nu \gamma^\nu{}_\alpha = \gamma^\mu{}_\alpha, \qquad \gamma^\mu{}_\mu = 3, \qquad \gamma^\mu{}_\nu u^\nu = 0. \tag{1.2.79}$$

The spacetime line element decomposes cleanly into temporal and spatial intervals:

$$\mathrm{d}s^2 = -\left(u_\mu \, \mathrm{d}x^\mu\right)^2 + \gamma_{\mu\nu} \, \mathrm{d}x^\mu \, \mathrm{d}x^\nu \, . \tag{1.2.80}$$

Any arbitrary 4-vector v^μ decomposes into a longitudinal component along u^μ and a purely spatial transverse component $v^\mu_\perp$:

$$v^\mu = v^\nu \delta^\mu{}_\nu = v^\nu U^\mu{}_\nu + v^\mu_\perp, \qquad v^\mu_\perp \equiv v^\nu \gamma^\mu{}_\nu. \tag{1.2.81}$$

Decomposition of vector

We work in units where $c = 1$. Decompose an arbitrary vector X^α into a component parallel to the four-velocity u^α and a component orthogonal to it:

$$X^\alpha = x^0 u^\alpha + X^\alpha_\perp, \qquad u_\alpha X^\alpha_\perp = 0, \qquad u_\alpha u^\alpha = -1. \tag{1}$$

Contracting both sides with u_α gives

$$u_\alpha X^\alpha = x^0 u_\alpha u^\alpha + u_\alpha X^\alpha_\perp = -x^0, \tag{2}$$

and hence

$$\boxed{x^0 = -u_\beta X^\beta}. \tag{3}$$

The orthogonal component is therefore

$$X^\alpha_\perp = X^\alpha - x^0 u^\alpha = X^\alpha + (u_\rho X^\rho)\, u^\alpha \tag{4}$$

$$= \left(\delta^\alpha_{\ \rho} + u^\alpha u_\rho\right) X^\rho. \tag{5}$$

Thus, defining the orthogonal projection operator

$$\boxed{\perp^\alpha_{\ \rho} \equiv \delta^\alpha_{\ \rho} + u^\alpha u_\rho,} \tag{6}$$

we have

$$\boxed{X^\alpha_\perp = \perp^\alpha_{\ \rho} X^\rho.} \tag{7}$$

Particle four-momentum and measured energy. For a particle of rest mass m with 4-velocity u^μ and 4-momentum $p^\mu = mu^\mu$ ($p^\mu p_\mu = -m^2$), the energy measured by an observer moving with 4-velocity u^μ is:

$$E = -p_\mu u^\mu \implies E^2 - p^\mu_\perp p_{\mu\perp} = m^2. \tag{1.2.82}$$

The spatial momentum orthogonal to u^μ is $p^\mu_\perp = p^\nu \gamma^\mu_{\ \nu}$.

This can be shown as follows:

$$\begin{aligned} p_{\perp\mu} p^\mu_\perp &= p^\mu \gamma^\nu_\mu p_\alpha \gamma^\alpha_\nu = p^\alpha \gamma^\nu_\alpha p_\mu \gamma^\mu_\nu = p_\mu p^\alpha \gamma^\mu_\alpha \\ &= p_\mu p^\alpha \left(\delta^\mu_\alpha + u^\mu u_\alpha\right) = p_\mu p^\mu + \left(p_\mu u^\mu\right)\left(p_\alpha u^\alpha\right) \\ &= -m^2 + E^2 \end{aligned}$$

The spatial metric $\gamma_{\mu\nu}$ acts as the intrinsic metric on the 3-hypersurface orthogonal to u^μ.

One can also find the intrinsic metric on the hypersurface as

$$^{(3)}h_{\mu\nu} = \gamma^\alpha{}_\mu \gamma^\beta{}_\nu g_{\alpha\beta} \tag{1}$$

$$= (\delta^\alpha{}_\mu + u^\alpha u_\mu)(\delta^\beta{}_\nu + u^\beta u_\nu) g_{\alpha\beta} = g_{\mu\nu} + u_\mu u_\nu \tag{2}$$

$$= \gamma_{\mu\nu} \tag{3}$$

Spatial and time derivatives

With the $1+3$ splitting defined, two distinct derivative operators naturally emerge:

1. **The Proper Time Derivative ($\dot{T}$):** The covariant derivative along the observer's worldline:
$$\dot{T}^{\alpha\beta\ldots}_{\mu\nu\ldots} \equiv u^\lambda \nabla_\lambda T^{\alpha\beta\ldots}_{\mu\nu\ldots}. \tag{1.2.83}$$

2. **The Projected Spatial Covariant Derivative ($\bar{\nabla}$):** The derivative projected orthogonally into the instantaneous local rest space Σ_t:
$$\bar{\nabla}_\lambda T^{\alpha\beta\ldots}_{\mu\nu\ldots} \equiv \gamma^{\lambda_1}{}_\lambda \gamma^\alpha{}_{\alpha_1} \gamma^\beta{}_{\beta_1} \ldots \gamma^{\mu_1}{}_\mu \gamma^{\nu_1}{}_\nu \ldots \nabla_{\lambda_1} T^{\alpha_1\beta_1\ldots}_{\mu_1\nu_1\ldots}. \tag{1.2.84}$$

The projected spatial derivative is metric-compatible with the spatial metric: $\bar{\nabla}_\alpha \gamma_{\mu\nu} = 0$.

Vorticity and the commutativity of spatial derivatives. In flat Euclidean space, partial derivatives with respect to different spatial coordinates commute unconditionally: $[\partial_i, \partial_j] = 0$. On a curved spacetime slice, does the projected spatial covariant derivative $\bar{\nabla}_\mu$ commute when acting on a scalar field ρ?

A direct computation of the commutator yields:

$$\left[\bar{\nabla}_\mu, \bar{\nabla}_\nu\right] \rho = 2\omega_{\mu\nu}\dot{\rho}, \tag{1.2.85}$$

where $\omega_{\mu\nu}$ is the vorticity tensor of the fluid congruence and $\dot{\rho} \equiv u^\lambda \nabla_\lambda \rho$. This reveals that projected spatial derivatives commute if and only if the fluid flow is irrotational ($\omega_{\mu\nu} = 0$). By Frobenius' theorem of differential topology, this is precisely the condition required for the family of local spatial rest planes to mesh together and form a foliation of smooth, global spatial hypersurfaces.

Distortion of the geodesic flow (Fluid Kinematics)

In physical cosmology, we frequently model the large-scale matter distribution as a continuous cosmic fluid—a swarm of galaxies flowing along their timelike worldlines. As the fluid evolves, neighboring worldlines deviate from one another: the swarm may expand, shear, spin, or accelerate under pressure gradients. To analyze these physical motions systematically, we decompose the velocity gradient into irreducible kinematic modes.

The kinematic decomposition. The covariant gradient of the four-velocity field $\nabla_\mu u_\nu$ decomposes uniquely into four irreducible pieces:

$$\nabla_\mu u_\nu = -u_\mu \dot{u}_\nu + \frac{1}{3}\Theta \gamma_{\mu\nu} + \sigma_{\mu\nu} + \omega_{\mu\nu}. \tag{1.2.86}$$

Taxonomy of fluid kinematic variables:

1. **Acceleration Vector ($\dot{u}^\nu \equiv u^\mu \nabla_\mu u^\nu$):** Purely spatial ($\dot{u}^\nu u_\nu = 0$). It measures the presence of non-gravitational forces (e.g., pressure gradients). For freely falling cosmic dust, $\dot{u}^\nu = 0$.

2. **Expansion Scalar ($\Theta \equiv \nabla_\mu u^\mu$):** The trace of the spatial velocity gradient. It quantifies the isotropic rate of fractional volume expansion of the fluid. In an expanding Friedmann universe, $\Theta = 3H(t)$, where $H \equiv \dot{a}/a$ is the Hubble parameter.

3. **Shear Tensor** ($\sigma_{\mu\nu} \equiv \left[\gamma^{\alpha}{}_{(\mu}\gamma^{\beta}{}_{\nu)} - \frac{1}{3}\Theta\gamma_{\mu\nu}\right]\nabla_\alpha u_\beta$)**:** Symmetric ($\sigma_{\mu\nu} = \sigma_{\nu\mu}$), traceless ($\sigma^{\mu}{}_{\mu} = 0$), and purely spatial ($\sigma_{\mu\nu}u^\mu = 0$). It describes volume-preserving anisotropic shape distortion—stretching a spherical fluid element into an ellipsoid. Its magnitude is quantified by the shear scalar $\sigma^2 \equiv \frac{1}{2}\sigma_{\mu\nu}\sigma^{\mu\nu} \geq 0$.

4. **Vorticity Tensor** ($\omega_{\mu\nu} \equiv \gamma^{\alpha}{}_{[\mu}\gamma^{\beta}{}_{\nu]}\nabla_\alpha u_\beta$)**:** Antisymmetric ($\omega_{\mu\nu} = -\omega_{\nu\mu}$) and purely spatial ($\omega_{\mu\nu}u^\mu = 0$). It characterizes the local rigid rotation of fluid elements relative to a non-rotating inertial frame. It can be dualized into a vorticity vector $\omega^\sigma \equiv \frac{1}{2}u_\alpha\epsilon^{\alpha\sigma\mu\nu}\omega_{\mu\nu} \iff \omega_{\mu\nu} = u^\beta\epsilon_{\beta\sigma\mu\nu}\omega^\sigma$, with vorticity scalar $\omega^2 \equiv \frac{1}{2}\omega_{\mu\nu}\omega^{\mu\nu} \geq 0$.

Defining a representative length scale $S(\tau)$ of a comoving fluid element along its worldline by:

$$\frac{\dot{S}}{S} \equiv \frac{1}{3}\Theta, \tag{1.2.87}$$

the volume of the fluid element scales as $V \propto S^3$. In cosmology, $S(\tau)$ plays the role of the generalized cosmic scale factor $a(t)$.

The Raychaudhuri equation

How does the expansion rate of the universe or a collapsing star evolve with time? Differentiating the expansion scalar along the flow, $\dot{\Theta} \equiv u^\nu\nabla_\nu(\nabla_\mu u^\mu)$, commuting covariant derivatives via the Ricci identity $[\nabla_\nu, \nabla_\mu]u^\mu = R_{\mu\nu}u^\mu u^\nu$, and substituting the kinematic decomposition (1.2.86), we obtain the celebrated **Raychaudhuri equation**:

$$\dot{\Theta} = -\frac{1}{3}\Theta^2 - 2\sigma^2 + 2\omega^2 - R_{\mu\nu}u^\mu u^\nu + \bar{\nabla}_\mu\dot{u}^\mu + \dot{u}^\mu\dot{u}_\mu. \tag{1.2.88}$$

Physical anatomy of the Raychaudhuri equation: Every term in Equation (1.2.88) carries an intuitive physical meaning:

- $-\frac{1}{3}\Theta^2$ **(Self-focusing):** An expanding flow ($\Theta > 0$) naturally decelerates due to inertia; a contracting flow ($\Theta < 0$) collapses faster and faster.
- $-2\sigma^2$ **(Shear assisted collapse):** Because $\sigma^2 \geq 0$, shear deformation *always* works to accelerate gravitational collapse.
- $+2\omega^2$ **(Centrifugal resistance):** Vorticity (rotation) provides a centrifugal barrier that opposes gravitational collapse.
- $-R_{\mu\nu}u^\mu u^\nu$ **(Direct gravitational attraction):** The direct spacetime curvature generated by matter.

Singularity theorems and energy conditions. For an irrotational ($\omega = 0$), freely falling ($\dot{u}^\mu = 0$) flow of normal matter satisfying the **Strong Energy Condition** ($R_{\mu\nu}u^\mu u^\nu \geq 0$), all terms on the right-hand side of Equation (1.2.88) are negative definite:

$$\dot{\Theta} + \frac{1}{3}\Theta^2 \leq 0 \implies \frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{1}{\Theta}\right) \geq \frac{1}{3} \implies \frac{1}{\Theta(\tau)} \geq \frac{1}{\Theta_0} + \frac{\tau}{3}. \tag{1.2.89}$$

This simple inequality is the mathematical origin of the Penrose–Hawking singularity theorems:

- **Gravitational Collapse:** If the flow is initially converging ($\Theta_0 < 0$), $\Theta(\tau) \to -\infty$ within a finite proper time $\tau \leq 3/|\Theta_0|$, guaranteeing that worldlines cross and form a focal point (caustic or gravitational singularity).
- **The Big Bang Singularity:** Reversing the arrow of time in an expanding universe ($\Theta_0 > 0$) where $\ddot{S}/S \leq 0$ guarantees that the scale factor must have vanished ($S \to 0$) at a finite time in the past: $\tau_{\text{sing}} \leq 3/\Theta_0$.

Table 1.1: Standard Energy Conditions in General Relativity

Name	Mathematical Definition	Physical Meaning
Null (NEC)	$T_{\mu\nu}k^\mu k^\nu \geq 0 \quad (\forall k^\mu, k^\mu k_\mu = 0)$	Light rays are focused by matter
Weak (WEC)	$T_{\mu\nu}u^\mu u^\nu \geq 0 \quad (\forall u^\mu, u^\mu u_\mu = -1)$	Energy density is non-negative for any observer
Strong (SEC)	$\left(T_{\mu\nu} - \frac{1}{2}T g_{\mu\nu}\right) u^\mu u^\nu \geq 0 \iff R_{\mu\nu}u^\mu u^\nu \geq 0$	Gravity is universally attractive ($\ddot{S} < 0$)
Dominant (DEC)	Causal energy flux: $-T^\mu{}_\nu u^\nu$ is time-like/null	Energy does not propagate faster than light

Case of vanishing vorticity ($\omega = 0$)

When the fluid vorticity vanishes identically ($\omega_{\mu\nu} = 0$), **Frobenius' theorem** of differential topology guarantees that the local rest spaces integrate globally into a foliation of smooth spacelike hypersurfaces Σ_t orthogonal to the fluid worldlines (Figure 1.7). Spacetime decomposes cleanly into a sequence of spatial slices parameterized by cosmic time t.

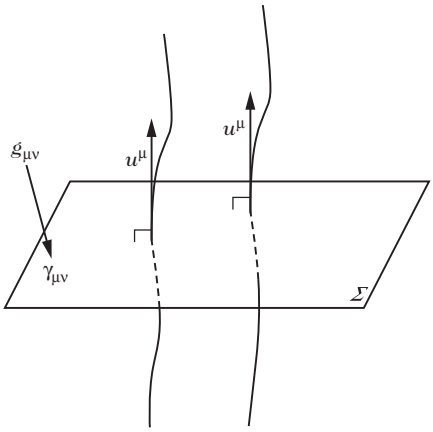


Figure 1.7: A timelike vector congruence u^μ defines an orthogonal rest space at each point. A global foliation of spacetime into spatial hypersurfaces Σ_t exists if and only if the flow vorticity vanishes ($\omega = 0$).

The **extrinsic curvature tensor** $K_{\mu\nu}$ quantifies how the spatial slice Σ_t curves and bends within the four-dimensional spacetime manifold $\mathcal{M}$:

$$K_{\mu\nu} \equiv \gamma^\alpha{}_\mu \nabla_\alpha u_\nu = \frac{1}{2}\mathcal{L}_u \gamma_{\mu\nu} = \frac{1}{3}\Theta\gamma_{\mu\nu} + \sigma_{\mu\nu}. \tag{1.2.90}$$

Its trace is the expansion scalar: $K \equiv K^\mu{}_\mu = \Theta$.

Projecting the four-dimensional spacetime Riemann tensor onto the spatial hypersurface Σ_t yields the **Gauss relation**:

$${}^{(3)}R^\alpha{}_{\beta\mu\nu} = \gamma^{\alpha'}{}_\alpha \gamma^{\beta'}{}_\beta \gamma^{\mu'}{}_\mu \gamma^{\nu}{}_{\nu'} R^{\alpha'}{}_{\beta'\mu'\nu'} - K^\alpha{}_\mu K^\beta{}_\nu + K^\beta{}_\mu K^\alpha{}_\nu. \tag{1.2.91}$$

Contraction of the Gauss relation yields the **scalar Gauss relation**:

$${}^{(3)}R + K^2 - K^{\alpha\beta}K_{\alpha\beta} = R + 2R_{\mu\nu}u^\mu u^\nu = 2G_{\mu\nu}u^\mu u^\nu, \tag{1.2.92}$$

which, written in terms of fluid kinematic variables, takes the transparent form:

$${}^{(3)}R = -\frac{2}{3}\Theta^2 + 2\sigma^2 + 2G_{\mu\nu}u^\mu u^\nu. \tag{1.2.93}$$

Projecting the Riemann tensor once along the normal u^μ and three times onto the slice yields the **Codazzi relation**:

$$\gamma^\mu{}_\rho u^\sigma \gamma^{\alpha'}{}_\alpha \gamma^{\beta'}{}_\beta R^\rho{}_{\sigma\alpha'\beta'} = \bar{\nabla}_\beta K^\mu{}_\alpha - \bar{\nabla}_\alpha K^\mu{}_\beta, \tag{1.2.94}$$

whose contraction gives:

$$\gamma^\mu{}_\alpha u^\nu R_{\mu\nu} = \bar{\nabla}_\mu K^\mu{}_\alpha - \bar{\nabla}_\alpha K. \tag{1.2.95}$$

Decomposition of rank-2 tensors

Relative to the timelike observer congruence u^μ, any symmetric rank-2 tensor $T_{\mu\nu}$ (such as the energy-momentum tensor) decomposes uniquely into four irreducible, physically intuitive parts:

$$T_{\mu\nu} = A u_\mu u_\nu + 2q_{(\mu}u_{\nu)} + B\gamma_{\mu\nu} + \pi_{\mu\nu}, \tag{1.2.96}$$

where:

- $A \equiv T_{\mu\nu}u^\mu u^\nu$: The **energy density** measured by an observer moving with four-velocity u^μ.
- $q_\mu \equiv -\gamma^\alpha{}_\mu T_{\alpha\beta}u^\beta$: The **spatial energy flux vector** ($q_\mu u^\mu = 0$).
- $B \equiv \frac{1}{3}T_{\mu\nu}\gamma^{\mu\nu}$: The **isotropic pressure**.
- $\pi_{\mu\nu} \equiv \left[\gamma^\alpha{}_{(\mu}\gamma^\beta{}_{\nu)} - \frac{1}{3}\gamma_{\mu\nu}\gamma^{\alpha\beta}\right] T_{\alpha\beta}$: The symmetric, traceless, purely spatial **anisotropic stress tensor** ($\pi_{\mu\nu}u^\mu = 0, \pi^\mu{}_\mu = 0$). It describes directional viscosity, shear stresses, and free-streaming radiation.

1.3 Equations of motion

1.3.1 Einstein's equations

In electrodynamics, Maxwell's equations follow from varying an action with respect to the gauge potential A_μ. In general relativity, the fundamental dynamical variable of spacetime is the metric tensor $g_{\mu\nu}$. To obtain second-order, generally covariant field equations for the metric, we seek an action constructed from geometric invariants.

The simplest non-trivial scalar invariant constructed from the metric and its derivatives up to second order is the Ricci scalar R. David Hilbert (1915) formulated the total action of gravity and matter as:

$$S = \frac{1}{2\kappa}\int (R - 2\Lambda)\sqrt{-g}\,\mathrm{d}^4x + \int \mathcal{L}_{\mathrm{mat}}\sqrt{-g}\,\mathrm{d}^4x, \tag{1.3.1}$$

where κ is the gravitational coupling constant (determined to be $\kappa = 8\pi G_\mathrm{N}/c^4$ by matching the Newtonian limit), Λ is the cosmological constant, and $\mathcal{L}_{\mathrm{mat}}$ is the matter Lagrangian density.

Variation of the Einstein–Hilbert action

To derive the field equations, we apply the principle of stationary action ($\delta S = 0$) under an arbitrary variation of the inverse metric $\delta g^{\mu\nu}$ that vanishes on the spacetime boundary $\partial\mathcal{V}$.

Varying the gravitational action:

$$\delta\int R\sqrt{-g}\,\mathrm{d}^4x = \int \left(R_{\mu\nu}\delta g^{\mu\nu}\sqrt{-g} + R\delta\sqrt{-g} + g^{\mu\nu}\delta R_{\mu\nu}\sqrt{-g}\right)\mathrm{d}^4x. \tag{1.3.2}$$

Using the determinant variation $\delta\sqrt{-g} = -\frac{1}{2}\sqrt{-g}g_{\mu\nu}\delta g^{\mu\nu}$, the second term becomes:

$$\int R\delta\sqrt{-g}\,\mathrm{d}^4x = -\frac{1}{2}\int R g_{\mu\nu}\delta g^{\mu\nu}\sqrt{-g}\,\mathrm{d}^4x. \tag{1.3.3}$$

For the third term, the variation of the Ricci tensor is given by the **Palatini identity**:

$$\delta R_{\mu\nu} = \nabla_\sigma(\delta\Gamma^\sigma_{\mu\nu}) - \nabla_\mu(\delta\Gamma^\sigma_{\sigma\nu}). \tag{1.3.4}$$

Because the difference between two connections $\delta\Gamma^\sigma_{\mu\nu}$ is a genuine rank-$(1,2)$ tensor, contracting with $g^{\mu\nu}$ produces a total covariant divergence:

$$g^{\mu\nu}\delta R_{\mu\nu}\sqrt{-g} = \partial_\sigma\left[\sqrt{-g}\left(g^{\mu\nu}\delta\Gamma^\sigma_{\mu\nu} - g^{\mu\sigma}\delta\Gamma^\rho_{\rho\mu}\right)\right]. \tag{1.3.5}$$

By Stokes' theorem, this boundary divergence integrates to zero under standard Dirichlet boundary conditions ($\delta g^{\mu\nu}|_{\partial\mathcal{V}} = 0$).

Combining these variations, we obtain the exact gravitational variation:

$$\delta S_{\text{grav}} = \frac{1}{2\kappa}\int\left(R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu}\right)\delta g^{\mu\nu}\sqrt{-g}\,\mathrm{d}^4x = \frac{1}{2\kappa}\int\left(G_{\mu\nu} + \Lambda g_{\mu\nu}\right)\delta g^{\mu\nu}\sqrt{-g}\,\mathrm{d}^4x. \tag{1.3.6}$$

The stress-energy tensor

The variation of the matter action with respect to the metric defines the **symmetric Hilbert stress-energy tensor** $T_{\mu\nu}$:

$$\delta S_{\text{mat}} = \delta\int\mathcal{L}_{\text{mat}}\sqrt{-g}\,\mathrm{d}^4x \equiv -\frac{1}{2}\int T_{\mu\nu}\delta g^{\mu\nu}\sqrt{-g}\,\mathrm{d}^4x, \tag{1.3.7}$$

yielding the functional derivative definition:

$$T_{\mu\nu} \equiv -\frac{2}{\sqrt{-g}}\frac{\delta\left(\sqrt{-g}\mathcal{L}_{\text{mat}}\right)}{\delta g^{\mu\nu}} = g_{\mu\nu}\mathcal{L}_{\text{mat}} - 2\frac{\delta\mathcal{L}_{\text{mat}}}{\delta g^{\mu\nu}}. \tag{1.3.8}$$

Einstein's field equations

Requiring the total action variation $\delta S = \delta S_{\text{grav}} + \delta S_{\text{mat}} = 0$ for arbitrary metric variations $\delta g^{\mu\nu}$ yields the celebrated **Einstein Field Equations**:

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = \kappa T_{\mu\nu}, \tag{1.3.9}$$

where $\kappa \equiv 8\pi G_{\text{N}}/c^4$. In the famous words of John Archibald Wheeler:

> *"Spacetime tells matter how to move; matter tells spacetime how to curve."*

1.3.2 Conservation equations

Taking the covariant divergence of Einstein's field equations (1.3.9), the contracted Bianchi identity ($\nabla^\mu G_{\mu\nu} = 0$) and metric compatibility ($\nabla^\mu g_{\mu\nu} = 0$) require:

$$\nabla_\nu T^{\mu\nu} = 0. \tag{1.3.10}$$

Local continuity versus global conservation. It is common in introductory discussions to assume that the vanishing divergence $\nabla_\nu T^{\mu\nu} = 0$ expresses a global conservation law for the total energy of matter in the universe, analogous to energy conservation in classical mechanics.

This assumption is physically incorrect. In special relativity, where spacetime is flat, the conservation equation is $\partial_\nu T^{\mu\nu} = 0$. By Gauss' divergence theorem, integrating this equation over a spatial volume yields a strictly constant total four-momentum vector $P^\mu = \int T^{0\mu}\mathrm{d}^3x$, satisfying $\mathrm{d}P^\mu/\,\mathrm{d}t = 0$. In general relativity, however, the partial derivative is replaced by the covariant derivative:

$$\nabla_\nu T^{\mu\nu} = \partial_\nu T^{\mu\nu} + \Gamma^\mu_{\nu\lambda}T^{\lambda\nu} + \Gamma^\nu_{\nu\lambda}T^{\mu\lambda} = 0. \tag{1.3.11}$$

The connection terms Γ represent the continuous, dynamic exchange of energy and momentum between matter and the gravitational field. Because the gravitational field itself carries energy, one might hope to define a total stress-energy tensor that includes gravity. However, by the Equivalence Principle, gravity can always be made to vanish locally in a freely falling frame, meaning that gravitational energy cannot be localized into a genuine, coordinate-independent tensor. Consequently, in a dynamic, expanding spacetime (such as the Friedmann universe), there exists no timelike Killing vector field and therefore no global conservation law for matter energy. As a concrete cosmological example, photons of the cosmic microwave background continuously lose energy as the universe expands, their wavelengths stretching with the cosmic scale factor ($\lambda \propto a(t) \implies E \propto 1/a$). Local energy-momentum conservation holds exactly in every local inertial frame, but global energy conservation does not apply to an expanding cosmos.

Upon performing the scalar–vector–tensor (SVT) decomposition, the conservation equation

$$\nabla_\mu T^{\mu\nu} = 0 \tag{1}$$

implies that only the scalar (longitudinal) part contributes to the divergence. The vector and tensor parts drop out by virtue of the transversality conditions used in their definitions. For example, consider the decomposition of a spatial vector,

$$V_i = \partial_i V + V_i^{\mathrm{T}}, \qquad \partial^i V_i^{\mathrm{T}} = 0, \tag{2}$$

where V is the scalar part and V_i^{T} is the transverse vector part. Taking the divergence gives

$$\partial^i V_i = \partial^i \partial_i V + \underbrace{\partial^i V_i^{\mathrm{T}}}_{=0} \tag{3}$$

$$= \nabla^2 V. \tag{4}$$

Thus, the transverse vector part makes no contribution to the divergence.
Likewise, in the SVT decomposition of a symmetric tensor, the vector and tensor pieces are defined to be transverse. Therefore, when the conservation equation

$$\nabla_\mu T^{\mu\nu} = 0 \tag{5}$$

is evaluated, the terms involving the divergence of these transverse pieces vanish identically. Consequently, the divergence equation selects only the longitudinal (scalar) contributions.

Covariant 1+3 projection of conservation equations

To understand the physical dynamics of matter in curved spacetime, we project the covariant conservation equation $\nabla_\mu T^{\mu\nu} = 0$ along and orthogonal to the fluid four-velocity u^μ. Using the general rank-2 decomposition:

$$T_{\mu\nu} = \rho u_\mu u_\nu + P\gamma_{\mu\nu} + 2q_{(\mu}u_{\nu)} + \pi_{\mu\nu}, \tag{1.3.12}$$

we obtain two exact dynamical equations:

1. **Energy Conservation Equation (Parallel Projection $-u_\nu \nabla_\mu T^{\mu\nu} = 0$):**

$$\dot{\rho} + \bar{\nabla}_\mu q^\mu + (\rho + P)\Theta + 2q_\mu \dot{u}^\mu + \sigma^\mu{}_\nu \pi^\nu{}_\mu = 0. \tag{1.3.13}$$

   This is the first law of relativistic thermodynamics:

- $\dot{\rho}$: Rate of change of comoving energy density.
- $(\rho + P)\Theta$: Dilution of energy and $P\,\mathrm{d}V$ work done by isotropic volume expansion Θ.
- $\bar{\nabla}_\mu q^\mu$: Divergence of the spatial heat/energy flux.
- $\sigma^\mu{}_\nu \pi^\nu{}_\mu$: Viscous heating generated by anisotropic shearing against viscous stresses.

2. **Momentum Conservation Equation (Orthogonal Projection $\gamma^\lambda{}_\nu \nabla_\mu T^{\mu\nu} = 0$):**

$$\gamma^\lambda{}_\nu \dot{q}^\nu + \bar{\nabla}^\lambda P + \bar{\nabla}_\alpha \pi^{\alpha\lambda} + \frac{4}{3}\Theta q^\lambda + \left(\sigma^\lambda{}_\alpha + \omega^\lambda{}_\alpha\right) q^\alpha + (\rho + P)\dot{u}^\lambda + \pi^{\alpha\lambda}\dot{u}_\alpha = 0. \quad (1.3.14)$$

This is the relativistic generalization of the Navier–Stokes equation, expressing spatial momentum balance.

The perfect fluid

In the standard cosmological model, cosmic matter and radiation are modeled on large scales as an isotropic, non-viscous **perfect fluid**. Dissipative heat fluxes and shear stresses vanish identically ($q^\mu = 0, \pi_{\mu\nu} = 0$), so the stress-energy tensor simplifies to:

$$T_{\mu\nu} = (\rho + P)u_\mu u_\nu + P g_{\mu\nu}. \quad (1.3.15)$$

The exact conservation equations (1.3.13) and (1.3.14) then reduce to:

$$\dot{\rho} + (\rho + P)\Theta = 0, \quad (1.3.16)$$

$$(\rho + P)\dot{u}^\mu + \bar{\nabla}^\mu P = 0. \quad (1.3.17)$$

Equation (1.3.16) is the **relativistic continuity equation**. Equation (1.3.17) is the **relativistic Euler equation**.

Physical insights from the fluid equations.

- **Pressureless Dust ($P = 0$):** For cold non-relativistic matter (baryons and cold dark matter), the pressure gradient vanishes ($\bar{\nabla}^\mu P = 0$). Equation (1.3.17) forces the acceleration to vanish: $\dot{u}^\mu = 0$. **Pressureless dust particles follow exact geodesics of spacetime!**
- **Pressure Contributes to Inertia:** In the Euler equation (1.3.17), the effective inertial mass density is $(\rho + P)$, not ρ alone! Pressure carries kinetic energy and momentum, so high pressure makes a relativistic fluid *harder* to accelerate.

For an irrotational fluid flow ($\omega = 0$), substituting Einstein's field equations into the scalar Gauss relation (1.2.93) yields the **generalized Friedmann equation**:

$${}^{(3)}R = 2\kappa\rho - \frac{2}{3}\Theta^2 + 2\sigma^2 + 2\Lambda. \quad (1.3.18)$$

In a homogeneous and isotropic FLRW universe where shear vanishes ($\sigma = 0$) and $\Theta = 3H = 3\dot{a}/a$, this reduces directly to the famous Friedmann equation governing cosmic expansion.

In terms of energy density ρ and pressure P, the standard energy conditions of Table 1.1 translate to:

- **Null Energy Condition (NEC):** $\rho + P \geq 0$.
- **Weak Energy Condition (WEC):** $\rho \geq 0$ and $\rho + P \geq 0$.
- **Strong Energy Condition (SEC):** $\rho + 3P \geq 0$ and $\rho + P \geq 0$.
- **Dominant Energy Condition (DEC):** $\rho \geq |P|$.

Electromagnetism in curved spacetime

Electromagnetism is described by the four-potential A_μ, whose antisymmetric curl defines the **Faraday electromagnetic field tensor**:

$$F_{\mu\nu} \equiv \nabla_\mu A_\nu - \nabla_\nu A_\mu = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{1.3.19}$$

The homogeneous Maxwell equations are geometric Bianchi identities ($3F_{[\mu\nu;\alpha]} = 0$):

$$\nabla_\alpha F_{\mu\nu} + \nabla_\mu F_{\nu\alpha} + \nabla_\nu F_{\alpha\mu} = 0. \tag{1.3.20}$$

The electromagnetic action in curved spacetime is:

$$S_{\rm em} = -\frac{1}{4}\int F_{\mu\nu}F^{\mu\nu}\sqrt{-g}\,\mathrm{d}^4x + \int A_\mu j^\mu \sqrt{-g}\,\mathrm{d}^4x, \tag{1.3.21}$$

where j^μ is the conserved electromagnetic four-current density ($\nabla_\mu j^\mu = 0$). Varying with respect to A_μ yields the inhomogeneous curved-spacetime Maxwell equations:

$$\nabla_\mu F^{\nu\mu} = j^\nu \iff \frac{1}{\sqrt{-g}}\partial_\mu\left(\sqrt{-g}F^{\nu\mu}\right) = j^\nu. \tag{1.3.22}$$

In the **Lorenz gauge** ($\nabla_\mu A^\mu = 0$), commuting covariant derivatives $[\nabla_\mu, \nabla_\nu]A^\mu = R^\mu{}_\nu A^\nu$ yields the curved-spacetime wave equation:

$$\Box A^\mu - R^\mu{}_\nu A^\nu = -j^\mu. \tag{1.3.23}$$

Notice the direct coupling between the vector potential and the spacetime Ricci curvature $R^\mu{}_\nu$.

Varying Equation (1.3.21) with respect to the metric yields the electromagnetic stress-energy tensor:

$$T^{\mu\nu}_{\rm em} = F^\mu{}_\lambda F^{\nu\lambda} - \frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}. \tag{1.3.24}$$

In four dimensions, $T^{\mu\nu}_{\rm em}$ is **identically traceless**:

$$T^\mu{}_\mu = F^{\mu\lambda}F_{\mu\lambda} - \frac{1}{4}(4)F_{\alpha\beta}F^{\alpha\beta} \equiv 0. \tag{1.3.25}$$

This tracelessness is the hallmark of conformal invariance in classical electrodynamics.

An observer with four-velocity u^μ measures physical electric and magnetic field vectors:

$$E^\mu \equiv F^{\mu\nu}u_\nu, \qquad B^\mu \equiv -\frac{1}{2}\epsilon^{\mu\nu\rho\sigma}u_\nu F_{\rho\sigma}, \tag{1.3.26}$$

which are purely spatial ($E^\mu u_\mu = 0, B^\mu u_\mu = 0$). The full Faraday tensor reconstructs as:

$$F_{\mu\nu} = 2u_{[\mu}E_{\nu]} - \epsilon_{\mu\nu\rho\sigma}u^\rho B^\sigma. \tag{1.3.27}$$

The electromagnetic stress-energy tensor decomposes as an effective fluid with:

$$\rho_{\rm em} = \frac{1}{2}\left(E^2 + B^2\right), \qquad P_{\rm em} = \frac{1}{3}\rho_{\rm em}, \qquad q^\mu = \epsilon^{\mu\nu\rho\sigma}u_\nu E_\rho B_\sigma, \tag{1.3.28}$$

and anisotropic stress tensor $\pi_{\mu\nu} = \frac{1}{3}(E^2 + B^2)\gamma_{\mu\nu} - E_\mu E_\nu - B_\mu B_\nu$.

The trajectory of a test particle of charge e and mass m is governed by the curved-spacetime Lorentz force law:

$$u^\mu \nabla_\mu u^\nu = \frac{e}{m}F^\nu{}_\mu u^\mu = \frac{e}{m}E^\nu. \tag{1.3.29}$$

Projecting Maxwell's equations along u^μ and onto Σ_t yields the $1+3$ relativistic Maxwell equations:

$$\bar{\nabla}_\mu E^\mu - 2\omega_\mu B^\mu = \epsilon_e, \tag{1.3.30}$$

$$\bar{\nabla}_\mu B^\mu + 2\omega_\mu B^\mu = 0, \tag{1.3.31}$$

$$\gamma^\lambda{}_\mu \dot{E}^\mu - \epsilon^{\alpha\lambda\rho\sigma} u_\alpha \bar{\nabla}_\rho B_\sigma = -\gamma^\lambda{}_\nu j^\nu - \frac{2}{3}\Theta E^\lambda + \sigma^\lambda{}_\nu E^\nu + \epsilon^{\alpha\lambda\rho\sigma} u_\alpha \left(\dot{u}_\rho B_\sigma + \omega_\rho E_\sigma\right), \tag{1.3.32}$$

$$\gamma^\lambda{}_\mu \dot{B}^\mu + \epsilon^{\alpha\lambda\rho\sigma} u_\alpha \bar{\nabla}_\rho E_\sigma = -\frac{2}{3}\Theta B^\lambda + \sigma^\lambda{}_\nu B^\nu - \epsilon^{\alpha\lambda\rho\sigma} u_\alpha \left(\dot{u}_\rho E_\sigma - \omega_\rho B_\sigma\right), \tag{1.3.33}$$

where $\epsilon_e \equiv -j^\mu u_\mu$ is the comoving charge density.

Scalar fields

In modern cosmology, scalar fields are the central mechanism driving early-universe inflation and late-universe dark energy. A minimally coupled real scalar field ϕ with potential $V(\phi)$ has action:

$$S_\phi = -\int \left[\frac{1}{2} g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi + V(\phi)\right]\sqrt{-g}\,\mathrm{d}^4x. \tag{1.3.34}$$

Varying with respect to ϕ yields the curved-spacetime **Klein–Gordon equation**:

$$\Box\phi - V'(\phi) = 0 \iff \frac{1}{\sqrt{-g}}\partial_\mu\left(\sqrt{-g}g^{\mu\nu}\partial_\nu\phi\right) - V'(\phi) = 0. \tag{1.3.35}$$

Varying the action with respect to $g^{\mu\nu}$ yields the scalar stress-energy tensor:

$$T_{\mu\nu} = \partial_\mu\phi\partial_\nu\phi - g_{\mu\nu}\left[\frac{1}{2}g^{\alpha\beta}\partial_\alpha\phi\partial_\beta\phi + V(\phi)\right]. \tag{1.3.36}$$

When the gradient $\partial_\mu\phi$ is timelike, the scalar field can be mapped directly to an effective perfect fluid with four-velocity:

$$u_\mu \equiv -\frac{\partial_\mu\phi}{\psi}, \qquad \psi \equiv \dot{\phi} = \sqrt{-\partial_\mu\phi\partial^\mu\phi}. \tag{1.3.37}$$

The effective energy density and isotropic pressure are:

$$\rho_\phi = \frac{1}{2}\dot{\phi}^2 + V(\phi), \qquad P_\phi = \frac{1}{2}\dot{\phi}^2 - V(\phi). \tag{1.3.38}$$

Negative pressure and cosmic inflation. If potential energy dominates over kinetic energy ($V(\phi) \gg \frac{1}{2}\dot{\phi}^2$), the equation of state approaches:

$$w_\phi \equiv \frac{P_\phi}{\rho_\phi} \approx -1 \implies P_\phi \approx -\rho_\phi. \tag{1.3.39}$$

This violates the Strong Energy Condition ($\rho + 3P = -2\rho < 0$), causing the Raychaudhuri equation to flip sign ($\ddot{S}/S > 0$) and producing accelerated, quasi-exponential cosmic expansion: **cosmological inflation**.

For this scalar fluid, the vorticity vanishes identically ($\omega_{\mu\nu} = 0$), and energy conservation $\dot{\rho}_\phi + (\rho_\phi + P_\phi)\Theta = 0$ is exactly equivalent to the Klein–Gordon equation:

$$\ddot{\phi} + \Theta\dot{\phi} + V'(\phi) = 0. \tag{1.3.40}$$

The term $\Theta\dot{\phi} = 3H\dot{\phi}$ acts as a cosmic friction (Hubble drag) slowing the field as it rolls down its potential.

1.3.3 ADM Hamiltonian formulation of general relativity

General relativity is formulated covariantly as four-dimensional geometry. However, to simulate gravitational collapse or the merger of binary black holes on a computer, or to quantize gravity canonically, we need an **initial-value (Cauchy) formulation**: space evolving through time.

The ADM formulation (1962) foliates a globally hyperbolic spacetime into a family of spacelike hypersurfaces Σ_t parameterized by a global time coordinate t. The spacetime metric decomposes into the **ADM line element**:

$$\mathrm{d}s^2 = -N^2\,\mathrm{d}t^2 + \gamma_{ij}\left(\mathrm{d}x^i + N^i\,\mathrm{d}t\right)\left(\mathrm{d}x^j + N^j\,\mathrm{d}t\right), \tag{1.3.41}$$

parameterized by three fundamental objects:

- **Lapse Function ($N(t,\mathbf{x})$):** Measures the rate of flow of proper time relative to coordinate time normal to the slice: $\mathrm{d}\tau = N\,\mathrm{d}t$.
- **Shift Vector ($N^i(t,\mathbf{x})$):** Measures the transverse coordinate drift of spatial grid lines between successive hypersurfaces Σ_t and $\Sigma_{t+\mathrm{d}t}$.
- **Spatial 3-Metric ($\gamma_{ij}(t,\mathbf{x})$):** Measures intrinsic physical distances within the spatial hypersurface Σ_t.

The unit timelike normal to Σ_t is $n_\alpha = (-N,0,0,0)$ with $n^\alpha = \frac{1}{N}(1,-N^i)$, satisfying $n^\alpha n_\alpha = -1$. The four-metric determinant satisfies $\sqrt{-g} = N\sqrt{\gamma}$, where $\gamma \equiv \det(\gamma_{ij})$.

The extrinsic curvature of the spatial slices is given in ADM variables by:

$$K_{ij} = \frac{1}{2N}\left(D_i N_j + D_j N_i - \dot{\gamma}_{ij}\right), \tag{1.3.42}$$

where D_i denotes the 3-dimensional spatial covariant derivative compatible with γ_{ij}.

In terms of ADM variables, the gravitational action takes the form:

$$S_{\mathrm{grav}} = \int \mathrm{d}t \int_{\Sigma_t} N\left({}^{(3)}R + K_{ij}K^{ij} - K^2\right)\sqrt{\gamma}\,\mathrm{d}^3x. \tag{1.3.43}$$

The Lagrangian density contains *no time derivatives* of the lapse N or shift N^i. Therefore, N and N^i are not dynamical variables; they are non-dynamical **Lagrange multipliers**. The only true dynamical degrees of freedom are the spatial metric components γ_{ij}.

The canonical momentum tensor conjugate to γ_{ij} is:

$$\pi^{ij} \equiv \frac{\partial \mathcal{L}}{\partial \dot{\gamma}_{ij}} = \sqrt{\gamma}\left(K\gamma^{ij} - K^{ij}\right). \tag{1.3.44}$$

Performing a Legendre transformation, the ADM Hamiltonian takes the form:

$$H = -\int_{\Sigma_t}\left(N\mathcal{C}_0 - 2N^i\mathcal{C}_i\right)\sqrt{\gamma}\,\mathrm{d}^3x, \tag{1.3.45}$$

where varying with respect to the Lagrange multipliers N and N^i produces the foundational **ADM constraint equations**:

$$\mathcal{C}_0 \equiv {}^{(3)}R + K^2 - K_{ij}K^{ij} = 16\pi G_{\mathrm{N}}\rho \qquad \text{(Hamiltonian / Energy Constraint)}, \tag{1.3.46}$$

$$\mathcal{C}_i \equiv D_j K^j{}_i - D_i K = 8\pi G_{\mathrm{N}} P_i \qquad \text{(Momentum Constraint)}, \tag{1.3.47}$$

with energy density $\rho \equiv T_{\mu\nu}n^\mu n^\nu$ and momentum density $P_i \equiv -\gamma^\mu{}_i T_{\mu\nu}n^\nu$. The constraints show that initial data (γ_{ij}, K_{ij}) cannot be chosen arbitrarily on an initial spatial slice: they must satisfy four non-linear elliptic partial differential constraint equations.

1.4 Geodesics in curved space-time

1.4.1 Conserved quantities along geodesics

Solving the non-linear geodesic equations $\ddot{x}^\mu + \Gamma^\mu_{\nu\lambda}\dot{x}^\nu\dot{x}^\lambda = 0$ directly is challenging. However, whenever spacetime possesses a continuous symmetry generated by a Killing vector field ξ^μ ($\nabla_{(\mu}\xi_{\nu)} = 0$), there exists a strictly conserved scalar along any geodesic:

$$u^\nu\nabla_\nu\left(u^\mu\xi_\mu\right) = \xi_\mu\left(u^\nu\nabla_\nu u^\mu\right) + u^\mu u^\nu\nabla_\nu\xi_\mu = 0. \tag{1.4.1}$$

The first term vanishes because the path is geodesic ($u^\nu\nabla_\nu u^\mu = 0$), and the second vanishes because the symmetric tensor $u^\mu u^\nu$ contracts with the antisymmetric Killing tensor $\nabla_{(\mu}\xi_{\nu)} = 0$.

Physical constants of motion.

- **Stationary Spacetime ($\xi = \partial_t$):** The metric is time-independent. The conserved quantity $E \equiv -p_\mu\xi^\mu = -p_0$ is the **conserved energy** of the particle.
- **Axisymmetric Spacetime ($\xi = \partial_\phi$):** The metric is rotationally symmetric around an axis. The conserved quantity $L \equiv p_\mu\xi^\mu = p_\phi$ is the **conserved angular momentum**.

Similarly, for any conserved matter stress-energy tensor $\nabla_\mu T^{\mu\nu} = 0$, the vector current $J^\mu \equiv T^{\mu\nu}\xi_\nu$ is covariantly conserved:

$$\nabla_\mu\left(T^{\mu\nu}\xi_\nu\right) = \xi_\nu\nabla_\mu T^{\mu\nu} + T^{\mu\nu}\nabla_\mu\xi_\nu = 0. \tag{1.4.2}$$

Integrating over a spatial volume yields a strictly constant conserved charge:

$$Q \equiv \int_V T^{0\nu}\xi_\nu\sqrt{-g}\,\mathrm{d}^3x, \qquad \frac{\mathrm{d}Q}{\mathrm{d}t} = 0. \tag{1.4.3}$$

1.4.2 Geodesic deviation and tidal forces

How can an observer inside an enclosed freely falling spaceship detect that spacetime is curved without looking outside? By tracking the relative acceleration between two adjacent freely falling particles.

Consider a one-parameter family of neighboring geodesics $x^\mu(\lambda, s)$, parameterized by affine parameter λ along each curve and labeled by s. The **separation vector** (deviation vector) between adjacent worldlines is $n^\mu \equiv \frac{\partial x^\mu}{\partial s}$, satisfying the commutation relation:

$$u^\mu\nabla_\mu n^\nu = n^\mu\nabla_\mu u^\nu. \tag{1.4.4}$$

The relative acceleration between neighboring geodesics is $a^\mu \equiv \frac{\mathrm{D}^2 n^\mu}{\mathrm{D}\lambda^2} = u^\alpha\nabla_\alpha(u^\beta\nabla_\beta n^\mu)$. Commuting covariant derivatives using Equation (1.4.4) and the Ricci identity yields the **equation of geodesic deviation** (the Jacobi equation):

$$\frac{\mathrm{D}^2 n^\mu}{\mathrm{D}\lambda^2} = R^\mu{}_{\nu\alpha\beta}u^\nu u^\alpha n^\beta = -R^\mu{}_{\nu\beta\alpha}u^\nu n^\beta u^\alpha. \tag{1.4.5}$$

The relative acceleration of test particles is directly proportional to the Riemann curvature tensor. Geodesic deviation is the exact general relativistic formulation of Newtonian tidal forces.

Covariant kinematic deviation and the generalized Hubble law

Decomposing the spatial separation vector as $n^\mu = \ell e^\mu$, where ℓ is the physical separation length and e^μ is a unit spatial vector ($e^\mu e_\mu = 1, e^\mu u_\mu = 0$), evaluating the rate of change of ℓ along the flow yields the **generalized Hubble law**:

$$\frac{\dot{\ell}}{\ell} = \frac{1}{3}\Theta + \sigma_{\mu\nu} e^\mu e^\nu. \tag{1.4.6}$$

The relative velocity between neighboring particles receives an isotropic expansion contribution $\frac{1}{3}\Theta$ and an anisotropic shear contribution $\sigma_{\mu\nu} e^\mu e^\nu$.

Projecting orthogonally describes the drift in the line-of-sight observation direction:

$$\gamma^\mu{}_\lambda \dot{e}^\lambda = \left[\sigma^\mu{}_\lambda - \left(\sigma_{\alpha\beta} e^\alpha e^\beta\right)\gamma^\mu{}_\lambda - \omega^\mu{}_\lambda\right] e^\lambda. \tag{1.4.7}$$

Geometric optics and cosmological redshift

In the geometric optics limit (where the photon wavelength is much smaller than the radius of curvature of spacetime), electromagnetic radiation propagates along **null geodesics**:

$$k_\mu k^\mu = 0, \qquad k^\mu \nabla_\mu k^\nu = 0, \qquad C_\mu k^\mu = 0, \tag{1.4.8}$$

where $k^\mu = \mathrm{d}x^\mu / \mathrm{d}p$ is the photon wavevector (with affine parameter p) and C_μ is the polarization vector.

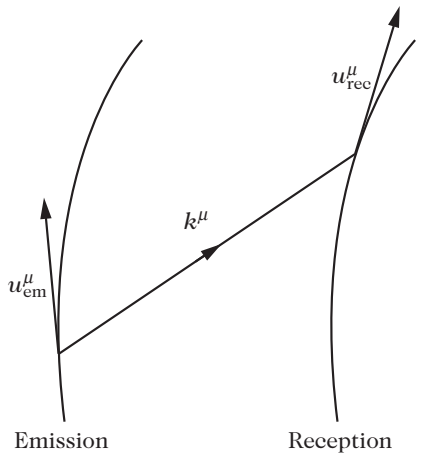


Figure 1.8: A photon emitted with wavevector k^μ by an emitter with four-velocity u^μ_{em} travels along a null geodesic through curved spacetime to a receiver with four-velocity u^μ_{rec}.

An observer with four-velocity u^μ measures a photon energy and frequency given by the scalar projection:

$$E = \hbar\omega = -k_\mu u^\mu. \tag{1.4.9}$$

If a photon is emitted at event em and received at event rec (Figure 1.8), the observed **cosmological redshift** z is defined by:

$$1 + z \equiv \frac{\lambda_{\mathrm{rec}}}{\lambda_{\mathrm{em}}} = \frac{\omega_{\mathrm{em}}}{\omega_{\mathrm{rec}}} = \frac{(k_\mu u^\mu)_{\mathrm{em}}}{(k_\mu u^\mu)_{\mathrm{rec}}}. \tag{1.4.10}$$

Decomposing the null wavevector as $k^\mu = E(u^\mu + e^\mu)$, where e^μ is the unit spatial direction vector ($e^\mu e_\mu = 1, e^\mu u_\mu = 0$), the rate of change of wavelength along the ray is given by the exact evolution formula:

$$\frac{1}{\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}p} = -\frac{1}{E}\frac{\mathrm{d}E}{\mathrm{d}p} = \left(\frac{1}{3}\Theta + \dot{u}_\mu e^\mu + \sigma_{\mu\nu} e^\mu e^\nu\right) E. \tag{1.4.11}$$

Anatomy of the cosmological redshift. Equation (1.4.11) reveals that cosmological redshift is a combination of three distinct physical phenomena:

- **Isotropic Expansion ($\frac{1}{3}\Theta = H$):** The stretching of light's wavelength caused by the overall volume expansion of the universe.
- **Kinematic Doppler Effect ($\dot{u}_\mu e^\mu$):** Frequency shifts caused by the acceleration or peculiar motion of the emitter and observer.
- **Gravitational Shear ($\sigma_{\mu\nu} e^\mu e^\nu$):** Anisotropic tidal distortions along the photon's line of sight.

1.5 The weak-field regime

1.5.1 The Newtonian limit

General relativity is a non-linear geometric theory. In our Solar System and across everyday astrophysics, however, Newton's inverse-square law $\mathbf{g} = -\boldsymbol{\nabla}\Phi_{\rm N}$ works with extraordinary precision. To understand how the non-linear curvature of spacetime reduces smoothly to Newtonian gravity, and to determine the numerical value of the gravitational coupling constant κ, we analyze the theory in the limit of weak fields and non-relativistic motion.

We expand the spacetime metric as a small perturbation around flat Minkowski spacetime:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \qquad |h_{\mu\nu}| \ll 1. \tag{1.5.1}$$

Weak fields versus the Newtonian limit. A frequent point of confusion is to conflate a "weak gravitational field" with the "Newtonian limit". The two concepts are physically distinct. A spacetime is said to be in the weak-field regime whenever the metric can be decomposed as a small perturbation around flat Minkowski spacetime, as in Equation (1.5.1). However, a weak field is not by itself sufficient to recover Newtonian gravity. Gravitational waves detected on Earth, for example, have microscopic amplitudes $|h_{\mu\nu}| \sim 10^{-21}$—representing an extraordinarily weak field—yet they propagate at the speed of light ($v = c$) and oscillate rapidly in time, phenomena completely foreign to Newtonian physics.

The true **Newtonian limit** requires three distinct physical conditions to hold simultaneously:

1. **Weak Gravitational Field:** The metric perturbation is much smaller than unity, $|h_{\mu\nu}| \sim |\Phi_{\rm N}/c^2| \ll 1$ (near Earth's surface, $\Phi_{\rm N}/c^2 \sim 10^{-9}$).
2. **Slow Motion:** The velocities of test particles and source matter are non-relativistic compared to light: $v \ll c$.
3. **Quasi-Static Fields:** Time derivatives of the gravitational field are negligibly small compared to spatial gradients: $|\partial_0 h_{\mu\nu}| \ll |\partial_i h_{\mu\nu}|$.

Linearized Einstein equations

To first order in the metric perturbation $h_{\mu\nu}$, the inverse metric is $g^{\mu\nu} = \eta^{\mu\nu} - h^{\mu\nu}$, where indices are raised and lowered using the flat background metric $\eta_{\mu\nu}$. The trace is $h \equiv h^\alpha{}_\alpha = \eta^{\alpha\beta} h_{\alpha\beta}$.

The linearized Christoffel symbols are:

$$\Gamma^\alpha_{\mu\nu} = \frac{1}{2}\eta^{\alpha\lambda}\left(\partial_\mu h_{\lambda\nu} + \partial_\nu h_{\lambda\mu} - \partial_\lambda h_{\mu\nu}\right). \tag{1.5.2}$$

The linearized Ricci tensor and Ricci scalar are:

$$R_{\mu\nu} = \partial_\alpha \Gamma^\alpha_{\mu\nu} - \partial_\mu \Gamma^\alpha_{\alpha\nu} = \frac{1}{2}\left(2\partial^\lambda \partial_{(\mu} h_{\nu)\lambda} - \Box h_{\mu\nu} - \partial_\mu \partial_\nu h\right), \tag{1.5.3}$$

$$R = \partial_\alpha \partial_\beta h^{\alpha\beta} - \Box h, \tag{1.5.4}$$

where $\Box \equiv \eta^{\alpha\beta}\partial_\alpha\partial_\beta = -\partial_0^2 + \nabla^2$.

Introducing the **trace-reversed metric perturbation**:

$$\bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \frac{1}{2} h \eta_{\mu\nu} \iff \bar{h} = -h, \tag{1.5.5}$$

the linearized Einstein equations $G_{\mu\nu} = \kappa T_{\mu\nu}$ take the form:

$$\Box \bar{h}_{\mu\nu} + \eta_{\mu\nu}\partial_\alpha\partial_\beta \bar{h}^{\alpha\beta} - \partial^\lambda \partial_\mu \bar{h}_{\nu\lambda} - \partial^\lambda \partial_\nu \bar{h}_{\mu\lambda} = -2\kappa T_{\mu\nu}. \tag{1.5.6}$$

Gauge freedom and the harmonic gauge

Under an infinitesimal coordinate transformation $x^\mu \to x'^\mu = x^\mu + \xi^\mu(x)$, the metric perturbation transforms as:

$$h_{\mu\nu} \longrightarrow h'_{\mu\nu} = h_{\mu\nu} - \partial_\mu \xi_\nu - \partial_\nu \xi_\mu \implies \bar{h}_{\mu\nu} \longrightarrow \bar{h}'_{\mu\nu} = \bar{h}_{\mu\nu} - \partial_\mu \xi_\nu - \partial_\nu \xi_\mu + \eta_{\mu\nu}\partial_\alpha \xi^\alpha. \tag{1.5.7}$$

This gauge freedom represents arbitrary coordinate re-labeling.

We can always choose the four gauge parameters ξ^μ satisfying $\Box \xi^\mu = \partial_\nu \bar{h}^{\mu\nu}$ to impose the **harmonic gauge condition** (also known as the **Lorenz gauge** or **de Donder gauge**):

$$\partial^\nu \bar{h}_{\mu\nu} = 0. \tag{1.5.8}$$

In the harmonic gauge, all gradient terms in Equation (1.5.6) drop out, decoupling the linearized Einstein equations into simple wave equations:

$$\Box \bar{h}_{\mu\nu} = -2\kappa T_{\mu\nu}. \tag{1.5.9}$$

Retarded potential solutions

Equation (1.5.9) is solved using the standard retarded Green's function for the d'Alembertian operator:

$$\bar{h}_{\mu\nu}(\mathbf{x}, t) = \frac{\kappa}{2\pi} \int \frac{T_{\mu\nu}\left(\mathbf{x}', t - |\mathbf{x} - \mathbf{x}'|/c\right)}{|\mathbf{x} - \mathbf{x}'|} \, \mathrm{d}^3 x'. \tag{1.5.10}$$

The general solution is the sum of the retarded particular solution $\bar{h}^{(\mathrm{ret})}_{\mu\nu}$ and a homogeneous solution $\bar{h}^{(\mathrm{hom})}_{\mu\nu}$ satisfying $\Box \bar{h}^{(\mathrm{hom})}_{\mu\nu} = 0$, which describes freely propagating **gravitational waves**.

Weak-field geodesic motion and determination of κ

Consider a non-relativistic test body moving with velocity $v^i = \mathrm{d}x^i / \mathrm{d}t \ll c$ in a weak, static gravitational field. The particle's four-velocity is $u^\mu \approx (c, v^i)$. The spatial component of the geodesic equation (1.2.33) reduces to leading order in v/c to:

$$\frac{\mathrm{d}^2 x^i}{\mathrm{d}t^2} = -c^2 \left[\Gamma^i_{00} + \left(\Gamma^0_{00}\delta^i_j - \frac{2}{c}\Gamma^i_{0j} \right) v^j \right]. \tag{1.5.11}$$

For a static field ($\partial_0 h_{\mu\nu} = 0$), the only non-vanishing Christoffel symbol at leading order is $\Gamma^i_{00} = -\frac{1}{2}\delta^{ij}\partial_j h_{00}$, yielding:

$$\frac{\mathrm{d}^2 x^i}{\mathrm{d}t^2} = \frac{c^2}{2}\delta^{ij}\partial_j h_{00}. \tag{1.5.12}$$

Comparing this directly with Newton's second law $\frac{\mathrm{d}^2 x^i}{\mathrm{d}t^2} = -\delta^{ij}\partial_j \Phi_\mathrm{N}$, we make the fundamental identification:

$$h_{00} = -\frac{2\Phi_\mathrm{N}}{c^2} \implies g_{00} = -\left(1 + \frac{2\Phi_\mathrm{N}}{c^2}\right), \qquad \bar{h}_{00} = -\frac{4\Phi_\mathrm{N}}{c^2}. \tag{1.5.13}$$

For a non-relativistic source ($P \ll \rho c^2$), the dominant stress-energy component is $T_{00} \approx \rho c^2$. In the static limit ($\Box \to \nabla^2$), the 00-component of the linearized field equation (1.5.9) becomes:

$$\nabla^2 \bar{h}_{00} = -2\kappa T_{00} \implies -\frac{4}{c^2}\nabla^2\Phi_{\rm N} = -2\kappa\rho c^2 \iff \nabla^2\Phi_{\rm N} = \frac{1}{2}\kappa c^4 \rho. \tag{1.5.14}$$

Comparing with Newton's Poisson equation $\nabla^2\Phi_{\rm N} = 4\pi G_{\rm N}\rho$, we determine the fundamental coupling constant of general relativity:

$$\kappa = \frac{8\pi G_{\rm N}}{c^4}. \tag{1.5.15}$$

Gravito-electromagnetism

The linearized Einstein equations display a striking structural analogy with Maxwell's electrodynamics. Defining the **gravito-magnetic vector potential** $A_i \equiv c\,\bar{h}_{0i}$, the non-relativistic geodesic equation takes the form of a generalized Lorentz force:

$$\mathbf{a} = -\boldsymbol{\nabla}\Phi_{\rm N} - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} + \mathbf{v} \times (\boldsymbol{\nabla} \times \mathbf{A}) + \frac{3}{c^2}\frac{\partial \Phi_{\rm N}}{\partial t}\mathbf{v}. \tag{1.5.16}$$

Defining the gravito-electric field $\mathbf{g} \equiv -\boldsymbol{\nabla}\Phi_{\rm N} - \frac{1}{c}\partial_t \mathbf{A}$ and gravito-magnetic field $\mathbf{B}_{\rm g} \equiv \boldsymbol{\nabla} \times \mathbf{A}$, the linearized field equations mirror Maxwell's equations (Table 1.2):

$$\boldsymbol{\nabla} \cdot \mathbf{B}_{\rm g} = 0, \qquad \boldsymbol{\nabla} \times \mathbf{g} = -\frac{1}{c}\frac{\partial \mathbf{B}_{\rm g}}{\partial t}, \tag{1.5.17}$$

$$\boldsymbol{\nabla} \cdot \mathbf{g} = -4\pi G_{\rm N}\rho, \qquad \boldsymbol{\nabla} \times \mathbf{B}_{\rm g} = -\frac{16\pi G_{\rm N}}{c^2}\mathbf{j}_m + \frac{1}{c}\frac{\partial \mathbf{g}}{\partial t}, \tag{1.5.18}$$

where $\mathbf{j}_m \equiv \rho\mathbf{v}$ is the mass-current density. A rotating mass generates a gravito-magnetic field that drags inertial frames—the **Lense–Thirring effect** (frame dragging), confirmed by the Gravity Probe B satellite experiment.

Table 1.2: Analogy between Electromagnetism and Linearized General Relativity

Property	**Electromagnetism**	**Linearized Gravity**
Potential Field	$A^\mu = (\Phi/c, \mathbf{A})$	$\bar{h}^{\mu\nu}$
Gauge Transformation	$A_\mu \to A_\mu + \partial_\mu \chi$	$\bar{h}_{\mu\nu} \to \bar{h}_{\mu\nu} - 2\partial_{(\mu}\xi_{\nu)} + \eta_{\mu\nu}\partial_\alpha \xi^\alpha$
Gauge Condition	$\partial_\mu A^\mu = 0$ (Lorenz)	$\partial_\mu \bar{h}^{\mu\nu} = 0$ (Harmonic / de Donder)
Wave Equations	$\Box A^\mu = -\mu_0 j^\mu$	$\Box \bar{h}^{\mu\nu} = -2\kappa T^{\mu\nu}$
Residual Gauge Invariance	$\Box \chi = 0$	$\Box \xi^\mu = 0$

Scalar-Vector-Tensor (SVT) decomposition

The metric perturbation $h_{\mu\nu}$ contains 10 degrees of freedom, but 4 of them are arbitrary coordinate choices (gauge freedom). In cosmology, different physical perturbations (density clumping, rotational eddies, gravitational waves) evolve completely independently at linear order. To analyze their evolution cleanly, we decompose the metric perturbations under the spatial rotation group SO(3) of the background flat space into irreducible scalar, vector, and tensor components:

$$h_{00} = -2A, \tag{1.5.19}$$

$$h_{0i} = \partial_i B + \bar{B}_i, \tag{1.5.20}$$

$$h_{ij} = 2C\delta_{ij} + 2\partial_i\partial_j E + 2\partial_{(i}\bar{E}_{j)} + \bar{E}_{ij}, \tag{1.5.21}$$

where $\bar{B}_i$ and $\bar{E}_i$ are transverse divergence-free vector fields ($\partial_i\bar{B}^i = 0, \partial_i\bar{E}^i = 0$), and $\bar{E}_{ij}$ is a symmetric, transverse, and traceless (TT) tensor field ($\partial^i\bar{E}_{ij} = 0, \bar{E}^i{}_i = 0$).

Under rotation $\mathcal{R}_i{}^j$,

$$\begin{aligned} h_{00} &= h_{00} && \text{(transforms like scalar)} \\ h_{0i} &= \mathcal{R}_i{}^j h_{0j} && \text{(ransforms like vector)} \\ h_{ij} &= \mathcal{R}_i{}^a\mathcal{R}_j{}^b h_{ab} && \text{(transforms like tensor)} \end{aligned}$$

The 10 independent components of the symmetric metric perturbation $h_{\mu\nu}$ decompose cleanly into three uncoupled sectors:

- **4 Scalar Degrees of Freedom (A, B, C, E):** Represent longitudinal potential wells and isotropic perturbations.
- **4 Vector Degrees of Freedom ($\bar{B}_i, \bar{E}_i$):** Two transverse vectors ($2 \times 2 = 4$ components), representing rotational velocity fields and frame dragging.
- **2 Tensor Degrees of Freedom ($\bar{E}_{ij}$):** One symmetric transverse-traceless matrix ($\frac{3\times4}{2} - 3 - 1 = 2$ components), representing true propagating gravitational waves.

Gauge invariance and the Bardeen potentials. Under an infinitesimal coordinate gauge transformation $x^0 \to x^0 + T$ and $x^i \to x^i + \partial^i L + \bar{L}^i$ (with $\partial_i\bar{L}^i = 0$), the metric perturbations transform as:

$$A \longrightarrow A + \dot{T}, \qquad B \longrightarrow B + \dot{L} - T, \qquad C \longrightarrow C, \qquad E \longrightarrow E + L, \tag{1.5.22}$$

$$\bar{B}_i \longrightarrow \bar{B}_i + \dot{\bar{L}}_i, \qquad \bar{E}_i \longrightarrow \bar{E}_i + \bar{L}_i, \qquad \bar{E}_{ij} \longrightarrow \bar{E}_{ij}. \tag{1.5.23}$$

The tensor perturbations $\bar{E}_{ij}$ are **automatically gauge-invariant**: gravitational waves cannot be created or destroyed by changing coordinates.

For the scalar and vector sectors, James Bardeen (1980) constructed gauge-invariant combinations that remain invariant under any coordinate re-labeling:

$$\Phi \equiv A + \dot{B} - \ddot{E} \qquad \text{(Bardeen Scalar Potential)}, \tag{1.5.24}$$

$$\Psi \equiv -C \qquad \text{(Bardeen Scalar Potential)}, \tag{1.5.25}$$

$$\bar{\Phi}_i \equiv \dot{\bar{E}}_i - \bar{B}_i \qquad \text{(Gauge-Invariant Vector Perturbation)}. \tag{1.5.26}$$

The four gauge parameters $(T, L, \bar{L}_i)$ eliminate 4 degrees of freedom, leaving $10 - 4 = \mathbf{6}$ physical degrees of freedom: 2 scalar modes (Φ, Ψ), 2 vector modes $(\bar{\Phi}_i)$, and 2 tensor modes $(\bar{E}_{ij})$. In the absence of anisotropic stress, Einstein's field equations enforce $\Phi = \Psi$, reducing the scalar sector to a single Newtonian potential.

1.5.2 Gravitational waves in vacuum

In vacuum ($T_{\mu\nu} = 0$), what propagating degrees of freedom does the gravitational field possess?

Evaluating the linearized vacuum Einstein equations for the SVT modes:

$$\nabla^2\Psi = 0, \qquad \partial_i\dot{\Psi} = 0, \qquad \Phi - \Psi = 0 \implies \Phi = \Psi = 0, \tag{1.5.27}$$

$$\nabla^2\bar{\Phi}_i = 0 \implies \bar{\Phi}_i = 0. \tag{1.5.28}$$

Subject to regular boundary conditions at infinity, the scalar and vector perturbations vanish identically. **There are no propagating scalar or vector gravitational waves in general relativity.**

The only propagating gravitational degrees of freedom in vacuum are the transverse-traceless tensor modes, which satisfy the sourceless wave equation:

$$\Box \bar{E}_{ij} = 0, \qquad \bar{E}^{i}_{\ i} = 0, \qquad \partial^{i} \bar{E}_{ij} = 0. \tag{1.5.29}$$

In the **Transverse-Traceless (TT) gauge**, the metric perturbation is purely spatial and traceless: $h^{\mathrm{TT}}_{0\mu} = 0, h^{\mathrm{TT}}_{ij} = \bar{E}_{ij}$.

Polarization modes of gravitational waves

For a monochromatic plane gravitational wave propagating in the $+z$ direction with frequency $\omega = ck$, the general solution in the TT gauge is a linear superposition of two orthogonal polarization states:

$$h^{\mathrm{TT}}_{ij}(t, z) = E_{+}(ct - z)e^{+}_{ij} + E_{\times}(ct - z)e^{\times}_{ij}, \tag{1.5.30}$$

with the orthogonal unit polarization tensors:

$$e^{+}_{ij} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad e^{\times}_{ij} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{1.5.31}$$

These two polarization states correspond to the two helicity states (± 2) of the massless spin-2 graviton.

Effect on test particles

How does a passing gravitational wave affect physical matter? Consider a circular ring of free test particles located in the xy-plane at $z = 0$. The relative tidal acceleration between neighboring particles is governed by the geodesic deviation equation (1.4.5):

$$\frac{\mathrm{d}^2 n^i}{\mathrm{d}t^2} = \frac{1}{2} \frac{\partial^2 h^{\mathrm{TT}}_{ij}}{\partial t^2} n^j. \tag{1.5.32}$$

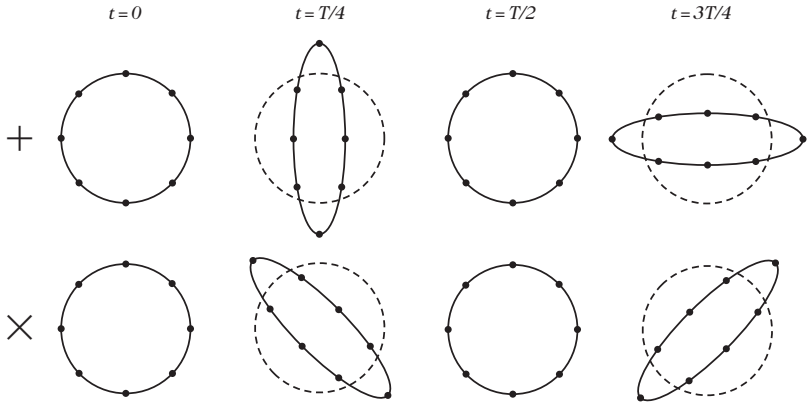


Figure 1.9: Quadrupolar deformation of a circular ring of test particles in the xy-plane caused by a gravitational plane wave propagating in the $+z$ direction. (Top): The $+$ polarization stretches along the x-axis while compressing along the y-axis, reversing half a period later. (Bottom): The $\times$ polarization produces an identical deformation rotated by 45°.

Integrating Equation (1.5.32) over one wave period $T = 2\pi/\omega$ yields the distinctive quadrupolar tidal deformations shown in Figure 1.9:

- **Plus (+) Polarization:** Stretches the ring along the x-axis while squeezing it along the y-axis during the first half-cycle, and reverses the distortion during the second half-cycle.
- **Cross (×) Polarization:** Produces the exact same quadrupolar deformation, rotated by 45°.

This quadrupolar distortion is the operational principle of laser interferometer gravitational wave detectors such as LIGO, Virgo, and KAGRA.

1.6 Experimental tests of general relativity

General relativity replaces Newtonian gravity with curved four-dimensional spacetime geometry. Clocks tick at different rates depending on gravitational potential, light bends around mass, and moving masses radiate ripples across spacetime. Because cosmological models extrapolate these laws across the entire observable Universe, every cornerstone of the theory must be subjected to multifaceted experimental scrutiny.

1.6.1 Tests of the Equivalence Principle

Universality of free fall (Weak Equivalence Principle)

The mass-energy of any physical body receives contributions from rest mass, electromagnetic atomic binding energy, strong nuclear interactions, and gravitational self-energy:

$$m_{\mathrm{G}} = m_{\mathrm{I}} + \sum_i \eta_i \frac{E_i}{c^2}. \tag{1.6.1}$$

If any fundamental interaction i couples differently to gravitational mass than to inertial mass ($\eta_i \neq 0$), bodies of different chemical compositions will fall with slightly different accelerations: $\mathbf{a} = (1 + \sum_i \eta_i \frac{E_i}{m_{\mathrm{I}}c^2})\mathbf{g}$.

The fractional difference in acceleration between two test bodies A and B is quantified by the dimensionless **Eötvös parameter** η:

$$\eta \equiv \frac{2|\mathbf{a}_A - \mathbf{a}_B|}{|\mathbf{a}_A + \mathbf{a}_B|} = \sum_i \eta_i \left[\left(\frac{E_i}{m_{\mathrm{I}}c^2} \right)_A - \left(\frac{E_i}{m_{\mathrm{I}}c^2} \right)_B \right]. \tag{1.6.2}$$

Experimental methods for WEP.

1. **Free Pendulums:** Measuring the period $T = 2\pi\sqrt{\frac{m_{\mathrm{I}}}{m_{\mathrm{G}}}\frac{L}{g}}$ for pendulums of different compositions yielded Newton's original bound $\eta < 10^{-3}$, improved to 10^{-5} by Bessel (1832).
2. **Torsion Balances (Figure 1.10):** Loránd Eötvös (1909) mounted two different materials on opposite ends of a beam suspended by a delicate torsion wire. The balance experiences gravitational acceleration $\mathbf{g}$ toward Earth's center and centrifugal acceleration $\mathbf{c} = \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{r})$ directed outward perpendicular to Earth's rotation axis. Any composition-dependent difference creates a net torque on the suspension fiber:

$$M = \eta \frac{L}{|\mathbf{g} + \mathbf{c}|} (\mathbf{e}_b \times \mathbf{c}) \cdot \mathbf{g}. \tag{1.6.3}$$

   Modern rotating torsion balances (such as the Eöt-Wash group at the University of Washington) push this precision to $\eta < 10^{-13}$, while the MICROSCOPE satellite experiment (2022) achieved the bound $\eta < 1.5 \times 10^{-15}$ in low-Earth orbit.

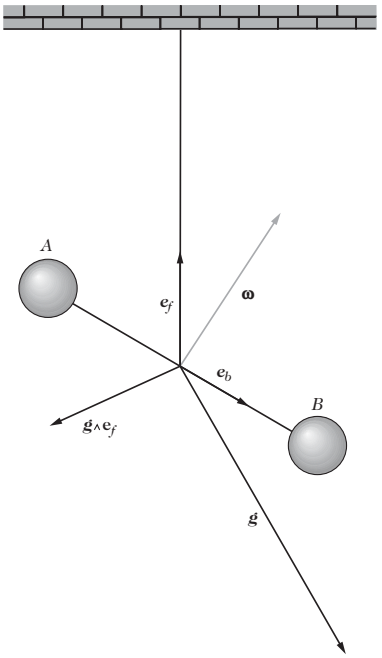


Figure 1.10: Principle of the Eötvös torsion balance. Two test bodies A and B of different materials are mounted at opposite ends of a beam suspended by a fine torsion fiber $\mathbf{e}_f$. If the bodies have different ratios of gravitational to inertial mass, the combination of gravitational acceleration $\mathbf{g}$ and Earth's centrifugal acceleration $\mathbf{c}$ exerts a net twisting torque on the fiber.

Lunar Laser Ranging and the Strong Equivalence Principle (SEP). In laboratory test masses, gravitational self-energy is negligible ($E_{\text{grav}}/mc^2 \sim 10^{-25}$). Testing the **Strong Equivalence Principle** requires planetary-sized bodies where self-gravity is significant ($E_{\text{grav}}/mc^2 \approx -4.6 \times 10^{-10}$ for Earth, -0.2×10^{-10} for the Moon).

Kenneth Nordtvedt (1968) showed that in alternative theories of gravity where the SEP is violated, the Earth and the Moon would fall toward the Sun with slightly different accelerations (the **Nordtvedt effect**), polarizing the lunar orbit toward the Sun:

$$\delta a \leq 2.6 \times 10^{-12} \eta_{\text{SEP}}\, \text{m} \cdot \text{s}^{-2} \implies \delta r_{\max} \leq -14 \eta_{\text{SEP}}\, \text{m}. \tag{1.6.4}$$

By firing laser pulses at corner-cube retroreflectors left on the Moon by Apollo astronauts and Soviet Lunokhod rovers, **Lunar Laser Ranging (LLR)** measures the Earth–Moon distance over decades with millimeter precision. The data confirms:

$$\eta_{\text{SEP}} = (-0.7 \pm 1.4) \times 10^{-13}, \tag{1.6.5}$$

verifying the Strong Equivalence Principle: gravitational binding energy falls at the exact same rate as ordinary matter.

Local Lorentz Invariance (LLI)

Local Lorentz Invariance requires that the outcome of any local non-gravitational experiment be independent of the velocity and orientation of the laboratory. The most sensitive tests are **Hughes–Drever nuclear magnetic resonance experiments**.

By monitoring the $J = 3/2$ nuclear ground state of ^{7}Li in an external magnetic field, any spatial anisotropy in the inertial mass tensor $\delta m_{\text{I}}^{ij} = \sum_A \delta_A \frac{E_A}{c^2}$ as the Earth rotates would induce an anomalous shift in the nuclear Zeeman splitting levels. These experiments constrain any violation of Local Lorentz Invariance to:

$$|\delta m^{ij} c^2| \lesssim 1.7 \times 10^{-16}\, \text{eV}. \tag{1.6.6}$$

Local Position Invariance (LPI)

Local Position Invariance requires that physical laws and fundamental dimensionless constants be independent of where and when in spacetime they are measured. The primary test is the **gravitational redshift of light**: a clock in a gravitational potential $\Phi_{\rm N}$ runs slower than a clock at infinity.

Comparing light emitted at one height and received at another, LPI predicts:

$$z = (1 + \alpha_{\rm LPI})\frac{\Delta\Phi_{\rm N}}{c^2}. \tag{1.6.7}$$

The Pound–Rebka experiment (1960) verified this to 10% using ^{57}Fe Mössbauer resonant absorption in a 22.5-meter tower at Harvard University. In 1976, the sub-orbital **Gravity Probe A** experiment launched a hydrogen maser clock to an altitude of $10,000\,$km, confirming LPI to $\alpha_{\rm LPI} = (1.4 \pm 70) \times 10^{-5}$.

Fifth force tests and Yukawa modifications

Many extensions of general relativity and quantum gravity theories predict the exchange of light scalar or vector bosons, producing a Yukawa-type modification to Newton's gravitational potential:

$$V(r) = -\frac{G_{\rm N} m_1 m_2}{r}\left(1 + \alpha_{12}\mathrm{e}^{-r/\lambda}\right), \tag{1.6.8}$$

where α is the coupling strength relative to gravity and $\lambda \equiv \hbar/(m_{\rm boson}c)$ is the characteristic interaction range.

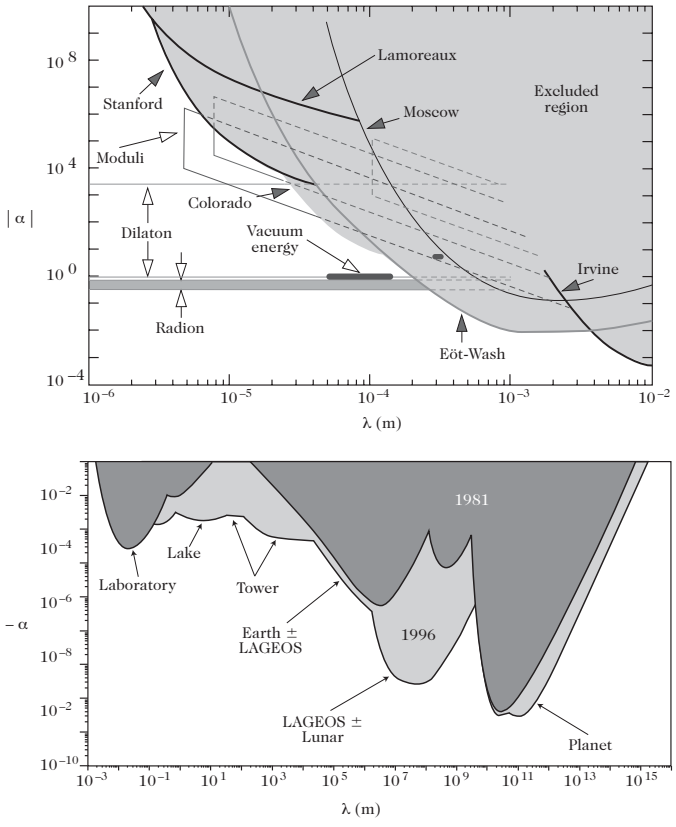


Figure 1.11: Experimental constraints on the strength $|\alpha|$ and range λ of a Yukawa fifth-force modification to Newton's law. Shaded regions are excluded by laboratory Casimir forces, torsion balances, satellite laser ranging (LAGEOS), lunar laser ranging, and planetary ephemerides.

As summarized in Figure 1.11, experimental constraints span from micrometers to Solar System scales (10^{11} m), severely bounding any deviation from the inverse-square law.

1.6.2 Solar System tests and the Parameterized Post-Newtonian (PPN) framework

The Schwarzschild metric

By Birkhoff's theorem, the unique static, spherically symmetric vacuum solution to Einstein's equations outside an isolated mass M is the **Schwarzschild metric**:

$$\mathrm{d}s^2 = -\left(1 - \frac{2G_\mathrm{N}M}{c^2 r}\right) c^2\,\mathrm{d}t^2 + \left(1 - \frac{2G_\mathrm{N}M}{c^2 r}\right)^{-1} \mathrm{d}r^2 + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\;\mathrm{d}\phi^2\right). \tag{1.6.9}$$

Parameterized Post-Newtonian (PPN) expansion

Rather than testing every competing metric theory of gravitation individually, Kenneth Nordtvedt and Clifford Will developed the **Parameterized Post-Newtonian (PPN) framework**. In the weak-field, slow-motion regime ($\frac{G_\mathrm{N}M}{c^2 r} \ll 1$), any metric theory of gravity can be expanded in terms of dimensionless post-Newtonian parameters:

$$\mathrm{d}s^2 = -\left[1 - \frac{2G_\mathrm{N}M}{c^2 r} + 2\left(\beta_\mathrm{PPN} - \gamma_\mathrm{PPN}\right)\left(\frac{G_\mathrm{N}M}{c^2 r}\right)^2\right] c^2\,\mathrm{d}t^2 + \left(1 + 2\gamma_\mathrm{PPN}\frac{G_\mathrm{N}M}{c^2 r}\right)\mathrm{d}r^2 + r^2\,\mathrm{d}\Omega^2\,. \tag{1.6.10}$$

Physical meaning of the PPN parameters.

- γ_PPN**:** Measures the amount of spatial curvature produced per unit mass. (In Newtonian gravity, space is flat, so $\gamma = 0$. In general relativity, $\gamma_\mathrm{PPN} \equiv 1$).
- β_PPN**:** Measures the degree of non-linearity in the gravitational superposition (the fact that gravity itself gravitates). (In general relativity, $\beta_\mathrm{PPN} \equiv 1$).

Deflection of light by a massive body

For a light ray propagating in the equatorial plane ($\theta = \pi/2$), time-translation and rotational Killing symmetries yield two constants of motion along the null geodesic $k^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}$:

$$\frac{\mathrm{d}ct}{\mathrm{d}\lambda} = E\,\mathrm{e}^{-\mu(r)}, \qquad r^2\frac{\mathrm{d}\phi}{\mathrm{d}\lambda} = L. \tag{1.6.11}$$

The null trajectory condition $k^\mu k_\mu = 0$ yields the orbital equation:

$$\left(\frac{1}{r^2}\frac{\mathrm{d}r}{\mathrm{d}\phi}\right)^2 = \left(\frac{E}{L}\right)^2 \mathrm{e}^{-\nu(r)-\mu(r)} - \frac{\mathrm{e}^{-\nu(r)}}{r^2}. \tag{1.6.12}$$

Defining the dimensionless variable $y \equiv \frac{G_\mathrm{N}M}{c^2 r}$ and impact parameter $b \equiv cL/E$, differentiating with respect to ϕ and expanding to first post-Newtonian order gives:

$$\frac{\mathrm{d}^2 y}{\mathrm{d}\phi^2} + y = 3\gamma_\mathrm{PPN}y^2 + \left(1 - \gamma_\mathrm{PPN}\right)\left(\frac{G_\mathrm{N}M}{bc^2}\right)^2. \tag{1.6.13}$$

Solving perturbatively to distance of closest approach $r_0 \simeq b$ (Figure 1.12) yields the total asymptotic deflection angle:

$$\Delta\phi = 2\left(1 + \gamma_\mathrm{PPN}\right)\frac{G_\mathrm{N}M}{r_0 c^2}. \tag{1.6.14}$$

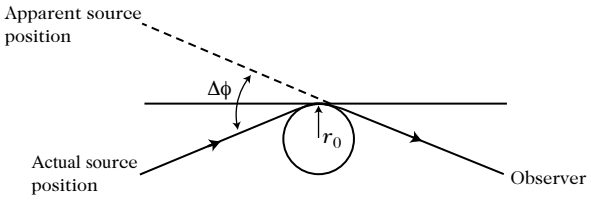


Figure 1.12: Geometry of light deflection by a central mass M. The apparent angular position of the distant source is shifted outward by the total deflection angle $\Delta\phi$ relative to its true position.

The factor of two. If light were a Newtonian corpuscle deflected purely by the Newtonian potential (g_{00} alone, $\gamma = 0$), the deflection would be $\Delta\phi_{\text{Newton}} = \frac{2G_{\text{N}}M}{r_0 c^2}$. In general relativity ($\gamma_{\text{PPN}} = 1$), the spatial curvature g_{rr} contributes an equal amount, **doubling** the prediction:

$$\Delta\phi_{\text{GR}} = \frac{4G_{\text{N}}M_\odot}{R_\odot c^2} = 1.751'' \quad \text{for light grazing the Sun.} \tag{1.6.15}$$

Arthur Eddington's 1919 solar eclipse expedition confirmed this doubled deflection. Modern **Very Long Baseline Interferometry (VLBI)** tracking quasar radio deflections confirms:

$$|\gamma_{\text{PPN}} - 1| < 4 \times 10^{-4}. \tag{1.6.16}$$

The Shapiro time delay effect

Discovered by Irwin Shapiro in 1964, the **Shapiro effect** (the fourth test of general relativity) is a relativistic time delay experienced by radar signals traversing a gravitational potential well. From Equation (1.6.12), the coordinate speed of light in the PPN metric is:

$$\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 = c^2 \mathrm{e}^{\mu(r)-\nu(r)} \left(1 - \mathrm{e}^{\mu(r)} \frac{b^2}{r^2}\right). \tag{1.6.17}$$

Because the coordinate speed is less than c, a radar signal sent from Earth (r_{E}) to a spacecraft or planet (r_{p}) grazing the Sun at closest approach r_0 takes an extra round-trip time:

$$\Delta t_{\text{excess}} = (1 + \gamma_{\text{PPN}}) \frac{2G_{\text{N}}M_\odot}{c^3} \ln\left(\frac{4r_{\text{E}}r_{\text{p}}}{r_0^2}\right). \tag{1.6.18}$$

In 2003, tracking radio signals to the **Cassini spacecraft** during solar occultation confirmed:

$$|\gamma_{\text{PPN}} - 1| = (2.1 \pm 2.3) \times 10^{-5}, \tag{1.6.19}$$

representing the most stringent constraint on spatial metric curvature to date.

Precession of the perihelion of Mercury

For a massive test planet orbiting on a timelike geodesic ($u^\mu u_\mu = -c^2$), setting $y \equiv 1/r$ yields the post-Newtonian radial orbital equation:

$$\frac{\mathrm{d}^2 y}{\mathrm{d}\phi^2} + y = \frac{G_{\text{N}}M}{L^2} + (2 + 2\gamma_{\text{PPN}} - \beta_{\text{PPN}}) \frac{G_{\text{N}}M}{c^2} y^2. \tag{1.6.20}$$

The non-linear term on the right-hand side prevents the Keplerian ellipse from closing, causing the perihelion to precess at each revolution by:

$$\Delta\phi = \frac{2\pi G_{\text{N}}M}{a\,(1 - e^2)\,c^2} (2 + 2\gamma_{\text{PPN}} - \beta_{\text{PPN}}), \tag{1.6.21}$$

where a is the semi-major axis and e is orbital eccentricity.

For the planet Mercury ($a = 5.79 \times 10^{10}\,\mathrm{m}, e = 0.2056$), general relativity ($\gamma_{\rm PPN} = \beta_{\rm PPN} = 1$) predicts an advance of:

$$\Delta\phi_{\rm GR} = \frac{6\pi G_{\rm N} M_\odot}{a(1-e^2)c^2} = 42.98''/\text{century}, \tag{1.6.22}$$

in breathtaking agreement with the observed anomalous residual of $(43.1 \pm 0.5)''$/century that had baffled astronomers for over sixty years.

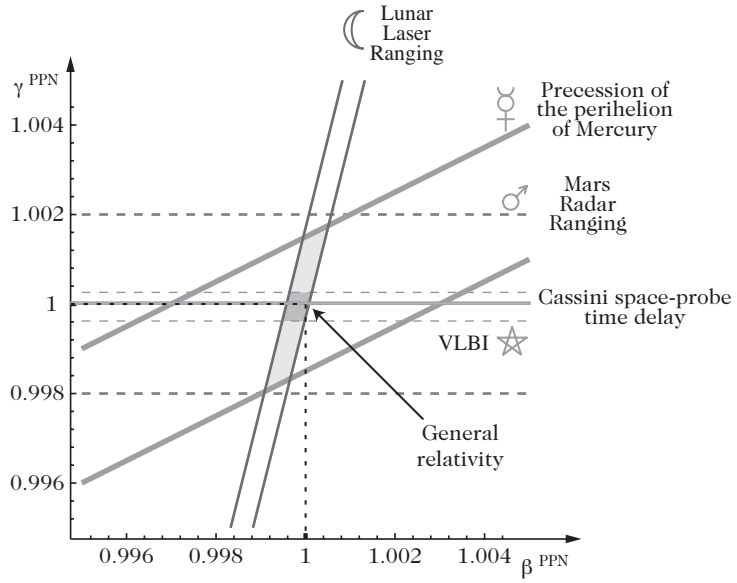


Figure 1.13: Summary of experimental constraints on the Parameterized Post-Newtonian parameters $(\beta_{\rm PPN}, \gamma_{\rm PPN})$ in the Solar System. The intersection of bounds from Mercury's perihelion precession, Lunar Laser Ranging, VLBI light deflection, and the Cassini radio time delay isolates general relativity ($\beta_{\rm PPN} = \gamma_{\rm PPN} = 1$) to high precision.

Table 1.3: Experimental Constraints on Gravitational Parameters in the Solar System

Parameter	**Bound / Measurement**	**Experimental Method**	**Source / System**
η (Weak WEP)	$< 10^{-3}$	Free pendulums	Newton (1686)
η (Weak WEP)	$< 5 \times 10^{-9}$	Torsion balance	Eötvös (1922)
η (Weak WEP)	$(-1.9 \pm 2.5) \times 10^{-12}$	Modern torsion balance	Be–Cu laboratory
η (Weak WEP)	$< 5.5 \times 10^{-13}$	Rotating torsion balance	Eöt-Wash group
$\eta_{\rm SEP}$ (Strong SEP)	$(-0.7 \pm 1.4) \times 10^{-13}$	Lunar Laser Ranging (LLR)	Earth–Moon in Sun field
$\alpha_{\rm LPI}$ (Grav. Redshift)	$< 10^{-2}$	^{57}Fe Mössbauer effect	Pound–Rebka (1960)
$\alpha_{\rm LPI}$ (Grav. Redshift)	$(1.4 \pm 70) \times 10^{-5}$	Suborbital H-maser clock	Gravity Probe A (1980)
$\lvert\delta m_{\rm I}^{ij} c^2\rvert$ (LLI)	$\lesssim 1.7 \times 10^{-16}\,\mathrm{eV}$	Hughes–Drever NMR	^{7}Li nuclear Zeeman
$\lvert\gamma_{\rm PPN} - 1\rvert$	$< 4 \times 10^{-4}$	Solar light deflection	VLBI astrometry
$\lvert\gamma_{\rm PPN} - 1\rvert$	$(2.1 \pm 2.3) \times 10^{-5}$	Radar time delay	Cassini spacecraft
$\lvert 2\gamma - \beta - 1\rvert$	$< 3 \times 10^{-3}$	Perihelion advance	Mercury orbit
$\lvert 4\beta - \gamma - 3\rvert$	$(-0.7 \pm 1.0) \times 10^{-3}$	Nordtvedt effect	LLR tracking

1.6.3 Gravitational radiation and binary pulsars

The quadrupole radiation formula

In linearized general relativity, accelerated masses emit gravitational waves. In the far-zone radiation field ($r \gg \lambda_{\rm GW}$), the transverse-traceless metric perturbation is governed by the **Einstein quadrupole formula**:

$$h_{ij}^{\rm TT}(\mathbf{x},t) = \frac{2G_{\rm N}}{c^4 r}\mathcal{P}_{ijkl}\frac{\mathrm{d}^2 Q_{kl}}{\mathrm{d}t^2}\left(t-\frac{r}{c}\right), \tag{1.6.23}$$

where $\mathcal{P}_{ijkl}$ is the transverse-traceless projection tensor, and $Q_{kl}(t)$ is the trace-free mass quadrupole moment of the source:

$$Q_{kl}(t) \equiv \int \rho(\mathbf{x},t)\left(x_k x_l - \frac{1}{3}r^2\delta_{kl}\right)\mathrm{d}^3x. \tag{1.6.24}$$

Why no monopole or dipole gravitational radiation?

- **No Monopole Radiation:** Conservation of total mass-energy ($\dot{M}=0$) forbids monopole pulsation.
- **No Dipole Radiation:** Conservation of total linear momentum ($\dot{\mathbf{P}}=0$) forbids electric dipole radiation, while conservation of total angular momentum ($\dot{\mathbf{J}}=0$) forbids magnetic dipole radiation.

Consequently, the lowest non-vanishing multipole is the **quadrupole moment**.

The total gravitational radiation luminosity (energy loss rate) is:

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\frac{G_{\rm N}}{5c^5}\dddot{Q}_{kl}\dddot{Q}_{kl}. \tag{1.6.25}$$

The Hulse–Taylor binary pulsar PSR B1913+16

Discovered in 1974 by Russell Hulse and Joseph Taylor, the relativistic binary pulsar PSR B1913+16 provided the first empirical proof of the existence of gravitational waves.

For two neutron stars of masses m_1 and m_2 in an eccentric orbit with semi-major axis a, orbital period P, and eccentricity e, the energy radiated away in gravitational waves causes the orbit to shrink and the orbital period to decay at the rate:

$$\left\langle\frac{\mathrm{d}P}{\mathrm{d}t}\right\rangle = -\frac{192\pi}{5c^5}\left(\frac{2\pi G_{\rm N}}{P}\right)^{5/3}\frac{m_1 m_2}{(m_1+m_2)^{1/3}}\frac{1+\frac{73}{24}e^2+\frac{37}{96}e^4}{(1-e^2)^{7/2}}. \tag{1.6.26}$$

For PSR B1913+16 ($m_1 = 1.441\,M_\odot, m_2 = 1.387\,M_\odot, e = 0.6171, P \approx 7.75\,\mathrm{h}$), general relativity predicts $\langle\dot{P}\rangle = -2.402\times10^{-12}\,\mathrm{s/s}$. Cumulative timing observations across four decades match this theoretical prediction to within 0.2%, proving that gravitational waves carry energy at the speed of light and earning Hulse and Taylor the 1993 Nobel Prize in Physics.

1.6.4 Need for tests at astrophysical and cosmological scales

While general relativity is tested to parts in 10^5 within the Solar System and binary pulsar systems, applying it to the Universe as a whole involves extrapolating its laws across 15 orders of magnitude in length scale. On these vast cosmological scales, several profound observational puzzles emerge:

1. **Galactic Rotation Curves and Dark Matter:** On galactic scales, stellar rotation curves remain flat at large radii ($v \approx \text{const} \implies \Phi_{\rm N} \propto \ln r$). This requires either massive halos of non-baryonic **cold dark matter** or infrared modifications of gravity.

2. **Galaxy Clusters and Gravitational Lensing:** In galaxy clusters ($\sim 2\,\mathrm{Mpc}$), gravitational lensing probes the metric sum $\Phi + \Psi$, while X-ray gas temperatures probe the Newtonian potential Φ. Concordance between the two confirms the relativistic relation $\Phi = \Psi$ up to dark matter modeling.

3. **Cosmic Acceleration and Dark Energy:** Supernovae Type Ia and CMB observations reveal that cosmic expansion is currently accelerating. In general relativity, acceleration requires an exotic negative-pressure **dark energy** component ($w \approx -1$, such as the cosmological constant Λ) or large-scale modifications of the gravitational action ($f(R)$, scalar-tensor theories).

4. **Cosmic Invariance of Fundamental Constants:** Quasar absorption spectra test whether the fine-structure constant α_{EM} and electron-to-proton mass ratio $\mu = m_p/m_e$ have varied over cosmological timescales, providing deep tests of Local Position Invariance across Hubble horizons.

Part II

The Modern Standard Cosmological Model

Chapter 2

The cosmological solution of Friedmann–Lemaître

The first relativistic cosmological solution of Einstein's field equations was formulated by Albert Einstein in 1917, who introduced a cosmological constant Λ to construct a static, spatially closed universe. The general dynamical cosmological solutions were discovered independently by Alexandre Friedmann (1922, 1924) and Georges Lemaître (1927).

This chapter details the mathematical derivation of the Friedmann–Lemaître spacetime from the symmetries of homogeneity and isotropy, analyzes its kinematics and relativistic dynamics, and establishes the observational relationships foundational to modern primordial cosmology.

2.1 The Friedmann–Lemaître spacetime

2.1.1 Constructing a cosmological model

Within the framework of general relativity, the central objective is to identify which exact solution of Einstein's field equations provides an accurate, idealized description of our Universe on the largest observable scales.

This task is constrained by fundamental observational limitations:

1. **Uniqueness of the Universe:** We observe a single realization of the Universe and cannot perform statistical ensembles over independent universes.

2. **Spacetime Localization:** Astronomical observations are gathered exclusively along our **past light cone**, while geophysical and local experiments probe a narrow worldtube around our worldline.

3. **Observable Universe vs. Global Universe:** We must distinguish between the *observable Universe* (the interior of our past particle horizon from which causal signals have reached us) and the *global Universe*, whose geometry beyond our horizon cannot be directly probed without theoretical extrapolation.

The standard cosmological model rests on four foundational hypotheses:

H1 **General Relativity as the Theory of Gravitation:** Gravity is governed by Einstein's general relativity, and the Einstein Equivalence Principle ensures that local non-gravitational physical laws remain invariant throughout cosmic history.

H2 **Cosmic Matter Content:** On large scales, the Universe is filled with a statistical mixture of pressureless baryonic and dark matter (dust), relativistic radiation (photons and neutrinos), and dark energy / cosmological constant Λ.

H3 Uniformity Principles (Homogeneity and Isotropy): On scales larger than galaxy clusters and superclusters ($\gtrsim 100\,\mathrm{Mpc}$), spacetime possesses spatial homogeneity and spatial isotropy.

H4 Global Spatial Topology: For a globally hyperbolic spacetime $\mathcal{M} = \mathbb{R} \times \Sigma$, the spatial slices Σ possess a simply connected or multi-connected topology compatible with the local constant curvature.

2.1.2 The Cosmological Principle and the Copernican Principle

Astronomical surveys of distant galaxies and the extreme angular uniformity of the Cosmic Microwave Background (CMB, isotropic to $\Delta T/T \sim 10^{-5}$) demonstrate that the observable Universe is highly isotropic around our vantage point.

Two distinct interpretations arise:

1. **Non-Copernican Centered Geometry:** Our Galaxy resides at the geometric center of a spherically symmetric, radially inhomogeneous universe (Figure 2.1, right).
2. **The Copernican Principle:** Humans do not occupy a privileged or special location in the Universe. If the Universe is isotropic around us, and our position is typical, then the Universe must be **isotropic around every point** (Figure 2.1, left).

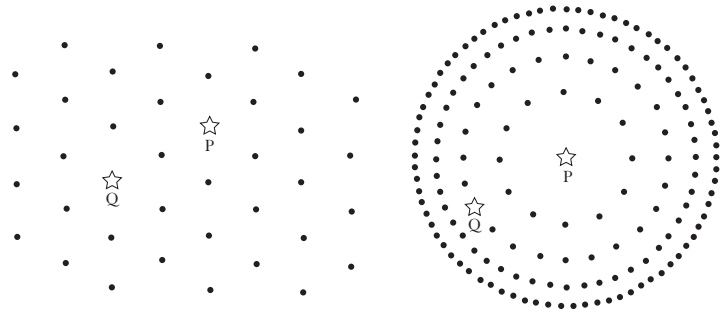


Figure 2.1: Point distributions illustrating uniformity: (Left) Statistically isotropic around every point, implying spatial homogeneity. (Right) Isotropic around a single privileged center P, where points P and Q are physically distinct. The Copernican principle excludes centered models.

The **Cosmological Principle** asserts that the Universe is both **spatially homogeneous** and **spatially isotropic** when averaged on scales greater than the largest bound structures ($\lambda_{\mathrm{smooth}} \gtrsim 100\,\mathrm{Mpc}$).

2.1.3 Spacetime symmetries and the Robertson–Walker metric

Homogeneity and isotropy are distinct geometric properties:

- **Spatial Homogeneity:** Spacetime $\mathcal{M}$ admits a foliation by a 1-parameter family of spacelike hypersurfaces Σ_t such that for any two points $P, Q \in \Sigma_t$, there exists an isometry of the spatial metric mapping P to Q (Figure 2.2, left).
- **Spatial Isotropy:** There exists a congruence of timelike worldlines with unit tangent 4-velocity u^μ (the **fundamental / comoving observers**) such that at any point P, any unit spatial tangent vector $e_1^\mu \in T_P\Sigma_t$ can be rotated into any other unit vector $e_2^\mu \in T_P\Sigma_t$ by an isometry that leaves P and $u^\mu(P)$ fixed (Figure 2.2, right).

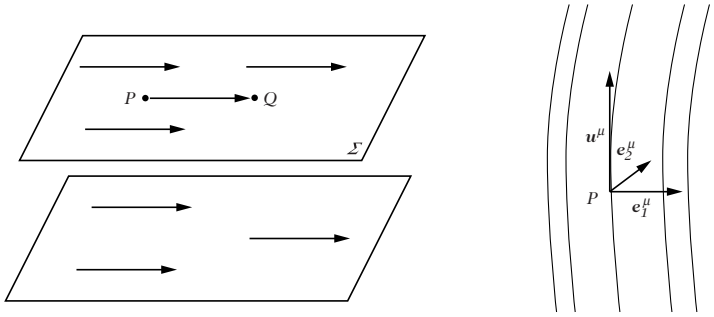


Figure 2.2: (Left) A spatially homogeneous spacetime foliated by spacelike slices Σ_t. (Right) Spatial isotropy ensures that any spatial tangent direction e_1^μ is isometric to e_2^μ about the fundamental observer u^μ.

Isotropy requires that the worldlines of fundamental observers be strictly orthogonal to the hypersurfaces of homogeneity Σ_t. If this were not the case, the non-vanishing projection of u^μ on Σ_t would single out a preferred spatial direction, violating isotropy.

Consequently, each spatial slice Σ_t is a 3-dimensional **maximally symmetric Riemannian space** of constant spatial curvature $K(t)$. The 3-dimensional Riemann curvature tensor of Σ_t satisfies:

$$ {}^{(3)}R_{ijkl} = \frac{K(t)}{a^2(t)} \left(\gamma_{ik}\gamma_{jl} - \gamma_{il}\gamma_{jk}\right), \tag{2.1.1} $$

where γ_{ij} is the dimensionless comoving spatial metric, $a(t)$ is the cosmic scale factor, and $K \in \{-1, 0, +1\}$ is the normalized curvature index:

1. **Spherical Geometry ($K = +1$):** Σ_t is a 3-sphere S^3 embedded in $\mathbb{R}^4$ via $x^2+y^2+z^2+w^2 = R^2(t)$:
$$ \mathrm{d}\ell^2_{(3)} = a^2(t)\left[\mathrm{d}\chi^2 + \sin^2\chi\left(\mathrm{d}\theta^2 + \sin^2\theta\ \mathrm{d}\phi^2\right)\right], \qquad \chi \in [0, \pi]. \tag{2.1.2} $$

2. **Flat Euclidean Geometry ($K = 0$):** $\Sigma_t = \mathbb{R}^3$:
$$ \mathrm{d}\ell^2_{(3)} = a^2(t)\left[\mathrm{d}\chi^2 + \chi^2\left(\mathrm{d}\theta^2 + \sin^2\theta\ \mathrm{d}\phi^2\right)\right], \qquad \chi \in [0, \infty). \tag{2.1.3} $$

3. **Hyperbolic Geometry ($K = -1$):** Σ_t is a 3-hyperboloid H^3 embedded in Minkowski space $\mathbb{R}^{3,1}$ via $-w^2 + x^2 + y^2 + z^2 = -R^2(t)$:
$$ \mathrm{d}\ell^2_{(3)} = a^2(t)\left[\mathrm{d}\chi^2 + \sinh^2\chi\left(\mathrm{d}\theta^2 + \sin^2\theta\ \mathrm{d}\phi^2\right)\right], \qquad \chi \in [0, \infty). \tag{2.1.4} $$

The Friedmann–Lemaître–Robertson–Walker (FLRW) line element

Setting the proper time of fundamental observers to cosmic time t ($u^\mu\,\mathrm{d}x_\mu \equiv -\,\mathrm{d}t$), the 4-dimensional spacetime metric takes the universal form:

$$ \mathrm{d}s^2 = -\,\mathrm{d}t^2 + a^2(t)\gamma_{ij}(\mathbf{x})\,\mathrm{d}x^i\,\mathrm{d}x^j\,. \tag{2.1.5} $$

Introducing **conformal time** η, defined by:

$$ \mathrm{d}t = a(t)\,\mathrm{d}\eta \implies \eta(t) = \int_{t_0}^{t} \frac{\mathrm{d}t'}{a(t')}, \tag{2.1.6} $$

the line element becomes conformally related to a static spacetime:

$$ \mathrm{d}s^2 = a^2(\eta)\left[-\,\mathrm{d}\eta^2 + \gamma_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j\right]. \tag{2.1.7} $$

In comoving spherical coordinates (χ, θ, ϕ), the spatial line element is:

$$\mathrm{d}\sigma^2 \equiv \gamma_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j = \mathrm{d}\chi^2 + f_K^2(\chi)\,\mathrm{d}\Omega^2\,, \qquad \mathrm{d}\Omega^2 \equiv \mathrm{d}\theta^2 + \sin^2\theta\;\mathrm{d}\phi^2\,, \tag{2.1.8}$$

where the coordinate distance function $f_K(\chi)$ is:

$$f_K(\chi) = \begin{cases} \frac{1}{\sqrt{K}}\sin\left(\sqrt{K}\chi\right) & (K>0,\ \text{Closed}),\\ \chi & (K=0,\ \text{Flat}),\\ \frac{1}{\sqrt{-K}}\sinh\left(\sqrt{-K}\chi\right) & (K<0,\ \text{Open}). \end{cases} \tag{2.1.9}$$

Alternatively, defining the radial coordinate $r \equiv f_K(\chi)$, the metric takes the standard curvature form:

$$\mathrm{d}s^2 = -\,\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-Kr^2} + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\;\mathrm{d}\phi^2\right)\right]. \tag{2.1.10}$$

Symmetry reduces the 10 metric functions of an arbitrary spacetime to a single dynamical function $a(t)$ and a discrete constant $K \in \{+1, 0, -1\}$.

Geometric tensors of the FLRW spacetime

From the metric (2.1.5), the non-vanishing Christoffel connection coefficients are:

$$\Gamma^0_{ij} = Ha^2\gamma_{ij}, \qquad \Gamma^i_{j0} = H\delta^i_j, \qquad \Gamma^i_{jk} = {}^{(3)}\Gamma^i_{jk}, \tag{2.1.11}$$

where $H(t) \equiv \frac{\dot{a}}{a} = \frac{c}{a}\frac{\mathrm{d}a}{\mathrm{d}(ct)}$ is the **Hubble expansion parameter**, and an overdot denotes differentiation with respect to cosmic time t.

The non-vanishing components of the Ricci tensor are:

$$R_{00} = -3\frac{\ddot{a}}{a}, \qquad R_{ij} = \left(2H^2 + \frac{\ddot{a}}{a} + \frac{2K}{a^2}\right)a^2\gamma_{ij}, \tag{2.1.12}$$

yielding the Ricci scalar curvature:

$$R = 6\left(H^2 + \frac{\ddot{a}}{a} + \frac{K}{a^2}\right) = 6\left[\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + \frac{K}{a^2}\right]. \tag{2.1.13}$$

The non-vanishing components of the Einstein tensor $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}$ are:

$$G_{00} = 3\left(H^2 + \frac{K}{a^2}\right), \tag{2.1.14}$$

$$G_{ij} = -\left(H^2 + 2\frac{\ddot{a}}{a} + \frac{K}{a^2}\right)a^2\gamma_{ij}. \tag{2.1.15}$$

Killing and Conformal Killing vectors

A maximally symmetric 3-dimensional spatial section admits $\frac{3(3+1)}{2} = \mathbf{6}$ independent spatial Killing vector fields satisfying $\nabla_{(\mu}\xi_{\nu)} = 0$:

1. **3 Spatial Translations:**

$$P^0_{[r]} = 0, \qquad P^i_{[r]} = \sqrt{1-Kr^2}\,\delta^i_r \qquad (r = 1,2,3). \tag{2.1.16}$$

2. **3 Spatial Rotations:**

$$R^0_{[rs]} = 0, \qquad R^i_{[rs]} = \delta^i_r x_s - \delta^i_s x_r \qquad (r,s = 1,2,3). \tag{2.1.17}$$

Because the Universe expands ($\dot{a} \neq 0$), time-translation invariance is broken; there is no timelike Killing vector. However, FLRW spacetime possesses **4 conformal Killing vectors** satisfying $\nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu = \frac{1}{2}(\nabla_\alpha \xi^\alpha) g_{\mu\nu}$:

$$T^0 = \sqrt{1 - Kr^2}, \qquad T^k = -Hx^k \sqrt{1 - Kr^2}, \tag{2.1.18}$$

$$L^0_{[r]} = x_r, \qquad L^k_{[r]} = H\left[\frac{1}{2}\delta^k_r\left(r^2 - \eta^2\right) - x^k x_r\right] \quad (K = 0). \tag{2.1.19}$$

2.2 Kinematics of the expanding Universe

2.2.1 Fundamental observers and the Hubble expansion law

In comoving coordinates $\mathbf{x}$, the worldline of a fundamental observer at rest with the expanding cosmological fluid is $\mathbf{x} =$ constant. Their 4-velocity is:

$$u^\mu = \delta^\mu_0 = (1, 0, 0, 0). \tag{2.2.1}$$

For two comoving galaxies at comoving separation $\Delta\mathbf{x}$, their physical distance at cosmic time t is:

$$\mathbf{r}(t) = a(t)\Delta\mathbf{x}. \tag{2.2.2}$$

Differentiating with respect to cosmic time yields the **Hubble Law**:

$$\mathbf{v}_{\rm rec} \equiv \dot{\mathbf{r}} = \dot{a}(t)\Delta\mathbf{x} = \frac{\dot{a}}{a}\mathbf{r} = H(t)\mathbf{r}. \tag{2.2.3}$$

The recession velocity $\mathbf{v}_{\rm rec}$ is the kinematic expansion of space itself.

For a non-comoving object with coordinate motion $\dot{\mathbf{x}}$, the total physical velocity splits as:

$$\mathbf{v}_{\rm tot} = \dot{\mathbf{r}} = H\mathbf{r} + a(t)\dot{\mathbf{x}} \equiv \mathbf{v}_{\rm rec} + \mathbf{v}_{\rm pec}, \tag{2.2.4}$$

where $\mathbf{v}_{\rm pec} \equiv a(t)\dot{\mathbf{x}}$ is the **peculiar velocity** relative to the comoving Hubble flow.

2.2.2 Propagation of light and cosmological redshift

In geometric optics, light rays travel along null geodesics ($k^\mu k_\mu = 0, k^\mu \nabla_\mu k^\nu = 0$). Decomposing the wavevector relative to comoving observers:

$$k^\mu = E u^\mu + p^\mu, \qquad E = -k_\mu u^\mu = \hbar\omega, \qquad p^\mu u_\mu = 0, \tag{2.2.5}$$

where $E^2 = a^2(t)\gamma_{ij}p^i p^j = p^2$.

Projecting the geodesic equation $k^\nu \nabla_\nu k^\mu = 0$ along u_μ:

$$u_\mu k^\nu \nabla_\nu k^\mu = 0 \implies E\dot{E} + \Gamma^0_{ij} p^i p^j = 0 \implies E\dot{E} + Ha^2\gamma_{ij}p^i p^j = 0. \tag{2.2.6}$$

Using $a^2\gamma_{ij}p^i p^j = E^2$, we obtain the exact energy damping law:

$$E\frac{\mathrm{d}E}{\mathrm{d}t} = -\frac{\dot{a}}{a}p^2 \implies \frac{\dot{E}}{E} = -H = -\frac{\dot{a}}{a} \implies E(t) \propto \frac{1}{a(t)}. \tag{2.2.7}$$

Because photon frequency $\nu \propto E$ and wavelength $\lambda = c/\nu$, the wavelength stretches in direct proportion to the cosmic scale factor:

$$\lambda(t) \propto a(t). \tag{2.2.8}$$

A photon emitted at cosmic time $t_{\rm em}$ and observed at $t_{\rm obs}$ undergoes an achromatic **cosmological redshift** z:

$$1 + z \equiv \frac{\lambda_{\rm obs}}{\lambda_{\rm em}} = \frac{\nu_{\rm em}}{\nu_{\rm obs}} = \frac{a(t_{\rm obs})}{a(t_{\rm em})}. \tag{2.2.9}$$

Setting $a(t_{\rm obs}) \equiv 1$ today, the scale factor at redshift z is simply $a(z) = \frac{1}{1+z}$.

2.2.3 Dynamics of massive test particles

For a massive test particle with 4-velocity v^μ ($v^\mu v_\mu = -1$), decomposing relative to the fundamental observers gives:

$$v^\mu = \alpha u^\mu + p^\mu, \qquad u_\mu p^\mu = 0, \qquad \alpha = \sqrt{1+p^2}, \tag{2.2.10}$$

where $p^2 \equiv a^2 \gamma_{ij} p^i p^j$.

Projecting the geodesic equation $v^\nu \nabla_\nu v^\mu = 0$ along u_μ yields:

$$\alpha\dot{\alpha} + Hp^2 = 0 \iff \alpha\dot{\alpha} + H\left(\alpha^2 - 1\right) = 0 \implies \alpha^2 - 1 \propto a^{-2}(t). \tag{2.2.11}$$

Consequently, the physical momentum and peculiar velocity of any freely moving particle decay as:

$$E^2 = p^2 \implies E\frac{\mathrm{d}E}{\mathrm{d}t} = p\frac{\mathrm{d}p}{\mathrm{d}t} \tag{2.2.12}$$

Then using $p\frac{\mathrm{d}p}{\mathrm{d}t} = -Hp^2$, we have

$$p(t) \propto \frac{1}{a(t)}, \qquad v_{\mathrm{pec}}(t) \propto \frac{1}{a(t)} \quad \text{(non-relativistic limit)}. \tag{2.2.13}$$

Any freely moving particle is decelerated by the cosmic expansion, coming to rest with respect to the comoving fundamental frame as $a \to \infty$.

2.3 Cosmological dynamics of the FLRW spacetime

2.3.1 Energy-momentum tensor of the cosmological fluid

The stress-energy tensor $T_{\mu\nu}$ is a symmetric rank-2 tensor. Under the requirement of spatial homogeneity and isotropy, its most general form relative to the fundamental comoving observers $u^\mu = (1,0,0,0)$ is that of a **perfect fluid**:

$$T_{\mu\nu} = \rho(t) u_\mu u_\nu + P(t)\gamma_{\mu\nu}(t) = \begin{pmatrix} \rho(t) & 0 & 0 & 0 \\ 0 & & & \\ 0 & & a^2(t)P(t)\gamma_{ij} & \\ 0 & & & \end{pmatrix}, \tag{2.3.1}$$

where $\rho(t) = T_{00}$ is the energy density and $P(t) = \frac{1}{3}T^i_{\ i}$ is the isotropic pressure, both depending exclusively on cosmic time t. Anisotropic stresses ($\pi_{\mu\nu} = 0$) and spatial energy fluxes ($q_\mu = 0$) vanish identically by isotropy.

2.3.2 The Friedmann equations and conservation laws

Substituting the Einstein tensor (2.1.14)–(2.1.15) and stress-energy tensor (2.3.1) into Einstein's field equations $G_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G_{\mathrm{N}} T_{\mu\nu}$ yields the two foundational **Friedmann equations**:

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G_{\mathrm{N}}}{3}\rho - \frac{K}{a^2} + \frac{\Lambda}{3}, \tag{2.3.2}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G_{\mathrm{N}}}{3}(\rho + 3P) + \frac{\Lambda}{3}. \tag{2.3.3}$$

Covariant conservation of stress-energy $\nabla_\mu T^{\mu\nu} = 0$ reduces to the **fluid continuity equation**:

$$\dot{\rho} + 3H(\rho + P) = 0 \iff \frac{\mathrm{d}}{\mathrm{d}t}\left(\rho a^3\right) + P\frac{\mathrm{d}a^3}{\mathrm{d}t} = 0. \tag{2.3.4}$$

The continuity equation is the first law of thermodynamics ($\mathrm{d}U = -P\,\mathrm{d}V$) applied to a comoving volume $V \propto a^3$. Due to the Bianchi identities, only two of the three equations (2.3.2), (2.3.3), and (2.3.4) are functionally independent with (2.3.2) being the Hamiltonian constraint of the theory. Since only two of them are independent, we only solve (2.3.2) and (2.3.4).

2.3.3 Cosmological equations in conformal time

Working in terms of conformal time $\mathrm{d}\eta = \mathrm{d}t\,/a(t)$, derivatives satisfy $X' \equiv \frac{\mathrm{d}X}{\mathrm{d}\eta} = a\dot{X}$ and $X'' - \mathcal{H}X' = a^2\ddot{X}$, where $\mathcal{H} \equiv \frac{a'}{a} = aH$ is the **comoving Hubble expansion rate**. The Friedmann and continuity equations become:

$$\mathcal{H}^2 = \frac{8\pi G_\mathrm{N}}{3}\rho a^2 - K + \frac{\Lambda}{3}a^2, \tag{2.3.5}$$

$$\mathcal{H}' = -\frac{4\pi G_\mathrm{N}}{3}a^2\,(\rho + 3P) + \frac{\Lambda}{3}a^2, \tag{2.3.6}$$

$$\rho' + 3\mathcal{H}\,(\rho + P) = 0. \tag{2.3.7}$$

Since (2.3.7) is $\nabla_{(\mu}T^{\mu\nu} = 0$ in disguise. This equation holds true for each fluid component individually and hence for constant equation of state (EoS) fluid, eq. (2.3.10) holds for them individually.

For a power-law scale factor in cosmic time $a(t) \propto t^n$ ($n \neq 1$), the conformal time dependence is:

$$a(t) \propto t^n \iff a(\eta) \propto \eta^{\frac{n}{1-n}}. \tag{2.3.8}$$

2.3.4 Equations of state and cosmic components

To close the system of dynamical equations, we specify a barotropic **equation of state**:

$$P = w\rho, \qquad w \equiv \frac{P}{\rho} = \text{constant}. \tag{2.3.9}$$

Integrating the continuity equation (2.3.4) for constant w yields:

$$\frac{\dot{\rho}}{\rho} = -3(1+w)\frac{\dot{a}}{a} \implies \rho(a) = \rho_0\left(\frac{a}{a_0}\right)^{-3(1+w)}. \tag{2.3.10}$$

Note that during reheating we do not expect w to be constant and thus this form will not work during that era.

Standard cosmological components scale as:

- **Non-relativistic Matter / Dust ($w = 0$):** Baryons and cold dark matter (CDM) have negligible thermal pressure ($P \ll \rho c^2$), scaling as comoving particle dilution:

$$\rho_m(a) = \rho_{m0}\left(\frac{a}{a_0}\right)^{-3} = \rho_{m0}(1+z)^3. \tag{2.3.11}$$

- **Relativistic Radiation ($w = 1/3$):** Photons and relativistic neutrinos experience both number density dilution $\propto a^{-3}$ and individual photon energy redshift $\propto a^{-1}$:

$$\rho_r(a) = \rho_{r0}\left(\frac{a}{a_0}\right)^{-4} = \rho_{r0}(1+z)^4. \tag{2.3.12}$$

- **Cosmological Constant / Vacuum Energy ($w = -1$):** Constant energy density:

$$\rho_\Lambda(a) = \rho_{\Lambda 0} = \frac{\Lambda}{8\pi G_{\rm N}} = \text{constant}. \tag{2.3.13}$$

- **Spatial Curvature ($w_{\rm eff} = -1/3$):** Curvature acts as an effective fluid with $\rho_K \propto a^{-2}$.

For a general fluid, the adiabatic sound speed $c_s^2 \equiv \frac{\partial P}{\partial \rho}$ governs the time variation of the equation of state:

$$w' = -3\mathcal{H}\,(1+w)\left(c_s^2 - w\right). \tag{2.3.14}$$

2.3.5 Dimensionless density parameters and the cosmic sum rule

Dividing the first Friedmann equation (2.3.2) by $H^2(t)$, we define the **critical energy density**:

$$\rho_{\rm crit}(t) \equiv \frac{3H^2(t)}{8\pi G_{\rm N}}, \qquad \rho_{\rm crit,0} = \frac{3H_0^2}{8\pi G_{\rm N}} = 1.878 \times 10^{-26} h^2\,{\rm kg\cdot m^{-3}}, \tag{2.3.15}$$

where $H_0 \equiv 100\,h\,{\rm km\cdot s^{-1}\cdot Mpc^{-1}}$.

We define dimensionless **density parameters**:

$$\Omega_X(t) \equiv \frac{\rho_X(t)}{\rho_{\rm crit}(t)} = \frac{8\pi G_{\rm N}\rho_X}{3H^2}, \qquad \Omega_\Lambda(t) \equiv \frac{\Lambda}{3H^2}, \qquad \Omega_K(t) \equiv -\frac{K}{H^2a^2}. \tag{2.3.16}$$

The first Friedmann equation (2.3.2) becomes the exact **cosmic sum rule** upon substituting (2.3.16):

$$\sum_X \Omega_X(t) + \Omega_\Lambda(t) + \Omega_K(t) = 1. \tag{2.3.17}$$

Using the solutions of the conservation equation, it follows that from (2.3.10)

$$\Omega_X = \Omega_{X0}\left(\frac{a}{a_0}\right)^{-3(1+w_{X0})}\left(\frac{H_0}{H}\right)^2 \tag{2.3.18}$$

Evaluating today (subscript 0), the normalized expansion rate $E(z) \equiv H(z)/H_0$ is:

$$E^2(z) = \Omega_{r0}(1+z)^4 + \Omega_{m0}(1+z)^3 + \Omega_{K0}(1+z)^2 + \Omega_{\Lambda 0}, \tag{2.3.19}$$

with spatial curvature parameter $\Omega_{K0} = 1 - \Omega_{m0} - \Omega_{r0} - \Omega_{\Lambda 0}$.

This form allows us to calculate the age of the universe with $\Omega_{m0} = 0.315$, $H_0 = 67.36 \times 1.02269 \times 10^{-3}{\rm Gyr}^{-1}$, $\Omega_{r0} = 9.22 \times 10^{-5}$, $\Omega_{\Lambda 0} = 0.685$ and $\Omega_{K0} = 0$,

$$\int_0^t {\rm d}t = \int_0^{a_0} \frac{{\rm d}a}{aH(a)} = \int_0^{a_0} \frac{{\rm d}a}{aH_0E(a)} = 13.8\,{\rm Gyr} \tag{1}$$

```
import numpy as np
from scipy.integrate import quad

Omega_m0 = 0.315
Omega_r0 = 9.22e-5
Omega_Lambda0 = 0.685
Omega_K0 = 0.0

H0 = 67.36 * 1.02269e-3

def E(a):
        return np.sqrt(Omega_r0/a**4 + Omega_m0/a**3+ Omega_K0/a**2 + Omega_Lambda0)
```

```
t0 = quad(lambda a: 1/(a*E(a)), 0, 1)[0]/H0

print(f"Age of the Universe = {t0:.4f} Gyr")
```

2.3.6 Exact analytical solutions of the Friedmann equations

Flat single-component universes ($K = 0, \Lambda = 0, w \neq -1$)

For a flat spacetime filled with a single fluid with equation of state w, $\rho \propto a^{-3(1+w)}$, the first Friedmann equation $\dot{a}/a \propto a^{-3(1+w)/2}$ integrates to:

$$a(t) = a_0 \left(\frac{t}{t_0}\right)^{\frac{2}{3(1+w)}}, \qquad a(\eta) = a_0 \left(\frac{\eta}{\eta_0}\right)^{\frac{2}{1+3w}} \quad (w \neq -1/3)\,. \tag{2.3.20}$$

- **Einstein–de Sitter Universe ($w = 0$, Dust):**

$$a(t) \propto t^{2/3}, \qquad a(\eta) \propto \eta^2, \qquad t_0 = \frac{2}{3H_0}. \tag{2.3.21}$$

- **Radiation-Dominated Universe ($w = 1/3$):**

$$a(t) \propto t^{1/2}, \qquad a(\eta) \propto \eta, \qquad t_0 = \frac{1}{2H_0}. \tag{2.3.22}$$

de Sitter spacetime ($K = 0, w = -1$)

When vacuum energy dominates ($\rho = \rho_\Lambda = \text{const}$), $H = H_\Lambda = \sqrt{\Lambda/3} = \text{const}$:

$$a(t) = a_0\, \mathrm{e}^{H_\Lambda t}, \qquad a(\eta) = -\frac{1}{H_\Lambda \eta} \quad (\eta < 0). \tag{2.3.23}$$

In global closed slicing ($K = +1$), de Sitter space is described by:

$$\mathrm{d}s^2 = -\,\mathrm{d}t^2 + \frac{\cosh^2(H_\Lambda t)}{H_\Lambda^2}\left[\mathrm{d}\chi^2 + \sin^2\chi\ \mathrm{d}\Omega^2\right]. \tag{2.3.24}$$

Curved single-component universes ($K = \pm 1, w \neq -1/3$)

In conformal time, the Friedmann equation $\mathcal{H}^2 = \frac{\Omega_0}{|1-\Omega_0|}(a/a_0)^{-(1+3w)} - K$ integrates to the parametric solution:

$$\frac{a(\eta)}{a_0} = \left(\frac{\Omega_0}{|1-\Omega_0|}\right)^{\frac{1}{1+3w}} \begin{cases} \left[\sin\left(\frac{1+3w}{2}\eta\right)\right]^{\frac{2}{1+3w}} & (K=+1), \\ \left[\sinh\left(\frac{1+3w}{2}\eta\right)\right]^{\frac{2}{1+3w}} & (K=-1). \end{cases} \tag{2.3.25}$$

For a closed matter-dominated universe ($K = +1, w = 0$), this yields the classical **cycloid solution**:

$$a(\eta) = a_{\max}\sin^2\left(\frac{\eta}{2}\right) = \frac{a_{\max}}{2}\left(1 - \cos\eta\right), \qquad t(\eta) = \frac{a_{\max}}{2}\left(\eta - \sin\eta\right). \tag{2.3.26}$$

Flat ΛCDM universe ($K = 0, w = 0, \Lambda > 0$)

For the standard cosmological model with cold matter and cosmological constant, the Friedmann equation integrates analytically to:

$$a(t) = a_0 \left(\frac{\Omega_{m0}}{\Omega_{\Lambda 0}}\right)^{1/3} \sinh^{2/3}\left(\frac{3}{2}\sqrt{\Omega_{\Lambda 0}} H_0 t\right). \tag{2.3.27}$$

At early times ($t \ll 1/H_0$), $a(t) \propto t^{2/3}$ (matter domination); at late times ($t \gg 1/H_0$), $a(t) \propto \mathrm{e}^{\sqrt{\Omega_{\Lambda 0}} H_0 t}$ (de Sitter acceleration).

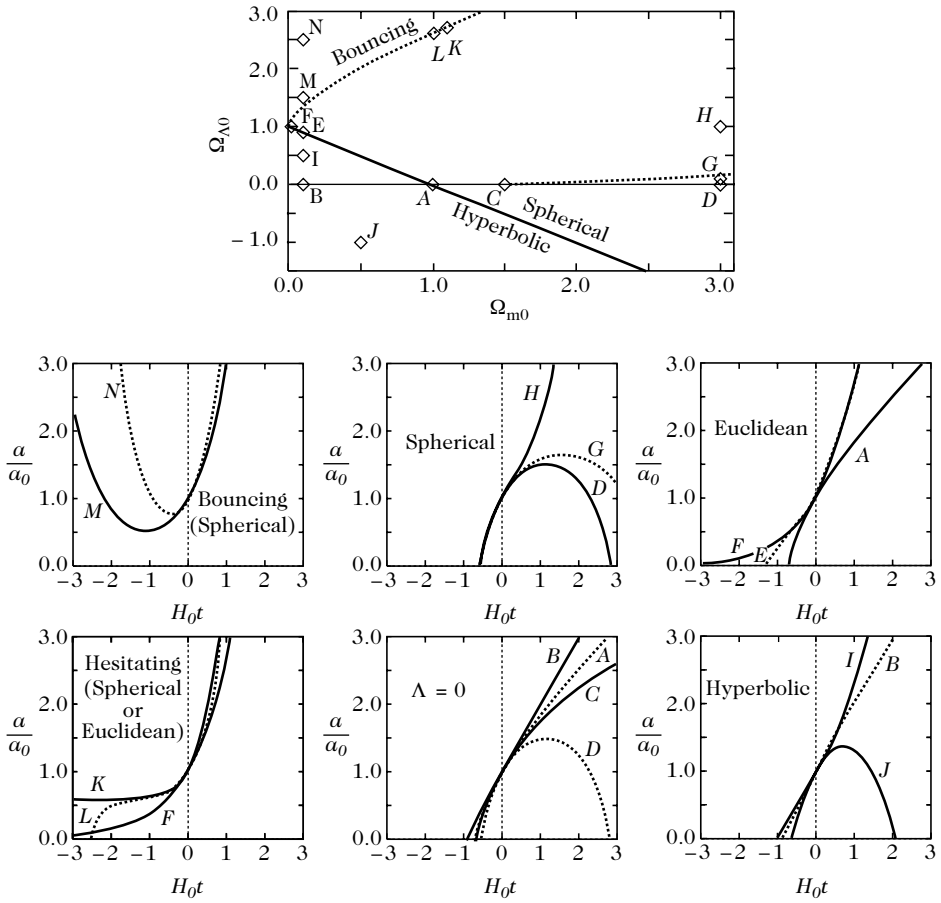


Figure 2.3: Evolution profiles of the cosmic scale factor $a(t)/a_0$ as a function of $H_0 t$ for various cosmological parameter combinations $(\Omega_{m0}, \Omega_{\Lambda 0}, K)$, illustrating Big Bang, Big Crunch, bouncing, hesitating, and de Sitter asymptotic regimes.

2.3.7 Dynamical system analysis in phase space

To analyze the global qualitative dynamics across all parameter regimes, we formulate the Friedmann equations as an autonomous dynamical system.

Defining the **deceleration parameter**:

$$q \equiv -\frac{\ddot{a}a}{\dot{a}^2} = -1 - \frac{\dot{H}}{H^2} = \frac{1}{2}(3\gamma - 2)(1 - \Omega_K) - \frac{3\gamma}{2}\Omega_\Lambda, \tag{2.3.28}$$

with $\gamma \equiv 1 + w$. Introducing the logarithmic e-fold time variable $p \equiv \ln(a/a_0)$, the derivative is $X' \equiv \frac{\mathrm{d}X}{\mathrm{d}p} = \frac{\dot{X}}{H}$.

where we used

$$\frac{\mathrm{d}}{\mathrm{d}\ln\left(\frac{a}{a_0}\right)} = \frac{\mathrm{d}t}{\mathrm{d}\ln\left(\frac{a}{a_0}\right)}\frac{\mathrm{d}}{\mathrm{d}t} = \frac{1}{\frac{\mathrm{d}\ln\left(\frac{a}{a_0}\right)}{\mathrm{d}t}}\frac{\mathrm{d}}{\mathrm{d}t} = \frac{1}{H}\frac{\mathrm{d}}{\mathrm{d}t} \tag{1}$$

The Friedmann equations transform into the 2-dimensional autonomous system:

$$\Omega_\Lambda' = 2(1+q)\Omega_\Lambda, \tag{2.3.29}$$

$$\Omega_K' = 2q\Omega_K, \tag{2.3.30}$$

with $\Omega_m = 1 - \Omega_\Lambda - \Omega_K$.

Fixed points and linear stability

Setting $\Omega'_\Lambda = 0$ and $\Omega'_K = 0$, we identify three fundamental equilibrium fixed points:

1. **Einstein–de Sitter (EdS) Point** $(\Omega_K, \Omega_\Lambda) = (0, 0)$**:** Flat, matter-dominated universe $(\Omega_m = 1)$.
2. **de Sitter (dS) Point** $(\Omega_K, \Omega_\Lambda) = (0, 1)$**:** Vacuum-dominated accelerating universe $(\Omega_\Lambda = 1)$.
3. **Milne (M) Point** $(\Omega_K, \Omega_\Lambda) = (1, 0)$**:** Empty hyperbolic universe $(\Omega_K = 1)$.

Table 2.1: Linear Stability of Cosmological Fixed Points as a Function of $\gamma = 1 + w$

Fixed Point	$\gamma \in (-\infty, 0)$	$\gamma = 0$	$\gamma \in (0, 2/3)$	$\gamma = 2/3$	$\gamma \in (2/3, +\infty)$
Einstein–de Sitter (EdS)	Attractor	—	Saddle	—	Repeller
de Sitter (dS)	Saddle	Attractor	Attractor	Attractor	Attractor
Milne (M)	Saddle	Repeller	Repeller	Repeller	Saddle

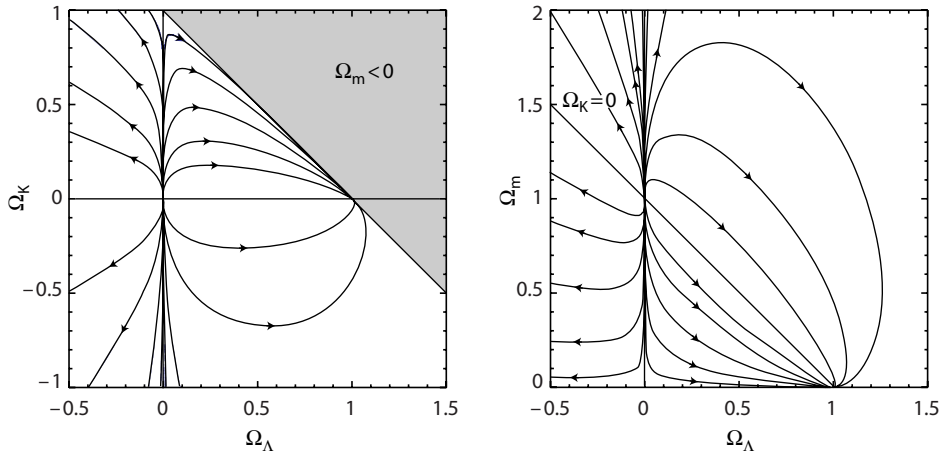


Figure 2.4: Phase space trajectories of the cosmological dynamical system in the $(\Omega_K, \Omega_\Lambda)$ plane (left) and $(\Omega_m, \Omega_\Lambda)$ plane (right) for dust $(\gamma = 1, w = 0)$. Arrows indicate the direction of cosmic time evolution.

Linear stability analysis (Table 2.1 and Figure 2.4) shows that for physical matter $(\gamma \geq 1)$, the de Sitter point is the unique global **future attractor**, while the Einstein–de Sitter state is an **unstable repeller** (the flatness problem of standard Big Bang cosmology).

The analytical boundary in the $(\Omega_{m0}, \Omega_{\Lambda 0})$ plane separating universes with eternal expansion from those collapsing into a Big Crunch is:

$$\Omega_{\Lambda 0} = \begin{cases} 0 & (\Omega_{m0} \leq 1), \\ 1 - \Omega_{m0} + \frac{2}{3} \cos\left(\arctan \Xi_{m0}^{-1} - \pi\right) & (\Omega_{m0} \geq 1), \end{cases} \tag{2.3.31}$$

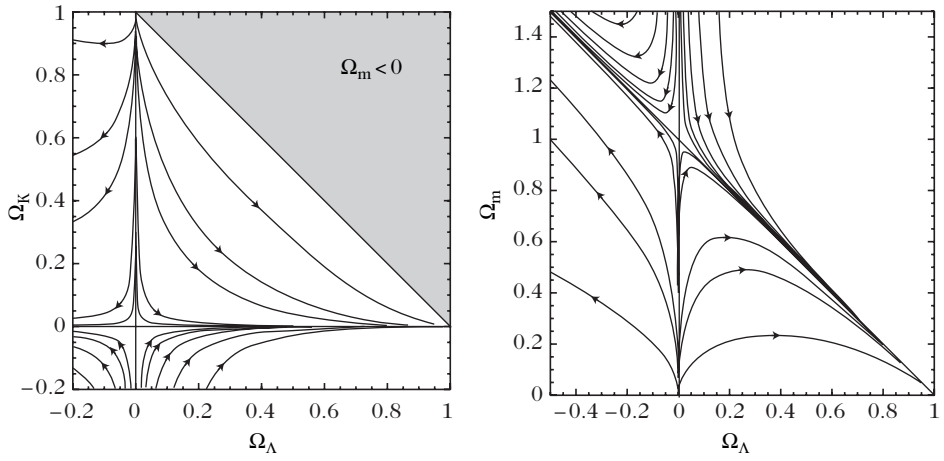


Figure 2.5: Phase space trajectories of the cosmological dynamical system in the $(\Omega_K, \Omega_\Lambda)$ plane (left) and $(\Omega_m, \Omega_\Lambda)$ plane (right) for radiation ($\gamma = 4/3, w = 1/3$).

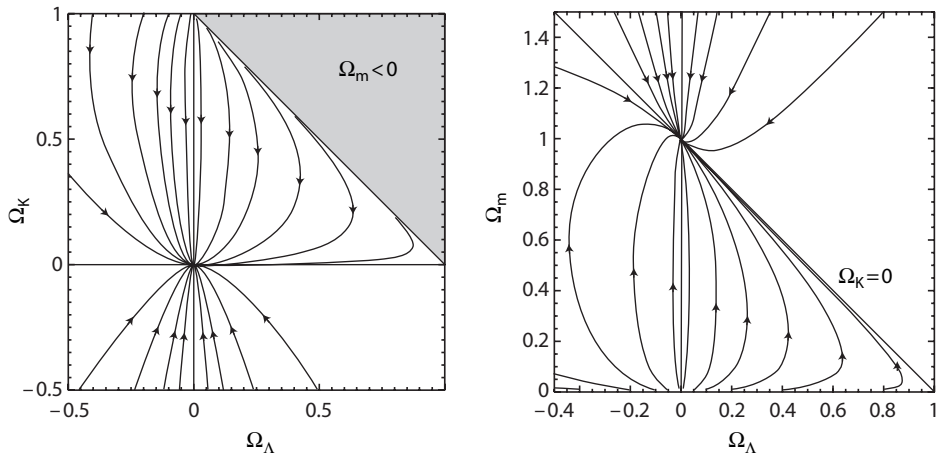


Figure 2.6: Phase space trajectories of the cosmological dynamical system in the $(\Omega_K, \Omega_\Lambda)$ plane (left) and $(\Omega_m, \Omega_\Lambda)$ plane (right) for a phantom / vacuum fluid ($\gamma = 0, w = -1$).

where $\Xi_{m0} \equiv \sqrt{|2\Omega_{m0} - 1|}$. Similarly, the boundary separating eternally expanding universes from bouncing cosmologies (which lack a past Big Bang singularity) is:

$$\Omega_{\Lambda 0} = \begin{cases} 1 - \Omega_{m0} + \frac{3}{2}\Omega_{m0}^{1/3}\left[(1 - \Omega_{m0} + \Xi_{m0})^{1/3} + (1 - \Omega_{m0} - \Xi_{m0})^{1/3}\right] & \left(0 \leq \Omega_{m0} \leq \frac{1}{2}\right), \\ 1 - \Omega_{m0} + \frac{2}{3}\cos\left(\arctan \Xi_{m0}^{-1} + \pi\right) & \left(\Omega_{m0} \geq \frac{1}{2}\right). \end{cases} \tag{2.3.32}$$

2.4 Cosmological timescales and distance measures

The characteristic distance and time scales of an expanding FLRW universe are set by the current Hubble expansion rate $H_0 \equiv 100\, h\, \mathrm{km \cdot s^{-1} \cdot Mpc^{-1}}$:

$$t_{H0} \equiv \frac{1}{H_0} = 9.778 \times 10^9\, h^{-1}\, \mathrm{yr} \approx 13.97\, \mathrm{Gyr} \quad (h \approx 0.70), \tag{2.4.1}$$

$$D_{H0} \equiv \frac{c}{H_0} = 2997.9\, h^{-1}\,\mathrm{Mpc} \approx 4280\,\mathrm{Mpc}. \tag{2.4.2}$$

The sphere of radius $D_H(t) \equiv c/H(t)$ is the **Hubble sphere**, outside of which recession velocities exceed the speed of light ($\mathbf{v}_{\rm rec} > c$).

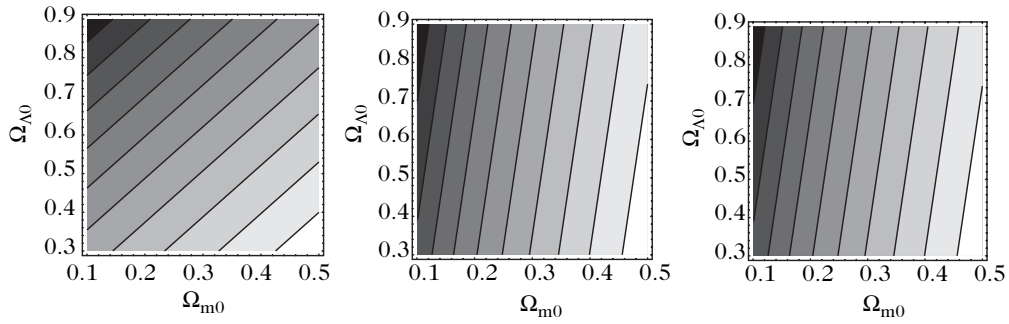


Figure 2.7: Isocontours of the normalized expansion rate function $E(z)$ in the $(\Omega_{m0}, \Omega_{\Lambda 0})$ plane at redshifts $z = 1$ (left), $z = 3$ (middle), and $z = 10$ (right). The degeneracy directions rotate with redshift, demonstrating how multi-epoch observations break cosmological parameter degeneracies.

2.4.1 Age of the Universe and look-back time

From the Hubble definition $H(a) = \frac{1}{a}\frac{\mathrm{d}a}{\mathrm{d}t} \implies \mathrm{d}t = \frac{\mathrm{d}a}{aH(a)} = \frac{\mathrm{d}a}{aH_0E(a)} = \frac{t_{H0}\,\mathrm{d}a}{aE(a)}$, the **cosmic age** of the Universe today is:

$$t_0 = t_{H0}\int_0^1 \frac{a_0}{a}\frac{\mathrm{d}a}{a_0}\frac{1}{E(a)} = t_{H0}\int_0^1 \frac{\mathrm{d}x}{xE(x)} = t_{H0}\int_0^\infty \frac{\mathrm{d}z}{(1+z)E(z)}. \tag{2.4.3}$$

The conformal age is:

$$d\eta = \frac{dt}{a} \implies \eta_0 = t_{H0}\int_0^1 \frac{\mathrm{d}x}{x^2E(x)} = t_{H0}\int_0^\infty \frac{\mathrm{d}z}{E(z)}. \tag{2.4.4}$$

For a light source observed at redshift z_*, the cosmic time at emission was:

$$t(z_*) = t_{H0}\int_{z_*}^\infty \frac{\mathrm{d}z}{(1+z)E(z)}. \tag{2.4.5}$$

The **look-back time** $\Delta t(z_*) \equiv t_0 - t(z_*)$ represents the elapsed time since the light departed the source:

$$\Delta t(z_*) = t_{H0}\int_0^{z_*} \frac{\mathrm{d}z}{(1+z)E(z)}. \tag{2.4.6}$$

2.4.2 Comoving radial distance

Along a radial null geodesic arriving at the observer at $\chi = 0$ and $t = t_0$, the FLRW metric gives

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + a^2(t)\,\mathrm{d}\chi^2 = 0 \quad \implies \quad \mathrm{d}\chi = \frac{c\,\mathrm{d}t}{a(t)}. \tag{2.4.7}$$

Integrating from emission at t_* to observation at t_0, we obtain

$$a_0\chi(z_*) = ca_0\int_{t_*}^{t_0}\frac{\mathrm{d}t}{a(t)} = c\int_{t_*}^{t_0}\frac{a_0}{a(t)}\frac{\mathrm{d}(\frac{a}{a_0})}{(\frac{a}{a_0})H(\frac{a}{a_0})} = \frac{c}{H_0}\int_{x_*}^1\frac{\mathrm{d}x}{x^2E(x)}. \tag{2.4.8}$$

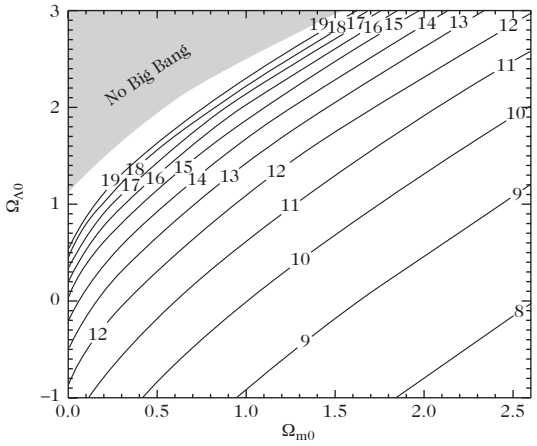


Figure 2.8: Isocontours of the present cosmic age t_0 (in units of 10^9 yr) in the $(\Omega_{m0}, \Omega_{\Lambda 0})$ plane. The shaded region denotes non-singular bouncing models lacking a past Big Bang.

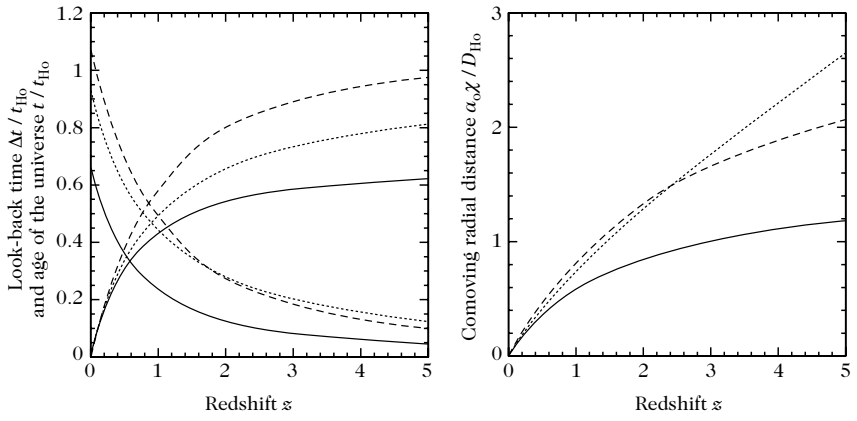


Figure 2.9: (Left) Cosmic age $t(z)/t_{H0}$ and look-back time $\Delta t(z)/t_{H0}$ as a function of redshift z for three cosmological models: Einstein–de Sitter ($\Omega_{m0} = 1, \Omega_{\Lambda 0} = 0$, solid), open ($\Omega_{m0} = 0.05, \Omega_{\Lambda 0} = 0$, dotted), and concordance flat ΛCDM ($\Omega_{m0} = 0.2, \Omega_{\Lambda 0} = 0.8$, dashed). (Right) Dimensionless comoving radial distance $a_0\chi(z)/D_{H0}$ for the same three models.

Since $x = (1+z)^{-1}$, we have $\mathrm{d}x\,/x^2 = -\,\mathrm{d}z$, and hence

$$a_0\chi(z_*) = \frac{c}{H_0}\int_0^{z_*}\frac{\mathrm{d}z}{E(z)} = D_{H0}\int_0^{z_*}\frac{\mathrm{d}z}{E(z)}, \tag{2.4.9}$$

where $D_{H0} \equiv c/H_0$ is the present-day Hubble distance. In a flat, matter-dominated universe $(\Omega_{m0} = 1, \Omega_{\Lambda 0} = 0, E(z) = (1+z)^{3/2})$:

$$a_0\chi(z) = 2D_{H0}\left(1 - \frac{1}{\sqrt{1+z}}\right). \tag{2.4.10}$$

2.4.3 Angular diameter distance and reciprocity theorem

Comoving angular diameter

The comoving cross-sectional area of a source subtending a solid angle $\mathrm{d}\Omega_{\rm obs}$ at comoving coordinate distance χ is $\mathrm{d}S_{\rm com} = R_{\rm ang}^2(\chi)\,\mathrm{d}\Omega_{\rm obs}$, where:

$$R_{\rm ang}(z) = f_K\left[\chi(z)\right], \tag{2.4.11}$$

with $f_K(\chi)$ given by Eq. (2.1.9).

Physical angular diameter distance

The **angular diameter distance** $D_A(z)$ is defined as the ratio of the source's proper transverse physical diameter $\mathrm{d}\ell_{\rm phys}$ to its observed angular size $\mathrm{d}\theta_{\rm obs}$:

$$D_A(z) \equiv \frac{\mathrm{d}\ell_{\rm phys}}{\mathrm{d}\theta_{\rm obs}} = \sqrt{\frac{\mathrm{d}S_{\rm phys}}{\mathrm{d}\Omega_{\rm obs}}} = a(t_{\rm em})R_{\rm ang}(z) = \frac{a_0 f_K\left[\chi(z)\right]}{1+z}. \tag{2.4.12}$$

Because $1+z$ grows in the denominator, $D_A(z)$ is **non-monotonic**: it increases at low redshift, reaches a maximum at $z \approx 1.5$, and then decreases at higher redshifts (Figure 2.10, left). Beyond this turnover redshift, distant objects appear with larger angular sizes on the sky.

Etherington's reciprocity theorem

If light travels along null geodesics and photons are conserved, the angular distance r_s measured from the source to the observer relates to the observer-measured distance $r_o = D_A$ by **Etherington's reciprocity theorem**:

$$r_s = r_o\,(1+z) = D_A(z)\,(1+z)\,. \tag{2.4.13}$$

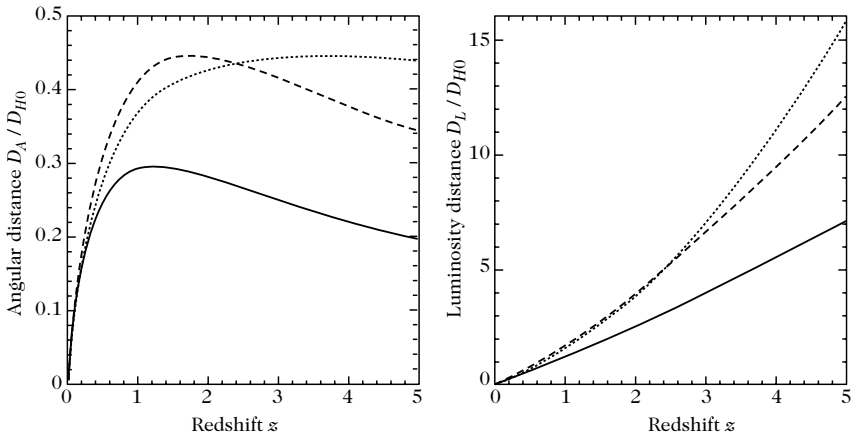


Figure 2.10: (Left) Angular diameter distance $D_A(z)/D_{H0}$ and (Right) Luminosity distance $D_L(z)/D_{H0}$ as a function of redshift z for three cosmological models: Einstein–de Sitter (solid), open universe (dotted), and flat ΛCDM (dashed).

2.4.4 Luminosity distance and the distance modulus

The luminosity distance

The **luminosity distance** $D_L(z)$ is defined by the inverse-square law relating the intrinsic bolometric luminosity L_{source} to the observed energy flux $\mathcal{F}_{\text{obs}}$:

$$\mathcal{F}_{\text{obs}} \equiv \frac{L_{\text{source}}}{4\pi D_L^2(z)}. \tag{2.4.14}$$

Two relativistic factors reduce the received bolometric flux:

1. **Photon Energy Redshift:** Individual photon energies are redshifted by $E_{\text{obs}} = E_{\text{em}}/(1+z)$.
2. **Time Dilation of Arrival Rate:** The arrival rate of photons is dilated by $\Delta t_{\text{obs}} = \Delta t_{\text{em}}(1+z)$.

Therefore, $L_{\text{obs}} = \frac{\Delta E_{\text{obs}}}{\Delta t_{\text{obs}}} = \frac{\Delta E_{\text{em}}}{\Delta t_{\text{em}}}\left(\frac{a}{a_0}\right)^2 = L_{\text{source}}/(1+z)^2$. Equating $\mathcal{F}_{\text{obs}} = \frac{L_{\text{source}}/(1+z)^2}{4\pi a_0^2 f_K^2(\chi)}$ with Eq. (2.4.14) yields:

$$D_L(z) = a_0\,(1+z)\,f_K\left[\chi(z)\right]. \tag{2.4.15}$$

Distance duality relation

Comparing Eq. (2.4.12) and Eq. (2.4.15), we obtain the exact **Etherington distance duality relation**:

$$D_L(z) = (1+z)^2\,D_A(z). \tag{2.4.16}$$

This relation is strictly geometric, depending only on Riemannian metric spacetime and photon conservation.

Differentiating $D_L(z)/(1+z)$ enables a direct model-independent reconstruction of the expansion rate $H(z)$:

$$\frac{1}{H(z)} = \frac{1}{c}\left[1 + \Omega_{K0}H_0^2\left(\frac{D_L(z)}{c(1+z)}\right)^2\right]^{-1/2}\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{D_L(z)}{1+z}\right]. \tag{2.4.17}$$

One can derive this by considering

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{D_L}{1+z}\right) = a_0\frac{\mathrm{d}f_K}{\mathrm{d}z} = a_0\frac{\mathrm{d}f_K}{\mathrm{d}\chi}\frac{\mathrm{d}\chi}{\mathrm{d}z} \tag{1}$$

$$= \frac{\mathrm{d}f_K}{\mathrm{d}\chi}\frac{1}{H(z)} = \sqrt{1 - Kf_K^2}\,\frac{1}{H(z)} \tag{2}$$

$$= \sqrt{1 - \frac{K}{a_0^2}\left(\frac{D_L}{1+z}\right)^2}\,\frac{1}{H(z)} \tag{3}$$

which using (2.3.16) becomes:

$$\frac{1}{H(z)} = \frac{1}{c}\left[1 + \Omega_{K0}H_0^2\left(\frac{D_L(z)}{c(1+z)}\right)^2\right]^{-1/2}\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{D_L(z)}{1+z}\right]. \tag{4}$$

Distance modulus and astronomical magnitudes

We usually express luminosity distance in terms of the *distance modulus*, $m - M$. This quantity is defined so that it equals zero when a light source is located at a standard distance of 10 pc:

$$m - M = -2.5\log\left[\frac{\phi(z)}{\phi(10\text{ pc})}\right]. \tag{2.4.18}$$

[1]

Weber–Fechner Law

Astronomical magnitudes use a logarithmic scale based on the Weber–Fechner law of human perception. This psychophysical law states that the minimum perceptible change in a stimulus, $\mathrm{d}S$, is directly proportional to the stimulus intensity S, such that $\mathrm{d}S \propto S$. Integrating this relationship yields the mathematical expression:

$$\rho(S) - \rho(S_0) = \alpha \ln\left(\frac{S}{S_0}\right), \tag{1}$$

where S_0 is a reference stimulus level, $\rho(S)$ represents the perceived intensity (sensation), and α is an empirical proportionality constant.

Using Eq. (2.4.18) and noting that at a distance of 10 pc, the luminosity distance is simply 10 pc, we obtain $m - M = 5\log[D_{\mathrm{L}}(z)/10\ \mathrm{pc}]$.[2] Substituting $10\ \mathrm{pc} = D_{\mathrm{H}_0}/(3\times 10^8 h)$ allows us to simplify this relation to:

$$m - M = 25 + 5\log\left[3000\frac{D_{\mathrm{L}}}{D_{\mathrm{H}_0}}\right] - 5\log h. \tag{2.4.19}$$

Here, m is the apparent magnitude, and the absolute magnitude M is defined relative to the Solar luminosity:

$$M - M_\odot = -2.5\log_{10}\left(\frac{L/4\pi D_L^2}{L_\odot/4\pi D_L^2}\right) \implies M = -2.5\log\left(\frac{L}{L_\odot}\right) + 4.76, \tag{2.4.20}$$

where $L_\odot \sim 3.8\times 10^{26}$ W is the Sun's luminosity, and $M_\odot = 4.76$ is its absolute magnitude. Under this definition, brighter objects have lower (or more negative) magnitudes.

Photometric filter systems and cosmological K-corrections

The luminosities derived so far represent total power emitted across all wavelengths, defining the *bolometric magnitude*. Because measuring full-spectrum bolometric magnitudes directly is challenging, astronomers use optical filters to isolate specific frequency bands. The magnitude M_X within a given filter band X is defined as:

$$M_X = -2.5\log\left(\frac{L_X}{L_{\odot X}}\right) + M_{\odot X}. \tag{2.4.21}$$

The standard astronomical filter bands and their corresponding properties are summarized below:

Light detected in a frequency band $\Delta\nu_X$ was emitted at a higher rest-frame frequency band $(1+z)\Delta\nu_X$ due to cosmic expansion. The observed flux per unit frequency is given by:

$$\phi_X = \frac{\mathrm{d}\phi}{\mathrm{d}\nu_0} = \frac{L_\nu d\nu_e}{4\pi D_L^2\,\mathrm{d}\nu_0} = \frac{L_\nu}{4\pi D_L^2}\frac{\mathrm{d}\nu_e}{\mathrm{d}\nu_o} \implies \phi_X(\nu) = (1+z)\frac{L_\nu[\nu(1+z)]}{4\pi D_{\mathrm{L}}^2}, \tag{2.4.22}$$

where L_ν is the source luminosity per unit frequency. Integrating over the bandwidth yields:

$$\phi_X(\nu) = \frac{(1+z)}{4\pi D_{\mathrm{L}}^2}\int_{\Delta\nu_X} L_\nu[\nu(1+z)]\,\mathrm{d}\nu. \tag{2.4.23}$$

[1] The factor of -2.5 is a historical convention chosen to match the magnitude scale originally established by Hipparcos.

[2] For same source such that L_{source} cancels.

Table 2.2: Standard Johnson–Cousins photometric filter bands and solar absolute magnitudes.

Band	**Central Wavelength (nm)**	**Width (nm)**	$M_\odot$
U (ultraviolet)	365	66	5.61
B (blue)	445	94	5.48
V (visible)	551	88	4.64
R (red)	658	138	4.42
I (infra-red)	806	149	4.08
J	1200	213	3.64
K	2190	390	3.28
Bolometric	—	∞	4.76

Because spectral deformation vanishes when a source is placed at the reference distance of 10 pc, the distance modulus in band X ($m_X - M_X = -2.5\log[\phi_X/\phi_X(10\text{ pc})]$) takes the form:

$$m_X - M_X = 5\log\left(\frac{D_{\rm L}}{10\text{ pc}}\right) + K, \tag{2.4.24}$$

where the K-correction accounts for spectral distortion caused by cosmological expansion:

$$K = -2.5\log(1+z) - 2.5\log\left\{\frac{\int_{\Delta\nu_X} L_\nu[\nu(1+z)]\,\mathrm{d}\nu}{\int_{\Delta\nu_X} L_\nu[\nu]\,\mathrm{d}\nu}\right\}. \tag{2.4.25}$$

Additional physical effects, such as interstellar or intergalactic extinction, can further modify the apparent magnitude.

2.4.5 Cosmic volume elements and galaxy number counts

The proper volume element $\mathrm{d}V_{\rm prop}$ subtended by a solid angle $\mathrm{d}\Omega$ in the redshift interval $[z, z+\mathrm{d}z]$ is the product of the transverse physical area $\mathrm{d}S_{\rm phys} = D_A^2(z)\,\mathrm{d}\Omega$ and proper radial line element $c\,\mathrm{d}t = \frac{D_{H0}\,\mathrm{d}z}{(1+z)E(z)}$:

$$\frac{\mathrm{d}V_{\rm prop}}{\mathrm{d}\Omega\,\mathrm{d}z} = D_A^2(z)\frac{c\,\mathrm{d}t}{\mathrm{d}z} = D_{H0}\frac{a_0^2 f_K^2\left[\chi(z)\right]}{(1+z)^3 E(z)}. \tag{2.4.26}$$

For a population of galaxies with proper number density $n_{\rm prop}(z)$, the differential number count is:

$$\frac{\mathrm{d}N}{\mathrm{d}\Omega\,\mathrm{d}z} = n_{\rm prop}(z)\frac{\mathrm{d}V_{\rm prop}}{\mathrm{d}\Omega\,\mathrm{d}z} = D_{H0}n_{\rm prop}(z)\frac{a_0^2 f_K^2\left[\chi(z)\right]}{(1+z)^3 E(z)}. \tag{2.4.27}$$

For non-evolving conserved galaxies with comoving number density $n_{\rm com}$, $n_{\rm prop}(z) = n_{\rm com}(1+z)^3$, simplifying the count to $\frac{\mathrm{d}N}{\mathrm{d}\Omega\mathrm{d}z} = D_{H0}n_{\rm com}\frac{a_0^2 f_K^2[\chi(z)]}{E(z)}$.

The cumulative **optical depth** $\tau(z)$ for a line of sight to intersect a galaxy of cross-sectional area $\sigma = \pi r_*^2$ up to redshift z is:

$$\tau(z) = \pi r_*^2 D_{H0}\int_0^z \frac{n_{\rm prop}(z')}{(1+z')E(z')}\,\mathrm{d}z'\,. \tag{2.4.28}$$

2.5 Kinematic expansions at small redshifts ($z \ll 1$)

2.5.1 Taylor expansion of the scale factor and deceleration parameter

Taylor expanding the scale factor $a(t)$ around the present epoch t_0:

$$a(t) = a_0 \left[1 + H_0 (t - t_0) - \frac{1}{2} q_0 H_0^2 (t - t_0)^2 + \frac{1}{6} j_0 H_0^3 (t - t_0)^3 + \dots\right], \tag{2.5.1}$$

where $H_0 \equiv \frac{\dot{a}_0}{a_0}$ is the Hubble constant, and q_0 is the dimensionless **deceleration parameter**:

$$q_0 \equiv -\frac{\ddot{a}_0 a_0}{\dot{a}_0^2} = -\left.\frac{\ddot{a}}{aH^2}\right|_{t=t_0}. \tag{2.5.2}$$

The sign convention is historical: $q_0 > 0$ corresponds to decelerating expansion, while $q_0 < 0$ signifies cosmic acceleration.

From the second Friedmann equation (2.3.3), q_0 is directly linked to the present cosmic energy content:

$$q_0 = \frac{1}{2} \sum_X (1 + 3w_X)\, \Omega_{X0} = \frac{1}{2}\Omega_{m0} + \Omega_{r0} - \Omega_{\Lambda 0}. \tag{2.5.3}$$

For a flat ΛCDM universe ($\Omega_{m0} \approx 0.3, \Omega_{\Lambda 0} \approx 0.7$), $q_0 = \frac{1}{2}(0.3) - 0.7 = -0.55 < 0$, confirming current cosmic acceleration.

2.5.2 Small-z expansions of cosmological observables

Expanding $E(z) = \sqrt{\Omega_{m0}(1+z)^3 + \Omega_{K0}(1+z)^2 + \Omega_{\Lambda 0}}$ to second order in z:

$$E(z) = 1 + (1 + q_0)\, z + \mathcal{O}\left(z^2\right). \tag{2.5.4}$$

Integrating the distance and time relations yields the classical low-redshift kinematic expansions:

$$\Delta t(z) = t_{H0} \left[z - \frac{1}{2}(q_0 + 2)\, z^2 + \mathcal{O}\left(z^3\right)\right], \tag{2.5.5}$$

$$a_0 \chi(z) = D_{H0} \left[z - \frac{1}{2}(1 + q_0)\, z^2 + \mathcal{O}\left(z^3\right)\right], \tag{2.5.6}$$

$$D_L(z) = D_{H0} \left[z + \frac{1}{2}(1 - q_0)\, z^2 + \mathcal{O}\left(z^3\right)\right], \tag{2.5.7}$$

$$D_A(z) = D_{H0} \left[z - \frac{1}{2}(3 + q_0)\, z^2 + \mathcal{O}\left(z^3\right)\right]. \tag{2.5.8}$$

The distance modulus at low redshift is:

$$\mu(z) = 25 + 5 \log_{10}\left(\frac{cz}{H_0\, \mathrm{Mpc}}\right) + 1.086\, (1 - q_0)\, z + \mathcal{O}\left(z^2\right), \tag{2.5.9}$$

providing the foundational formula for measuring H_0 and q_0 using Type Ia supernovae.

2.5.3 The cosmic redshift drift (Sandage–Loeb effect)

If an astronomical source is observed at epoch t_0 with redshift z and monitored over a time interval δt_0, the cosmic expansion induces a continuous secular **redshift drift**:

$$1 + z + \delta z = \frac{a(t_0 + \delta t_0)}{a(t_{\mathrm{em}} + \delta t_{\mathrm{em}})}. \tag{2.5.10}$$

Using $\delta t_{\mathrm{em}} = \delta t_0/(1+z)$, the exact time rate of change of redshift is:

$$\frac{\delta z}{\delta t_0} = H_0 \left[(1 + z) - E(z)\right]. \tag{2.5.11}$$

At small redshift ($z \ll 1$), this reduces to:

$$\frac{\delta z}{\delta t_0} = -q_0 H_0 z + \mathcal{O}\left(z^2\right). \tag{2.5.12}$$

Though tiny ($\delta z/\delta t_0 \sim 10^{-8}\,\text{century}^{-1}$), measuring this spectroscopic drift provides a direct, non-geometric probe of cosmic acceleration free of source astrophysical evolution.

2.6 Cosmological horizons and global conformal structure

An **horizon** is a null or spacelike hypersurface that divides spacetime into causally accessible and inaccessible regions (Figures 2.11 and 2.12).

and 3.12.

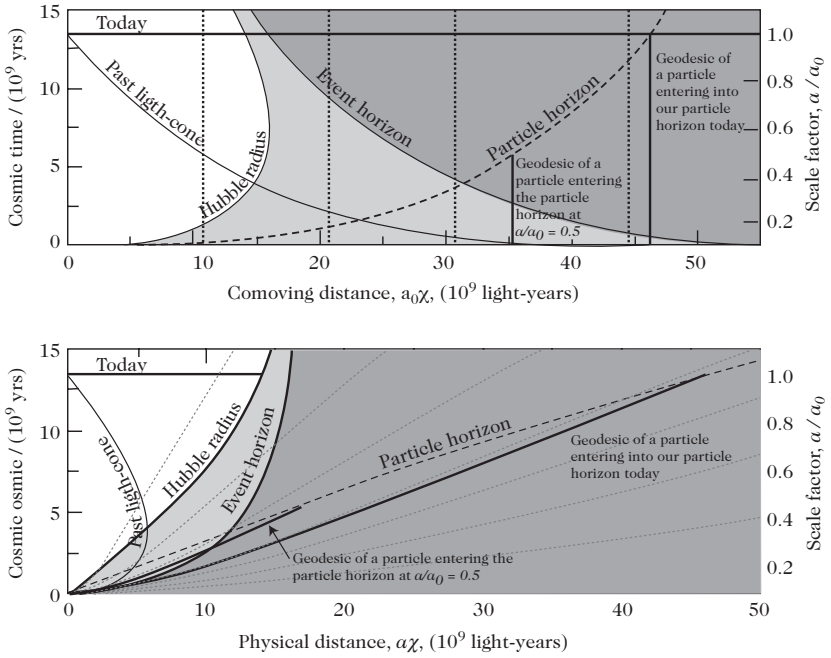


Figure 2.11: Particle horizon geometry in comoving coordinates (top) and physical coordinates (bottom) for a ΛCDM universe ($\Omega_{m0} = 0.3, \Omega_{\Lambda 0} = 0.7, h = 0.7$). The particle horizon $\chi = \varphi(t)$ expands monotonically, engulfing more comoving fundamental observers over cosmic time.

2.6.1 The cosmological event horizon

For a fundamental observer O, the **event horizon** is the boundary separating events that can ever emit a signal that reaches O from events that remain forever causally disconnected from O.

A necessary and sufficient condition for the existence of an event horizon is the convergence of the conformal time integral as $t \to \infty$:

$$\chi_{\text{event}}(t_0) = \int_{t_0}^{\infty} \frac{c\ \mathrm{d}t}{a(t)} < \infty. \tag{2.6.1}$$

Any photon emitted at epoch t_0 at comoving distance $\chi > \chi_{\text{event}}(t_0)$ will never reach the observer.

A photon on the $t = t_0$ hypersurface at $\chi = \chi_0$ reaches the observer at origin at future infinity. If $\chi_0 = \infty$ then all of $t = t_0$ hypersurface is observable otherwise the observer can only observe events on $t = t_0$ hypersurface from $\chi = 0$ to $\chi = \chi_0$.

For single-fluid power-law universes $a(t) \propto t^n$, the integral converges if and only if $n > 1$, which translates to:

$$w < -\frac{1}{3} \iff \rho + 3P < 0 \quad \text{(Violation of Strong Energy Condition)}. \tag{2.6.2}$$

In de Sitter spacetime ($a = \mathrm{e}^{Ht}$), the physical event horizon radius is static:

$$R_{\text{event}}(t_0) = a(t_0)\chi_{\text{event}}(t_0) = \frac{c}{H} = \text{constant}. \tag{2.6.3}$$

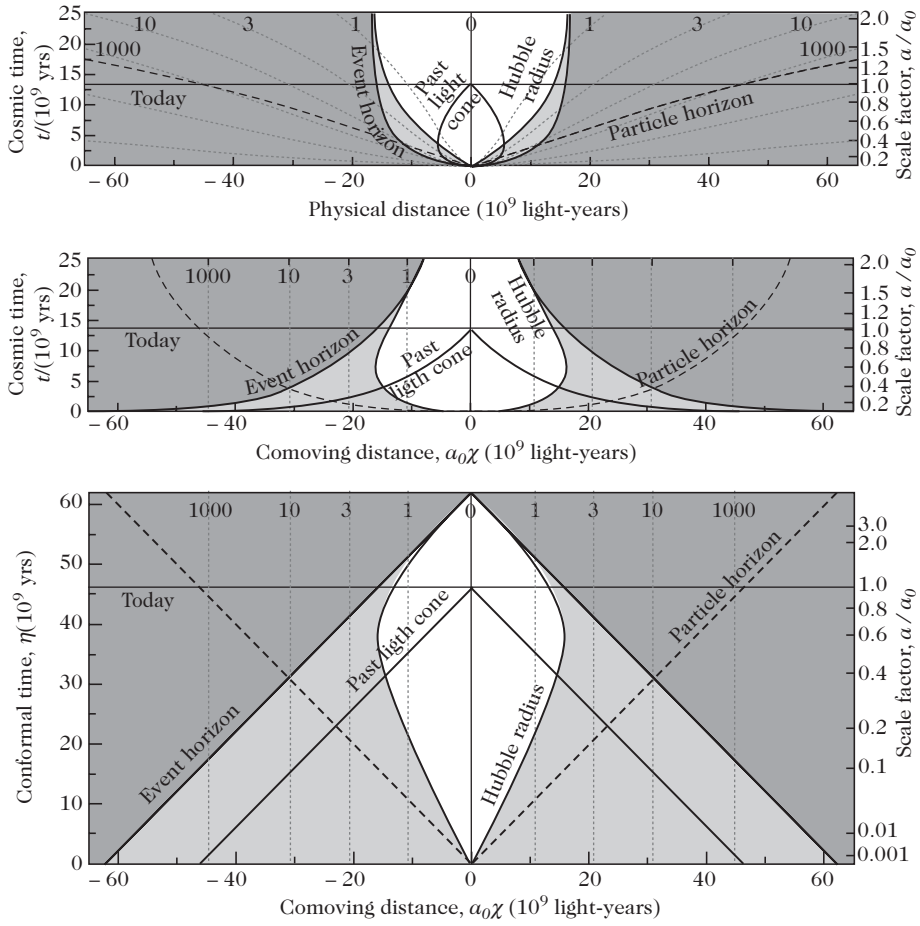


Figure 2.12: Spacetime diagrams for a concordance ΛCDM model ($\Omega_{m0} = 0.3, \Omega_{\Lambda 0} = 0.7, h = 0.7$) plotted in: (Top) physical coordinates $(a\chi, t)$, (Middle) comoving coordinates (χ, t), and (Bottom) conformal coordinates (χ, η). Shown are comoving galaxy worldlines (dotted), past light cone (bold line), event horizon (bold curve), Hubble sphere (solid line), and particle horizon (dashed line).

2.6.2 The cosmological particle horizon

The **particle horizon** at epoch t_0 is the 2-sphere dividing all particles into those that have been observable by O since the initial singularity ($t = 0$) and those that have not yet entered causal contact.

A particle horizon exists if and only if the past conformal time integral converges as $t \to 0$:

$$\chi_{\text{particle}}(t_0) = \int_0^{t_0} \frac{c\,\mathrm{d}t}{a(t)} < \infty. \tag{2.6.4}$$

For a power-law scale factor $a(t) \propto t^n$, a particle horizon exists if and only if $n < 1$:

$$w > -\frac{1}{3} \iff \rho + 3P > 0 \quad \text{(Standard decelerating cosmology)}. \tag{2.6.5}$$

In single-component power-law cosmologies, the event horizon and particle horizon are **mutually exclusive**:

- $w > -1/3$ (matter, radiation): Particle horizon exists; Event horizon does not.
- $w < -1/3$ (dark energy, de Sitter inflation): Event horizon exists; Particle horizon does not.

The physical diameter of the particle horizon at cosmic time t_2 for an initial emission epoch $t_1 \ll t_2$ is:

$$D_{\text{p.h.}}(t_2) = 2a(t_2) \int_0^{t_2} \frac{c\,\mathrm{d}t}{a(t)} = \frac{4}{1+3w} \frac{c}{H(t_2)} = \frac{4}{1+3w} D_H(t_2). \tag{2.6.6}$$

For matter domination ($w = 0$), $D_{\text{p.h.}}(t) = 4D_H(t) = 6ct$; for radiation domination ($w = 1/3$), $D_{\text{p.h.}}(t) = 2D_H(t) = 4ct$.

2.6.3 Global conformal structure and Penrose–Carter diagrams

To analyze the global asymptotic causal structure of cosmological spacetimes, we perform a conformal transformation $\tilde{g}_{\mu\nu} = \Omega^2 g_{\mu\nu}$ mapping infinite physical spacetime onto a compact manifold $\widetilde{\mathcal{M}}$ with boundary $\partial\widetilde{\mathcal{M}} = \mathscr{I}$ (scri), where light cones remain at $\pm 45°$.

Minkowski spacetime as a conformal archetype

In advanced and retarded null coordinates $v = t + r$ and $u = t - r$, defining $V = \arctan v$ and $U = \arctan u$, followed by $T = V + U$ and $R = V - U$, the Minkowski metric is conformally embedded in the **Einstein static cylinder** $\mathbb{R} \times S^3$:

$$\mathrm{d}\bar{s}^2 = -\,\mathrm{d}T^2 + \mathrm{d}R^2 + \sin^2 R\,\mathrm{d}\Omega^2\,, \qquad -\pi < T \pm R < \pi, \quad R \geq 0. \tag{2.6.7}$$

The boundary consists of:

- $\mathscr{I}^+$ (Future Null Infinity) and $\mathscr{I}^-$ (Past Null Infinity): 3D null hypersurfaces where outgoing/ingoing null rays terminate and originate.
- i^+ (Future Timelike Infinity) and i^- (Past Timelike Infinity): Vertices where timelike geodesics end and begin.
- i^0 (Spatial Infinity): Endpoint of all spacelike geodesics.

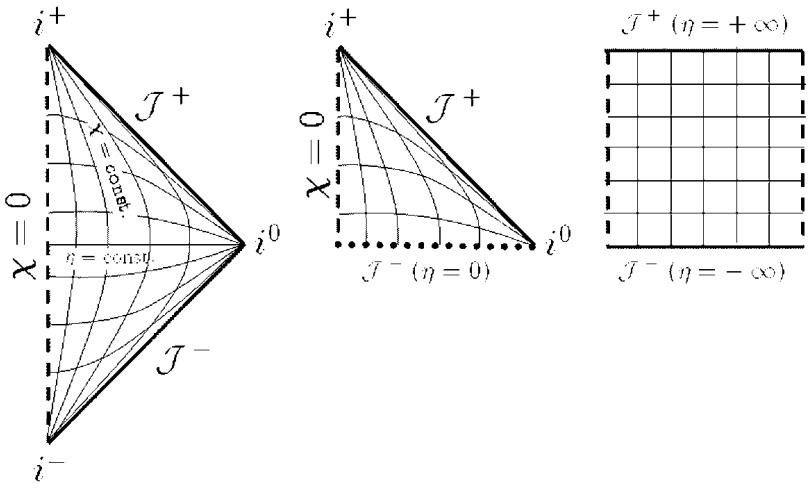


Figure 2.13: Conformal Penrose diagrams: (Left) Minkowski spacetime diamond, (Middle) Spatially flat FLRW universe with $\Lambda = 0, P > 0$ featuring an initial spacelike Big Bang singularity, and (Right) de Sitter spacetime slab with spacelike past and future conformal infinities $\mathscr{I}^{\pm}$.

Conformal diagrams of FLRW and de Sitter spacetimes

Because FLRW spacetimes are conformally related to static manifolds ($\mathrm{d}s^2 = a^2(\eta)[-\mathrm{d}\eta^2 + \mathrm{d}\sigma_K^2]$), their conformal diagrams correspond to sub-regions of the Einstein cylinder (Figures 2.13 and 2.14):

1. **Flat FLRW ($K = 0, \Lambda = 0, w \geq 0$):** Conformal time range $\eta \in (0, \infty)$ maps to the upper half of the Minkowski diamond ($T > 0$). The Big Bang at $\eta = 0$ is a **spacelike initial singularity boundary** (Figure 2.13, middle).

2. **de Sitter Spacetime:** The conformal time runs over $\eta \in (-\infty, 0)$ in flat coordinates, or $T \in (0, \pi)$ in global closed slicing, forming a square strip whose past and future boundaries $\mathscr{I}^{\pm}$ are **spacelike** (Figure 2.13, right).

3. **Closed FLRW ($K = +1$):** Conformal time spans $\eta \in (0, 2\pi)$ for cycloid dust collapse (spacelike Big Bang and Big Crunch), $\eta \in (0, \infty)$ for hesitating/Lemaître universes, and $\eta \in (-\infty, \infty)$ for non-singular bouncing models (Figure 2.14).

Theorem 2.1 (Conformal Horizon Criterion)**.** The existence of cosmological horizons is determined by the causal nature of the conformal boundaries:

$$\mathscr{I}^- \text{ is spacelike} \iff \text{Existence of a Particle Horizon,} \tag{2.6.8}$$

$$\mathscr{I}^+ \text{ is spacelike} \iff \text{Existence of an Event Horizon.} \tag{2.6.9}$$

2.7 Beyond the cosmological principle: anisotropic and inhomogeneous spacetimes

While the Friedmann–Lemaître model assumes maximal spatial symmetry (6 Killing vectors), testing the robustness of standard cosmological predictions requires examining exact solutions of Einstein's field equations with reduced symmetry groups.

If $\mathcal{J}^+$ is space-like, the worldline of any fundamental terminates on a point O of $\mathcal{J}^+$ and the past light cone of O divides the Universe into events that can be seen by the observer and events that he can never see; there is en event horizon. If $\mathcal{J}^+$ is null, all the worldlines will pass through the vertex i^+, so that $0 = i^+$ and its past light cone is $\mathcal{J}^+$ and there is no event horizon.

In conclusion,

$$\begin{aligned} \mathcal{J}^- \text{ space-like} &\iff \text{existence of a particle horizon,} \\ \mathcal{J}^+ \text{ space-like} &\iff \text{existence of an event horizon.} \end{aligned}$$

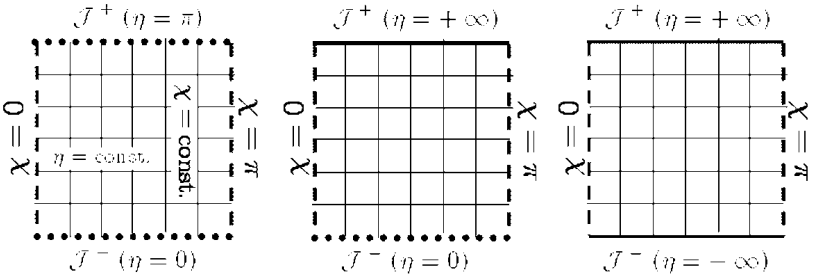


Figure 2.14: Conformal diagrams of FLRW spacetimes with spherical spatial sections ($K=+1$): (Left) Big Bang to Big Crunch cycloid ($\Lambda=0$), (Middle) Hesitating Lemaître universe with asymptotic de Sitter future ($\Lambda>0$), and (Right) Non-singular bouncing cosmology ($\Lambda>0$). Dashed horizontal boundaries represent curvature singularities.

2.7.1 Classification of spacetime isometries and Lie algebras

Continuous symmetries are generated by Killing vector fields satisfying Killing's equation $\mathcal{L}_\xi g_{\mu\nu} = \nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu = 0$.

The vector space of Killing fields forms a finite-dimensional Lie algebra of dimension $r \leq n(n+1)/2 = 10$ with Lie bracket defined by the commutator:

$$[\xi_a, \xi_b]^\alpha \equiv \xi_a^\beta \partial_\beta \xi_b^\alpha - \xi_b^\beta \partial_\beta \xi_a^\alpha = C^c{}_{ab} \xi_c^\alpha, \tag{2.7.1}$$

where $C^c{}_{ab} = -C^c{}_{ba}$ are the **structure constants** of the isometry group, satisfying the Jacobi identity:

$$C^e{}_{[ab} C^d{}_{c]e} = 0. \tag{2.7.2}$$

2.7.2 Classification of cosmological models

The isometry group of a 4-dimensional spacetime manifold $\mathcal{M}$ is characterized by:

- $s = \dim(\text{surfaces of transitivity / orbits}) \leq 4$,
- $q = \dim(\text{isotropy subgroup at a point}) \leq 3$,
- $r = s + q \leq 10$, the total dimension of the Lie algebra of Killing vector fields.

For physical cosmological fluids satisfying the null energy condition ($\rho + P > 0$), the 4-velocity u^μ is an invariant timelike vector, restricting spatial isotropy to subgroups of SO(3), so that $q \in \{0, 1, 3\}$.

Table 2.3: Classification of Cosmological Spacetimes by Homogeneity (s) and Isotropy (q)

Isotropy Group	**Spacetime Homogeneous** ($s = 4$)	**Spatially Homogeneous** ($s = 3$)	**Inhomogeneous** ($s = 2$)
$q = 6$ (Maximal 4D)	Minkowski, de Sitter, Anti-de Sitter	—	—
$q = 3$ (Maximal 3D)	Einstein static universe	**Friedmann–Lemaître (FLRW)**	—
$q = 1$ (Axial Symmetry)	LRS Bianchi types	LRS Bianchi, Kantowski–Sachs	**Lemaître–Tolman–Bondi**
$q = 0$ (Anisotropic)	—	**General Bianchi Types I–IX**	Szekeres models

2.7.3 Spatially inhomogeneous spacetimes

The Lemaître–Tolman–Bondi (LTB) metric

The simplest exact inhomogeneous cosmological solution is the **Lemaître–Tolman–Bondi (LTB)** spacetime ($s = 2, q = 1, r = 3$), describing a spherically symmetric, pressureless dust fluid ($P = 0$):

$$\mathrm{d}s^2 = -\,\mathrm{d}t^2 + S^2(t,r)\,\mathrm{d}r^2 + R^2(t,r)\,\mathrm{d}\Omega^2\,. \tag{2.7.3}$$

Einstein's field equations reduce to three coupled equations:

$$S(t,r) = \frac{R'(t,r)}{\sqrt{1-K(r)r^2}}, \tag{2.7.4}$$

$$\dot{R}^2(t,r) = \frac{2G_{\mathrm{N}}M(r)}{R(t,r)} - K(r)r^2, \tag{2.7.5}$$

$$4\pi\rho(t,r) = \frac{M'(r)}{R^2(t,r)R'(t,r)}, \tag{2.7.6}$$

where $R' \equiv \frac{\partial R}{\partial r}$ and $\dot{R} \equiv \frac{\partial R}{\partial t}$.

The general solution is specified by three arbitrary radial functions:

1. $M(r)$: The gravitational mass enclosed within the comoving shell r.
2. $E(r) \equiv -\frac{1}{2}K(r)r^2$: The local spatial curvature profile determining whether shells are locally hyperbolic, flat, or elliptic.
3. $t_0(r)$: The inhomogeneous "bang time" function describing the local epoch at which $R(t_0(r), r) = 0$.

The FLRW spacetime is recovered by imposing that the energy density be spatially homogeneous:

$$\rho'(t,r) = 0. \tag{2.7.7}$$

The mass equation (2.7.6) then gives:

$$M(r) = \frac{4\pi}{3}\rho(t)R^3(t,r), \tag{2.7.8}$$

where we have imposed the regularity condition:

$$R(t,0) = 0, \tag{2.7.9}$$

so that the integration constant vanishes at the symmetry centre. Introducing a reference density ρ_0, we can write:

$$R(t,r) = \left(\frac{3M(r)}{4\pi\rho_0}\right)^{1/3}\left(\frac{\rho_0}{\rho(t)}\right)^{1/3} = ra(t), \tag{2.7.10}$$

with the definitions:

$$r = \left(\frac{3M(r)}{4\pi\rho_0}\right)^{1/3}, \qquad a(t) = \left(\frac{\rho_0}{\rho(t)}\right)^{1/3}. \tag{2.7.11}$$

By using above in (2.7.5), we recover $K(r) = \text{const} \equiv K$. Thus, the FLRW spacetime is recovered as the homogeneous and isotropic special case:

$$R(t,r) = a(t)r, \qquad M(r) = \frac{4\pi}{3}\rho(t)a^3(t)r^3, \qquad K(r) = K. \tag{2.7.12}$$

The difference between the conventional Friedmann equation (2.3.2) and its LTB generalization, Eq. (2.7.5), is that all the quantities in the LTB case depend on the r-coordinate. In FLRW, one has rotational symmetry about every point, which is broken here and we only have rotational symmetry about the origin. LTB models are widely used to analyze nonlinear cosmic voids, local Hubble bubbles, and backreaction effects.

The Swiss-cheese model

Inhomogeneous Swiss-cheese models are constructed by cutting spherical cavities out of a background FLRW universe and inserting concentric spherically symmetric Schwarzschild mass distributions. By Darmois–Israel junction conditions, matching across the boundary requires:

$$M_{\text{Schw}} = \frac{4\pi}{3}\left[\rho a^3 R^3\right]_{\text{FLRW}}. \tag{2.7.13}$$

Because the cavity mass equals the missing FLRW mass, Birkhoff's theorem ensures that the gravitational field outside the cavity is completely unperturbed: the local inhomogeneity is gravitationally screened.

2.7.4 Spatially homogeneous anisotropic spacetimes (Bianchi models)

Spatially homogeneous models ($s = 3$) possess a 3-dimensional isometry group acting transitively on spatial slices Σ_t. The Lie algebra structure constants $C^c{}_{ab}$ can be decomposed as:

$$C^c{}_{ab} = \varepsilon_{dab}n^{cd} + \delta^c_b A_a - \delta^c_a A_b, \tag{2.7.14}$$

where $n^{ab} = \text{diag}(n_1, n_2, n_3)$ with $n_i \in \{0, \pm 1\}$, $A_b = (a, 0, 0)$, and the Jacobi identity requires $an_1 = 0$.

This classifies 3D Lie algebras into the **9 Bianchi types** (I through IX), partitioned into **Class A** ($a = 0$) and **Class B** ($a \neq 0$).

The Bianchi type I universe and shear dominance

The simplest anisotropic cosmology is the **Bianchi I** spacetime with flat spatial sections:

$$\mathrm{d}s^2 = -\,\mathrm{d}t^2 + X^2(t)\,\mathrm{d}x^2 + Y^2(t)\,\mathrm{d}y^2 + Z^2(t)\,\mathrm{d}z^2 = -\,\mathrm{d}t^2 + S^2(t)\mathrm{e}^{2\beta_i(t)}\,\mathrm{d}x^i\,\mathrm{d}x^i\,, \tag{2.7.15}$$

where $S(t) \equiv (XYZ)^{1/3}$ is the mean scale factor and $\sum_{i=1}^3 \beta_i(t) = 0$.

Table 2.4: Bianchi Classification of Spatially Homogeneous Anisotropic Spacetimes

Class	Bianchi Type	a	n_1	n_2	n_3	Isotropic FLRW Limit
Class A ($a = 0$)	**I**	0	0	0	0	Flat FLRW ($K = 0$)
	II	0	0	0	1	—
	VII$_0$	0	0	1	1	Flat FLRW ($K = 0$)
	VI$_0$	0	0	1	-1	—
	IX	0	1	1	1	Closed FLRW ($K = +1$, Mixmaster)
	VIII	0	1	1	-1	—
Class B ($a \neq 0$)	**V**	1	0	0	0	Open FLRW ($K = -1$)
	IV	1	0	0	1	—
	III	1	0	1	-1	—
	VII$_a$	a	0	1	1	Open FLRW ($K = -1$)
	VI$_a$	a	0	1	-1	—

The kinematic shear tensor $\sigma_{ij} = \frac{1}{2}\dot{\gamma}_{ij}$ evolves as:

$$\sigma^i{}_j = \frac{\Sigma^i{}_j}{S^3(t)}, \qquad \sigma^2 \equiv \frac{1}{2}\sigma_{ij}\sigma^{ij} = \frac{\Sigma^2}{2S^6(t)}, \tag{2.7.16}$$

where $\Sigma^i{}_j$ is a constant trace-free matrix.

The generalized Friedmann equation is:

$$3\left(\frac{\dot{S}}{S}\right)^2 = 8\pi G_{\rm N}\rho(t) + \frac{\Sigma^2}{2S^6(t)}. \tag{2.7.17}$$

Because the shear energy density scales as $\rho_\sigma \propto S^{-6}$, it **strictly dominates all matter and radiation components** ($\rho_m \propto S^{-3}, \rho_r \propto S^{-4}$) as $S \to 0$ near the initial singularity:

$$\lim_{S\to 0} \frac{\rho_\sigma}{\rho_{\rm fluid}} \to \infty. \tag{2.7.18}$$

In the vacuum shear-dominated limit (the **Kasner metric**):

$$X(t) \propto t^{p_1}, \quad Y(t) \propto t^{p_2}, \quad Z(t) \propto t^{p_3}, \qquad \sum_{i=1}^{3} p_i = 1, \quad \sum_{i=1}^{3} p_i^2 = 1. \tag{2.7.19}$$

Except for the flat static case $(1, 0, 0)$, one index is always negative ($p_1 < 0 < p_2 \leq p_3$): space contracts along one spatial axis while expanding along the other two.

As the universe expands ($S \to \infty$), the shear decays rapidly ($\rho_\sigma \propto S^{-6} \to 0$), and the spacetime undergoes natural **isotropization**, converging asymptotically to the isotropic Friedmann–Lemaître solution.

Chapter 3

The Standard Cosmological Model: Observational Tests and Paradigm

The transformation of cosmology from a speculative mathematical branch of general relativity into a rigorous empirical science is among the crowning achievements of modern physics. In the early decades of the twentieth century, the Friedmann–Lemaître–Robertson–Walker (FLRW) metric offered an elegant mathematical framework for a dynamic, homogeneous, and isotropic universe, but it remained detached from direct observational validation. Over the ensuing century, four independent observational pillars converged to elevate this theoretical architecture into the concordance ΛCDM model of cosmology:

1. **The Hubble Expansion of Spacetime:** The systematic, linear recession velocity of distant galaxies proportional to their distance ($v = H_0 d$), demonstrating that space itself is expanding dynamically.

2. **The Cosmic Microwave Background (CMB):** An omnipresent, isotropic relic radiation field with a Planck blackbody spectrum at $T_0 = 2.7255\,\mathrm{K}$, preserving a pristine snapshot of the hot, dense early plasma at the moment when neutral atoms first formed.

3. **Primordial Big-Bang Nucleosynthesis (BBN):** The synthesis of light element isotopes (^{4}He, D,^{3} He,^{7} Li) during the first twenty minutes of cosmic expansion, establishing that the early Universe was once a nuclear furnace operating at temperatures exceeding $10^9\,\mathrm{K}$.

4. **The Growth of Cosmological Large-Scale Structure:** The gravitational growth and clustering of tiny primordial matter density perturbations into the vast cosmic web of galaxies, filaments, and clusters observed across the contemporary cosmos.

This chapter traces the observational and thermodynamic foundations of the standard Big Bang paradigm. We begin by examining cosmic kinematics through the Hubble diagram, the standardization of Type Ia supernovae that revealed the present accelerated expansion, and the astrophysical resolution of the historical cosmological age crisis. We then turn to the microphysics of the expanding cosmic plasma: relativistic equilibrium thermodynamics, the non-equilibrium Boltzmann transport equation governing thermal relic freeze-out, the light-element synthesis network of BBN, and the atomic physics of recombination and the surface of last scattering.

3.1 The Hubble diagram and cosmic age constraints

3.1.1 The Hubble expansion law

In an expanding FLRW spacetime, the observed frequency of light emitted by a distant astronomical source is shifted towards longer wavelengths. A photon emitted with wavelength λ_{em} at

scale factor $a(t_{\rm em})$ is detected today at t_0 with wavelength $\lambda_{\rm obs}$ at scale factor $a(t_0) = 1$. The cosmological redshift z is defined kinematically as:

$$1 + z \equiv \frac{\lambda_{\rm obs}}{\lambda_{\rm em}} = \frac{a(t_0)}{a(t_{\rm em})} = \frac{1}{a(t_{\rm em})}. \tag{3.1.1}$$

A common conceptual pitfall for the beginner is to interpret the cosmological redshift as a conventional kinematic Doppler shift caused by galaxies moving through a pre-existing static space. In general relativity, galaxies are essentially stationary with respect to their local comoving coordinates; the redshift arises because the spacetime fabric itself stretches beneath the propagating light wave over its cosmic journey.

To connect this general relativistic stretching to local astronomical observations, consider a source at a relatively small physical distance d. In the nearby Universe ($z \ll 1$), light was emitted at a time $t_{\rm em}$ very close to the present epoch t_0. Taylor-expanding the cosmological scale factor about t_0 yields:

$$a(t) = a(t_0) + \dot{a}(t_0)(t - t_0) + \frac{1}{2}\ddot{a}(t_0)(t - t_0)^2 + \mathcal{O}\left[(t - t_0)^3\right]. \tag{3.1.2}$$

Setting $a(t_0) = 1$ and noting that the lookback time for a photon traveling along a radial null geodesic is $t_0 - t_{\rm em} \approx d/c + \mathcal{O}(d^2)$, the redshift simplifies to:

$$z = \frac{1}{a(t_{\rm em})} - 1 \approx \frac{\dot{a}(t_0)}{a(t_0)}\,(t_0 - t_{\rm em}) + \mathcal{O}\left[(t_0 - t_{\rm em})^2\right] = \frac{H_0}{c}d + \mathcal{O}(d^2), \tag{3.1.3}$$

where $H_0 \equiv \dot{a}(t_0)/a(t_0)$ is the present expansion rate, known as the **Hubble constant**. Interpreting the apparent recession velocity as $v \approx cz$, we recover the famous **Hubble Law**:

$$v = H_0 d \iff z = \frac{H_0}{c} D_L(z), \tag{3.1.4}$$

where $D_L(z)$ is the luminosity distance, and the proportionality constant is conventionally parameterized as $H_0 \equiv 100\,h\,\mathrm{km \cdot s^{-1} \cdot Mpc^{-1}}$.

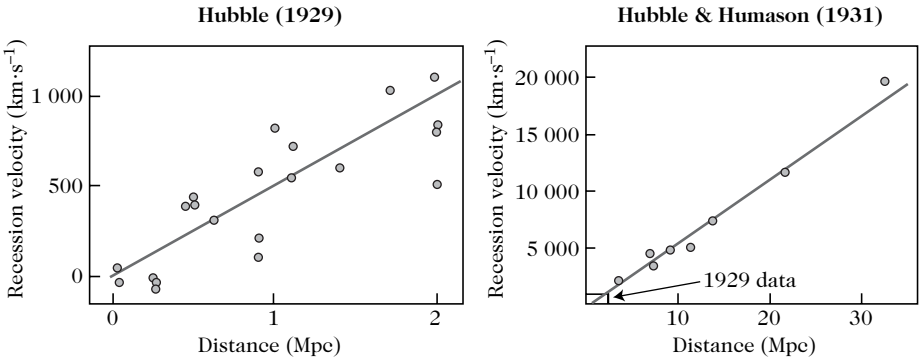


Figure 3.1: Historical Hubble diagrams: (Left) Edwin Hubble's landmark 1929 discovery using 18 galaxies out to $v \approx 1000\,\mathrm{km \cdot s^{-1}}$, yielding an initial overestimate $H_0 \approx 500\,\mathrm{km \cdot s^{-1} \cdot Mpc^{-1}}$. (Right) Hubble and Humason's 1931 extension out to $v \approx 20\,000\,\mathrm{km \cdot s^{-1}}$ confirming linearity across larger cosmic volumes.

The observational challenge in measuring H_0 lies in decoupling the pristine cosmological expansion from local astrophysical noise. Real galaxies do not move solely with the unperturbed Hubble flow; they experience local gravitational forces induced by neighboring clusters and cosmic filaments, generating **peculiar velocities** $\mathbf{v}_{\rm pec}$ of order 300–600 $\mathrm{km \cdot s^{-1}}$. The total observed

line-of-sight velocity is $v_{\rm obs} = H_0 d + v_{\rm pec}$. If one observes a galaxy at $v_{\rm obs} \sim 1000\,\mathrm{km\cdot s^{-1}}$, the peculiar motion can introduce a devastating systematic error of 30–50%. To determine H_0 reliably, astronomical observations must push beyond the local gravitational neighborhood to distances where the cosmic recession velocity dwarfs peculiar motions ($v \gtrsim 10^4\,\mathrm{km\cdot s^{-1}}$, corresponding to $z \gtrsim 0.03$), where peculiar velocity contamination drops below 3%.

3.1.2 The extragalactic distance ladder

Measuring distances in astronomy is notoriously difficult because we observe only the two-dimensional sky and the incoming bolometric energy flux F. The flux from a source of intrinsic luminosity L falls off according to the inverse-square law:

$$F = \frac{L}{4\pi D_L^2}. \tag{3.1.5}$$

To determine the luminosity distance D_L, one must either measure an object of known intrinsic luminosity (a **standard candle**) or an object of known physical transverse size (a **standard ruler**).

Because direct geometric trigonometric parallaxes were historically restricted to stars in the solar neighborhood (within $\lesssim 100\,\mathrm{pc}$ prior to space-based astrometry missions like Hipparcos and Gaia), astronomers built the **cosmic distance ladder**—a bootstrap chain where each overlapping rung calibrates the next:

1. **Geometric Primary Calibrators:** Direct physical methods that require no astrophysical modeling. These include stellar trigonometric parallax measured with milliarcsecond to microarcsecond precision by the Gaia satellite, detached eclipsing binaries in the Large Magellanic Cloud (LMC), and geometric orbital water megamasers orbiting supermassive black holes in active galactic nuclei (such as NGC 4258).

2. **Pulsating Standard Candles:** Classical **Cepheid variable stars**, which undergo radial pulsations driven by the opacity-driven kappa mechanism in their partially ionized helium layers. Henrietta Leavitt discovered that the pulsation period P correlates tightly with absolute visual magnitude M_V:

$$M_V = -\alpha \log_{10} P + \beta, \tag{3.1.6}$$

   providing a robust standard candle out to distances of $\sim 30\,\mathrm{Mpc}$. Complementing Cepheids are Tip of the Red Giant Branch (TRGB) stars, whose core helium flash produces a sharp, standardizable cutoff in the stellar luminosity function.

3. **Secondary Standard Candles in the Hubble Flow:** Bright extragalactic tracers calibrated in nearby galaxies hosting Cepheids or TRGB stars. These include Type Ia supernovae (SNe Ia), the Tully–Fisher relation for spiral galaxies (relating total optical luminosity to rotational velocity, $L \propto v_{\rm rot}^4$), and the Fundamental Plane for elliptical galaxies. These secondary tracers are luminous enough to reach deep into the unperturbed Hubble flow ($z \sim 0.01$–0.1).

The landmark Hubble Space Telescope (HST) Key Project on the Extragalactic Distance Scale anchored the modern measurement of H_0 by discovering Cepheid variables in 18 galaxies out to 25 Mpc, tying the primary geometric rungs directly to Type Ia supernovae and other secondary methods (Figure 3.2). Table 3.1 summarizes the individual results, which converged to a combined determination $H_0 = 72 \pm 8\,\mathrm{km\cdot s^{-1}\cdot Mpc^{-1}}$.

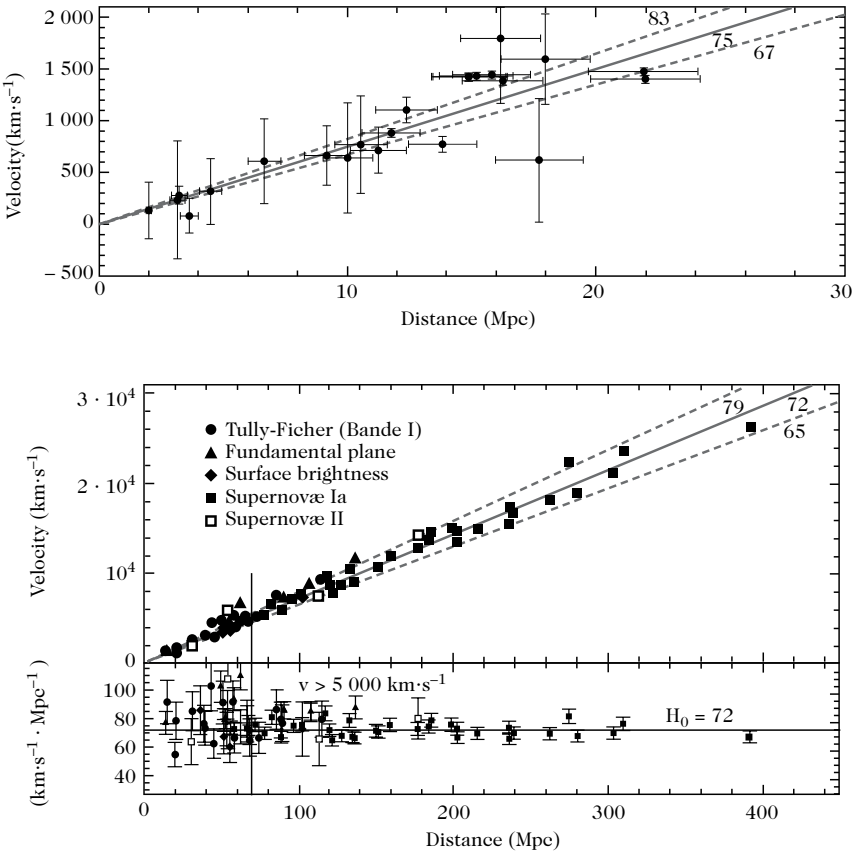


Figure 3.2: (Top) Hubble diagram for Cepheids obtained by the HST Key Project out to $\sim 20\,\mathrm{Mpc}$. (Bottom) Combined Hubble diagram extending out to $\sim 300\,\mathrm{Mpc}$ utilizing secondary standard candles (Type Ia supernovae, Tully–Fisher, Fundamental Plane, surface brightness fluctuations, and Type II supernovae).

Table 3.1: Measurements of the Hubble Constant H_0 from the Cosmic Distance Ladder (HST Key Project)

Astronomical Method / Tracer	**Measured** H_0 $(\mathrm{km \cdot s^{-1} \cdot Mpc^{-1}})$
Type Ia Supernovae (SN Ia)	$71 \pm 2_{\mathrm{stat}} \pm 6_{\mathrm{sys}}$
Tully–Fisher Relation	$71 \pm 3_{\mathrm{stat}} \pm 7_{\mathrm{sys}}$
Surface Brightness Fluctuations (SBF)	$70 \pm 5_{\mathrm{stat}} \pm 6_{\mathrm{sys}}$
Type II Supernovae (SN II)	$72 \pm 9_{\mathrm{stat}} \pm 7_{\mathrm{sys}}$
Fundamental Plane of Ellipticals	$82 \pm 2_{\mathrm{stat}} \pm 6_{\mathrm{sys}}$
HST Key Project Combined	$\mathbf{72 \pm 8}$

3.1.3 Physical, non-ladder methods for measuring H_0

The fundamental weakness of any distance ladder is the cumulative propagation of systematic errors: an error in the primary geometric anchor propagates into the Cepheid zero-point, which in turn offsets the calibration of Type Ia supernovae. To guard against systematic bias, cosmologists developed **physical, non-ladder methods** that measure cosmological distances in a single step using first-principles physics:

1. **The Expanding Photosphere Method (EPM) of Type II Supernovae:** When a

core-collapse massive star explodes, its ejecta form an optically thick, expanding spherical photosphere. Measuring the Doppler expansion velocity $v_{\rm phot}$ of spectral lines determines the physical radius $R(t) = v_{\rm phot}(t - t_{\rm exp})$. Photometric measurements of the blackbody angular diameter $\theta = 2R(t)/D_L$ then directly yield the luminosity distance $D_L = 2v_{\rm phot}(t - t_{\rm exp})/\theta$.

2. **Gravitational Lensing Time Delays:** When light from a time-variable background quasar passes by a foreground galaxy, general relativity predicts that light rays bend, creating multiple images. Because the light paths travel along different spacetime geodesics and through different gravitational potentials, photons arrive with differential time delays Δt_{AB}:

$$\Delta t_{AB} = \frac{1+z_L}{c}\frac{D_L D_S}{D_{LS}}\left[\frac{1}{2}\left(\boldsymbol{\theta}_A - \boldsymbol{\beta}\right)^2 - \psi(\boldsymbol{\theta}_A) - \frac{1}{2}\left(\boldsymbol{\theta}_B - \boldsymbol{\beta}\right)^2 + \psi(\boldsymbol{\theta}_B)\right] \propto H_0^{-1}, \quad (3.1.7)$$

where D_L, D_S, D_{LS} are the angular diameter distances to the lens, to the source, and between lens and source, and ψ is the two-dimensional lensing potential. Measuring the time lag between light curves of image pairs measures the absolute physical size of the system, determining H_0 independently of any distance ladder.

3. **Sunyaev–Zel'dovich (SZ) Effect and X-Ray Emission in Galaxy Clusters:** The hot ionized intracluster gas ($T_e \sim 10^7$–$10^8\,\mathrm{K}$) in rich galaxy clusters produces two distinct observational signatures:

   - Thermal inverse Compton scattering of CMB photons off cluster electrons causes a spectral shift (the thermal SZ effect), producing a temperature distortion $\Delta T_{\rm SZ} \propto \int n_e T_e \, \mathrm{d}\ell$ that is independent of cluster redshift.
   - Thermal bremsstrahlung generates X-ray surface brightness $S_X \propto D_A^{-1} \int n_e^2 T_e^{1/2} \, \mathrm{d}\ell$.

   Combining $\Delta T_{\rm SZ}^2/S_X$ isolates the physical depth of the cluster $\ell_{\rm cluster}$. Assuming spherical symmetry, $\ell_{\rm cluster} = D_A\theta$, which directly determines the angular diameter distance $D_A \propto H_0^{-1}$.

Table 3.2: Independent Physical Determinations of the Hubble Constant H_0

Physical Method	**Value of H_0 ($\mathrm{km \cdot s^{-1} \cdot Mpc^{-1}}$)**	**Reference**
SN II Expanding Photosphere	73 ± 15	Schmidt et al. (1994)
Gravitational Lensing Delay	$72 \pm 7_{\rm stat} \pm 15_{\rm sys}$	Tonry & Franx (1998)
Gravitational Lensing (B1608)	60 ± 20	Fassnacht et al. (2000)
Thermal Sunyaev–Zel'dovich	60 ± 4	Reese et al. (2002)
WMAP 1-Year Primary CMB	72 ± 5	Spergel et al. (2003)

Remarkably, as summarized in Table 3.2, these completely independent physical methods converged within their statistical uncertainties to $H_0 \approx 70\,\mathrm{km \cdot s^{-1} \cdot Mpc^{-1}}$, confirming the overall reliability of the distance ladder and setting the physical expansion timescale of the contemporary Universe.

3.2 Type Ia supernovae and the discovery of cosmic acceleration

3.2.1 Standardizing Type Ia supernovae

While the Hubble expansion at low redshifts probes the current rate of cosmic expansion H_0, probing the dynamic contents of the Universe requires measuring whether the expansion is speeding up or slowing down. In general relativity, attractive gravitational forces from ordinary matter and radiation are expected to decelerate the expansion over time. Detecting this cosmic deceleration requires observing standard candles at cosmological redshifts ($z \gtrsim 0.3$), where the lookback time corresponds to a substantial fraction of the age of the Universe.

Type Ia supernovae (SNe Ia) are the premier standard candles for high-redshift cosmology. The standard astrophysical model posits that a Type Ia supernova is the thermonuclear disruption of a carbon-oxygen white dwarf in a binary system. As the white dwarf accretes mass from a companion or merges with another white dwarf, its central density and temperature rise until runaway thermonuclear carbon and oxygen burning ignites near the Chandrasekhar mass limit:

$$M_{\rm Ch} \approx 1.44\, M_\odot. \tag{3.2.1}$$

Because the exploding progenitor mass is governed by fundamental quantum electron degeneracy pressure, the total radioactive nickel (^{56}Ni) synthesized and the resulting peak absolute magnitude are remarkably uniform across the cosmos ($M_B \approx -19.3$ mag, luminous enough to outshine an entire host galaxy).

Nonetheless, raw Type Ia supernovae exhibit an intrinsic peak luminosity scatter of roughly 0.3–0.5 mag—too large to distinguish between subtle cosmological models. Mark Phillips (1993) discovered that this intrinsic dispersion can be dramatically reduced: supernovae with intrinsically brighter peak luminosities have broader, slower-decaying light curves, whereas fainter supernovae fade more rapidly. Parameterizing the decline in B-band magnitude during the first 15 days following maximum light by $\Delta m_{15}(B)$, the **Phillips relation** standardizes the absolute magnitude to an astonishing precision of ~ 0.12–0.15 mag, translating to distance uncertainties of only ~ 6–7%.

3.2.2 The 1998 discovery of cosmic acceleration

In 1998, two competing teams of astronomers—the **Supernova Cosmology Project (SCP)**, led by Saul Perlmutter, and the **High-Z Supernova Search Team (HZT)**, led by Brian Schmidt and Adam Riess—sought to measure the cosmic deceleration parameter q_0 by discovering dozens of high-redshift Type Ia supernovae ($z \approx 0.2$–0.8).

At second order in redshift, the luminosity distance in an FLRW universe is given by:

$$D_L(z) = \frac{cz}{H_0}\left[1 + \frac{1}{2}\,(1 - q_0)\, z + \mathcal{O}(z^2)\right], \tag{3.2.2}$$

where the dimensionless **deceleration parameter** is defined as:

$$q_0 \equiv -\frac{\ddot{a}(t_0)a(t_0)}{\dot{a}^2(t_0)} = -\frac{\ddot{a}_0}{a_0 H_0^2}. \tag{3.2.3}$$

In the Friedmann equations with pressureless matter Ω_{m0} and vacuum energy $\Omega_{\Lambda 0}$:

$$q_0 = \frac{1}{2}\Omega_{m0} - \Omega_{\Lambda 0}. \tag{3.2.4}$$

If gravity decelerates the expansion, $\ddot{a} < 0 \implies q_0 > 0$. In that case, distant supernovae should appear brighter than in an empty universe ($q_0 = 0$), because deceleration implies that the universe expanded faster in the past, meaning the light took less time to reach us.

The observational findings shocked the scientific community: high-redshift supernovae were systematically ~ 0.25 mag *fainter* (more distant) than expected in any decelerating matter-dominated universe (Figure 3.3). The data demanded that the deceleration parameter be

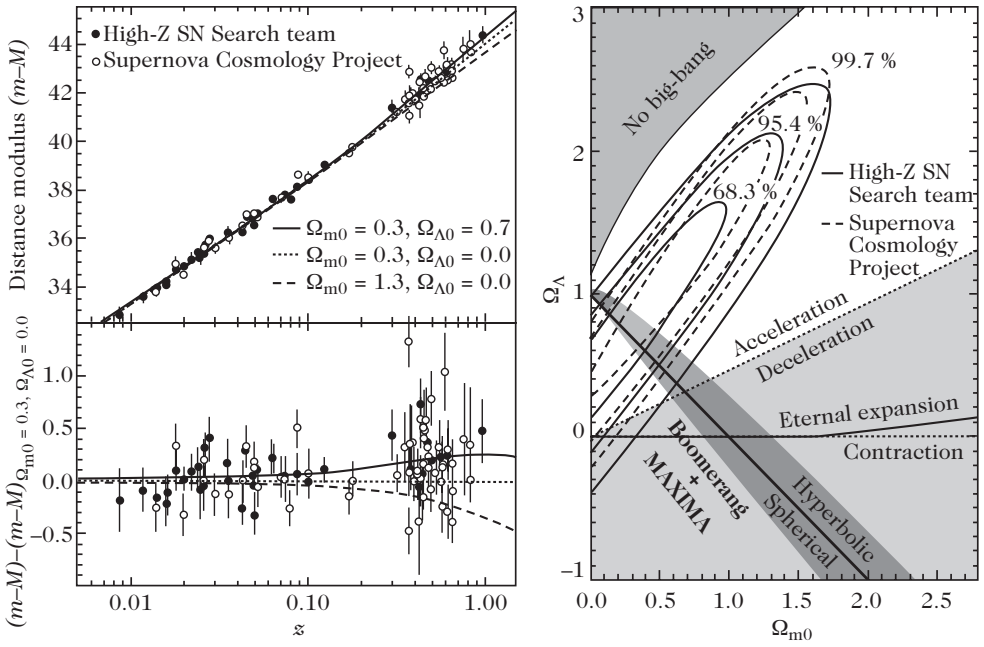


Figure 3.3: (Left) Hubble diagram of high-z Type Ia supernovae from SCP and HZT compared with FLRW models. The lower panel shows distance modulus residuals relative to an empty universe ($\Omega_{m0} = 0.3, \Omega_{\Lambda 0} = 0$). (Right) Joint statistical likelihood contours in the $(\Omega_{m0}, \Omega_{\Lambda 0})$ plane, demonstrating $q_0 < 0$ and isolating $\Omega_{\Lambda 0} \approx 0.7, \Omega_{m0} \approx 0.3$ when combined with CMB balloon data.

negative:

$$q_0 = \frac{1}{2}\Omega_{m0} - \Omega_{\Lambda 0} < 0 \quad (\text{at } 2.8\sigma \text{ confidence}), \tag{3.2.5}$$

demonstrating that cosmic expansion is currently **accelerating** ($\ddot{a} > 0$). The supernovae data established the linear degeneracy relation:

$$8\Omega_{m0} - 6\Omega_{\Lambda 0} \simeq -2 \pm 1. \tag{3.2.6}$$

When combined with early CMB measurements from BOOMERanG and MAXIMA indicating spatial flatness ($\Omega_K \approx 0 \implies \Omega_{m0} + \Omega_{\Lambda 0} \approx 1$), the parameters intersected at the concordance values:

$$\Omega_{\Lambda 0} \approx 0.70, \qquad \Omega_{m0} \approx 0.30. \tag{3.2.7}$$

Could this unexpected dimming be caused by mundane astrophysics, such as hypothetical gray intergalactic dust or progenitor metallicity evolution? The definitive resolution came with the discovery of the highest-redshift supernova SN 1997ff at $z \approx 1.7$ by the HST GOODS survey (Figure 3.4). In a universe with $\Omega_m \approx 0.3$ and $\Omega_\Lambda \approx 0.7$, matter density scales as $(1+z)^3$ while vacuum energy remains constant. At redshifts $z > 0.5$, matter density dominated over dark energy, meaning the Universe must have *decelerated* in the past before transitioning to acceleration at $z \approx 0.5$. In contrast to hypothetical dust (which would dim supernovae monotonically with redshift), the data at $z > 1$ showed the predicted cosmic turnaround: supernovae at $z > 1$ appeared brighter than the dust extrapolation, confirming true general relativistic cosmic acceleration.

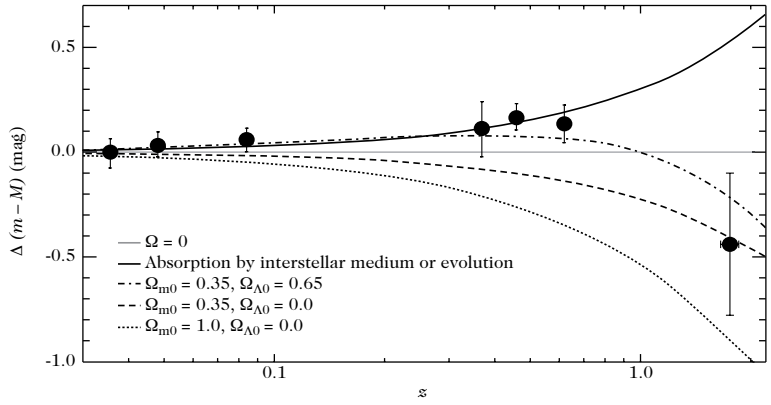


Figure 3.4: Residual Hubble diagram $\Delta(m - M)$ relative to an empty universe $\Omega = 0$. The detection of very high-redshift SNe Ia at $z \approx 1.7$ (HST GOODS survey) rules out monotonic astrophysical dimming (grey intergalactic dust or progenitor metallicity evolution) and confirms the transition from past deceleration ($z > 0.5$) to present acceleration ($z < 0.5$).

3.2.3 Dark energy equation of state

Accelerated expansion requires a dominant energy component with negative pressure. In general relativity, the second Friedmann acceleration equation is:

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + 3P\right). \tag{3.2.8}$$

Accelerated expansion ($\ddot{a} > 0$) occurs if and only if the cosmic fluid violates the strong energy condition, $\rho + 3P < 0$. Parameterizing the fluid by its equation of state parameter $w \equiv P/\rho$, acceleration demands:

$$w < -\frac{1}{3}. \tag{3.2.9}$$

For Einstein's cosmological constant Λ, $P_\Lambda = -\rho_\Lambda$, so $w = -1$ exactly.

More generally, dynamical scalar field models (known as quintessence) allow $w(t)$ to vary dynamically over cosmic time. Assuming first a constant effective equation of state w_X in a flat universe, joint analyses of SNe Ia, the CMB, and galaxy clustering strongly constrain $w_X < -0.6$ and center tightly on $w_X = -1$ (Figure 3.5).

To test for dynamical evolution, cosmologists parameterize the equation of state using the two-parameter Chevallier–Polarski–Linder (CPL) ansatz:

$$w(a) = w_0 + w_a\,(1 - a) = w_0 + w_a\frac{z}{1+z}, \tag{3.2.10}$$

where w_0 represents the present equation of state and $w_a = -\,\mathrm{d}w\,/\,\mathrm{d}a\,|_{a=1}$ characterizes its time derivative. Joint cosmological probes constrain:

$$w_0 = -0.98 \pm 0.19, \qquad w_a = 0.05^{+0.83}_{-0.65}, \tag{3.2.11}$$

consistent with a time-independent cosmological constant ($w_0 = -1, w_a = 0$) at all observed redshifts.

3.3 Cosmic age determinations and the age crisis resolution

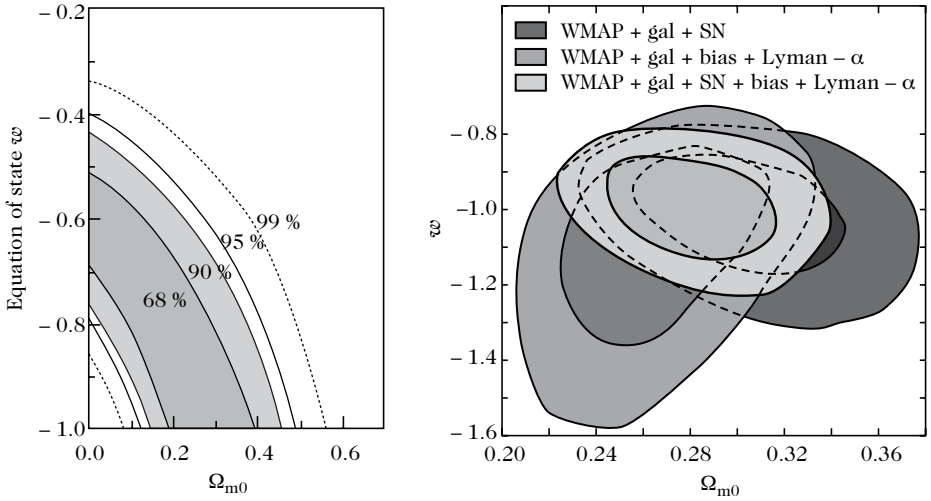


Figure 3.5: (Left) Joint constraints on constant dark energy equation of state w and Ω_{m0} in a flat universe. (Right) 68% and 95% confidence regions combining CMB (WMAP), SDSS galaxy clustering, SNe Ia, and Lyman-α forest data.

3.3.1 Astrophysical age indicators in the Milky Way

A fundamental consistency test of any cosmological model is that the age of the Universe t_u must strictly exceed the ages of its oldest constituent objects:

$$t_u > t_{\rm star}. \tag{3.3.1}$$

Long before precision CMB cosmology, astronomers developed three independent stellar chronometers to date the oldest structures in our Galaxy:

1. **Nuclear Cosmochronology:** Dating ultra-metal-poor halo stars (such as CS 22892-052 and CS 31082-0018) via radioactive decay of heavy isotopes produced in rapid neutron-capture (r-process) nucleosynthesis during stellar mergers or supernovae. Comparing observed isotopic ratios with theoretical nucleosynthetic production ratios provides an absolute radioactive clock:
   - Thorium-232 (^{232}Th, half-life $\tau_{1/2} = 14.05\,$Gyr): yields stellar age $t_* = 15.2 \pm 3.7\,$Gyr.
   - Uranium-238 (^{238}U, half-life $\tau_{1/2} = 4.47\,$Gyr): yields stellar age $t_* = 12.5 \pm 3.0\,$Gyr.
2. **White Dwarf Cooling Sequences:** White dwarfs are electron-degenerate stellar remnants that have ceased thermonuclear burning and cool strictly through radiative and neutrino emission. Modeling the fading luminosity function of degenerate white dwarfs in the solar neighborhood yields a lower bound on the Galactic disk age:
$$t_{\rm disk} = 9.8^{+1.1}_{-0.8}\,{\rm Gyr}. \tag{3.3.2}$$
3. **Globular Cluster Main-Sequence Turnoff (MSTO):** Globular clusters are dense, gravitationally bound systems containing up to 10^6 coeval, chemically homogeneous stars formed in a single primordial burst. Because more massive stars burn hydrogen faster, stellar luminosity on the main sequence scales as $L \propto M^{3.5}$, giving main-sequence lifetime:
$$\tau_{\rm MS} \propto \frac{M}{L} \propto M^{-2.5} \propto L_{\rm MSTO}^{-0.71}. \tag{3.3.3}$$

The position of the main-sequence turnoff in the Hertzsprung–Russell color-magnitude diagram serves as a precision chronometer. Advanced stellar evolutionary models calibrated with Hipparcos and Gaia parallaxes yield globular cluster ages:

$$t_{\rm GC} \approx 11.5\text{–}14.0\,\text{Gyr}. \tag{3.3.4}$$

Table 3.3: Summary of Astrophysical Age Constraints on the Universe

Chronometer / Method	**Stellar / Cluster Age (Gyr)**	**Reference**
Nucleochronology (^{232}Th, CS 22892)	15.2 ± 3.7	Sneden et al. (1996)
Nucleochronology (^{238}U, CS 31082)	12.5 ± 3.0	Cayrel et al. (2001)
Globular Clusters (MSTO Isochrones)	11.5 ± 1.3	Chaboyer et al. (1998)
Hipparcos-calibrated MSTO	11.8 ± 1.2	Gratton et al. (1997)
Eclipsing Binaries in Clusters	14.0 ± 1.2	Pont et al. (1998)
WMAP + Large-Scale Structure	$\mathbf{13.7 \pm 0.2}$	Spergel et al. (2003)

Accounting for an incubation time of $\Delta t_{\rm form} \approx 0.5\text{–}1.5\,\text{Gyr}$ for the primordial gas to collapse and form the first stars and clusters ($z \approx 6\text{–}20$), the empirical age of the Universe is firmly bounded by:

$$9.6\,\text{Gyr} \leq t_u \leq 15.4\,\text{Gyr}. \tag{3.3.5}$$

3.3.2 Resolution of the cosmological age crisis

During the 1980s and 1990s, the empirical lower bounds on stellar ages collided with cosmological theory, creating the historic **cosmological age crisis**. Theoretical physicists favored the simple, spatially flat Einstein–de Sitter (EdS) universe with $\Omega_m = 1$ and $\Omega_\Lambda = 0$. In this model, integrating the Friedmann equation gives the cosmic age:

$$t_0^{\rm EdS} = \int_0^1 \frac{{\rm d}a}{aH(a)} = \frac{1}{H_0}\int_0^1 a^{1/2}\,{\rm d}a = \frac{2}{3H_0} = 6.52\,h^{-1}\,\text{Gyr}. \tag{3.3.6}$$

For the empirically measured Hubble parameter $h \approx 0.70$, this prediction yields:

$$t_0^{\rm EdS} \approx 9.3\,\text{Gyr}. \tag{3.3.7}$$

This theoretical age was in sharp conflict with globular cluster stellar ages of 12–14 Gyr—a flat matter-dominated universe was younger than the oldest stars within it!

The resolution of the crisis arrived with the discovery of dark energy. In a flat ΛCDM universe with non-relativistic matter Ω_{m0} and vacuum energy $\Omega_{\Lambda 0} = 1 - \Omega_{m0}$, the Friedmann equation is:

$$H(a) = H_0\sqrt{\Omega_{m0}a^{-3} + \Omega_{\Lambda 0}}. \tag{3.3.8}$$

The expansion age of the Universe is computed by the integral:

$$t_0^{\Lambda\rm CDM} = \int_0^1 \frac{{\rm d}a}{aH_0\sqrt{\Omega_{m0}a^{-3} + \Omega_{\Lambda 0}}} = \frac{1}{H_0\sqrt{\Omega_{\Lambda 0}}}\int_0^1 \frac{a^{1/2}\,{\rm d}a}{\sqrt{a^3 + \frac{\Omega_{m0}}{\Omega_{\Lambda 0}}}}. \tag{3.3.9}$$

Substituting $u = a^{3/2} \implies \mathrm{d}u = \frac{3}{2}a^{1/2}\,\mathrm{d}a$, the integral converts to standard elementary form:

$$t_0^{\Lambda\mathrm{CDM}} = \frac{2}{3H_0\sqrt{\Omega_{\Lambda 0}}}\int_0^1 \frac{\mathrm{d}u}{\sqrt{u^2 + \frac{\Omega_{m0}}{\Omega_{\Lambda 0}}}} = \frac{2}{3H_0\sqrt{\Omega_{\Lambda 0}}}\,\mathrm{arcsinh}\left(\sqrt{\frac{\Omega_{\Lambda 0}}{\Omega_{m0}}}\right). \tag{3.3.10}$$

Evaluating Eq. (3.3.10) for concordance parameters $\Omega_{m0} \approx 0.30$ and $\Omega_{\Lambda 0} \approx 0.70$ yields:

$$t_0^{\Lambda\mathrm{CDM}} \approx \frac{0.96}{H_0} \approx 13.8\,\mathrm{Gyr}, \tag{3.3.11}$$

comfortably accommodating the oldest globular clusters and nuclear chronometer ages, and resolving the cosmological age crisis completely.

3.4 Thermodynamics of the expanding cosmic plasma

3.4.1 The radiation-dominated early Universe

In the contemporary Universe, non-relativistic matter (cold dark matter and baryons) dominates over radiation by several orders of magnitude. However, because the physical energy density of non-relativistic matter dilutes purely with cosmic volume as $\rho_m(a) = \rho_{m0}a^{-3}$, whereas relativistic radiation redshifts in both number density (a^{-3}) and individual photon momentum ($p \propto a^{-1}$) as $\rho_r(a) = \rho_{r0}a^{-4}$, the energy density of radiation grows much more rapidly as we look back in time toward $a \to 0$. In the early cosmic epochs, the total energy density was completely dominated by relativistic radiation.

To determine when radiation relinquished its dominance to matter, we must establish the precise radiation energy budget of the Universe today. The radiation bath consists of two distinct components: the cosmic microwave background (CMB) photons and the relic cosmic neutrino background (CNB). The photon energy density is fixed directly by the measured CMB temperature $T_{\gamma 0} = 2.7255\,\mathrm{K}$:

$$n_{\gamma 0} = \frac{2\zeta(3)}{\pi^2}T_{\gamma 0}^3 = 410.73\,\mathrm{photons}\cdot\mathrm{cm}^{-3}, \tag{3.4.1}$$

$$\rho_{\gamma 0} = \frac{\pi^2}{15}T_{\gamma 0}^4 = 4.645\times 10^{-31}\,\mathrm{kg}\cdot\mathrm{m}^{-3} = 0.2606\,\mathrm{eV}\cdot\mathrm{cm}^{-3}. \tag{3.4.2}$$

Dividing by the critical density $\rho_{\mathrm{crit},0} = 3H_0^2/(8\pi G) = 1.878\times 10^{-26}\,h^2\,\mathrm{kg}\cdot\mathrm{m}^{-3}$, the photon density parameter is:

$$\Omega_{\gamma 0} \equiv \frac{\rho_{\gamma 0}}{\rho_{\mathrm{crit},0}} = 2.473\times 10^{-5}\,h^{-2}. \tag{3.4.3}$$

For neutrinos, standard electroweak decoupling and e^+e^- annihilation predict a relic temperature $T_{\nu 0} = (4/11)^{1/3}T_{\gamma 0} \approx 1.945\,\mathrm{K}$. In a reference universe where neutrinos remain ultra-relativistic, the neutrino energy density is parameterized by the effective number of relativistic neutrino species $N_{\mathrm{eff}} \approx 3.044$:

$$\rho_{\nu 0}^{*} = 3\times 2\times\frac{7}{8}\left(\frac{\pi^2}{30}\right)T_{\nu 0}^4 = \frac{7}{8}N_{\mathrm{eff}}\left(\frac{4}{11}\right)^{4/3}\rho_{\gamma 0} \approx 0.692\,\rho_{\gamma 0}. \tag{3.4.4}$$

The neutrino radiation density parameter is therefore:

$$\Omega_{\nu 0}^{*} = 1.711\times 10^{-5}\,h^{-2}, \tag{3.4.5}$$

giving a total relativistic radiation density parameter today of:

$$\Omega_{r0}^{*} = \Omega_{\gamma 0} + \Omega_{\nu 0}^{*} = 4.184\times 10^{-5}\,h^{-2}. \tag{3.4.6}$$

Matter-radiation equality. Equating the matter and radiation energy densities, $\rho_m(a_{\rm eq}) = \rho_r(a_{\rm eq})$, defines the epoch of **matter-radiation equality**. In terms of the reference radiation density, the scale factor at equality is:

$$a_{\rm eq} = \frac{\Omega^*_{r0}}{\Omega^*_{m0}} = \frac{4.1837 \times 10^{-5}}{\Omega^*_{m0} h^2}, \tag{3.4.7}$$

which corresponds to the equality redshift:

$$1 + z_{\rm eq} = \frac{1}{a_{\rm eq}} = 23902\, \Omega^*_{m0} h^2 \simeq 3612\, \Theta_{2.7}^{-4} \left(\frac{\Omega_{m0} h^2}{0.15} \right), \tag{3.4.8}$$

where $\Theta_{2.7} \equiv T_{\rm CMB}/2.7255\,{\rm K}$. For concordance parameters $\Omega_{m0} h^2 \approx 0.143$, $z_{\rm eq} \approx 3400$, and the plasma temperature at equality is:

$$T_{\rm eq} = T_{\gamma 0}\,(1 + z_{\rm eq}) \approx 0.825\,{\rm eV} \approx 9570\,{\rm K}. \tag{3.4.9}$$

Expansion timescale in the radiation era. Throughout the radiation-dominated era ($a \ll a_{\rm eq}$), the spatial curvature and dark energy terms in the Friedmann equation are completely negligible. The expansion is governed solely by the relativistic plasma:

$$H^2 = \left(\frac{\dot a}{a} \right)^2 = \frac{8\pi G}{3} \rho_r(T) = \frac{8\pi G}{3} \frac{\pi^2}{30} g_*(T) T^4, \tag{3.4.10}$$

where $g_*(T)$ is the effective number of relativistic degrees of freedom. Integrating $\dot a \propto a^{-1}$ yields the characteristic radiation-era power-law expansion:

$$a(t) \propto t^{1/2} \implies H(t) = \frac{\dot a}{a} = \frac{1}{2t}. \tag{3.4.11}$$

Substituting $H(t) = 1/(2t)$ into Eq. (3.4.10) establishes the fundamental relationship between the physical age of the Universe t and the plasma temperature T:

$$t = \frac{1}{2H} = \left(\frac{90}{32\pi^3 G g_*} \right)^{1/2} \frac{1}{T^2} = 0.301\, g_*^{-1/2} \frac{M_{\rm Pl}}{T^2} \approx \frac{2.42}{\sqrt{g_*}} \left(\frac{1\,{\rm MeV}}{T} \right)^2 {\rm s}, \tag{3.4.12}$$

where $M_{\rm Pl} \equiv G^{-1/2} \approx 1.22 \times 10^{19}\,{\rm GeV}$ is the Planck mass. Over epochs where $g_*(T)$ changes slowly, $T \propto a^{-1} \propto t^{-1/2}$.

Cosmic time at matter-radiation equality. To determine the exact cosmic age $t_{\rm eq}$ at matter-radiation equality, we integrate the Friedmann equation accounting for both radiation and matter:

$$\left(\frac{\dot a}{a} \right)^2 = H_0^2 \left(\Omega^*_{r0} a^{-4} + \Omega^*_{m0} a^{-3} \right) = H_0^2 \Omega^*_{m0} a^{-3} \left(1 + \frac{a_{\rm eq}}{a} \right). \tag{3.4.13}$$

Separating variables gives:

$$\frac{a\, {\rm d}a}{\sqrt{a_{\rm eq} + a}} = H_0 \sqrt{\Omega^*_{m0}}\, {\rm d}t\,. \tag{3.4.14}$$

Integrating from the Big Bang ($a = 0$) to matter-radiation equality ($a = a_{\rm eq}$):

$$t_{\rm eq} = \frac{H_0^{-1}}{\sqrt{\Omega^*_{m0}}} \int_0^{a_{\rm eq}} \frac{a\, {\rm d}a}{\sqrt{a_{\rm eq} + a}}. \tag{3.4.15}$$

Substituting $u = a_{\rm eq} + a \implies a = u - a_{\rm eq}$, the integral evaluates analytically:

$$\int_0^{a_{\rm eq}} \frac{(u - a_{\rm eq})\, {\rm d}u}{\sqrt{u}} = \left[\frac{2}{3} u^{3/2} - 2 a_{\rm eq} u^{1/2} \right]_{a_{\rm eq}}^{2a_{\rm eq}}$$

$$= a_{\rm eq}^{3/2}\left[\frac{2}{3}\left(2\sqrt{2}-1\right)-2\left(\sqrt{2}-1\right)\right] = \frac{2}{3}\left(2-\sqrt{2}\right)a_{\rm eq}^{3/2}. \tag{3.4.16}$$

Thus, the exact time of matter-radiation equality is:

$$t_{\rm eq} = \frac{2}{3}\frac{H_0^{-1}}{\sqrt{\Omega_{m0}^*}}\left(2-\sqrt{2}\right)a_{\rm eq}^{3/2} \approx 0.3905\,\frac{(\Omega_{r0}^*)^{3/2}}{(\Omega_{m0}^*)^2}H_0^{-1} \approx \left(\Omega_{m0}^* h^2\right)^{-2} \times 1024\,{\rm yr}. \tag{3.4.17}$$

For concordance parameters ($\Omega_{m0}^* \approx 0.30, h \approx 0.70$), this yields $t_{\rm eq} \approx 47{,}400\,{\rm yr}$ (or $\approx 51{,}000\,{\rm yr}$ for $\Omega_m h^2 = 0.143$), marking the watershed transition from radiation domination to the era of gravitational structure growth.

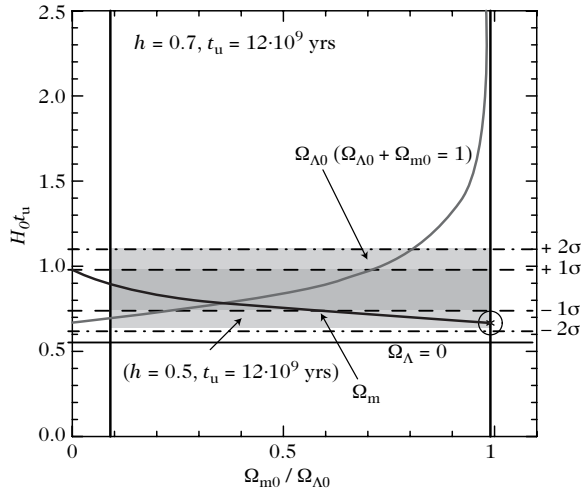


Figure 3.6: Comparison between the empirical constraint on $H_0 t_u$ (from stellar ages $t_u \approx 12\,{\rm Gyr}$ and $h \approx 0.7$, shaded bands) and the theoretical dynamical expansion age $H_0 t_0(\Omega_{m0}, \Omega_{\Lambda 0})$. The dark curve denotes flat universes ($\Omega_{m0} + \Omega_{\Lambda 0} = 1$) vs. $\Omega_{\Lambda 0}$; the light curve denotes $\Omega_{\Lambda 0} = 0$ vs. Ω_{m0}. The Einstein–de Sitter point ($H_0 t_0 = 2/3$) is excluded at $> 1.5\sigma$.

3.4.2 Equilibrium relativistic thermodynamics

Phase-space distribution functions

Under spatial homogeneity and isotropy, the phase-space distribution function $f_i(\mathbf{x}, \mathbf{p}, t)$ of a particle species i depends only on the particle energy $E = \sqrt{p^2 + m_i^2}$ and cosmic time t through the local temperature $T_i(t)$.

For a gas in local thermodynamic equilibrium (kinetic and chemical), $f_i(E, T)$ is given by the **Fermi–Dirac** $(+)$ or **Bose–Einstein** $(-)$ distribution:

$$f_i(E,T) = \frac{1}{\exp\left[\frac{E-\mu_i}{T_i(t)}\right] \pm 1}, \tag{3.4.18}$$

where g_i is the internal spin/polarization degeneracy factor, and μ_i is the chemical potential.

The covariant definition of distribution function take the form of

$$f_i(U,T) = \frac{1}{\exp\left[\frac{-p_\mu U^\mu - \mu_i}{T_i(t)}\right] \pm 1}, \tag{1}$$

Free-streaming invariance of relativistic distributions. A profound and elegant feature of cosmological expansion is how the phase-space distribution of decoupled particles evolves. Suppose a species drops out of thermal contact at time t_D when it is ultra-relativistic ($T_D \gg m_i$, so $E \approx p$). After decoupling, the particles free-stream collisionlessly along spacetime geodesics, and general relativity dictates that their physical momentum redshifts strictly as:

$$p(t) = p(t_D)\frac{a(t_D)}{a(t)}. \tag{3.4.19}$$

By Liouville's theorem, phase-space density is conserved along particle trajectories in the absence of collisions: $f(p,t) = f(pa/a_D, t_D)$. Substituting into the relativistic Fermi–Dirac or Bose–Einstein distribution:

$$f(p,t) = \frac{1}{\exp\left[\frac{p\,a(t)/a(t_D)}{T_D}\right] \pm 1} = \frac{1}{\exp\left[\frac{p}{T(t)}\right] \pm 1}, \qquad \text{where } T(t) \equiv T_D\frac{a(t_D)}{a(t)}. \tag{3.4.20}$$

Remarkably, *the functional form of the equilibrium distribution is preserved exactly*, even though the particles are completely decoupled and undergo no collisions whatsoever! The only effect of cosmic expansion is that the effective temperature redshifts as $T \propto a^{-1}$.

Crucially, this preservation holds only as long as the particles remain relativistic ($E \approx p$). If the decoupled particles have a non-zero rest mass m, they eventually become non-relativistic ($T(t) < m$). In this non-relativistic regime, their energy becomes $E \approx m + p^2/(2m)$, but their momenta continue to redshift as $p \propto a^{-1}$. Consequently, their phase-space distribution does *not* evolve into a non-relativistic Maxwell–Boltzmann distribution at temperature T; instead, it retains the frozen relativistic profile with an effective kinetic temperature cooling as $T_{\rm kin} \propto a^{-2}$. This distinction is essential for cosmic relics such as massive neutrinos and warm dark matter.

Thermal Decoupling and Freeze-Out Criteria

In the early Universe, cosmic expansion acts as a competitive dynamical process against particle interactions. A particle species i maintains local thermodynamic equilibrium (both thermal and chemical) with the background plasma if its interaction rate, $\Gamma_i \equiv n_i\langle\sigma v\rangle$, significantly exceeds the Hubble expansion rate $H(t) \equiv \dot{a}/a$:

$$\Gamma_i(T) \gg H(T) \iff \tau_{\rm coll} \ll \tau_H \quad \text{(Thermal Equilibrium)}, \tag{3.4.21}$$

where $\tau_{\rm coll} = \Gamma_i^{-1}$ is the mean free time between scattering events and $\tau_H = H^{-1}$ represents the characteristic timescale of space expansion.

Thermal Evolution and Cosmic Expansion. As the Universe undergoes adiabatic expansion, entropy conservation in a comoving volume ($S \propto g_{*S}T^3a^3 = \text{const.}$) implies that the temperature of the cosmic plasma scales inversely with the scale factor:

$$T \propto \frac{1}{a(t)}. \tag{3.4.22}$$

Consequently, physical volume elements expand as $V \propto a^3 \propto T^{-3}$. During the radiation-dominated era, the total energy density is dominated by relativistic species ($\rho \approx \rho_{\rm rad} = \frac{\pi^2}{30}g_*T^4$). The Friedmann equation then dictates the temperature dependence of the expansion rate:

$$H^2 = \frac{8\pi G}{3}\rho_{\rm rad} \implies H(T) = \left(\frac{8\pi^3 g_*}{90}\right)^{1/2}\frac{T^2}{M_{\rm Pl}} \propto T^2 \propto \frac{1}{a^2}, \tag{3.4.23}$$

where $M_{\rm Pl} \equiv G^{-1/2} \approx 1.22 \times 10^{19}$ GeV is the Planck mass.

Interaction Rate Scaling. The interaction rate $\Gamma_i = n_i \langle \sigma v \rangle$ depends strongly on whether the species is relativistic ($T \gg m_i$) or non-relativistic ($T \ll m_i$). For a relativistic species, the number density scales purely with the thermal background as $n_i \propto T^3 \propto a^{-3}$. Considering an interaction cross-section that scales thermally as $\sigma \sim G^2 E^n \sim G^2 T^n$ (where G is the coupling constant with dimensions of mass^{-2}):

$$\Gamma_i(T) \sim n_i \sigma v \propto T^3 \cdot T^n = T^{n+3}. \tag{3.4.24}$$

Comparing the interaction rate to the Hubble parameter yields the dimensionless ratio governing equilibrium:

$$\frac{\Gamma_i(T)}{H(T)} \propto \frac{T^{n+3}}{T^2/M_{\rm Pl}} \propto M_{\rm Pl}\, T^{n+1} \propto a^{-(n+1)}. \tag{3.4.25}$$

Freeze-Out Condition. For weak interactions mediated by massive gauge bosons below the electroweak scale (such as standard light neutrino decoupling), $n = 2$ and the cross-section scales as $\sigma \sim G_{\rm F}^2 T^2$, where $G_{\rm F}$ is the Fermi constant. Thus:

$$\frac{\Gamma}{H} \propto T^3 \propto \frac{1}{a^3}. \tag{3.4.26}$$

Because $\Gamma(T)$ falls faster than $H(T)$ as $a(t)$ grows, the interaction rate eventually drops below the expansion rate. The system undergoes **thermal freeze-out / decoupling** at a characteristic decoupling temperature T_D (or scale factor a_D):

$$\Gamma_i(T_D) \approx H(T_D) \implies T_D = \text{Decoupling Temperature}. \tag{3.4.27}$$

For $T < T_D$ (or $a > a_D$), $\Gamma_i < H$, meaning the time required for particles to interact becomes longer than the age of the Universe. The species drops out of chemical and thermal equilibrium, maintaining a frozen comoving number density ($Y_i \equiv n_i/s = \text{const.}$) and free-streaming along null or timelike geodesics.

Thermodynamic macroscopic integrals

Integrating over isotropic momentum space $\mathrm{d}^3p = 4\pi p^2\, \mathrm{d}p = 4\pi\sqrt{E^2 - m_i^2}E\, \mathrm{d}E$, the particle number density n_i, energy density ρ_i, and isotropic pressure P_i are:

$$n_i(T) = \frac{g_i}{(2\pi)^3} \int f_i(\mathbf{p})\, \mathrm{d}^3p = \frac{g_i}{2\pi^2} \int_{m_i}^{\infty} \frac{\sqrt{E^2 - m_i^2}E\, \mathrm{d}E}{\exp\left(\frac{E-\mu_i}{T}\right) \pm 1}, \tag{3.4.28}$$

$$\rho_i(T) = \frac{g_i}{(2\pi)^3} \int E(p) f_i(\mathbf{p})\, \mathrm{d}^3p = \frac{g_i}{2\pi^2} \int_{m_i}^{\infty} \frac{\sqrt{E^2 - m_i^2}E^2\, \mathrm{d}E}{\exp\left(\frac{E-\mu_i}{T}\right) \pm 1}, \tag{3.4.29}$$

$$P_i(T) = \frac{g_i}{(2\pi)^3} \int \frac{p^2}{3E} f_i(\mathbf{p})\, \mathrm{d}^3p = \frac{g_i}{6\pi^2} \int_{m_i}^{\infty} \frac{\left(E^2 - m_i^2\right)^{3/2}\, \mathrm{d}E}{\exp\left(\frac{E-\mu_i}{T}\right) \pm 1}. \tag{3.4.30}$$

Energy Density

We will now, explicitly find the energy density, pressure and number density in the early universe.

$$\rho^{\rm eq} = \sum_{\text{all species}} \frac{g_i}{(2\pi\hbar)^3} \int_0^{\infty} \frac{\sqrt{p^2c^2 + m^2c^4}}{\exp\left(\sqrt{p^2c^2+m^2c^4} - \mu/k_BT\right) \pm 1} 4\pi p^2 dp$$

Table 3.4: Asymptotic Limits of Thermodynamic Quantities for Relativistic and Non-Relativistic Species

Regime	**Statistics**	**Number Density** n	**Energy Density** ρ	**Pressure** P
Relativistic $(T \gg m, \lvert\mu\rvert \ll T)$	**Bosons**	$g\frac{\zeta(3)}{\pi^2}T^3$	$g\frac{\pi^2}{30}T^4$	$\frac{1}{3}\rho$
	Fermions	$\frac{3}{4}g\frac{\zeta(3)}{\pi^2}T^3$	$\frac{7}{8}g\frac{\pi^2}{30}T^4$	$\frac{1}{3}\rho$
Degenerate Fermi $(\mu \gg T \gg m)$	**Fermions**	$g\frac{\mu^3}{6\pi^2}$	$g\frac{\mu^4}{8\pi^2}$	$\frac{1}{3}\rho$
Non-Relativistic $(T \ll m)$	**Both**	$g\left(\frac{mT}{2\pi}\right)^{3/2}\mathrm{e}^{\frac{\mu-m}{T}}$	$mn + \frac{3}{2}nT \approx mn$	$nT \ll \rho$

$$= \sum_{\text{boson}} \frac{g_i}{2\pi^2\hbar^3}\int_0^\infty dp \frac{p^2\sqrt{p^2c^2+m^2c^4}}{\exp\left(\sqrt{p^2c^2+m^2c^4}-\mu/k_BT\right)-1} + \sum_{\text{fermion}} \frac{g_i}{2\pi^2\hbar^3}\int_0^\infty dp \frac{p^2\sqrt{p^2c^2+m^2c^4}}{\exp\left(\sqrt{p^2c^2+m^2c^4}-\mu/k_BT\right)+1}$$

defining $x \equiv {}^{mc^2}/_{k_BT}$ and $\xi \equiv {}^{pc}/_{k_BT}$, we get

$$= \sum_{\text{boson}} \frac{g_i}{2\pi^2\hbar^3}\int_0^\infty \frac{k_BT}{c} d\xi \frac{\frac{k_B^2T^2}{c^2}\xi^2\sqrt{k_B^2T^2\xi^2+k_B^2T^2x^2}}{\exp\left(\sqrt{\xi^2+x^2}-\mu/k_BT\right)-1} + \sum_{\text{fermion}} \frac{g_i}{2\pi^2\hbar^3}\int_0^\infty \frac{k_BT}{c} d\xi \frac{\frac{k_B^2T^2}{c^2}\xi^2\sqrt{k_B^2T^2\xi^2+k_B^2T^2}}{\exp\left(\sqrt{\xi^2+x^2}-\mu/k_BT\right)+1}$$

$$= \frac{k_B^4T^4}{2\pi^2\hbar^3c^3}\left[\sum_{\text{boson}} g_i\int_0^\infty d\xi \frac{\xi^2\sqrt{\xi^2+x^2}}{\exp\left(\sqrt{\xi^2+x^2}-\mu/k_BT\right)-1} + \sum_{\text{fermion}} g_i\int_0^\infty d\xi \frac{\xi^2\sqrt{\xi^2+x^2}}{\exp\left(\sqrt{\xi^2+x^2}-\mu/k_BT\right)+1}\right]$$

Ultra-Relativistic Limit In the early universe, $\mu \approx 0$. However, if we assume that $x \to 0$, then the integrals can be evaluated exactly and this condition is referred to as ultra-relativistic limit.

$$\begin{aligned}
\rho^{\text{eq}} &= \frac{k_B^4T^4}{2\pi^2\hbar^3c^3}\left[\sum_{\text{boson}} g_i\int_0^\infty d\xi\frac{\xi^3}{e^\xi-1} + \sum_{\text{fermion}} g_i\int_0^\infty d\xi\frac{\xi^3}{e^\xi+1}\right] \\
&= \frac{k_B^4T^4}{2\pi^2\hbar^3c^3}\left[\sum_{\text{boson}} g_i\int_0^\infty d\xi\frac{\xi^3}{e^\xi-1} + \sum_{\text{fermion}} g_i\int_0^\infty d\xi\frac{\xi^3}{e^\xi-1} - \sum_{\text{fermion}} g_i\int_0^\infty d\xi\frac{2\xi^3}{e^{2\xi}-1}\right] \\
&= \frac{k_B^4T^4}{2\pi^2\hbar^3c^3}\left\{\sum_{\text{boson}} g_i\zeta(3+1)\Gamma(3+1) + \sum_{\text{fermion}} g_i\zeta(3+1)\Gamma(3+1) - \frac{g_i}{8}\zeta(3+1)\Gamma(3+1)\right\} \\
&= \frac{k_B^4T^4}{2\pi^2\hbar^3c^3}\left\{\sum_{\text{boson}} g_i\frac{\pi^4}{15} + \sum_{\text{fermion}} g_i\left(1-\frac{1}{8}\right)\frac{\pi^4}{15}\right\}
\end{aligned}$$

$$= \frac{k_B^4 T^4}{2\pi^2\hbar^3 c^3}\frac{\pi^4}{15}\underbrace{\left\{\sum_{\text{boson}} g_i + \frac{7}{8}\sum_{\text{fermion}} g_i\right\}}_{g^\star}$$

$$= \frac{\pi^2 k_B^4 T^4}{30\hbar^3 c^3} g^\star$$

where $g^\star$ is referred as effective degree of freedom. Note that the above derivation only holds if $k_B T_i >> m_i c^2$ for the species of particle under consideration therefore only relativistic bosons and fermions contribute to it. As the universe was cooling down due to adiabatic expansion. Temperature of universe would eventually fall below the energy scale associated with the mass of particle, in that condition the relativistic particle approximation would start getting violated. Thus those non-relativistic species of particle would stop contributing to the effective degree of freedom.

Non-Relativistic Limit Here, we have to note that, when we convert the integral from momentum space to integral over E via $E^2 = (mc^2)^2 + (Pc)^2$, the lower limit can't be set to $E = 0$ since, the smallest value of E is mc^2. Here we assume $\xi << x$, so $\xi/x << 1$.

$$\begin{aligned}
\rho^{\text{eq}} &= \sum_{\text{all species}} g_i \frac{k_B^4 T^4}{2\pi^2\hbar^3 c^3}\int_0^\infty d\xi \frac{\xi^2\sqrt{\xi^2+x^2}}{e^{\sqrt{\xi^2+x^2}} \pm 1} \\
&= \sum_{\text{all species}} g_i \frac{k_B^4 T^4}{2\pi^2\hbar^3 c^3}\int_0^\infty d\xi \frac{\xi^2 x\sqrt{1+\xi^2/x^2}}{e^{x(1+\xi^2/x^2)^{1/2}}} \\
&= \sum_{\text{all species}} g_i \frac{k_B^4 T^4}{2\pi^2\hbar^3 c^3}\left[\int_0^\infty d\xi\xi^2 x e^{-x-\xi^2/2x} + \int_0^\infty d\xi \frac{\xi^4}{2x} e^{-x-\xi^2/2x}\right] \\
&\approx \sum_{\text{all species}} g_i \frac{k_B^4 T^4}{2\pi^2\hbar^3 c^3} x e^{-x}\int_0^\infty d\xi\xi^2 e^{-\xi^2/2x} \\
&= \sum_{\text{all species}} g_i mc^2 \frac{k_B^3 T^3}{2\pi^2\hbar^3 c^3} e^{-mc^2/k_B T}\int_0^\infty d\xi\xi^2 e^{-\xi^2/2x} \qquad (\text{using } x = {}^{mc^2}/_{k_B T}) \\
&= \sum_{\text{all species}} mc^2 g_i \left(\frac{mk_B T}{2\pi\hbar^2}\right)^{3/2} e^{-mc^2/k_B T}
\end{aligned}$$

The Number Density

The number density, we will define using $n_+ - n_-$, but first do the calculation without assuming the particle and anti-particle nature:

$$\begin{aligned}
n_i^{\text{eq}} &= \frac{g_i}{(2\pi\hbar)^3}\int_0^\infty \frac{1}{\exp\left(\sqrt{p^2c^2+m^2c^4}-\mu/k_B T\right)\pm 1} 4\pi p^2 dp \\
&= \frac{g_i}{2\pi^2\hbar^3}\int_0^\infty dp \frac{p^2}{\exp\left(\sqrt{p^2c^2+m^2c^4}-\mu/k_B T\right)\pm 1}
\end{aligned}$$

defining $x \equiv {}^{mc^2}/_{k_B T}$ and $\xi \equiv {}^{pc}/_{k_B T}$, we get

$$\begin{aligned}
&= \frac{g_i}{2\pi^2\hbar^3}\int_0^\infty \frac{k_B T}{c} d\xi \frac{\frac{k_B^2 T^2}{c^2}\xi^2}{\exp\left(\sqrt{\xi^2+x^2}-\mu/k_B T\right)\pm 1} \\
&= g_i \frac{k_B^3 T^3}{2\pi^2\hbar^3 c^3}\int_0^\infty d\xi \frac{\xi^2}{\exp\left(\sqrt{\xi^2+x^2}-\mu/k_B T\right)\pm 1}
\end{aligned}$$

Ultra-relativistic Limit In the early universe, $\mu \approx 0$. At temperatures above about 10^{16} K, all particles of the Standard Model are highly relativistic, so that their masses can be neglected. Therefore we assume that $x \to 0$, the integrals can then be evaluated exactly and this condition is referred to as ultra-relativistic limit.

$$\begin{aligned} n^{\mathrm{eq}} &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \int_0^\infty d\xi \frac{\xi^2}{\exp\left(\sqrt{\xi^2 + x^2} - \mu/k_B T\right) \pm 1} \\ &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \int_0^\infty d\xi \frac{\xi^2}{e^{\xi - \mu/k_B T} \pm 1} \end{aligned}$$

for $\mu/k_B T << 1$

$$= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \begin{cases} \frac{3}{2}\zeta(3) & \text{for fermions} \\ \\ 2\zeta(3) & \text{for bosons} \end{cases}$$

we will now, evaluate the net number density:

$$n_+^{\mathrm{eq}} - n_-^{\mathrm{eq}} = g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \left[\int_0^\infty d\xi \frac{\xi^2}{e^{\xi - \mu/k_B T} \pm 1} - \int_0^\infty d\xi \frac{\xi^2}{e^{\xi + \mu/k_B T} \pm 1} \right]$$

doing shift of variable $y = \xi - \mu/k_B T$ in first integral and $y = \xi + \mu/k_B T$ in the second integral, we have

$$\begin{aligned} &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \left[\int_{-\mu/k_B T}^\infty dy \frac{(y + \mu/k_B T)^2}{e^y \pm 1} - \int_{\mu/k_B T}^\infty dy \frac{(y - \mu/k_B T)^2}{e^y \pm 1} \right] \\ &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \left[\int_{-\mu/k_B T}^\infty dy \frac{(y + \mu/k_B T)^2}{e^y \pm 1} - \int_{\mu/k_B T}^\infty dy \frac{(y - \mu/k_B T)^2}{e^y \pm 1} \right] \\ &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \left[\int_{-\mu/k_B T}^0 dy \frac{(y + \mu/k_B T)^2}{e^y \pm 1} + \int_0^\infty dy \frac{(y + \mu/k_B T)^2}{e^y \pm 1} \right. \\ &\qquad \left. - \left\{ \int_0^\infty dy \frac{(y - \mu/k_B T)^2}{e^y \pm 1} - \int_0^{\mu/k_B T} dy \frac{(y - \mu/k_B T)^2}{e^y \pm 1} \right\} \right] \\ &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \left[\int_{-\mu/k_B T}^0 dy \frac{(y + \mu/k_B T)^2}{e^y \pm 1} + \int_0^{\mu/k_B T} dy \frac{(y - \mu/k_B T)^2}{e^y \pm 1} \right. \\ &\qquad \left. + \int_0^\infty dy \frac{(y + \mu/k_B T)^2 - (y - \mu/k_B T)^2}{e^y \pm 1} \right] \end{aligned}$$

redefining $y \to -y$ in the first integral and using $a^2 - b^2 = (a+b)(a-b)$ in the last integral

$$\begin{aligned} &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \left[\int_0^{\mu/k_B T} dy \frac{(-y + \mu/k_B T)^2}{e^{-y} \pm 1} + \int_0^{\mu/k_B T} dy \frac{(y - \mu/k_B T)^2}{e^y \pm 1} \right. \\ &\qquad \left. + \int_0^\infty dy \frac{(\mu/k_B T + \mu/k_B T)(y + y)}{e^y \pm 1} \right] \\ &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \left[\int_0^{\mu/k_B T} dy (y - \mu/k_B T)^2 + 4\frac{\mu}{k_B T} \int_0^\infty dy \frac{y}{e^y \pm 1} \right] \\ &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \left[\frac{(y - \mu/k_B T)^3}{3} \Big|_{y=0}^{y=\mu/k_B T} + 4\frac{\mu}{k_B T} \int_0^\infty dy \frac{y}{e^y \pm 1} \right] \end{aligned}$$

$$= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \left[\frac{\mu^3}{3k_B^3 T^3} + \mathcal{I} \right]$$

We will now, explictely evaluate the last integral for fermion and boson. We use the following identity:

$$\begin{aligned} \frac{1}{e^y + 1} &= \frac{1}{e^y - 1} + \text{something} \\ \text{something} &= \frac{1}{e^y + 1} - \frac{1}{e^y - 1} \\ &= \frac{e^y - 1 - e^y - 1}{(e^y + 1)(e^y - 1)} \\ &= -\frac{2}{e^{2y} - 1} \end{aligned}$$

hence,

$$\frac{1}{e^y + 1} = \frac{1}{e^y - 1} - \frac{2}{e^{2y} - 1}$$

therefore,

$$\begin{aligned} 4\frac{\mu}{k_B T} \int_0^\infty dy \frac{y}{e^y + 1} &= 4\frac{\mu}{k_B T} \left\{ \int_0^\infty dy \frac{y}{e^y - 1} - \int_0^\infty dy \frac{2y}{e^{2y} - 1} \right\} \\ &= 4\frac{\mu}{k_B T} \left\{ \Gamma(2)\zeta(2) - \frac{1}{2}\Gamma(2)\zeta(2) \right\} \\ &= 2\frac{\mu}{k_B T}\Gamma(2)\zeta(2) = \frac{\pi^2 \mu}{3k_B T} \end{aligned}$$

and for bosons,

$$4\frac{\mu}{k_B T} \int_0^\infty dy \frac{y}{e^y - 1} = 4\frac{\mu}{k_B T}\Gamma(2)\zeta(2) = \frac{2\pi^2 \mu}{3k_B T}$$

Since, the matter surrounding us are mostly fermions, we get the net fermion density as

$$n^{\text{eq}}_{\text{fermion}} = g_i \frac{k_B^3 T^3}{6\pi^2 \hbar^3 c^3} \left[\frac{\mu^3}{k_B^3 T^3} + \frac{\pi^2 \mu}{k_B T} \right]$$

and

$$n^{\text{eq}}_{\text{boson}} = g_i \frac{k_B^3 T^3}{6\pi^2 \hbar^3 c^3} \left[\frac{\mu^3}{k_B^3 T^3} + \frac{2\pi^2 \mu}{k_B T} \right]$$

Non-Relativistic Limit Here we assume $\xi << x$, so $\xi/x << 1$.

$$\begin{aligned} n^{\text{eq}} &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \int_0^\infty d\xi \frac{\xi^2}{e^{\sqrt{\xi^2 + x^2} + \mu/k_B T} \pm 1} \\ &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \int_0^\infty d\xi \frac{\xi^2}{e^{x(1 + \xi^2/x^2)^{1/2} + \mu/k_B T}} \\ &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \int_0^\infty d\xi \xi^2 e^{-x - \xi^2/2x + \mu/k_B T} \\ &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} e^{-x + \mu/k_B T} \int_0^\infty d\xi \xi^2 e^{-\xi^2/2x} \\ &= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \sqrt{\frac{\pi}{2}} x^{3/2} e^{-x + \mu/k_B T} \qquad (\text{where } x = {}^{mc^2}/_{k_B T}) \end{aligned}$$

$$= g_i \frac{k_B^3 T^3}{2\pi^2 \hbar^3 c^3} \sqrt{\frac{\pi}{2}} \left(\frac{mc^2}{k_B T}\right)^{3/2} e^{-mc^2/k_B T + \mu/k_B T}$$
$$= g_i \frac{1}{\hbar^3} \left(\frac{m k_B T}{2\pi}\right)^{3/2} e^{-mc^2 - \mu/k_B T}$$

Thus we arrive at, with $\mu = 0$

$$\boxed{n^{\text{eq}} = g_i \left(\frac{m k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-mc^2/k_B T}}$$

Pressure

The quantity of interest is pressure:

$$P^{\text{eq}} = \sum_{\text{all species}} \frac{g_i}{(2\pi\hbar)^3} \int \frac{|\vec{p}|^2 c^2}{3E} f(\vec{p}) d^3p$$
$$= \sum_{\text{all species}} \frac{g_i}{2\pi^2\hbar^3} \int_0^\infty \frac{p^2 c^2}{3\sqrt{p^2c^2 + m^2c^4}} \frac{1}{e^{\sqrt{p^2c^2+m^2c^4} - \mu/k_B T} \pm 1} p^2 dp$$

defining $x \equiv {}^{mc^2}/_{k_B T}$ and $\xi = {}^{pc}/_{k_B T}$, we get

$$= \sum_{\text{all species}} \frac{g_i}{2\pi^2\hbar^3} \int_0^\infty \frac{k_B^2 T^2 \xi^2}{3 k_B T \sqrt{\xi^2 + x^2}} \frac{1}{\exp\left(\sqrt{\xi^2 + x^2} - \mu/k_B T\right) \pm 1} \left(\frac{k_B T}{c}\xi\right)^2 \frac{k_B T}{c} d\xi$$
$$= \sum_{\text{all species}} \frac{g_i}{2\pi^2\hbar^3} \int_0^\infty \frac{k_B T \xi^2}{3\sqrt{\xi^2 + x^2}} \frac{1}{\exp\left(\sqrt{\xi^2 + x^2} - \mu/k_B T\right) \pm 1} \left(\frac{k_B T}{c}\xi\right)^2 \frac{k_B T}{c} d\xi$$
$$= \frac{1}{3} \sum_{\text{all species}} g_i \frac{k_B^4 T^4}{2\pi^2\hbar^3 c^3} \int_0^\infty \frac{\xi^4}{\sqrt{\xi^2 + x^2}} \frac{1}{\exp\left(\sqrt{\xi^2 + x^2} - \mu/k_B T\right) \pm 1} d\xi$$

Ultra-relativistic limit In the ultra-relativistic limit $x \to 0$ and $\mu \approx 0$, we observe that

$$P^{\text{eq}} = \frac{1}{3} \sum_{\text{all species}} g_i \frac{k_B^4 T^4}{2\pi^2\hbar^3 c^3} \int_0^\infty \frac{\xi^4}{\sqrt{\xi^2}} \frac{1}{\exp\left(\sqrt{\xi^2} - \mu/k_B T\right) \pm 1} d\xi$$
$$= \frac{1}{3} \underbrace{\frac{k_B^4 T^4}{2\pi^2\hbar^3 c^3} \left[\sum_{\text{boson}} g_i \int_0^\infty d\xi \frac{\xi^3}{e^{\xi - \mu/k_B T} - 1} + \sum_{\text{fermion}} g_i \int_0^\infty d\xi \frac{\xi^3}{e^{\xi - \mu/k_B T} + 1}\right]}_{\rho^{\text{eq}}}$$
$$= \frac{1}{3} \rho^{\text{eq}}$$

Non-relativistic Limit In the limit $\xi << x \implies {}^{\xi}/_{x} << 1$

$$P^{\text{eq}} = \frac{1}{3} \sum_{\text{all species}} g_i \frac{k_B^4 T^4}{2\pi^2\hbar^3 c^3} \int_0^\infty \frac{\xi^4}{\sqrt{\xi^2 + x^2}} \frac{1}{\exp\left(\sqrt{\xi^2 + x^2} - \mu/k_B T\right) \pm 1} d\xi$$
$$= \frac{1}{3} \sum_{\text{all species}} g_i \frac{k_B^4 T^4}{2\pi^2\hbar^3 c^3} \int_0^\infty \frac{\xi^4}{x\sqrt{1 + \frac{\xi^2}{x^2}}} \frac{1}{\exp\left(x\sqrt{1 + \frac{\xi^2}{x^2}} - \mu/k_B T\right) \pm 1} d\xi$$
$$\approx \frac{1}{3} \sum_{\text{all species}} g_i \frac{k_B^4 T^4}{2\pi^2\hbar^3 c^3} \int_0^\infty \frac{\xi^4}{x + \frac{\xi^2}{2x}} \frac{1}{\exp\left(x + \frac{\xi^2}{2x} - \mu/k_B T\right) \pm 1} d\xi$$

$$
\begin{aligned}
&= \frac{k_B T}{3}\left[\sum_{\text{all species}} g_i \frac{k_B^3 T^3}{2\pi^2\hbar^3 c^3}\int_0^\infty \frac{\xi^4}{x+\frac{\xi^2}{2x}}\frac{1}{\exp\left(x+\frac{\xi^2}{2x}-\mu/k_BT\right)\pm 1}d\xi\right]\\
&\approx \frac{k_B T}{3}\left[\sum_{\text{all species}} g_i \frac{k_B^3 T^3}{2\pi^2\hbar^3 c^3}\int_0^\infty \frac{\xi^4}{x} e^{-x-\xi^2/2x+\mu/k_BT} d\xi\right]\\
&= \frac{k_B T}{3}\left[\sum_{\text{all species}} g_i \frac{k_B^3 T^3}{2\pi^2\hbar^3 c^3}\int_0^\infty \frac{\xi^4}{x} e^{-x-\xi^2/2x+\mu/k_BT} d\xi\right]\\
&= \frac{k_B T}{3}\left[\sum_{\text{all species}} g_i \frac{k_B^3 T^3}{2\pi^2\hbar^3 c^3}\frac{1}{x}\times 3\sqrt{\frac{\pi}{2}}x^{5/2} e^{-x-\mu/k_BT}\right]\\
&= k_B T \underbrace{\left[\sum_{\text{all species}} g_i \left(\frac{mk_BT}{2\pi\hbar^2}\right)^{3/2} e^{-mc^2-\mu/k_BT}\right]}_{n^{\text{eq}}}\\
&= n^{\text{eq}} k_B T
\end{aligned}
$$

Using Maxwell-Boltzmann distribution

Now that we have seen how keeping the explicit $\hbar$ and c, makes the calculation look somewhat clustered, we will switch to natural units for this part. Following the previous section

Equilibrium Number density

The equilibrium number density is given by,

$$n^{\text{eq}} = \frac{g_i}{(2\pi\hbar)^3}\int_0^\infty d^3p e^{-\frac{\sqrt{p^2c^2+m^2c^4}}{k_BT}} = \frac{g}{(2\pi)^3}\int d^3p\, e^{-\frac{\sqrt{p^2+m^2}}{T}},$$

where $p = |\mathbf{p}|$ and m is the mass of the particle. Since the integrand depends only on the magnitude of the momentum, it is convenient to switch to spherical coordinates in momentum space. In three dimensions, the volume element in spherical coordinates is

$$d^3p = p^2\, dp\, d\Omega = 4\pi p^2\, dp$$

Thus, the integral reduces to

$$n^{\text{eq}} = \frac{g}{(2\pi)^3} 4\pi \int_0^\infty p^2 e^{-\frac{\sqrt{p^2+m^2}}{T}}\, dp.$$

To proceed, we use the substitution

$$p = m\sinh t, \quad \text{so that} \quad dp = m\cosh t\, dt.$$

With this substitution, the energy becomes

$$\sqrt{p^2+m^2} = \sqrt{m^2\sinh^2 t + m^2} = m\cosh t,$$

and the factor p^2dp transforms as

$$p^2 dp = (m\sinh t)^2 \cdot (m\cosh t\, dt) = m^3 \sinh^2 t \cosh t\, dt.$$

Substituting these expressions into the integral, we obtain

$$n^{\text{eq}} = \frac{g}{2\pi^2} m^3 \int_0^\infty \sinh^2 t \cosh t\, e^{-\frac{m\cosh t}{T}}\, dt.$$

Using the identity $\sinh^2 t = \cosh^2 t - 1$, we have

$$\sinh^2 t \cosh t = (\cosh^2 t - 1)\cosh t = \cosh^3 t - \cosh t.$$

Hence, the integral becomes

$$n^{\rm eq} = \frac{g}{2\pi^2} m^3 \int_0^\infty (\cosh^3 t - \cosh t)\, e^{-\frac{m\cosh t}{T}}\, dt.$$

We can relate powers of $\cosh t$ to $\cosh(3t)$ using the identity

$$\cosh 3t = 4\cosh^3 t - 3\cosh t \quad \Rightarrow \quad \cosh^3 t - \cosh t = \frac{1}{4}(\cosh 3t - \cosh t).$$

Thus, the integral can be rewritten as

$$I = \frac{g}{2\pi^2} m^3 \int_0^\infty \frac{1}{4}(\cosh 3t - \cosh t)\, e^{-\frac{m\cosh t}{T}}\, dt = I = \frac{g}{8\pi^2} m^3 \int_0^\infty (\cosh 3t - \cosh t)\, e^{-\frac{m\cosh t}{T}}\, dt.$$

Next, we recall the integral representation of the modified Bessel function of the second kind:

$$K_\nu(z) = \int_0^\infty e^{-z\cosh t}\cosh(\nu t)\, dt, \quad \mathrm{Re}(z) > 0.$$

Comparing with our integral, we identify

$$\int_0^\infty \cosh 3t\, e^{-m\cosh t/T}\, dt = K_3\left(\frac{m}{T}\right), \quad \int_0^\infty \cosh t\, e^{-m\cosh t/T}\, dt = K_1\left(\frac{m}{T}\right).$$

Hence, we can write

$$I = \frac{g}{8\pi^2} m^3 \left[K_3\left(\frac{m}{T}\right) - K_1\left(\frac{m}{T}\right)\right].$$

Finally, using the recurrence relation for modified Bessel functions

$$K_{\nu+1}(z) - K_{\nu-1}(z) = \frac{2\nu}{z} K_\nu(z),$$

with $\nu = 2$, we get

$$K_3(z) - K_1(z) = \frac{4}{z} K_2(z).$$

Therefore, the final result of the integral is

$$n^{\rm eq} = \frac{g}{8\pi^2} m^3 \cdot \frac{4T}{m} K_2\left(\frac{m}{T}\right) = \frac{g}{(2\pi)^3} 4\pi m^2 T\, K_2\left(\frac{m}{T}\right)$$

Equilibrium Energy Density

We consider a particle species of mass m and internal degrees of freedom g, in thermal equilibrium at temperature T, obeying Maxwell–Boltzmann statistics. The equilibrium energy density is defined as

$$\rho_{\rm eq}(T) = g \int \frac{d^3p}{(2\pi)^3}\, E\, e^{-E/T}, \qquad E = \sqrt{p^2 + m^2}. \tag{1}$$

Using spherical symmetry in momentum space,

$$d^3p = 4\pi p^2 dp, \tag{2}$$

we obtain

$$\rho_{\rm eq} = \frac{g}{2\pi^2}\int_0^\infty dp\, p^2\sqrt{p^2+m^2}\; e^{-\sqrt{p^2+m^2}/T}. \tag{3}$$

Introduce the change of variables

$$p = m\sinh u, \qquad E = \sqrt{p^2+m^2} = m\cosh u. \tag{4}$$

Then

$$dp = m\cosh u\, du, \tag{5}$$

and

$$p^2 E\, dp = m^4 \sinh^2 u \cosh^2 u\, du. \tag{6}$$

Substituting into (3),

$$\rho_{\rm eq} = \frac{gm^4}{2\pi^2}\int_0^\infty du\;\sinh^2 u\cosh^2 u\; e^{-m\cosh u/T}. \tag{7}$$

Using the identity

$$\sinh^2 u = \cosh^2 u - 1, \tag{8}$$

we find

$$\sinh^2 u\cosh^2 u = \cosh^4 u - \cosh^2 u. \tag{9}$$

Thus,

$$\rho_{\rm eq} = \frac{gm^4}{2\pi^2}\int_0^\infty du\,\left(\cosh^4 u - \cosh^2 u\right)e^{-m\cosh u/T}. \tag{10}$$

We use the identities

$$\cosh^2 u = \frac{1}{2}\left(\cosh 2u + 1\right), \tag{11}$$

$$\cosh^4 u = \frac{1}{8}\left(\cosh 4u + 4\cosh 2u + 3\right). \tag{12}$$

Subtracting,

$$\cosh^4 u - \cosh^2 u = \frac{1}{8}\cosh 4u - \frac{1}{8}. \tag{13}$$

Substituting into (10),

$$\rho_{\rm eq} = \frac{gm^4}{16\pi^2}\int_0^\infty du\,\Big[\cosh 4u - 1\Big]e^{-m\cosh u/T}. \tag{14}$$

The modified Bessel function of the second kind is defined by

$$K_\nu(z) = \int_0^\infty du\, e^{-z\cosh u}\cosh(\nu u). \tag{15}$$

Using this definition, and noting $\cosh(0\cdot u) = 1$ so that the constant term is just K_0, equation (14) becomes

$$\rho_{\rm eq} = \frac{gm^4}{16\pi^2}\left[K_4(x) - K_0(x)\right], \qquad x \equiv \frac{m}{T}. \tag{16}$$

The modified Bessel functions satisfy the same recursion relation already used for the number density above,

$$K_{\nu+1}(x) - K_{\nu-1}(x) = \frac{2\nu}{x}K_\nu(x). \tag{17}$$

Applying this with $\nu = 3$ and $\nu = 1$:

$$K_4 - K_2 = \frac{6}{x} K_3, \tag{18}$$

$$K_2 - K_0 = \frac{2}{x} K_1. \tag{19}$$

Adding these two relations gives $K_4 - K_0$ directly,

$$K_4 - K_0 = \frac{6}{x} K_3 + \frac{2}{x} K_1. \tag{20}$$

Using the identity (also already found for the number density, with $\nu = 2$)

$$K_3 = K_1 + \frac{4}{x} K_2, \tag{21}$$

we obtain

$$K_4 - K_0 = \frac{8}{x} K_1 + \frac{24}{x^2} K_2. \tag{22}$$

Substituting into (16),

$$\rho_{\rm eq} = \frac{g m^4}{16\pi^2} \left(\frac{8}{x} K_1 + \frac{24}{x^2} K_2 \right). \tag{23}$$

Restoring $x = m/T$, i.e. $1/x = T/m$ and $1/x^2 = T^2/m^2$,

$$\boxed{\rho_{\rm eq}(T) = \frac{g m^2 T}{2\pi^2} \left[m K_1\!\left(\frac{m}{T}\right) + 3T K_2\!\left(\frac{m}{T}\right) \right].} \tag{24}$$

Equilibrium Pressure

We now have two of the three quantities needed to describe a species in equilibrium: how many particles there are ($n^{\rm eq}$) and how much energy they carry ($\rho^{\rm eq}$). The third one, the pressure $P^{\rm eq}$, is what enters the Friedmann acceleration equation and the equation-of-state parameter $w = P/\rho$, which tells us whether a species behaves like radiation, like matter, or like something in between as the Universe expands and cools. We derive it here directly from the Maxwell–Boltzmann distribution, the same way as $n^{\rm eq}$ and $\rho^{\rm eq}$ above, rather than by taking the $\xi \ll x$ limit of the general Bose–Einstein/Fermi–Dirac integral as we did earlier.

Starting from the general isotropic-pressure integral (3.4.30), with $\mu = 0$ and $f = e^{-E/T}$,

$$P^{\rm eq} = \frac{g}{(2\pi)^3} \int d^3p \, \frac{p^2}{3E} \, e^{-\sqrt{p^2+m^2}/T} = \frac{g}{6\pi^2} \int_0^\infty dp \, \frac{p^4}{E} \, e^{-E/T}, \qquad E = \sqrt{p^2 + m^2}.$$

We use the same substitution as for $n^{\rm eq}$ and $\rho^{\rm eq}$,

$$p = m \sinh u, \qquad E = m \cosh u, \qquad dp = m \cosh u \, du.$$

Then

$$p^4 \, dp = (m \sinh u)^4 (m \cosh u \, du) = m^5 \sinh^4 u \cosh u \, du,$$

so that

$$\frac{p^4}{E} \, dp = \frac{m^5 \sinh^4 u \cosh u}{m \cosh u} \, du = m^4 \sinh^4 u \, du.$$

Hence

$$P^{\rm eq} = \frac{g m^4}{6\pi^2} \int_0^\infty \sinh^4 u \, e^{-m \cosh u / T} \, du.$$

To write $\sinh^4 u$ as a sum of $\cosh(nu)$ terms, use $\sinh^2 u = \cosh^2 u - 1$, so that

$$\sinh^4 u = (\cosh^2 u - 1)^2 = \cosh^4 u - 2\cosh^2 u + 1.$$

Using the same identities as before, $\cosh^2 u = \frac{1}{2}(\cosh 2u + 1)$ and $\cosh^4 u = \frac{1}{8}(\cosh 4u + 4\cosh 2u + 3)$,

$$\begin{aligned}\sinh^4 u &= \frac{1}{8}\big(\cosh 4u + 4\cosh 2u + 3\big) - \big(\cosh 2u + 1\big) + 1 \\ &= \frac{1}{8}\cosh 4u + \frac{1}{2}\cosh 2u + \frac{3}{8} - \cosh 2u \\ &= \frac{1}{8}\cosh 4u - \frac{1}{2}\cosh 2u + \frac{3}{8} \\ &= \frac{1}{8}\Big(\cosh 4u - 4\cosh 2u + 3\Big).\end{aligned}$$

Substituting back,

$$P^{\mathrm{eq}} = \frac{gm^4}{48\pi^2}\int_0^\infty \Big(\cosh 4u - 4\cosh 2u + 3\Big)e^{-m\cosh u/T}\,du.$$

Using the Bessel-function integral representation (with $\cosh(0\cdot u) = 1$ giving K_0),

$$P^{\mathrm{eq}} = \frac{gm^4}{48\pi^2}\Big[K_4(x) - 4K_2(x) + 3K_0(x)\Big], \qquad x \equiv \frac{m}{T}.$$

We reduce this combination using the same recursion relation as before, $K_{\nu+1}(x) - K_{\nu-1}(x) = \frac{2\nu}{x}K_\nu(x)$, applied with $\nu = 3, 2, 1$:

$$K_4 = K_2 + \frac{6}{x}K_3, \qquad K_3 = K_1 + \frac{4}{x}K_2, \qquad K_0 = K_2 - \frac{2}{x}K_1.$$

Substituting the middle relation into the first,

$$K_4 = K_2 + \frac{6}{x}\left(K_1 + \frac{4}{x}K_2\right) = K_2 + \frac{6}{x}K_1 + \frac{24}{x^2}K_2.$$

Then

$$\begin{aligned}K_4 - 4K_2 + 3K_0 &= \left(K_2 + \frac{6}{x}K_1 + \frac{24}{x^2}K_2\right) - 4K_2 + 3\left(K_2 - \frac{2}{x}K_1\right) \\ &= \underbrace{(1 - 4 + 3)}_{=0}K_2 + \frac{24}{x^2}K_2 + \underbrace{\left(\frac{6}{x} - \frac{6}{x}\right)}_{=0}K_1 \\ &= \frac{24}{x^2}K_2(x).\end{aligned}$$

Every bare K_1 term and every bare K_2 term cancels, leaving the remarkably simple result

$$K_4(x) - 4K_2(x) + 3K_0(x) = \frac{24}{x^2}K_2(x).$$

Therefore

$$P^{\mathrm{eq}} = \frac{gm^4}{48\pi^2}\cdot\frac{24}{x^2}K_2(x) = \frac{gm^4}{2\pi^2 x^2}K_2(x).$$

Restoring $x = m/T$, so that $m^4/x^2 = m^2T^2$,

$$\boxed{P^{\mathrm{eq}}(T) = \frac{gm^2T^2}{2\pi^2}\,K_2\!\left(\frac{m}{T}\right).} \tag{25}$$

Check: the ideal gas law falls out for free. Compare this with the number density result found above, $n^{\rm eq} = \dfrac{gm^2T}{2\pi^2}K_2(m/T)$. The two expressions differ by exactly one factor of T:

$$P^{\rm eq}(T) = n^{\rm eq}(T)\,T. \tag{26}$$

This is nothing but the classical ideal-gas law $PV = Nk_BT$, and here it holds *exactly*, for any value of m/T, not just in some limit. This should not be a surprise: throwing away the quantum ±1 in the Bose–Einstein/Fermi–Dirac denominator is precisely what makes a gas "classical" in the first place, and $P = nT$ is the defining property of a classical ideal gas. It is also exactly the relation we already found by brute-force expanding the general relativistic integral in the limit $\xi \ll x$ at the end of §3.4.2; here the same result comes out directly and exactly, with no expansion needed, because Maxwell–Boltzmann statistics were assumed from the very first line.

Limits, and how they connect to the rest of the chapter.

- **Non-relativistic limit ($x = m/T \gg 1$, i.e. $T \ll m$).** Using the large-argument behaviour $K_2(x) \approx K_1(x) \approx \sqrt{\pi/2x}\,e^{-x}$, we get $P^{\rm eq} \approx n^{\rm eq}T$ as always, while the energy density is dominated by the rest mass, $\rho^{\rm eq} \approx m\,n^{\rm eq}$. Hence

$$w \equiv \frac{P^{\rm eq}}{\rho^{\rm eq}} \approx \frac{T}{m} \ll 1,$$

  reproducing the last row of Table 3.4 ($P = nT \ll \rho$): a cold, non-relativistic species carries essentially no pressure. This is exactly the statement used later whenever we treat a massive relic species (a dark-matter candidate, or the baryons after nucleosynthesis) as pressureless matter.

- **Ultra-relativistic limit ($x \ll 1$, i.e. $T \gg m$).** Using the small-argument behaviour $K_2(x) \approx 2/x^2$, both $n^{\rm eq}$ and $P^{\rm eq}$ grow as gT^3 and gT^4 respectively, i.e. $P^{\rm eq} \approx \rho^{\rm eq}/3$: the familiar radiation equation of state. The precise numerical coefficients ($\pi^2/30$ for bosons, $\frac{7}{8} \times \pi^2/30$ for fermions, as in Table 3.4) do not come out right, because Maxwell–Boltzmann statistics discard the quantum ±1 that distinguishes bosons from fermions. This is exactly why the precision calculation of $g_*(T)$ in §3.4.3 must go back to the full Bose–Einstein/Fermi–Dirac integrals: the Maxwell–Boltzmann approximation used in this section is only ever *exact* in the non-relativistic limit, and only an approximation ($E - \mu \gg T$) in the relativistic one.

With $n^{\rm eq}$, $\rho^{\rm eq}$, and $P^{\rm eq}$ in hand, we can describe a species completely *while it stays in equilibrium*. What is still missing, to know *when* a species falls out of equilibrium in the first place, is the interaction rate $\Gamma \sim n^{\rm eq}\langle\sigma v\rangle$ compared with the Hubble rate H (recall the freeze-out criterion of §3.4.2), and for that we need a practical, computable expression for the thermally averaged cross section $\langle\sigma v\rangle$ itself. We derive that next, once the Boltzmann equation has been set up in §3.4.7.

3.4.3 Effective relativistic degrees of freedom $g_*(T)$

Summing over all relativistic species ($m_i \ll T$), the total radiation energy density is parameterized by $g_*(T)$:

$$\rho_r(T) = g_*(T)\frac{\pi^2}{30}T^4, \tag{3.4.31}$$

where the **effective number of relativistic degrees of freedom** for energy is:

$$g_*(T) \equiv \sum_{i=\text{bosons}} g_i \left(\frac{T_i}{T}\right)^4 + \frac{7}{8} \sum_{i=\text{fermions}} g_i \left(\frac{T_i}{T}\right)^4 . \tag{3.4.32}$$

The factor 7/8 originates from the integral ratio $\int_0^\infty \frac{x^3\,\mathrm{d}x}{\mathrm{e}^x+1} = \frac{7}{8}\int_0^\infty \frac{x^3\,\mathrm{d}x}{\mathrm{e}^x-1} = \frac{7}{8}\frac{\pi^4}{15}$.

In the radiation era ($H^2 = \frac{8\pi G_\mathrm{N}}{3}\rho_r$), the Hubble expansion rate and cosmic time are:

$$H(T) = \sqrt{\frac{8\pi G_\mathrm{N}}{3}\frac{\pi^2}{30} g_*(T) T^4} = \left(\frac{4\pi^3 G_\mathrm{N}}{45} g_*\right)^{1/2} T^2 \simeq 1.66\, g_*^{1/2} \frac{T^2}{M_\mathrm{Pl}}, \tag{3.4.33}$$

$$t(T) = \frac{1}{2H(T)} \simeq 0.301\, g_*^{-1/2} \frac{M_\mathrm{Pl}}{T^2} \simeq 2.42\, g_*^{-1/2} \left(\frac{T}{1\,\mathrm{MeV}}\right)^{-2} \mathrm{s}, \tag{3.4.34}$$

where $M_\mathrm{Pl} \equiv 1/\sqrt{G_\mathrm{N}} = 1.22 \times 10^{19}\,\mathrm{GeV}$ is the Planck mass.

3.4.4 Chemical potentials and matter–antimatter asymmetry

Inelastic processes such as double Compton scattering ($e^- + \gamma \leftrightarrow e^- + 2\gamma$) and Bremsstrahlung ($e^- + p \leftrightarrow e^- + p + \gamma$) create and destroy photons freely, demanding:

$$\mu_\gamma = 0. \tag{3.4.35}$$

For any particle-antiparticle pair A and $\bar{A}$ in thermal equilibrium via rapid annihilation $A + \bar{A} \leftrightarrow 2\gamma$:

$$\mu_A + \mu_{\bar{A}} = 2\mu_\gamma = 0 \implies \mu_{\bar{A}} = -\mu_A. \tag{3.4.36}$$

In the relativistic regime ($T \gg m_A$), the net particle number asymmetry is:

$$n_A - n_{\bar{A}} = \frac{g_A T^3}{6\pi^2}\left[\pi^2\left(\frac{\mu_A}{T}\right) + \left(\frac{\mu_A}{T}\right)^3\right] \approx \frac{g_A T^3}{6}\left(\frac{\mu_A}{T}\right) \quad (\mu_A \ll T). \tag{3.4.37}$$

Baryon asymmetry and conservation of η. Cosmic electric charge neutrality ($n_p = n_e - n_{\bar{e}}$) and the empirical baryon-to-photon ratio:

$$\eta \equiv \frac{n_B}{n_\gamma} = \frac{n_p + n_n}{n_\gamma} \approx 6 \times 10^{-10} \tag{3.4.38}$$

impose that the chemical potentials are tiny:

$$\frac{\mu_e}{T} \approx \frac{\mu_p}{T} \sim 10^{-9} \ll 1, \tag{3.4.39}$$

justifying setting $\mu_i \approx 0$ for all Standard Model species in early thermodynamic calculations. Furthermore, because baryon number is strictly conserved at temperatures below the electroweak sphaleron freeze-out ($T \lesssim 100\,\mathrm{GeV}$), the total comoving baryon number $N_B = n_B a^3$ is conserved. After e^+e^- annihilation, comoving photon number is likewise conserved, guaranteeing that η remains constant throughout the subsequent evolution of the Universe. In terms of cosmological density parameters today:

$$\eta = \frac{\rho_{b0}/m_p}{n_{\gamma 0}} = \frac{\Omega_{b0}\rho_{\mathrm{crit},0}/m_p}{\frac{2\zeta(3)}{\pi^2} T_{\gamma 0}^3} \approx 2.739 \times 10^{-8}\, \Omega_{b0} h^2. \tag{3.4.40}$$

3.4.5 Comoving entropy conservation and $g_{*S}(T)$

From the thermodynamic relation $\mathrm{d}U = T\,\mathrm{d}S - P\,\mathrm{d}V + \mu\,\mathrm{d}N$, the relativistic **entropy density** is:

$$s \equiv \frac{\rho + P - \sum_i \mu_i n_i}{T} \approx \frac{\rho + P}{T}. \tag{3.4.41}$$

Combining with the fluid continuity equation $\mathrm{d}(\rho a^3) = -P\,\mathrm{d}a^3$, the total comoving entropy is strictly conserved:

$$\mathrm{d}(sa^3) = 0 \iff S \equiv s(T)a^3(t) = \text{constant}. \tag{3.4.42}$$

The entropy density is parameterized by $g_{*S}(T)$ (also denoted $q_*(T)$):

$$s(T) = g_{*S}(T)\frac{2\pi^2}{45}T^3, \tag{3.4.43}$$

where:

$$g_{*S}(T) \equiv \sum_{i=\text{bosons}} g_i \left(\frac{T_i}{T}\right)^3 + \frac{7}{8}\sum_{i=\text{fermions}} g_i \left(\frac{T_i}{T}\right)^3. \tag{3.4.44}$$

When all relativistic species share a common temperature $T_i = T$, $g_{*S}(T) = g_*(T)$.

The **comoving abundance (yield)** of any species i is defined by:

$$Y_i \equiv \frac{n_i}{s}, \tag{3.4.45}$$

which remains constant under cosmic expansion in the absence of creation/annihilation.

3.4.6 Thermal history: Neutrino decoupling and e^+e^- annihilation

Cosmic neutrino decoupling ($T \approx 1\,\mathrm{MeV}$)

Neutrinos (ν_e, ν_μ, ν_τ) are kept in thermal equilibrium with charged leptons via electroweak weak neutral- and charged-current interactions ($\nu\bar{\nu} \leftrightarrow e^+e^-$, $\nu e^- \leftrightarrow \nu e^-$). The weak interaction rate is:

$$\Gamma_{\text{weak}} = n_e\langle\sigma_{\text{weak}}v\rangle \approx G_{\text{F}}^2 T^5. \tag{3.4.46}$$

Comparing with the Hubble rate $H \approx 1.66 g_*^{1/2} T^2/M_{\text{Pl}}$:

$$\frac{\Gamma_{\text{weak}}}{H} \approx \left(\frac{T}{1\,\mathrm{MeV}}\right)^3. \tag{3.4.47}$$

At $T_{\nu,D} \approx 1\,\mathrm{MeV}$ ($t \approx 1\,\mathrm{s}$), $\Gamma_{\text{weak}} \lesssim H$, and the three neutrino flavors decouple, free-streaming with temperature $T_\nu \propto a^{-1}$.

Electron-positron annihilation and photon reheating

Shortly after neutrino decoupling, the temperature drops below the electron mass ($T < m_e = 0.511\,\mathrm{MeV}$). Electrons and positrons annihilate ($e^+ + e^- \to 2\gamma$), but because neutrinos are already decoupled, the entropy of the $e^\pm$ pairs is dumped **exclusively into the photon bath**:

- Before annihilation ($T > m_e$): The electromagnetic plasma contains photons ($g_\gamma = 2$) and $e^\pm$ pairs ($g_e = 4$):

$$g_{*S,\text{em}}(T > m_e) = 2 + \frac{7}{8}(2+2) = \frac{11}{2}. \tag{3.4.48}$$

- After annihilation ($T < m_e$): Only photons remain relativistic ($g_\gamma = 2$):

$$g_{*S,\text{em}}(T < m_e) = 2. \tag{3.4.49}$$

By comoving entropy conservation, $S_{\rm em} = g_{*S,\rm em}(aT)^3 = \text{constant}$:

$$(aT_\gamma)^3_{\rm after} = \frac{g_{*S}(T > m_e)}{g_{*S}(T < m_e)} (aT_\gamma)^3_{\rm before} = \frac{11}{4} (aT_\nu)^3 . \tag{3.4.50}$$

This yields the exact temperature ratio between photons and the Cosmic Neutrino Background (CNB):

$$T_\nu = \left(\frac{4}{11}\right)^{1/3} T_\gamma \approx 0.7137\, T_\gamma. \tag{3.4.51}$$

Given $T_{\gamma 0} = 2.7255\,\mathrm{K}$ today, the relic neutrino temperature is $T_{\nu 0} = 1.945\,\mathrm{K}$ ($k_{\rm B} T_{\nu 0} = 1.676 \times 10^{-4}\,\mathrm{eV}$), with total number density:

$$n_\nu = 3 \times \frac{3}{4} \left(\frac{4}{11}\right) n_\gamma = \frac{9}{11} n_\gamma \approx 336\,\mathrm{cm}^{-3} \quad (112\,\mathrm{cm}^{-3} \text{ per flavor}). \tag{3.4.52}$$

Present degrees of freedom g_{*0} and g_{*S0}

Evaluating today for photons and 3 relativistic neutrino species:

$$g_{*0} = 2 + 3 \times \frac{7}{8}(2) \left(\frac{T_\nu}{T_\gamma}\right)^4 = 2 + \frac{21}{4} \left(\frac{4}{11}\right)^{4/3} = 3.363, \tag{3.4.53}$$

$$g_{*S0} = 2 + 3 \times \frac{7}{8}(2) \left(\frac{T_\nu}{T_\gamma}\right)^3 = 2 + \frac{21}{4} \left(\frac{4}{11}\right) = \frac{43}{11} = 3.909. \tag{3.4.54}$$

Table 3.5: Particle Content and Effective Relativistic Degrees of Freedom $g_*(T)$ in the Standard Model

Temperature T	**Mass Threshold**	**Relativistic Particle Species**	$g_*(T)$
$T < m_e$	$< 0.511\,\mathrm{MeV}$	γ (+3 decoupled ν)	3.36
$m_e < T < T_{\nu,D}$	0.511–1.0 MeV	$\gamma, e^\pm, 3\nu$	10.75
$T_{\nu,D} < T < m_\mu$	1–106 MeV	$\gamma, e^\pm, \mu^\pm, 3\nu$	14.25
$m_\mu < T < m_\pi$	106–135 MeV	$\gamma, e^\pm, \mu^\pm, \pi^\pm, \pi^0, 3\nu$	17.25
$T_{\rm QCD} < T < m_s$	150–400 MeV	$\gamma, e^\pm, \mu^\pm, 3\nu, u, d, 8g$ (QGP)	61.75
$m_s < T < m_c$	0.5–1.3 GeV	$+s, \bar{s}$	75.75
$m_c < T < m_\tau$	1.3–1.8 GeV	$+c, \bar{c}$	86.25
$m_\tau < T < m_b$	1.8–4.2 GeV	$+\tau^\pm$	92.25
$m_b < T < m_W$	4.2–80.4 GeV	$+b, \bar{b}$	96.25
$m_W < T < m_t$	80.4–173 GeV	$+W^\pm, Z^0$	106.75
$T > m_H, m_t$	$> 173\,\mathrm{GeV}$	Full Standard Model ($+t, \bar{t}, H^0$)	**106.75**
$T > T_{\rm GUT}$ (Minimal SU(5))	$> 10^{15}\,\mathrm{GeV}$	Full GUT Multiplets	161.75
$T > T_{\rm SUSY}$ (MSSM)	$> 1\,\mathrm{TeV}$	Full Supersymmetric Spectrum	228.75

Neutrino masses, oscillations, and cosmological bounds

The baseline instantaneous decoupling model assumes that neutrinos are strictly massless. However, neutrino oscillation experiments (Super-Kamiokande, SNO, KamLAND) have conclusively established that neutrino flavor eigenstates (ν_e, ν_μ, ν_τ) are quantum superpositions of distinct mass eigenstates (ν_1, ν_2, ν_3). The measured mass-squared differences are:

$$\Delta m^2_{21} \equiv m_2^2 - m_1^2 \approx 7.5 \times 10^{-5}\,\mathrm{eV}^2 \quad \text{(solar oscillations)}, \tag{3.4.55}$$

$$|\Delta m^2_{31}| \equiv |m_3^2 - m_1^2| \approx 2.5 \times 10^{-3}\,\mathrm{eV}^2 \quad \text{(atmospheric oscillations)}. \tag{3.4.56}$$

Because oscillations measure only squared mass differences, two mass orderings are possible:

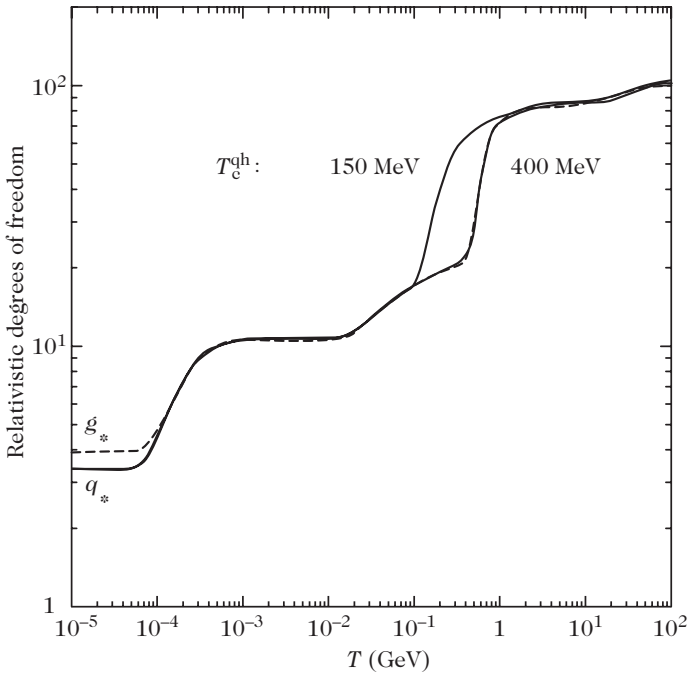


Figure 3.7: Evolution of the effective relativistic degrees of freedom $g_*(T)$ and $g_{*S}(T)$ as a function of temperature T in the Standard Model $\mathrm{SU}(3)_c \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$, showing the electroweak, quark-hadron QCD ($T_c \approx 150$–$400\,\mathrm{MeV}$), and e^+e^- annihilation steps.

- **Normal Hierarchy (NH):** $m_1 < m_2 \ll m_3 \implies \sum m_\nu \geq \sqrt{\Delta m^2_{21}} + \sqrt{|\Delta m^2_{31}|} \approx 0.059\,\mathrm{eV}$.
- **Inverted Hierarchy (IH):** $m_3 \ll m_1 < m_2 \implies \sum m_\nu \geq 2\sqrt{|\Delta m^2_{31}|} \approx 0.10\,\mathrm{eV}$.

Comparing these lower bounds to the present neutrino temperature $T_{\nu 0} \approx 1.945\,\mathrm{K} = 1.676 \times 10^{-4}\,\mathrm{eV}$, we discover that *at least two of the three neutrino mass eigenstates have rest masses far exceeding their thermal energy today* ($m_\nu \gg T_{\nu 0}$). Consequently, relic neutrinos have transitioned from relativistic radiation in the early Universe to non-relativistic matter today!

The present energy density of non-relativistic relic neutrinos is:

$$\rho_{\nu 0} = \sum_{i=1}^{3} m_i n_{\nu i,0} = \left(\sum_{i=1}^{3} m_i \right) \frac{3}{11} n_{\gamma 0}. \tag{3.4.57}$$

Dividing by the critical density $\rho_{\mathrm{crit},0} = 3H_0^2/(8\pi G) = 1.054 \times 10^4\, h^2\,\mathrm{eV} \cdot \mathrm{cm}^{-3}$ yields the celebrated **Cowsik–McClelland relation**:

$$\Omega_{\nu 0} h^2 = \frac{\sum m_\nu}{94.14\,\mathrm{eV}}. \tag{3.4.58}$$

Cosmological upper bounds from the Planck CMB and Large Scale Structure survey ($\sum m_\nu < 0.12\,\mathrm{eV}$) place the neutrino contribution to cosmic matter at $\Omega_\nu \lesssim 0.003$, meaning relic neutrinos behave as hot dark matter that suppresses the growth of small-scale matter perturbations.

Instantaneous decoupling versus precision thermal history. Under the idealized approximation that neutrino decoupling is strictly instantaneous and occurs before e^+e^- annihilation commences, neutrinos receive none of the released entropy, yielding $T_\nu = (4/11)^{1/3} T_\gamma$ and

$N_{\rm eff} \equiv 3.000$ exactly [Eq. (3.4.118)]. In reality, because weak interaction cross sections scale quadratically with energy ($\sigma \propto G_{\rm F}^2 E^2$), high-energy neutrinos in the tail of the Fermi–Dirac distribution remain partially coupled to the plasma down to $T \sim 0.5\,\text{MeV}$, absorbing a small fraction of the entropy released during e^+e^- annihilation. Accurately tracking this out-of-equilibrium energy transfer requires solving the relativistic Boltzmann transport equation alongside neutrino flavor oscillations and radiative plasma QED corrections. After establishing the general Boltzmann transport theory and thermal relic freeze-out kinetics in Section 3.4.7, we return to this problem in Section 3.4.10, presenting the complete first-principles quantum kinetic derivation that establishes the precision Standard Model benchmark $N_{\rm eff} = 3.044$.

3.4.7 Out-of-equilibrium relativistic Boltzmann equation

To compute the relic abundance of decoupled species (dark matter, light nuclei in BBN, and recombination), we formulate the **covariant Boltzmann transport equation**:

$$\hat{\mathcal{L}}[f] = \mathcal{C}[f], \tag{3.4.59}$$

where $\mathcal{C}[f]$ is the collision integral, and $\hat{\mathcal{L}}$ is the Liouville operator along the phase-space geodesic congruence:

$$\hat{\mathcal{L}}[f] = p^\alpha \frac{\partial f}{\partial x^\alpha} - \Gamma^\alpha_{\beta\gamma} p^\beta p^\gamma \frac{\partial f}{\partial p^\alpha}. \tag{3.4.60}$$

In a spatially homogeneous and isotropic FLRW metric, $f = f(E,t)$, reducing the Liouville operator to:

$$\hat{\mathcal{L}}[f] = E\frac{\partial f}{\partial t} - Hp^2 \frac{\partial f}{\partial E}. \tag{3.4.61}$$

Integrating over momentum space $\frac{g_i}{(2\pi)^3}\int \frac{\mathrm{d}^3p}{E}$, the left-hand side yields the macroscopic rate equation:

$$\frac{g_i}{(2\pi)^3}\int \hat{\mathcal{L}}[f]\frac{\mathrm{d}^3p}{E} = \dot{n}_i + 3Hn_i = \frac{1}{a^3}\frac{\mathrm{d}}{\mathrm{d}t}\left(n_i a^3\right). \tag{3.4.62}$$

Boltzmann Transport Equation

Boltzmann Equation, in theory, can be used to derive the distribution function and its time evolution. However, the problem is the need for boundary condition. Which means, we need to know the distribution at some time t. But, in the framework where the system begins to move away from the equilibrium, we already know the distribution at some time t from the equilibrium statistical mechanics. Thus, it can be used to inquire more about the system.

$$\frac{\partial f}{\partial t} + \frac{\vec{p}}{m}\cdot\nabla_{\vec{q}} f + \vec{F}\cdot\nabla_{\vec{p}} f = \left(\frac{\partial f}{\partial t}\right)_{\text{collision}}$$

Example

Let us consider Maxwell-Boltzmann distribution[a]:

$$f(\vec{p},\vec{q}) = Ae^{-\beta\left(\frac{|\vec{p}|^2}{2m}+U(\vec{q})\right)}, \text{ where } \beta = \frac{1}{k_B T}.$$

Since this distribution describes equilibrium distribution the collision term would vanish i.e., $\partial_t f = 0$, whereas the gradients are:

$$\nabla_{\vec{q}} f = -\beta f \nabla_{\vec{q}} U = \beta \vec{F} f, \qquad \nabla_{\vec{p}} f = -\beta \frac{\vec{p}}{m} f,$$

that is the sum of the two gradient term in the Boltzmann Transport Equation is zero. The density distribution under this distribution is not uniform, unless the potential is a

constant, in which the distribution becomes Maxwell distribution

$$f(\vec{p}) = Ae^{-\beta\frac{|\vec{p}|^2}{2m}},$$

that gives a non-zero gradient $\nabla_{\vec{p}} f = -\beta\frac{\vec{p}}{m} f$, but still satisfies the Boltzmann equation, since now $\vec{F} = -\nabla_{\vec{q}} U = 0$.

Derivation

We assume binary collisions throughout (justified below in assumption a) and that any external force does not act appreciably during the brief interval of the collision itself (it still appears in the streaming term below).
In the absence of collisions, all particles that start out at $(\vec{x}, \vec{p}, t)$ end up at $(\vec{x}+\vec{v}\,dt, \vec{p}, t+dt)$. In the presence of collisions, however, some particles leave this stream while others join it: let the number of j-particles lost due to collision be $\Gamma^-_{ji}\, d^3x\, d^3p_j\, dt$, and the number of j-particles that join the stream due to collision with an i-particle be $\Gamma^+_{ji}\, d^3x\, d^3p_j\, dt$, where i, j label the colliding species.

$$\left(\frac{\partial f_j}{\partial t}\right)_{\text{collision}} = \sum_i \Gamma^+_{ji} - \Gamma^-_{ji}$$

Consider a $12 \to 1'2'$ scattering process, with particle 1 the particle whose distribution f_1 we are tracking and particle 2 its collision partner. The construction below follows Cercignani & Kremer, *The Relativistic Boltzmann Equation*, Ch. 2. Three assumptions – together called the *Stoßzahlansatz* ("supposition of the number of collisions") – underlie everything that follows:

a) only binary collisions are taken into account (reasonable for a dilute gas, where the volume occupied by the particles is much smaller than the volume of the gas);

b) the momenta of two particles about to collide are uncorrelated – this is the *molecular chaos* assumption, and it applies equally to $(\vec{p}_1, \vec{p}_2)$ before a collision and to $(\vec{p}_{1'}, \vec{p}_{2'})$ before the inverse collision that produces $(\vec{p}_1, \vec{p}_2)$;

c) $f(\vec{x}, \vec{p}, t)$ does not vary appreciably over a time longer than the duration of a collision but shorter than the time between collisions (and likewise in space, over the range of the interaction).

Before setting up the collision itself, we first check that f can consistently be taken to be a Lorentz scalar, i.e. that the ordinary phase-space count $dN = f\, d^3x\, d^3p$ used to define f is a frame-independent statement. The number of particles crossing a spacelike hypersurface element $d\Sigma_\mu$ is computed exactly as one computes charge crossing a surface from a current: as the flux of a 4-vector, $dN = dJ^\mu\, d\Sigma_\mu$. For particles with momentum in d^3p around p, all moving with the same ordinary velocity $\vec{v} = \vec{p}/p^0$, the current they contribute is (density)×(velocity 4-vector),

$$dJ^\mu(x,p) = \underbrace{f(x,p)\, d^3p}_{\text{density in this frame}} \times \underbrace{\frac{p^\mu}{p^0}}_{=(1,\vec{v})} = p^\mu f(x,p)\,\frac{d^3p}{p^0} \equiv p^\mu f(x,p)\, d\Pi, \qquad d\Pi \equiv \frac{d^3p}{p^0},$$

manifestly a 4-vector provided f is a scalar, since $d\Pi$ is the standard invariant momentum measure (the same combination as $d^4p\,\delta(p^2 - m^2)$ used in §3.4.7, up to the conventional factor of 2). The number of worldlines crossing $d\Sigma_\mu$ is then the scalar

$$dN = dJ^\mu\, d\Sigma_\mu = f(x,p)\, p^\mu\, d\Sigma_\mu\, d\Pi.$$

Choosing an equal-time slice, $d\Sigma_\mu = (d^3x, \vec{0})$, so that $p^\mu d\Sigma_\mu = p^0 d^3x$, gives

$$dN = f(x,p)\, p^0\, d\Pi\, d^3x = f(x,p)\, d^3p\, d^3x,$$

recovering exactly the ordinary phase-space count we started from. Since dN is a pure particle count – hence a scalar – and f is a scalar by this construction, $d^3x\, d^3p$ is itself Lorentz invariant, even though neither factor is separately.

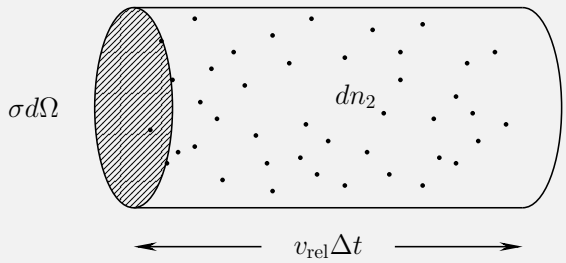


Figure 3.8: The collision cylinder swept out by particle 1 (at rest) through the density dn_2 of particle-2 targets in time Δt: base $\sigma\, d\Omega$, height $v_{\rm rel}\Delta t$.

Now set up the collision itself. Along particle 1's trajectory, define the affine parameter λ_1 via $p_1^\mu \equiv dx_1^\mu/d\lambda_1$, so that $dt = E_1\, d\lambda_1$; since λ_1 is proportional to particle 1's own proper time, it is a Lorentz scalar regardless of which frame we use to compute it. Promoting the general definition above ($\Gamma^+_{ji}\, d^3x\, d^3p_j\, dt$) by trading the frame-dependent dt for the invariant $d\lambda_1$, let

$$\Gamma^+ \equiv \frac{\text{number of collisions}}{d^3x\, d^3p_1\, \Delta\lambda_1}.$$

Since $d^3x\, d^3p_1$ was already shown to be Lorentz invariant, and $\Delta\lambda_1$ is invariant by construction, Γ^+ is a genuine Lorentz scalar *by definition*.
To evaluate it, work in the rest frame of particle 1, $\vec{v}_1 = \vec{0}$; there $\Delta t = E_1\, \Delta\lambda_1 = m_1\, \Delta\lambda_1$ is (up to the constant m_1) particle 1's own proper time. The number of particle-1's about $\vec{x}$ is $dn_1\, d^3x$ with $dn_1 \equiv f_1\, d^3p_1$ (valid in this frame, by the result derived above).
The particle-2's about to collide with them, occupying a volume dV_2, number $dn_2\, dV_2$; since these particles move at the relative speed $v_{\rm rel} = |\vec{v}_2|$ in this frame, their density here is $dn_2/\sqrt{1-v_{\rm rel}^2}$, where $dn_2 \equiv f_2\, d^3p_2$ is their density in *their own* rest frame. The number of collisions taking place in Δt is, simply,

$$\text{Number of collisions} = \underbrace{dn_1\, d^3x}_{\text{initial number of colliding particles}} \overbrace{\frac{dn_2}{\sqrt{1-v_{\rm rel}^2}}\, dV_2}^{\text{number of target particle-2's swept up}}. \tag{1}$$

Writing the swept volume as the collision cylinder of base $\sigma\, d\Omega$ and height $v_{\rm rel}\Delta t$ (Fig. 3.8), i.e. $dV_2 = \sigma\, d\Omega\, v_{\rm rel}\, \Delta t$, Eq. (1) becomes, in terms of the cross section,

$$\text{Number of collisions} = \underbrace{dn_1\, d^3x}_{\text{initial number of colliding particles}} \overbrace{\frac{dn_2}{\sqrt{1-v_{\rm rel}^2}}\, \sigma\, d\Omega\, v_{\rm rel}}^{\text{rate of collision}}\, \Delta t. \tag{2}$$

Dividing by the phase-space volume $d^3x\,d^3p_1$ of the particle we are tracking connects this to Γ^+ as just defined:

$$\frac{\text{number of collisions}}{\text{volume of phase space}} = \frac{\text{net change in the number of particle-1's}}{\text{volume of phase space}} = (\Gamma^+ - \Gamma^-)\,\Delta\lambda_1,$$

where Γ^+ is the rate just defined above (for $12 \to 1'2'$) and Γ^- is the analogous rate for the inverse process $1'2' \to 12$, computed further below. Working out Γ^+ first, and trading the geometric Δt above for the invariant $\Delta\lambda_1$ via $\Delta t = E_1\,\Delta\lambda_1$,

$$\Gamma^+ \cancel{d^3x\,d^3p_1\,\Delta\lambda_1} = \underbrace{f_1\,\cancel{d^3p_1}\,\cancel{d^3x}}_{dn_1\,d^3x}\,\frac{dn_2}{\sqrt{1-v_{\rm rel}^2}}\,\sigma\,d\Omega\,v_{\rm rel}\,E_1\cancel{\Delta\lambda_1}$$

$$\Gamma^+ = E_1\,f_1\,\frac{dn_2}{\sqrt{1-v_{\rm rel}^2}}\,\sigma\,d\Omega\,v_{\rm rel}.$$

To write Γ^+ covariantly, multiply through by d^3p_1 and use the general (frame-independent) definitions $dn_1/\sqrt{1-v_1^2} \equiv f_1\,d^3p_1$ and $dn_2/\sqrt{1-v_2^2} \equiv f_2\,d^3p_2$ (here $v_1 = 0$, $v_2 = v_{\rm rel}$); the density product appearing in $\Gamma^+\,d^3p_1$ becomes

$$\Gamma^+\,d^3p_1 = E_1\,\frac{dn_1\,dn_2}{\sqrt{1-v_{\rm rel}^2}}\,\sigma\,d\Omega\,v_{\rm rel} = E_1\,\frac{dn_1}{\sqrt{1-v_1^2}}\,\frac{dn_2}{\sqrt{1-v_2^2}}\,\frac{\sqrt{1-v_1^2}\sqrt{1-v_2^2}}{\sqrt{1-v_{\rm rel}^2}}\,\sigma\,d\Omega\,v_{\rm rel}$$
$$= E_1\,f_1\,d^3p_1\,f_2\,d^3p_2\,\frac{p_1\cdot p_2}{E_1E_2}\,v_{\rm rel}\,\sigma\,d\Omega, \tag{3}$$

using $\sqrt{1-v_1^2}\sqrt{1-v_2^2}/\sqrt{1-v_{\rm rel}^2} = 1 - \vec{v}_1\cdot\vec{v}_2 = p_1\cdot p_2/(E_1E_2)$ – a standard kinematic identity, trivially true here since both sides equal 1 when $\vec{v}_1 = \vec{0}$, and manifestly frame-independent once written in terms of dot products. Dividing (3) by d^3p_1, the explicit E_1 cancels the one hiding inside $p_1\cdot p_2/(E_1E_2)$:

$$\Gamma^+ = f_1 f_2\,d^3p_2\,v_{\rm rel}\,\frac{p_1\cdot p_2}{E_2}\,\sigma\,d\Omega. \tag{4}$$

In this frame $p_1\cdot p_2 = E_1E_2 - \vec{p}_1\cdot\vec{p}_2 = E_1E_2$ (since $\vec{p}_1 = \vec{0}$), so $p_1\cdot p_2/E_2 = E_1$ and $v_{\rm rel} = |\vec{v}_1 - \vec{v}_2|$ trivially, and Eq. (4) becomes $\Gamma^+ = f_1f_2\,d^3p_2\,(E_1|\vec{v}_1 - \vec{v}_2|)\,\sigma\,d\Omega$. The combination $E_1|\vec{v}_1 - \vec{v}_2|$ is not itself a Lorentz scalar, but multiplying and dividing by E_2 reveals one: since $E_1 = m_1$ here,

$$E_1E_2|\vec{v}_1 - \vec{v}_2| = E_1E_2|\vec{v}_2| = E_1|\vec{p}_2| = \sqrt{E_1^2E_2^2 - E_1^2m_2^2} = \sqrt{(p_1\cdot p_2)^2 - m_1^2m_2^2}, \tag{5}$$

built purely from dot products and masses, hence a genuine Lorentz scalar. Writing $F_{12} \equiv 4\sqrt{(p_1\cdot p_2)^2 - m_1^2m_2^2}$, so that $E_1|\vec{v}_1 - \vec{v}_2| = F_{12}/(4E_2)$, Γ^+ becomes

$$\Gamma^+ = \frac{f_1f_2}{2}\,\frac{d^3p_2}{2E_2}\,F_{12}\,\sigma\,d\Omega. \tag{6}$$

Integrating over the target momenta $\vec{p}_2$ and scattering angles (matching $(\partial f_j/\partial t)_{\rm collision} = \sum_i \Gamma^+_{ji} - \Gamma^-_{ji}$ above, with the sum over i now a continuum integral over $\vec{p}_2$),

$$\left(\frac{{\rm d}f_1}{{\rm d}\lambda_1}\right)_{\rm loss} = \int\Gamma^+ = \frac{1}{2}\int\frac{d^3p_2}{2E_2}\,F_{12}\,d\sigma\;f_1f_2.$$

The same argument for the inverse process $1'2' \to 12$ gives the gain term with primed quantities and $\Delta\lambda_1'$ (particle $1'$'s own affine parameter, in general a different frame). Liouville's theorem – phase-space volume is conserved along the collision map $(p_1, p_2) \to (p_{1'}, p_{2'})$ – guarantees the loss and gain measures agree, so loss and gain share exactly the same phase-space measure and can be written under one integral sign. Molecular chaos – the assumption, also known as *Stoßzahlansatz*, that the momenta of colliding partners are uncorrelated – then gives the net rate:

$$E_1\left[\frac{\partial f_1}{\partial t} + \frac{\vec{p}_1}{E_1}\cdot\nabla_{\vec{q}} f_1 + \vec{F}\cdot\nabla_{\vec{p}_1} f_1\right] = \frac{\mathrm{d}f_1}{\mathrm{d}\lambda_1} = \frac{1}{2}\int \frac{d^3p_2}{2E_2} F_{12}\; \underbrace{d\sigma}_{\text{integrated over all final states}}\; (f_{1'}f_{2'} - f_1 f_2) \qquad (7)$$

Every factor entering the right-hand side is manifestly Lorentz invariant; σ itself is promoted to a fully quantum, relativistic cross section below. The above equation holds true in flat spacetime, however in curved spacetime the LHS of Boltzmann Transport Equation needs modification as follows:

$$\begin{aligned}
\frac{\mathrm{d}f}{\mathrm{d}\lambda} &= \frac{\partial f}{\partial t}\frac{\mathrm{d}t}{\mathrm{d}\lambda} + \frac{\partial f}{\partial x^i}\frac{\mathrm{d}x^i}{\mathrm{d}\lambda} + \frac{\partial f}{\partial E}\frac{\mathrm{d}E}{\mathrm{d}\lambda} + \frac{\partial f}{\partial p^i}\frac{\mathrm{d}p^i}{\mathrm{d}\lambda} \\
&= p^0\frac{\partial f}{\partial t} + p^i\frac{\partial f}{\partial x^i} + \frac{\mathrm{d}p^\mu}{\mathrm{d}\lambda}\frac{\partial f}{\partial p^\mu} \\
&= p^\mu\frac{\partial f}{\partial x^\mu} + \frac{\mathrm{d}p^\mu}{\mathrm{d}\lambda}\frac{\partial f}{\partial p^\mu} \\
&= p^\mu\frac{\partial f}{\partial x^\mu} - \Gamma^\mu_{\alpha\beta}p^\alpha p^\beta\frac{\partial f}{\partial p^\mu} = \mathcal{L}[f]
\end{aligned}$$

We have substituted $dp^\mu/d\lambda$ from the geodesic equation. Now, we will simplify the above in case of FLRW geometry.

$$\begin{aligned}
\frac{\mathrm{d}f}{\mathrm{d}\lambda} &= p^\mu\frac{\partial f}{\partial x^\mu} + \frac{p^\mu}{p^\mu}\frac{\mathrm{d}p^\mu}{\mathrm{d}\lambda}\frac{\partial f}{\partial p^\mu} \\
&= p^\mu\frac{\partial f}{\partial x^\mu} + \frac{\mathrm{d}\ln p^\mu}{\mathrm{d}\lambda}\frac{\partial f}{\partial \ln p^\mu} \\
&= p^\mu\frac{\partial f}{\partial x^\mu} - \frac{\mathrm{d}\ln a}{\mathrm{d}\lambda}\frac{\partial f}{\partial \ln p^\mu} \qquad \text{(used } p^\mu a = \text{constant)} \\
\frac{\mathrm{d}t}{\mathrm{d}\lambda}\frac{\mathrm{d}f}{\mathrm{d}t} &= p^\mu\frac{\partial f}{\partial x^\mu} - \frac{\mathrm{d}t}{\mathrm{d}\lambda}\frac{\mathrm{d}\ln a}{\mathrm{d}t}\frac{\partial f}{\partial \ln p^\mu}
\end{aligned}$$

assuming $f \equiv f(t, p)$

$$\begin{aligned}
p^0\frac{\mathrm{d}f}{\mathrm{d}t} &= p^0\frac{\partial f}{\partial t} - p^0\frac{\mathrm{d}\ln a}{\mathrm{d}t}\frac{\partial f}{\partial \ln p^\mu} = \mathcal{C}[f] \\
\frac{\mathrm{d}f}{\mathrm{d}t} &= \frac{\mathcal{C}[f]}{p^0}
\end{aligned}$$

We can express the Boltzmann Equation simply as:[b]

$$\frac{\mathrm{d}f}{\mathrm{d}\lambda} = \mathcal{L}[f] = \mathcal{C}[f]$$

Now, we will integrate f with respect to d^3p

$$g\int \frac{\mathrm{d}f}{\mathrm{d}t}\frac{d^3p}{(2\pi\hbar)^3} = g\int\left[\frac{\partial f}{\partial t} - \frac{\mathrm{d}\ln a}{\mathrm{d}t}\frac{\partial f}{\partial \ln p^\mu}\right]\frac{d^3p}{(2\pi\hbar)^3}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial t}\frac{g}{(2\pi\hbar)^3}\int f(t,p)d^3p - \frac{\mathrm{d}\ln a}{\mathrm{d}t}\frac{g}{(2\pi\hbar)^3}\int \frac{\partial f}{\partial \ln p^\mu}d^3p \\
&= \frac{\mathrm{d}n}{\mathrm{d}t} - \frac{\mathrm{d}\ln a}{\mathrm{d}t}\frac{g}{(2\pi\hbar)^3}\left\{ f(t,p)\times 4\pi p^3\Big|_{p=0}^{p=\infty} - \int 3f(t,p)d^3p\right\} \\
&= \frac{\mathrm{d}n}{\mathrm{d}t} + \frac{\mathrm{d}\ln a}{\mathrm{d}t}\times 3n = \frac{1}{a^3}\frac{\mathrm{d}na^3}{\mathrm{d}t}
\end{aligned}
\tag{8}
$$

where we evaluated the second integral as follows:

$$
\begin{aligned}
\int \frac{\partial f}{\partial \ln p^\mu}d^3p &= \int \frac{\partial f}{\partial \ln p^\mu}4\pi p^2\frac{p}{p}dp \\
&= \int_0^\infty \frac{\partial f}{\partial \ln p^\mu}d\ln p\times 4\pi p^3 - \int_0^\infty f d(4\pi p^3) \\
&= f4\pi p^3\Big|_{p=0}^{p=\infty} - 3\int_0^\infty f4\pi p^2dp
\end{aligned}
$$

This gives $\dot{n}+3Hn = \frac{1}{a^3}\frac{\mathrm{d}(na^3)}{\mathrm{d}t}$, matching Eq. (3.4.62).

The collision term. Above, $\mathcal{L}[f] = \mathcal{C}[f]$ means precisely $p^0\frac{\mathrm{d}f}{\mathrm{d}t} = \mathcal{C}[f]$, i.e. (using $\frac{\mathrm{d}t}{\mathrm{d}\lambda_1} = p_1^0 = E_1$) $\frac{\mathrm{d}f_1}{\mathrm{d}\lambda_1} = \mathcal{C}[f_1]$. Eq. (7) already gives $\frac{\mathrm{d}f_1}{\mathrm{d}\lambda_1}$ explicitly and covariantly, so $\mathcal{C}[f_1]$ is exactly its right-hand side:

$$
\mathcal{C}[f_1] = \frac{1}{2}\int \frac{d^3p_2}{2E_2}F_{12}\,d\sigma\,(f_{1'}f_{2'} - f_1f_2). \tag{9}
$$

Integrating with $g_1\int d^3p_1/E_1$, as in Eq. (3.4.62):

$$
C \equiv \frac{g_1}{(2\pi)^3}\int \frac{d^3p_1}{E_1}\mathcal{C}[f_1] = g_1\int \frac{d^3p_1}{(2\pi)^3\,2E_1}\int \frac{d^3p_2}{2E_2}F_{12}\,d\sigma\,(f_{1'}f_{2'} - f_1f_2), \tag{10}
$$

where the $\frac{1}{2}$ of Eq. (9) combines with $1/E_1$ into $d\Pi_1 \equiv d^3p_1/[(2\pi)^3 2E_1]$.
By $d\sigma \equiv d\Gamma/F$, $F_{12}\,d\sigma_{12\to 1'2'}$ is $\mathrm{d}\Gamma_{12\to 1'2'}$ of Eq. (1), relabelled $i,j,k,l \to 1,2,1',2'$:

$$
F_{12}\,d\sigma_{12\to 1'2'} = (2\pi)^4\delta^{(4)}(p_1+p_2-p_{1'}-p_{2'})\,|\mathcal{M}|^2\,\frac{d^3p_{1'}}{(2\pi)^3 2E_{1'}}\frac{d^3p_{2'}}{(2\pi)^3 2E_{2'}}.
$$

Supplying the same internal-degeneracy and $(2\pi)^3$ normalization for particle 2, $d^3p_2 \to g_2\,d^3p_2$, from one-particle states normalized as $\langle \mathbf{p}|\mathbf{p}'\rangle = 2E(2\pi)^3\delta^{(3)}(\mathbf{p}-\mathbf{p}')$, Eq. (10) becomes

$$
C = (2\pi)^4\int g_1\mathrm{d}\Pi_1\,g_2\mathrm{d}\Pi_2\,g_{1'}\mathrm{d}\Pi_{1'}\,g_{2'}\mathrm{d}\Pi_{2'}\;\delta^{(4)}(p_1+p_2-p_{1'}-p_{2'})\,|\mathcal{M}|^2\,(f_{1'}f_{2'}-f_1f_2),
$$

the relativistic collision integral of Eq. (3.4.65) (relabel $1,2,1',2' \to i,j,k,l$).

[a] It can be derived for dilute gases using Boltzmann Transport Equation, see chapter 4 of Kerson Huang's Statistical Mechanics

[b] better derivation for collision term can be found in "Boltzmann equation and its cosmological applications" by Seishi Enomoto

Collision integral for $2 \leftrightarrow 2$ scatterings

The fundamental challenge in solving the Boltzmann equation lies in evaluating the collision functional $\mathcal{C}[f_i]$. Consider a two-body scattering and annihilation process:

$$i + j \longleftrightarrow k + l. \tag{3.4.63}$$

The macroscopic collision term C_i decomposes into creation ($k + l \to i + j$) and destruction ($i + j \to k + l$) components:

$$C_i = C_{kl\to ij} - C_{ij\to kl}, \tag{3.4.64}$$

where the destruction rate per unit volume is:

$$\begin{aligned} C_{ij\to kl} = (2\pi)^4 \int \frac{g_i \mathrm{d}^4 p_i}{(2\pi)^3}\, \delta(E_i^2 - \mathbf{p}_i^2 - m_i^2)\, \ldots\, \frac{g_l \mathrm{d}^4 p_l}{(2\pi)^3}\, \delta(E_l^2 - \mathbf{p}_l^2 - m_l^2) \\ \times\, \delta^{(4)}(p_i + p_j - p_k - p_l)\, |\mathcal{M}|^2_{ij\to kl}\, f_i f_j (1 \pm f_k)(1 \pm f_l). \end{aligned} \tag{3.4.65}$$

Here, the $+$ sign corresponds to Bose enhancement and the $-$ sign to Fermi blocking for the final-state particles k and l. The matrix element $|\mathcal{M}|^2_{ij\to kl}$ is the spin-averaged quantum transition amplitude, and the four-dimensional Dirac delta enforces energy-momentum conservation $p_i + p_j = p_k + p_l$.

Derivation of the collision integral

Consider the reaction

$$i + j \leftrightarrow k + l.$$

The collision term in the Boltzmann equation measures the *net* change in the number of particles of species i due to this reaction. There are therefore two contributions:

$$\text{net rate} = \text{rate of } k + l \to i + j - \text{rate of } i + j \to k + l.$$

The derivation naturally separates into three parts. First, we determine how the occupation numbers of the quantum states enter the transition probability. Second, we determine the transition rate associated with a given set of momenta. Finally, we sum over all available momentum and internal states to obtain the collision integral.

We use natural units, $\hbar = c = k_B = 1$, and work in a local inertial frame. The four species are taken to be distinct, so no additional identical-particle symmetry factors are required. Correlations between distinct momentum and internal-state modes are neglected. The occupation number N_a refers to a single quantum mode, while

$$f_a(t, \mathbf{p}_a) \equiv \langle N_a \rangle$$

denotes its mean occupation. The quantity n_a will denote the corresponding particle number density.

1. Occupation-number factors

The first question is: given a particular set of occupation numbers, what is the amplitude for the transition $i + j \to k + l$?

The operator responsible for this transition has the structure

$$a_k^\dagger a_l^\dagger a_j a_i.$$

The two annihilation operators remove the incoming particles, whereas the two creation operators add the outgoing particles.

For a bosonic or fermionic mode, define

$$\eta_a = \begin{cases} +1, & \text{boson}, \\ -1, & \text{fermion}. \end{cases}$$

The canonical (anti)commutation relations imply

$$a_a a_a^\dagger = 1 + \eta_a a_a^\dagger a_a.$$

Since

$$\widehat{N}_a = a_a^\dagger a_a,$$

we have

$$\begin{aligned} \| a_a \left| N_a \right\rangle \|^2 &= \left\langle N_a \right| a_a^\dagger a_a \left| N_a \right\rangle = N_a, \\ \| a_a^\dagger \left| N_a \right\rangle \|^2 &= \left\langle N_a \right| a_a a_a^\dagger \left| N_a \right\rangle = 1 + \eta_a N_a. \end{aligned}$$

Thus

$$a_a \left| N_a \right\rangle = \sqrt{N_a} \left| N_a - 1 \right\rangle,$$

whereas

$$a_a^\dagger \left| N_a \right\rangle = \sqrt{1 + \eta_a N_a} \left| N_a + 1 \right\rangle.$$

The physical interpretation is important. An incoming particle can be removed with an amplitude proportional to $\sqrt{N_a}$, since there are N_a particles available in the mode. The creation of an outgoing particle gives the factor $\sqrt{1+\eta_a N_a}$. For bosons this becomes $\sqrt{1+N_a}$, giving Bose enhancement. For fermions it becomes $\sqrt{1-N_a}$, which vanishes when the state is already occupied and therefore gives Pauli blocking.
For the complete transition,

$$\begin{aligned} & a_k^\dagger a_l^\dagger a_j a_i \left| \ldots, N_i, \ldots, N_j, \ldots, N_k, \ldots, N_l, \ldots \right\rangle \\ & \quad = \sqrt{N_i N_j (1+\eta_k N_k)(1+\eta_l N_l)} \\ & \qquad \times \left| \ldots, N_i - 1, \ldots, N_j - 1, \ldots, N_k + 1, \ldots, N_l + 1, \ldots \right\rangle, \end{aligned}$$

up to an irrelevant phase and the appropriate ordering convention for fermionic operators. Consequently, if $\mathcal{A}_0$ denotes the transition amplitude with the occupation-number dependence removed, then

$$|\mathcal{A}_{\{N\}}|^2 = |\mathcal{A}_0|^2 N_i N_j (1+\eta_k N_k)(1+\eta_l N_l).$$

Thus the statistical factors appearing in the collision integral are not inserted by hand. They follow directly from the matrix elements of the creation and annihilation operators.

2. Statistical averaging
A macroscopic gas is not described by one particular occupation-number state. Instead, we average over an ensemble described by a density operator ρ:

$$\langle O \rangle \equiv \mathrm{Tr}(\rho O).$$

In particular,

$$f_a = \langle a_a^\dagger a_a \rangle = \langle N_a \rangle.$$

Using the operator identity above,

$$\langle a_a a_a^\dagger \rangle = 1 + \eta_a f_a.$$

If correlations between the four participating modes are neglected, the occupation-number factor therefore factorizes:

$$\begin{aligned}\langle N_i N_j (1+\eta_k N_k)(1+\eta_l N_l)\rangle \\ \simeq f_i f_j (1+\eta_k f_k)(1+\eta_l f_l).\end{aligned}$$

Thus the forward reaction is weighted by

$$f_i f_j (1+\eta_k f_k)(1+\eta_l f_l),$$

while the inverse reaction is weighted by

$$f_k f_l (1+\eta_i f_i)(1+\eta_j f_j).$$

No assumption of thermal equilibrium has been made. In equilibrium one would have

$$f_a^{\mathrm{eq}}(E_a) = \frac{1}{\exp[(E_a-\mu_a)/T]-\eta_a},$$

but the collision term is valid for arbitrary nonequilibrium distributions $f_a(t,\mathbf{p}_a)$.

3. Continuum normalization

We now derive the rate directly in the continuum, without introducing a finite box. This is the cleanest way to obtain the final form of the collision integral.

Use relativistically normalized one-particle states,

$$\langle \mathbf{p}, r | \mathbf{p}', r' \rangle = 2E(2\pi)^3 \delta^{(3)}(\mathbf{p}-\mathbf{p}')\delta_{rr'},$$

where

$$E = \sqrt{\mathbf{p}^2 + m^2}.$$

For fixed momenta and internal states, the S-matrix element is written as

$$\langle f|S|i\rangle = \mathrm{i}(2\pi)^4 \delta^{(4)}(q)\, \mathcal{M}_{ij\to kl},$$

with

$$q = p_i + p_j - p_k - p_l.$$

The invariant amplitude $\mathcal{M}$ contains the microscopic dynamics of the scattering process. The remaining factors arise from the normalization of states and from the statistical occupation of those states.

For a fixed incoming pair, the probability per unit time to produce particles with momenta in d^3p_k and d^3p_l is proportional to the final-state phase-space measure:

$$\begin{aligned}\mathrm{d}\Gamma_{ij\to kl} = (2\pi)^4 \delta^{(4)}(q) |\mathcal{M}_{ij\to kl}|^2 \\ \times \frac{\mathrm{d}^3p_k}{(2\pi)^3\, 2E_k} \frac{\mathrm{d}^3p_l}{(2\pi)^3\, 2E_l}.\end{aligned} \tag{1}$$

To describe a gas, rather than one particular incoming pair, we must sum over the available incoming particles. This introduces the same invariant phase-space measure for each incoming particle:

$$\mathrm{d}\Pi_a \equiv \frac{\mathrm{d}^3p_a}{(2\pi)^3\, 2E_a}.$$

Hence the destruction rate per unit volume for $i+j\to k+l$ is

$$C_{ij\to kl} = (2\pi)^4 \int \mathrm{d}\Pi_i \mathrm{d}\Pi_j \mathrm{d}\Pi_k \mathrm{d}\Pi_l\, \delta^{(4)}(q) |\mathcal{M}_{ij\to kl}|^2$$

$$\times f_i f_j (1+\eta_k f_k)(1+\eta_l f_l).$$

The delta function enforces energy-momentum conservation, the matrix element describes the microscopic interaction, and the occupation factors determine how many particles are available initially and how the quantum statistics of the final states modify the transition.

4. The invariant phase-space measure
The one-particle measure

$$\mathrm{d}\Pi_a = \frac{\mathrm{d}^3 p_a}{(2\pi)^3\, 2E_a}$$

is Lorentz invariant. Its appearance follows from the relativistic normalization of one-particle states.
The factor $2E_a$ comes from the normalization

$$\langle \mathbf{p}|\mathbf{p}'\rangle = 2E(2\pi)^3\delta^{(3)}(\mathbf{p}-\mathbf{p}'),$$

while $(2\pi)^3$ is the conventional momentum-space normalization. Thus the invariant phase-space measure is simply the continuum form of the quantum-state normalization.

5. Internal-state degeneracies
If a species has g_a internal states, such as spin or color states, the internal-state labels must also be summed. Explicitly, one has

$$\sum_{r_i,r_j,r_k,r_l} |\mathcal{M}_{r_i r_j \to r_k r_l}|^2.$$

For an unpolarized gas, it is conventional to define a matrix element averaged over the initial internal states and summed over the final ones:

$$\overline{|\mathcal{M}|^2} \equiv \frac{1}{g_i g_j} \sum_{r_i,r_j,r_k,r_l} |\mathcal{M}_{r_i r_j \to r_k r_l}|^2.$$

The important point is that the convention must be used consistently: the internal-state sums should not be counted both in $|\mathcal{M}|^2$ and again through the phase-space factors.

6. Gain and loss
We can now combine the forward and inverse reactions. The forward reaction

$$i + j \to k + l$$

removes particles of species i and therefore gives a loss term:

$$C^{\mathrm{loss}}_{ij\to kl} = (2\pi)^4 \int \mathrm{d}\Pi_i \mathrm{d}\Pi_j \mathrm{d}\Pi_k \mathrm{d}\Pi_l\, \delta^{(4)}(q)|\mathcal{M}|^2 \times f_i f_j (1+\eta_k f_k)(1+\eta_l f_l).$$

The inverse reaction

$$k + l \to i + j$$

creates particles of species i and therefore gives a gain term:

$$C^{\mathrm{gain}}_{kl\to ij} = (2\pi)^4 \int \mathrm{d}\Pi_i \mathrm{d}\Pi_j \mathrm{d}\Pi_k \mathrm{d}\Pi_l\, \delta^{(4)}(q)|\mathcal{M}|^2 \times f_k f_l (1+\eta_i f_i)(1+\eta_j f_j).$$

Assuming microscopic reversibility, the same invariant matrix element may be used for the forward and inverse reactions. Hence

$$\begin{aligned} C_i &= C^{\text{gain}}_{kl\to ij} - C^{\text{loss}}_{ij\to kl} \\ &= (2\pi)^4 \int \mathrm{d}\Pi_i \mathrm{d}\Pi_j \mathrm{d}\Pi_k \mathrm{d}\Pi_l \, \delta^{(4)}(q) |\mathcal{M}|^2 \\ &\quad \times \Big[f_k f_l (1+\eta_i f_i)(1+\eta_j f_j) \\ &\qquad - f_i f_j (1+\eta_k f_k)(1+\eta_l f_l) \Big]. \end{aligned}$$

This is the desired collision integral.
Expanding the phase-space measures gives

$$\begin{aligned} C_i = (2\pi)^4 \int \prod_{a=i,j,k,l} \frac{\mathrm{d}^3 p_a}{(2\pi)^3 \, 2E_a} \\ \times \delta^{(4)}(p_i + p_j - p_k - p_l) |\mathcal{M}|^2 \\ \times \Big[f_k f_l (1+\eta_i f_i)(1+\eta_j f_j) \\ - f_i f_j (1+\eta_k f_k)(1+\eta_l f_l) \Big]. \end{aligned}$$

Writing

$$1 + \eta_a f_a = \begin{cases} 1 + f_a, & \text{bosons}, \\ 1 - f_a, & \text{fermions}, \end{cases}$$

gives the familiar quantum-statistical form of the collision integral.
Finally, the collision term enters the Boltzmann equation as

$$\dot{n}_i + 3Hn_i = C_i.$$

The term $3Hn_i$ describes dilution of the number density due to the expansion of the universe, while C_i describes the local creation and destruction of particles through microscopic reactions.
In the dilute classical limit,

$$f_a \ll 1,$$

the Bose-enhancement and Pauli-blocking factors approach unity, and the statistical part of the collision term becomes

$$f_k f_l - f_i f_j.$$

The usual classical Boltzmann collision term is therefore recovered.
For reactions involving identical particles, additional symmetry and particle-counting factors must be included. These factors are distinct from the quantum-statistical factors derived above.

7. The symmetry factor, made explicit
Concretely, once the number density

$$n = \frac{g}{(2\pi)^3} \int d^3p \, f(\vec{p})$$

is integrated against $\dot{n}_i + 3Hn_i = C_i$, the collision term C_i of Eq. (3.4.67) becomes, for a generic $1 + 2 \to 3 + 4$ process,

$$\frac{1}{a^3}\frac{\mathrm{d}(na^3)}{\mathrm{d}t} = \prod_{i=1}^{4}\int \frac{g_i\, d^3p_i}{(2\pi)^3\, 2E_i} \left|\tilde{\mathcal{M}}_{12\to 34}\right|^2 (2\pi)^4\delta^{(4)}(p_1+p_2-p_3-p_4)$$
$$\times \frac{1}{S}\Big[f_3 f_4(1\pm f_1)(1\pm f_2) - f_1 f_2(1\pm f_3)(1\pm f_4)\Big], \quad (2)$$

where S is exactly the symmetry factor promised above: it counts identical particles among the initial or final states, so that summing over all momentum configurations does not overcount final states that differ only by permuting identical particles, and

$$\left|\tilde{\mathcal{M}}_{12\to 34}\right|^2 = \frac{1}{\prod_i g_i}\sum_{\text{spins } i} |\mathcal{M}_i|^2$$

is the matrix element already averaged over the initial and final spins and internal degrees of freedom, which is why it can sit outside the g_i-weighted phase-space measure. Equation (2) is nothing but Eq. (3.4.67) integrated once more against $d^3p_i/2E_i$ and written with S made explicit; we will use it, with a concrete S, when we specialize to a definite annihilation process in §3.4.8.

Assuming $\hat{\mathcal{C}}\hat{\mathcal{P}}$ invariance (or $\hat{\mathcal{T}}$ invariance / microscopic reversibility), the forward and backward transition amplitudes are identical:

$$|\mathcal{M}|^2_{ij\to kl} = |\mathcal{M}|^2_{kl\to ij} \equiv |\mathcal{M}|^2. \quad (3.4.66)$$

Using the on-shell identity $\mathrm{d}^4p\,\delta(E^2-\mathbf{p}^2-m^2) = \frac{\mathrm{d}^3p}{2E}$, the total net collision rate takes the compact form:

$$C_i = (2\pi)^4\int \delta^{(4)}(p_i+p_j-p_k-p_l)\left(\prod_{a=i,j,k,l}\frac{g_a\mathrm{d}^3p_a}{2(2\pi)^3E_a}\right)|\mathcal{M}|^2$$
$$\times\left[f_k f_l(1\pm f_i)(1\pm f_j) - f_i f_j(1\pm f_k)(1\pm f_l)\right]. \quad (3.4.67)$$

The Boltzmann equation $\dot{n}_i + 3Hn_i = C_i$ thus captures three physical effects:

1. **Hubble dilution** ($3Hn_i$), reducing particle density due to volume expansion;
2. **Particle creation** ($C_{kl\to ij}$), populating state i via inverse reactions;
3. **Particle destruction** ($C_{ij\to kl}$), depleting state i via forward annihilations.

Maxwell–Boltzmann approximation and chemical equilibrium

In most cosmological applications (such as freeze-out and nucleosynthesis), particle energies satisfy $E_a - \mu_a \gg T$, allowing quantum statistical factors to be neglected:

$$1\pm f_a \approx 1, \qquad f_a(p_a) \approx \mathrm{e}^{(\mu_a - E_a)/T}. \quad (3.4.68)$$

Applying this directly to the statistical bracket of Eq. (3.4.67), every $1\pm f$ factor is set to 1 and the collision term collapses to its familiar classical form,

$$f_k f_l(1\pm f_i)(1\pm f_j) - f_i f_j(1\pm f_k)(1\pm f_l) \;\approx\; f_k f_l - f_i f_j, \quad (3.4.69)$$

which is exactly the dilute-gas limit already anticipated in the classical treatment of the collision integral above. Under this approximation, the actual particle number density n_a relates to its fictitious zero-chemical-potential equilibrium density $\bar{n}_a \equiv n_a(\mu_a = 0)$ via:

$$n_a = \mathrm{e}^{\mu_a/T}\bar{n}_a. \quad (3.4.70)$$

This is immediate once f_a from Eq. (3.4.68) is integrated over momentum: the chemical potential is momentum-independent, so it simply factors out of the integral,

$$n_a = g_a e^{\mu_a/T}\int \frac{d^3p}{(2\pi)^3}\, e^{-E_a/T} \;\equiv\; e^{\mu_a/T}\,\bar{n}_a, \qquad \bar{n}_a \equiv g_a\int \frac{d^3p}{(2\pi)^3}\, e^{-E_a/T}, \tag{3.4.71}$$

which is precisely Eq. (3.4.70) written out in full. Energy conservation $E_k + E_l = E_i + E_j$ enables the phase-space distribution product of Eq. (3.4.69) to factorize as:

$$\begin{aligned} f_k f_l - f_i f_j &= e^{-(E_k+E_l)/T}\left[e^{(\mu_k+\mu_l)/T} - e^{(\mu_i+\mu_j)/T}\right] \\ &= e^{-(E_i+E_j)/T}\left[\frac{n_k n_l}{\bar{n}_k\bar{n}_l} - \frac{n_i n_j}{\bar{n}_i\bar{n}_j}\right]. \end{aligned} \tag{3.4.72}$$

Substituting Eq. (3.4.72) into Eq. (3.4.67) yields the **integrated Boltzmann rate equation**:

$$\dot{n}_i + 3Hn_i = -\langle\sigma v\rangle\left(n_i n_j - \frac{\bar{n}_i\bar{n}_j}{\bar{n}_k\bar{n}_l} n_k n_l\right), \tag{3.4.73}$$

where the **thermally averaged cross section** $\langle\sigma v\rangle$ is defined covariantly by:

$$\bar{n}_i\bar{n}_j\langle\sigma v\rangle \equiv \int \frac{d^3p_i}{2E_i}\frac{d^3p_j}{2E_j}\frac{d^3p_k}{2E_k}\frac{d^3p_l}{2E_l}\frac{\delta^{(4)}(p_i+p_j-p_k-p_l)}{(2\pi)^8}\,|\mathcal{M}|^2\, e^{-(E_i+E_j)/T}. \tag{3.4.74}$$

[1]

Dividing Eq. (3.4.73) through by $n_i a^3$ and writing $\dot{n}_i + 3Hn_i = \frac{1}{a^3}\frac{d(n_i a^3)}{dt}$ (as in Eq. (3.4.62)) together with $H = \frac{d\ln a}{dt}$ makes the size of each term manifest:

$$\frac{1}{n_i a^3}\frac{d(n_i a^3)}{d\ln a} = -\frac{n_j\langle\sigma v\rangle}{H}\left[1 - \frac{n_k n_l}{n_i n_j}\left(\frac{\bar{n}_i\bar{n}_j}{\bar{n}_k\bar{n}_l}\right)\right] \equiv -\frac{\Gamma}{H}\left[1 - \frac{n_k n_l}{n_i n_j}\left(\frac{\bar{n}_i\bar{n}_j}{\bar{n}_k\bar{n}_l}\right)\right], \tag{3.4.75}$$

with the interaction rate $\Gamma \equiv n_j\langle\sigma v\rangle$. Written this way, the comoving abundance $n_i a^3$ changes appreciably, per Hubble time, only when the bracket is of order unity *and* $\Gamma/H \gtrsim 1$. When the reaction rate per particle $\Gamma \sim n_j\langle\sigma v\rangle$ exceeds the expansion rate H ($\Gamma \gg H$), Eq. (3.4.75) forces the bracketed term to vanish, establishing **chemical equilibrium**:

$$\frac{n_i n_j}{\bar{n}_i\bar{n}_j} = \frac{n_k n_l}{\bar{n}_k\bar{n}_l} \iff \mu_i + \mu_j = \mu_k + \mu_l. \tag{3.4.76}$$

As the Universe expands and cools, H eventually overtakes Γ, breaking chemical equilibrium and freezing out particle abundances.

Evaluating the thermally averaged cross section $\langle\sigma v\rangle$

The covariant definition in Eq. (3.4.74) is exact, but as it stands it is an integral over eight momentum components (the two three-momenta of the incoming particles, with the four-momentum-conserving delta function still to be used), which is not something one can simply hand to a numerical integrator. Following Mirco Cannoni, we now reduce it to a single integral over the Mandelstam invariant $s = (p_i + p_j)^2$. This is precisely the object that enters the freeze-out equation for Y_X in §3.4.8 below and the cold/hot relic abundance formulas of §3.4.9, so it is worth deriving carefully and in full, rather than just quoting the result. The derivation has two parts: first we need a manifestly Lorentz-invariant expression for the relative velocity $v_{\rm rel}$ of the two incoming particles (since $\langle\sigma v\rangle$ is only meaningful once we know what "v" means for

[1] The thermal averaging comes entirely from the integral over p_i and p_j, the momenta of the two particles sitting in the heat bath; p_k and p_l, the outgoing momenta, are already summed over and stored inside σ itself.

two particles that are not, in general, moving collinearly), and then we carry out the phase-space integral itself.

Part 1: Relative velocity between two particles

This derivation follows *Landau & Lifshitz, The Classical Theory of Fields*, Section 12, where we obtain a Lorentz-invariant expression for the relative velocity between two particles in special relativity. Let the four-momenta of the two particles be

$$p_1^\mu = (E_1, \mathbf{p}_1) = (\gamma_1 m_1, \gamma_1 m_1 \mathbf{v}_1), \qquad p_2^\mu = (E_2, \mathbf{p}_2) = (\gamma_2 m_2, \gamma_2 m_2 \mathbf{v}_2), \tag{3.4.77}$$

where natural units $c = 1$ are used and $\gamma_i = (1 - v_i^2)^{-1/2}$.

To determine the relative velocity between the two particles, we consider the scalar product of their four-momenta in the rest frame of the second particle. In this frame, the momentum of particle 2 is

$$p_2^\mu = (m_2, \mathbf{0}),$$

while particle 1 moves with velocity $\mathbf{v}_{\rm rel}$ and Lorentz factor $\gamma_{\rm rel} = (1 - v_{\rm rel}^2)^{-1/2}$. Thus,

$$p_1^\mu = (\gamma_{\rm rel} m_1, \gamma_{\rm rel} m_1 \mathbf{v}_{\rm rel}).$$

The Lorentz scalar product between the two momenta is

$$p_1 \cdot p_2 = E_1 E_2 - \mathbf{p}_1 \cdot \mathbf{p}_2 = \gamma_1\gamma_2 m_1 m_2 (1 - \mathbf{v}_1 \cdot \mathbf{v}_2) = E_1 E_2 (1 - \mathbf{v}_1 \cdot \mathbf{v}_2) \tag{3.4.78}$$

where we substituted the definitions $E_i = \gamma_i m_i$ and $\mathbf{p}_i = \gamma_i m_i \mathbf{v}_i$. Upon squaring, we have

$$(p_1 \cdot p_2)^2 = (m_1 m_2 \gamma_1 \gamma_2)^2 (1 - \mathbf{v}_1 \cdot \mathbf{v}_2)^2. \tag{3.4.79}$$

Expanding the square of the scalar product, we find

$$(p_1 \cdot p_2)^2 = (m_1 m_2 \gamma_1 \gamma_2)^2 (1 - 2\mathbf{v}_1 \cdot \mathbf{v}_2 + (\mathbf{v}_1 \cdot \mathbf{v}_2)^2). \tag{3.4.80}$$

Recall the vector identity

$$(\mathbf{a} \times \mathbf{b})^2 = |\mathbf{a}|^2 |\mathbf{b}|^2 - (\mathbf{a} \cdot \mathbf{b})^2,$$

We may write

$$(\mathbf{v}_1 \cdot \mathbf{v}_2)^2 = v_1^2 v_2^2 - (\mathbf{v}_1 \times \mathbf{v}_2)^2.$$

Substituting this, we obtain

$$(p_1 \cdot p_2)^2 = (m_1 m_2 \gamma_1 \gamma_2)^2 \left[1 - 2\mathbf{v}_1 \cdot \mathbf{v}_2 + v_1^2 v_2^2 - (\mathbf{v}_1 \times \mathbf{v}_2)^2\right]. \tag{3.4.81}$$

We also have

$$2\mathbf{v}_1 \cdot \mathbf{v}_2 = v_1^2 + v_2^2 - (\mathbf{v}_1 - \mathbf{v}_2)^2$$

Substituting this form back, we have

$$\begin{aligned}(p_1 \cdot p_2)^2 &= (m_1 m_2 \gamma_1 \gamma_2)^2 \left[(\mathbf{v}_1 - \mathbf{v}_2)^2 + 1 - v_1^2 - v_2^2 + v_1^2 v_2^2 - (\mathbf{v}_1 \times \mathbf{v}_2)^2\right] \\ &= (m_1 m_2 \gamma_1 \gamma_2)^2 \left[(\mathbf{v}_1 - \mathbf{v}_2)^2 + (1 - v_1^2)(1 - v_2^2) - (\mathbf{v}_1 \times \mathbf{v}_2)^2\right]\end{aligned}$$

Using $1 - v_i^2 = 1/\gamma_i^2$, this becomes

$$(p_1 \cdot p_2)^2 = (m_1 m_2 \gamma_1 \gamma_2)^2 \left[(\mathbf{v}_1 - \mathbf{v}_2)^2 + \frac{1}{\gamma_1^2 \gamma_2^2} - (\mathbf{v}_1 \times \mathbf{v}_2)^2\right]. \tag{3.4.82}$$

Subtracting $m_1^2 m_2^2$ from both sides gives

$$(p_1 \cdot p_2)^2 - m_1^2 m_2^2 = (m_1 m_2 \gamma_1 \gamma_2)^2 \left[(\mathbf{v}_1 - \mathbf{v}_2)^2 - (\mathbf{v}_1 \times \mathbf{v}_2)^2\right]. \tag{3.4.83}$$

Diving by (3.4.79), we get our expression for the magnitude of relative velocity $v_{\rm rel}$:

$$v_{\rm rel} = \frac{\sqrt{(p_1 \cdot p_2)^2 - m_1^2 m_2^2}}{p_1 \cdot p_2} = \frac{\sqrt{(\mathbf{v}_1 - \mathbf{v}_2)^2 - (\mathbf{v}_1 \times \mathbf{v}_2)^2}}{1 - \mathbf{v}_1 \cdot \mathbf{v}_2}. \tag{3.4.84}$$

This is the general Lorentz-invariant expression for the relative velocity between two particles moving with arbitrary velocities $\mathbf{v}_1$ and $\mathbf{v}_2$. We can express the above using the mandelstam variable

$$p_1 \cdot p_2 = \frac{s - m_1^2 - m_2^2}{2}$$

Then,

$$\begin{aligned} v_{\rm rel} &= \frac{\sqrt{\left(\frac{s-m_1^2-m_2^2}{2}\right)^2 - m_1^2 m_2^2}}{\frac{s-(m_1^2+m_2^2)}{2}} = \frac{\sqrt{(s - m_1^2 - m_2^2)^2 - 4m_1^2 m_2^2}}{[s - (m_1^2 + m_2^2)]} \\ &= \frac{\sqrt{[s - (m_1 + m_2)^2][s - (m_1 - m_2)^2]}}{[s - (m_1^2 + m_2^2)]} \end{aligned} \tag{3.4.85}$$

For colinear motion, $\mathbf{v}_1 \parallel \mathbf{v}_2$, the cross product vanishes and the expression reduces to the familiar relativistic law of velocity addition,

$$v_{\rm rel} = \frac{|\mathbf{v}_1 - \mathbf{v}_2|}{1 - \mathbf{v}_1 \mathbf{v}_2} = \frac{E_1 E_2}{p_1 \cdot p_2} |\mathbf{v}_1 - \mathbf{v}_2| \tag{3.4.86}$$

Part 2: Reducing $\langle \sigma v \rangle$ to a single integral over s

The thermal average of $\sigma v_{\rm rel}$ following Mirco Cannoni is

$$\langle \sigma |\mathbf{v}_1 - \mathbf{v}_2| \rangle = \frac{1}{(4\pi)^2 T^2 m_1^2 m_2^2 K_2(m_1/T) K_2(m_2/T)} \int d^3\mathbf{p}_1 d^3\mathbf{p}_2 \frac{(p_1 \cdot p_2)}{E_1 E_2} v_{\rm rel}\, \sigma(s)\, e^{-(E_1+E_2)/T}. \tag{3.4.87}$$

This is a complicated integral which can be simplified by transforming it into a single integral over the invariant mass squared $s = (p_1 + p_2)^2$. All calculations are performed in the rest frame of the heat bath, $u^\mu = (1, \mathbf{0})$. The integration measure could be simplified as:

$$d^3\mathbf{p}_i = p_i^2 dp_i\, d\Omega_i, \qquad E_i = \sqrt{p_i^2 + m_i^2}, \qquad i = 1, 2.$$

Let us perform the following transformation $\theta_1 \to \theta_1$ and $\theta_2 \to (\theta_2 - \theta_1) \equiv \theta$ which is equivalent to choosing $\mathbf{p}_1$ along the z-axis and allowing θ be the angle between $\mathbf{p}_1$ and $\mathbf{p}_2$. Then $d\Omega_1 = 4\pi$ and $d\Omega_2 = 2\pi\, d(\cos\theta)$, giving

$$d^3\mathbf{p}_1\, d^3\mathbf{p}_2 = 8\pi^2\, |\mathbf{p_1}|^2 |\mathbf{p_2}|^2 dp_1 dp_2\, d(\cos\theta).$$

Also, using $d\mathbf{p}_i = (E_i/\mathbf{p}_i)\, dE_i$, we have

$$\frac{d^3\mathbf{p}_1}{E_1} \frac{d^3\mathbf{p}_2}{E_2} = 8\pi^2\, |\mathbf{p_1}||\mathbf{p_2}| dE_1\, dE_2\, d(\cos\theta).$$

We recall

$$p_{1\mu} p^{2\mu} = E_1 E_2 - |\mathbf{p_1}||\mathbf{p_2}| \cos\theta, \qquad s = m_1^2 + m_2^2 + 2(E_1 E_2 - |\mathbf{p_1}||\mathbf{p_2}| \cos\theta).$$

Hence,

$$p_{1\mu} p^{2\mu} = \frac{s - (m_1^2 + m_2^2)}{2}.$$

Let's perform the change of variable as suggested by P. Gondolo and G. Gelmini, Nucl. Phys. B 360

$$Y = E_1 + E_2, \qquad Z = E_1 - E_2, \qquad s = (p_1 + p_2)^2.$$

The associated Jacobian for the transformation can be found as:

$$\frac{\partial(Y, Z, s)}{\partial(E_1, E_2, \cos\theta)} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ \dfrac{\partial s}{\partial E_1} & \dfrac{\partial s}{\partial E_2} & \dfrac{\partial s}{\partial \cos\theta} \end{pmatrix}.$$

The remaining derivatives can be computed from $s = m_1^2 + m_2^2 + 2(E_1 E_2 - p_1 p_2 \cos\theta)$:

$$\frac{\partial s}{\partial E_1} = 2\left(E_2 - \cos\theta\,\frac{E_1|\mathbf{p_2}|}{|\mathbf{p_1}|}\right)$$

$$\frac{\partial s}{\partial E_2} = 2\left(E_1 - \cos\theta\,\frac{E_2|\mathbf{p_1}|}{|\mathbf{p_2}|}\right),$$

$$\frac{\partial s}{\partial \cos\theta} = 2|\mathbf{p_1}||\mathbf{p_2}|$$

Now the determinant:

$$\det \frac{\partial(Y, Z, s)}{\partial(E_1, E_2, \cos\theta)} = -4|\mathbf{p_1}||\mathbf{p_2}|$$

Hence,

$$dE_1\, dE_2\, d(\cos\theta) = \frac{1}{4|\mathbf{p_1}||\mathbf{p_2}|}\, dY\, dZ\, ds$$

Plugging this into the momentum-space measure:

$$\frac{d^3\mathbf{p}_1}{E_1}\frac{d^3\mathbf{p}_2}{E_2} = 8\pi^2|\mathbf{p_1}||\mathbf{p_2}| dE_1\, dE_2\, d(\cos\theta) = 2\pi^2\, dY\, dZ\, ds.$$

Inserting the above results into (3.4.87), we obtain

$$\langle\sigma|\mathbf{v}_1 - \mathbf{v}_2|\rangle = \frac{\pi^2}{(4\pi)^2 T^2\, m_1^2 m_2^2 K_2(m_1/T) K_2(m_2/T)} \int dY\, dZ\, ds\, \sigma(s)\,[s - (m_1^2 + m_2^2)] v_{\rm rel}\, e^{-Y/T}. \tag{3.4.88}$$

The variables Y and Z are restricted by kinematics: for fixed s, not every pair (Y, Z) is possible. The restriction arises from the condition $|\cos\theta| \le 1$. From $s = m_1^2 + m_2^2 + 2(E_1 E_2 - p_1 p_2 \cos\theta)$ we isolate $\cos\theta$:

$$\cos\theta = \frac{2E_1 E_2 + m_1^2 + m_2^2 - s}{2p_1 p_2}.$$

Substitute $E_1 = (Y + Z)/2$, $E_2 = (Y - Z)/2$, and $p_i = \sqrt{E_i^2 - m_i^2}$. Then

$$\cos\theta = \frac{Y^2 - Z^2 + 2(m_1^2 + m_2^2) - 2s}{4\sqrt{\left(\frac{Y+Z}{2}\right)^2 - m_1^2}\sqrt{\left(\frac{Y-Z}{2}\right)^2 - m_2^2}}.$$

The condition $|\cos\theta| \le 1$ determines the allowed range of Z for each (Y, s). Squaring and solving for Z gives the two roots $Z_\pm$ satisfying

$$Z_\pm = \frac{Y(m_1^2 - m_2^2) \pm 2\sqrt{s}\, p'(s)\, \sqrt{Y^2 - s}}{s},$$

where

$$p'(s) = \frac{\sqrt{[s-(m_1+m_2)^2][s-(m_1-m_2)^2]}}{2\sqrt{s}} \tag{3.4.89}$$

Hence the integration interval in Z is $Z \in [Z_-, Z_+]$, and its length is

$$Z_+ - Z_- = \frac{4p'(s)}{\sqrt{s}}\sqrt{Y^2-s}.$$

We note that the integrand in (3.4.88) is independent of Z, so the Z-integration is simply:

$$\int_{Z_-}^{Z_+} dZ\, e^{-Y/T} = e^{-Y/T}\,(Z_+ - Z_-) = e^{-Y/T}\,\frac{4p'(s)}{\sqrt{s}}\sqrt{Y^2-s}.$$

Substitute this back into (3.4.88):

$$\langle\sigma v\rangle = \frac{1}{4T\, m_1^2 m_2^2 K_2(m_1/T)K_2(m_2/T)} \int_{(m_1+m_2)^2}^{\infty} ds\,\sigma(s)\,[s-(m_1^2+m_2^2)]v_{\rm rel}\,\frac{p'(s)}{\sqrt{s}}\int_{\sqrt{s}}^{\infty} dY\, e^{-Y/T}\sqrt{Y^2-s}. \tag{3.4.90}$$

Let $y = Y/\sqrt{s}$, so $dY = \sqrt{s}\,dy$ and $Y^2 - s = s(y^2-1)$. Then

$$\int_{\sqrt{s}}^{\infty} dY\, e^{-Y/T}\sqrt{Y^2-s} = s\int_1^{\infty} dy\, e^{-(\sqrt{s}/T)y}\sqrt{y^2-1}.$$

The standard integral representation of the modified Bessel function K_1 is

$$K_1(a) = a\int_1^{\infty} dy\, e^{-ay}\sqrt{y^2-1}, \qquad \mathrm{Re}\, a > 0.$$

Thus

$$\int_{\sqrt{s}}^{\infty} dY\, e^{-Y/T}\sqrt{Y^2-s} = T\sqrt{s}\,K_1(\sqrt{s}/T)$$

Using above with (3.4.85) and (3.4.89), we find

$$\langle\sigma|\mathbf{v}_1-\mathbf{v}_2|\rangle = \frac{1}{8T\, m_1^2 m_2^2 K_2(m_1/T)K_2(m_2/T)} \times \int_{(m_1+m_2)^2}^{\infty} ds\,\frac{[s-(m_1+m_2)^2][s-(m_1-m_2)^2]}{\sqrt{s}}\sigma(s)\,K_1(\sqrt{s}/T). \tag{3.4.91}$$

Equation (3.4.91) is the punchline: given only the masses m_1, m_2 of the annihilating particles and their annihilation cross section $\sigma(s)$ as a function of the centre-of-mass energy squared, $\langle\sigma v\rangle$ reduces to a single, one-dimensional integral over s that can be evaluated numerically (or, for simple enough $\sigma(s)$, in closed form) for any bath temperature T. This is exactly the ingredient the Boltzmann rate equation (3.4.73) needs, and it is what we now plug into the freeze-out calculation.

3.4.8 Frozen interactions and thermal relic freeze-out

Consider a stable massive particle species X and its antiparticle $\bar{X}$ undergoing pair annihilation into light Standard Model bath particles $l, \bar{l}$:

$$X + \bar{X} \longleftrightarrow l + \bar{l}. \tag{3.4.92}$$

Because the light particles $l, \bar{l}$ remain tightly coupled to the thermal plasma at zero chemical potential ($\mu_l = \mu_{\bar{l}} = 0$), their distribution satisfies $f_l f_{\bar{l}} = \mathrm{e}^{-(E_l+E_{\bar{l}})/T} = \mathrm{e}^{-(E_X+E_{\bar{X}})/T} = \bar{f}_X \bar{f}_{\bar{X}}$. For a symmetric relic ($n_X = n_{\bar{X}}$), the rate equation simplifies to:

$$\dot{n}_X + 3Hn_X = -\langle\sigma v\rangle\left(n_X^2 - \bar{n}_X^2\right). \tag{3.4.93}$$

Let us make Eq. (3.4.93) concrete by specializing the generic collision integral of Eq. (2) to the process $X\bar{X} \to l\bar{l}$ of Eq. (3.4.92), labelling $1 \equiv X,\ 2 \equiv \bar{X},\ 3 \equiv l,\ 4 \equiv \bar{l}$. For the symmetric relic considered above ($n_X = n_{\bar{X}}$) the chemical potentials satisfy $\mu_X = \mu_{\bar{X}} \equiv \mu$, and since the light bath particles have $\mu_l = \mu_{\bar{l}} = 0$, chemical equilibrium of the annihilation reaction forces $\mu = 0$ as well.

$$\begin{aligned}\tilde{C} &= \prod_{i=1}^{4} \int \frac{g_i d^3p_i}{(2\pi)^3 2E_i} \left|\tilde{\mathcal{M}}_{12\to 34}\right|^2 (2\pi)^4 \delta^{(4)}(p_1 + p_2 - p_3 - p_4)\, \frac{1}{S}\, (f_3 f_4 - f_1 f_2) \\ &= \int \frac{g_1 d^3p_1}{(2\pi)^3} \frac{g_2 d^3p_2}{(2\pi)^3} \underbrace{\left[\frac{1}{4E_1E_2} \int \frac{d^3p_3}{(2\pi)^3 2E_3} \frac{d^3p_4}{(2\pi)^3 2E_4} (2\pi)^4 \delta^{(4)}(p_1 + p_2 - p_3 - p_4) \left|\tilde{\mathcal{M}}\right|^2\right]}_{\equiv\ \sigma v_{\rm rel}} (f_3 f_4 - f_1 f_2) \\ &= \int \frac{g_1 d^3p_1}{(2\pi)^3} \frac{g_2 d^3p_2}{(2\pi)^3}\, (f_3 f_4 - f_1 f_2)\, \sigma v_{\rm rel},\end{aligned}$$

where $S = \left(\frac{1}{2} \times 2\right)^{-1} = 1$ is the identical-particle symmetry factor for this process (already introduced as S in Eq. (2)), and the underbraced quantity is exactly the standard relation between the invariant amplitude, the flux factor $4E_1E_2$ of Eq. (5), and the (un-averaged) cross section times relative velocity.
Now set $f_3 f_4 = f_l f_{\bar{l}} = (f_X f_{\bar{X}})_{\rm eq} \equiv \bar{f}_X \bar{f}_{\bar{X}}$ – using chemical equilibrium of the light bath particles – and $f_1 f_2 = f_X f_{\bar{X}}$, and integrate $\sigma v_{\rm rel}$ over the incoming momenta to form the thermally averaged cross section:

$$\begin{aligned}\dot{n}_X + 3Hn_X = \frac{1}{a^3}\frac{\mathrm{d}(n_X a^3)}{\mathrm{d}t} &= \int \frac{g_1 d^3p_1}{(2\pi)^3} \frac{g_2 d^3p_2}{(2\pi)^3} \left(\bar{f}_X \bar{f}_{\bar{X}} - f_X f_{\bar{X}}\right) \sigma v_{\rm rel} \\ &= \bar{n}_X \bar{n}_{\bar{X}} \langle\sigma v\rangle - n_X n_{\bar{X}} \langle\sigma v\rangle \\ &= \langle\sigma v\rangle \left(\bar{n}_X^2 - n_X^2\right),\end{aligned}$$

where the definition of the thermally averaged cross section, Eq. (3.4.74), was used to turn the two momentum integrals into equilibrium and non-equilibrium number densities, and $n_X = n_{\bar{X}}$, $\bar{n}_X = \bar{n}_{\bar{X}}$ (symmetric relic) in the last step. This reproduces exactly Eq. (3.4.93) above.

To eliminate the Hubble dilution term, we express the abundance in terms of the **comoving yield** (number of particles per comoving entropy volume):[2]

$$Y_X \equiv \frac{n_X}{s}, \qquad \bar{Y}_X \equiv \frac{\bar{n}_X}{s}. \tag{3.4.94}$$

Using comoving entropy conservation $\frac{\mathrm{d}(sa^3)}{\mathrm{d}t} = 0 \implies \dot{n}_X + 3Hn_X = s\dot{Y}_X$, we get:

$$\dot{Y}_X = \frac{\dot{n}_X}{s} - \frac{n_X \dot{s}}{s^2} = \frac{1}{s}\left[\dot{n}_X - \frac{n_X}{s}(-3Hs)\right] = -\langle\sigma v\rangle s \left(Y_X^2 - \bar{Y}_X^2\right). \tag{3.4.95}$$

where in the last step we used (3.4.93) and (3.4.94). Introducing the dimensionless inverse temperature parameter:

$$x \equiv \frac{m_X}{T}, \qquad \frac{\mathrm{d}x}{\mathrm{d}t} = m_X \frac{\mathrm{d}(1/T)}{\mathrm{d}t} = -\frac{m_X}{T}\frac{\dot{T}}{T} = Hx, \tag{3.4.96}$$

[2] Since $n_X \propto \frac{1}{a^3}$ and $s \propto \frac{1}{a^3}$, we find that Y does not get diluted.

the evolution equation becomes:

$$\frac{\mathrm{d}Y_X}{\mathrm{d}x} = -\frac{\langle\sigma v\rangle s}{Hx}\left(Y_X^2 - \bar{Y}_X^2\right). \tag{3.4.97}$$

During the radiation-dominated era, the expansion rate and entropy density are:

$$H(T) = \sqrt{\frac{4\pi^3 g_*}{45}}\frac{T^2}{M_{\rm Pl}} = \sqrt{\frac{4\pi^3 g_*}{45}}\frac{m_X^2}{M_{\rm Pl}x^2}, \qquad s(T) = \frac{2\pi^2}{45}q_*(T)\frac{m_X^3}{x^3}, \tag{3.4.98}$$

where $q_* \equiv g_{*S}$ denotes the entropy relativistic degrees of freedom. Defining the dimensionless coupling parameter:

$$\lambda \equiv \langle\sigma v\rangle\sqrt{\frac{\pi}{45}}\frac{q_*}{\sqrt{g_*}}m_X M_{\rm Pl}, \tag{3.4.99}$$

and the deviation from equilibrium $\Delta \equiv Y_X - \bar{Y}_X$, Eq. (3.4.97) takes the standard Lee–Weinberg form:

$$\frac{\mathrm{d}\Delta}{\mathrm{d}x} = -\frac{\mathrm{d}\bar{Y}_X}{\mathrm{d}x} - \lambda x^{-2}\Delta\left(\Delta + 2\bar{Y}_X\right). \tag{3.4.100}$$

3.4.9 Asymptotic solutions: Cold and hot relics

This is interaction based mechanism for producing dark matter. Observeration does not tell us which mechanism was used for the production, only the final abundance Y.

Cold thermal relics ($x_f \gg 1$)

A **cold relic** freezes out when it is non-relativistic ($T_f \ll m_X$, i.e., $x_f \gtrsim 20$). The non-relativistic equilibrium yield is:

$$\bar{Y}_X(x) = \frac{\bar{n}_X}{s} = \frac{g_X\left(\frac{m_X T}{2\pi}\right)^{3/2}\mathrm{e}^{-m_X/T}}{\frac{2\pi^2}{45}q_* T^3} = \frac{45}{2\pi^4}\sqrt{\frac{\pi}{8}}\frac{g_X}{q_*}x^{3/2}\mathrm{e}^{-x}. \tag{3.4.101}$$

The thermally averaged cross section is parameterized by a velocity expansion $\langle\sigma v\rangle = \sigma_0 f(x) = \sigma_0 x^{-n}$, where $n = 0$ corresponds to s-wave annihilation and $n = 1$ to p-wave annihilation. Defining:

$$\lambda_0 \equiv \sigma_0\sqrt{\frac{\pi}{45}}\frac{q_*}{\sqrt{g_*}}m_X M_{\rm Pl}, \tag{3.4.102}$$

we analyze the two asymptotic regimes of Eq. (3.4.100):

1. **Pre-decoupling regime ($1 < x \ll x_f$):** The abundance tracks equilibrium ($Y_X \approx \bar{Y}_X$, $\Delta \ll \bar{Y}_X$). Setting $\frac{\mathrm{d}\Delta}{\mathrm{d}x} \approx 0$ yields:

$$\frac{\mathrm{d}\Delta}{\mathrm{d}x} = -\frac{\mathrm{d}\bar{Y}}{\mathrm{d}x} - \lambda\underbrace{\frac{\Delta^2}{x^2}}_{\approx 0} - \frac{2\lambda\Delta\bar{Y}}{x^2} \approx 0 \tag{3.4.103}$$

   using (3.4.99) and (3.4.102), we can write

$$\Delta \simeq -\frac{x^2}{2\lambda_0 f(x)}\frac{\mathrm{d}\ln\bar{Y}_X}{\mathrm{d}x} \simeq \frac{x^2}{2\lambda_0 f(x)}, \tag{3.4.104}$$

   since $\frac{\mathrm{d}\ln\bar{Y}_X}{\mathrm{d}x} = -1 + \frac{3}{2x} \approx -1$ for $x \gg 1$.

2. **Post-decoupling regime ($x \gg x_f$):** The equilibrium yield drops exponentially ($\bar{Y}_X \to 0$), leaving $Y_X \approx \Delta$. The Boltzmann equation reduces to:

$$\frac{\mathrm{d}\Delta}{\mathrm{d}x} \simeq -\lambda_0 \frac{f(x)}{x^2}\Delta^2 \implies \Delta_\infty^{-1} - \Delta_f^{-1} = \lambda_0 \int_{x_f}^{\infty} \frac{f(x)}{x^2}\mathrm{d}x. \tag{3.4.105}$$

Since $\Delta_f^{-1} \ll \Delta_\infty^{-1}$, the asymptotic relic yield is:

$$Y_\infty \simeq \left[\lambda_0 \int_{x_f}^{\infty} \frac{f(x)}{x^2}\mathrm{d}x\right]^{-1} = \frac{(n+1)x_f^{n+1}}{\lambda_0}. \tag{3.4.106}$$

Freeze-out temperature determination: The freeze-out point x_f is defined where $\Delta(x_f) = c\bar{Y}_X(x_f)$ with $c \sim \mathcal{O}(1)$ (numerical calibration gives $c(c+2) \approx 1$). Equating Eq. (3.4.104) and Eq. (3.4.101) yields the transcendental equation:

$$x_f^{1/2}\mathrm{e}^{x_f} f(x_f) = \frac{45}{\pi^4}\sqrt{\frac{\pi}{8}}\frac{g_X}{q_*}c\lambda_0. \tag{3.4.107}$$

Equivalently, using the intuitive condition $\Gamma(x_f) = H(x_f)$ gives:

$$\frac{g_X}{\sqrt{g_*}}x_f^{1/2}\mathrm{e}^{-x_f} f(x_f)\sigma_0 m_X M_{\mathrm{Pl}} = 4\pi^3\sqrt{\frac{2}{45}}. \tag{3.4.108}$$

Iterative solution gives $x_f \approx \ln(m_X M_{\mathrm{Pl}}\sigma_0) \approx 20$–$30$ over a vast range of particle masses ($10\,\mathrm{GeV} \le m_X \le 10\,\mathrm{TeV}$).

Present relic density and the WIMP miracle: The present relic number density $n_{X0} = s_0 Y_\infty$ is:

$$n_{X0} = 1.09 \times 10^4 \frac{\sqrt{g_*}}{q_*}\left(\frac{q_{*0}}{3.91}\right)\Theta_{2.7}^3\left[\int_{x_f}^{\infty}\frac{f(x)}{x^2}\mathrm{d}x\right]^{-1}\frac{1}{m_X M_{\mathrm{Pl}}\sigma_0}\,\mathrm{cm}^{-3}. \tag{3.4.109}$$

The cosmological density parameter $\Omega_{X0} \equiv \rho_{X0}/\rho_{\mathrm{crit},0} = m_X n_{X0}/\rho_{\mathrm{crit},0}$ evaluates to:

$$\Omega_{X0}h^2 \simeq 0.31\left(\frac{g_*(x_f)}{100}\right)^{-1/2}(n+1)x_f^{n+1}\left(\frac{q_{*0}}{3.91}\right)\Theta_{2.7}^3\left(\frac{\sigma_0}{10^{-38}\,\mathrm{cm}^2}\right)^{-1}. \tag{3.4.110}$$

Remarkably, a typical electroweak-scale cross section $\sigma_0 \approx 10^{-38}\,\mathrm{cm}^2 \approx 1\,\mathrm{pb} \sim \alpha_W^2/m_{\mathrm{weak}}^2$ naturally yields $\Omega_{X0}h^2 \approx 0.12$, matching the observed cold dark matter density. This striking coincidence is known as the **WIMP miracle**.

Numerical implementation of the freeze-out Boltzmann equation. To solve the non-linear Riccati rate equation (3.4.97) numerically, standard explicit integration algorithms fail due to extreme numerical stiffness: before freeze-out ($x < x_f$), the equilibrium curve $\bar{Y}(x)$ falls exponentially, while the coupling λ/x^2 is enormous ($\sim 10^{14}$–10^{16}), forcing the solution to track $\bar{Y}(x)$ within machine precision. Implicit stiff integrators (such as the Radau IIA algorithm in SciPy) are required to transition smoothly across the stiff equilibrium regime into the frozen asymptotic plateau.

The following Python code solves the Boltzmann freeze-out equation for three representative annihilation cross sections $\sigma_1 > \sigma_0 > \sigma_2$, reproducing the trajectories shown in Figure 3.9:

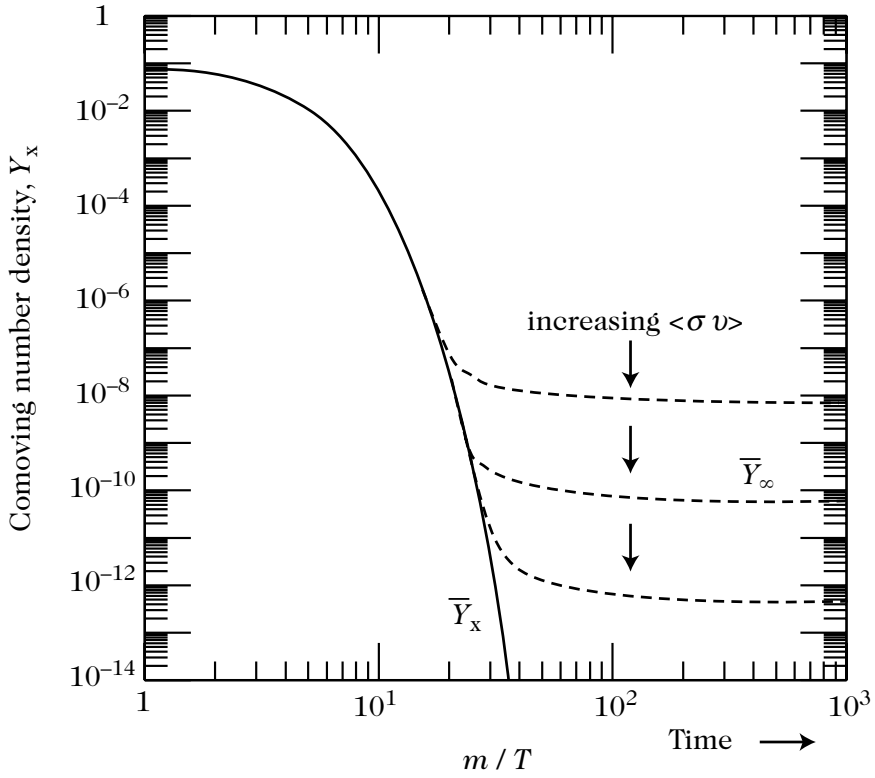


Figure 3.9: Numerical integration of the Boltzmann equation (3.4.100) for comoving abundance $Y_X(x)$ as a function of $x = m/T$. The solid curves display freeze-out departures from the equilibrium curve $\bar{Y}_X(x)$ for increasing annihilation cross sections $\langle\sigma v\rangle$, showing that larger cross sections maintain equilibrium longer and result in smaller residual relic abundances $Y_\infty \propto 1/\langle\sigma v\rangle$.

```python
import numpy as np
from scipy.integrate import solve_ivp

# Thermal parameters of the relativistic early Universe
gs0 = 100.0    # Effective entropy relativistic degrees of freedom g_{*S}
g0  = 2.0      # Internal spin degeneracy of dark matter particle
m0  = 1000.0   # Mass of relic particle (in GeV)

# Cross-section couplings c, c1, c2 (with sigma = c^2 / m0^2)
c, c1, c2 = 0.1, 0.5, 0.05
sigma0 = c**2  / m0**2
sigma1 = c1**2 / m0**2
sigma2 = c2**2 / m0**2

# Dimensionless coupling parameters lambda = 1.32 * sqrt(g_{*S}) * m_0 * M_Pl * sigma
# (M_Pl = 1.22e19 GeV, with 2.44e18 GeV representing the effective expansion scale)
lam0 = 1.32 * np.sqrt(gs0) * m0 * 2.44e18 * sigma0
lam1 = 1.32 * np.sqrt(gs0) * m0 * 2.44e18 * sigma1
lam2 = 1.32 * np.sqrt(gs0) * m0 * 2.44e18 * sigma2

# Non-relativistic equilibrium yield Y_eq(z) = 0.145 * (g/g_{*S}) * z^(3/2) * exp(-z)
def yeq(z):
    return 0.145 * (g0 / gs0) * (z**1.5) * np.exp(-z)
```

```python

def solve_boltzmann_freezeout(lam):
    # Riccati differential rate equation: dY/dz = - (lam / z^2) * [Y(z)^2 - Y_eq(z)^2]
    def ode(z, y):
        return [- (lam / (z**2)) * (y[0]**2 - yeq(z)**2)]

    # Solve across z in [0.01, 500.0] using the stiff Radau IIA integrator
    z_span = (0.01, 500.0)
    y_init = [yeq(0.01)]
    sol = solve_ivp(ode, z_span, y_init, method='Radau', rtol=1e-5, atol=1e-18, max_step=0.1)
    return sol

# Execute solver for base, enhanced, and suppressed cross sections
sol0 = solve_boltzmann_freezeout(lam0)  # Intermediate: c = 0.1
sol1 = solve_boltzmann_freezeout(lam1)  # Enhanced:     c = 0.5  -> lowest relic yield
sol2 = solve_boltzmann_freezeout(lam2)  # Suppressed:   c = 0.05 -> highest relic yield
```

Listing 3.1: Numerical integration of the Lee–Weinberg Boltzmann freeze-out Riccati equation for varying annihilation cross sections.

Executing this solver yields asymptotic abundances $Y_\infty^{(1)} \approx 3.8\times10^{-15} < Y_\infty^{(0)} \approx 8.4\times10^{-14} < Y_\infty^{(2)} \approx 3.2\times10^{-13}$, directly verifying the inverse scaling $Y_\infty \propto 1/\lambda \propto 1/\sigma_0$ illustrated in Figure 3.9.

Hot thermal relics ($x_f \ll 1$)

A **hot relic** decouples while still highly relativistic ($T_f \gg m_X$, $x_f \ll 1$). Its equilibrium comoving yield is constant:

$$\bar{Y}_X = \frac{45\zeta(3)}{2\pi^4}\varepsilon_{\rm FB}\frac{g_X}{q_*}, \tag{3.4.111}$$

where $\varepsilon_{\rm FB} = 1$ for bosons and $3/4$ for fermions. The asymptotic relic yield simply equals the equilibrium yield at decoupling:

$$Y_\infty = \bar{Y}_X(x_f) = \frac{45\zeta(3)}{2\pi^4}\varepsilon_{\rm FB}\frac{g_X}{q_*(x_f)}. \tag{3.4.112}$$

Assuming subsequent expansion remains adiabatic, today's number density is:

$$n_{X0} = s_0 Y_\infty = 776.8\left(\frac{q_{*0}}{3.91}\right)\frac{\varepsilon_{\rm FB}g_X}{q_*(x_f)}\Theta_{2.7}^3\,{\rm cm}^{-3}. \tag{3.4.113}$$

If the particle is non-relativistic today ($m_X > T_0 \approx 1.7\times10^{-4}\,{\rm eV}$), its energy density is $\rho_{X0} = m_X n_{X0} = m_X s_0 Y_\infty$, yielding:

$$\Omega_{X0}h^2 = 0.739\left(\frac{q_{*0}}{3.91}\right)\left(\frac{m_X}{10\,{\rm eV}}\right)\frac{\varepsilon_{\rm FB}g_X}{q_*(x_f)}. \tag{3.4.114}$$

The requirement $\Omega_{X0}h^2 \le 1$ establishes the Lee–Weinberg / Cowsik–McClelland mass upper bound:

$$m_X \le 13.5\,\frac{q_*(x_f)}{\varepsilon_{\rm FB}g_X}\,{\rm eV}. \tag{3.4.115}$$

Light neutrinos and warm relics: For standard left-handed neutrinos decoupling at $T_{\nu,D} \approx 1\,\mathrm{MeV}$ ($g_X = 2$, $\varepsilon_{\mathrm{FB}} = 3/4$, $q_*(T_{\nu,D}) = 10.75$), Eq. (3.4.114) yields the celebrated **Gershtein–Zeldovich / Cowsik–McClelland formula**:

$$\Omega_{\nu 0} h^2 = \frac{\sum m_\nu}{91.5\,\mathrm{eV}}. \tag{3.4.116}$$

For particles that decouple much earlier (e.g., at the electroweak scale $T \sim 300\,\mathrm{GeV}$ with $q_* \approx 106.75$), the relic density is suppressed by entropy injection from subsequent particle annihilations:

$$\Omega_{X0} h^2 = \frac{m_X}{910\,\mathrm{eV}}. \tag{3.4.117}$$

Such species have lower present temperatures and abundances relative to CMB photons and are designated as **warm relics**.

3.4.10 Precision neutrino decoupling and the Standard Model benchmark $N_{\mathrm{eff}} = 3.044$

In precision modern cosmology, the radiation energy density of the Universe after the completion of e^+e^- annihilation is defined universally by:

$$\rho_{\mathrm{rad}} = \left[1 + \frac{7}{8}\left(\frac{4}{11}\right)^{4/3} N_{\mathrm{eff}}\right] \rho_\gamma. \tag{3.4.118}$$

Under the idealized assumption that neutrinos decouple instantaneously as a step function at $T \approx 1\,\mathrm{MeV}$, the entropy of e^+e^- pairs is dumped exclusively into photons, yielding $N_{\mathrm{eff}} = 3.000$ exactly (as derived in Section 3.4.6).

However, precision cosmological probes (Planck, CMB-S4, LiteBIRD) measure the relativistic expansion rate at the sub-percent level, where the instantaneous decoupling approximation fails. Three distinct physical effects modify the relic neutrino energy density:

1. **Non-instantaneous decoupling:** Weak interaction cross sections scale quadratically with energy ($\sigma \propto G_{\mathrm{F}}^2 E^2$). High-energy neutrinos in the tail of the Fermi–Dirac distribution remain in thermal contact with electrons and positrons down to $T \sim 0.5\,\mathrm{MeV}$, absorbing a small fraction of the entropy released during e^+e^- annihilation.

2. **Finite-temperature QED corrections:** Coulomb interactions and thermal photon/-electron self-energies modify the equation of state of the electromagnetic plasma, reducing the effective degrees of freedom of the photon bath.

3. **Neutrino flavor oscillations:** As non-thermal spectral distortions are generated primarily in the electron neutrino sector via charged-current interactions ($\nu_e e^\pm \leftrightarrow \nu_e e^\pm$), quantum neutrino oscillations redistribute energy between ν_e and the muon and tau neutrino sectors (ν_μ, ν_τ).

Connection to the standard cosmological Boltzmann equation. Now that we have established the relativistic Boltzmann equation (3.4.62) for thermal relics, we can see clearly how neutrino decoupling requires an essential quantum generalization. For a classical, single-particle species i with phase-space distribution function $f_i(p,t)$ in a spatially homogeneous and isotropic FLRW universe, our standard Boltzmann transport equation reads:

$$\hat{\mathcal{L}}[f_i] = \mathcal{C}_i[f] \quad \Longleftrightarrow \quad \left(\frac{\partial}{\partial t} - Hp\frac{\partial}{\partial p}\right) f_i(p,t) = \frac{1}{E_i}\mathcal{C}_i[f], \tag{3.4.119}$$

where the left-hand side operator $\hat{\mathcal{L}}_{\mathrm{FLRW}} \equiv \partial_t - Hp\partial_p$ is the comoving Liouville streaming operator reflecting the continuous cosmological redshifting of physical momentum ($\dot{p} = -Hp$).

Integrating Eq. (3.4.119) over momentum space $\frac{g_i}{(2\pi)^3}\int \frac{\mathrm{d}^3p}{E_i}$ yields the standard macroscopic number density rate equation:

$$\dot{n}_i + 3Hn_i = C_i. \tag{3.4.120}$$

If neutrinos were classical, distinguishable, and non-oscillating particles, their decoupling from the electromagnetic plasma would simply be described by three uncoupled copies of Eq. (3.4.119), one for each flavor:

$$\left(\frac{\partial}{\partial t} - Hp\frac{\partial}{\partial p}\right) f_{\nu_\alpha}(p,t) = \frac{1}{E}\mathcal{C}_\alpha[f], \qquad \alpha \in \{e,\mu,\tau\}.$$

However, quantum mechanics introduces a fundamental complication: active neutrino flavor eigenstates $(\nu_e, \nu_\mu, \nu_\tau)$ do not coincide with the mass eigenstates (ν_1, ν_2, ν_3). In the hot early Universe, neutrinos are produced and interact in flavor eigenstates, but propagate and oscillate in mass eigenstates. Because the oscillation length is much shorter than the cosmological horizon and collision mean free path, the neutrino state during decoupling is not an incoherent statistical mixture of flavors, but a **coherent quantum superposition**.

The quantum density matrix generalization. To treat coherent quantum flavor oscillations and incoherent electroweak scatterings simultaneously within an expanding spacetime, the classical scalar distribution functions $f_{\nu_\alpha}(p,t)$ must be generalized to a 3×3 **Hermitian density matrix in flavor space**, $\varrho(p,t)$ (and $\bar{\varrho}(p,t)$ for antineutrinos):

$$\varrho(p,t) = \begin{pmatrix} \varrho_{ee}(p,t) & \varrho_{e\mu}(p,t) & \varrho_{e\tau}(p,t) \\ \varrho_{\mu e}(p,t) & \varrho_{\mu\mu}(p,t) & \varrho_{\mu\tau}(p,t) \\ \varrho_{\tau e}(p,t) & \varrho_{\tau\mu}(p,t) & \varrho_{\tau\tau}(p,t) \end{pmatrix}. \tag{3.4.121}$$

The physical interpretation of the matrix elements is direct:

- The **diagonal elements** $\varrho_{\alpha\alpha}(p,t) \equiv f_{\nu_\alpha}(p,t)$ are the standard single-particle occupation numbers for each flavor $\alpha \in \{e,\mu,\tau\}$, matching the classical distribution functions of our standard Boltzmann equation.
- The **off-diagonal elements** $\varrho_{\alpha\beta}(p,t)$ $(\alpha \neq \beta)$ encode the quantum phase coherences between different flavor states that generate neutrino oscillations.

The non-equilibrium evolution of this quantum ensemble is governed by the **quantum Boltzmann transport equation**:

$$\underbrace{\left(\frac{\partial}{\partial t} - Hp\frac{\partial}{\partial p}\right)\varrho(p,t)}_{\text{FLRW Liouville streaming}} = \underbrace{-i\,[\mathcal{H}(p,t),\,\varrho(p,t)]}_{\text{von Neumann flavor precession}} + \underbrace{\mathcal{C}[\varrho,\bar{\varrho}]}_{\text{Inelastic collision integral}}. \tag{3.4.122}$$

Comparing Eq. (3.4.122) with our standard Boltzmann equation (3.4.119) clarifies the physical role of every term:

1. **Liouville streaming (LHS):** The kinematic operator $(\partial_t - Hp\partial_p)$ is identical to that of our standard FLRW Boltzmann equation, ensuring that the momentum of every coherent mode redshifts as $p \propto a^{-1}$ with cosmic expansion.
2. **Quantum flavor precession (Commutator):** The commutator term $-i[\mathcal{H},\varrho]$ is the von Neumann operator generated by the effective in-medium Hamiltonian $\mathcal{H}(p,t)$ (incorporating vacuum mass mixing and forward-scattering thermal matter potentials). This term drives unitary flavor oscillations and quantum coherence transfer.
3. **Electroweak collisions (RHS):** The non-Abelian collision operator $\mathcal{C}[\varrho,\bar{\varrho}]$ describes real, inelastic $2 \leftrightarrow 2$ scatterings and pair annihilations with $e^\pm$ and other neutrinos, driving the system toward kinetic and chemical equilibrium.

Recovery of the standard Boltzmann equation. In the classical limit where neutrino flavor mixing is absent or neglected ($\mathcal{H} \to 0$ and vanishing off-diagonal coherences $\varrho_{\alpha\beta} = 0$), the commutator vanishes identically:

$$[\mathcal{H}, \varrho] = 0.$$

Equation (3.4.122) then uncouples along the diagonal into three independent equations:

$$\left(\frac{\partial}{\partial t} - Hp\frac{\partial}{\partial p}\right) f_{\nu_\alpha}(p,t) = \mathcal{C}_\alpha[f], \qquad \alpha \in \{e,\mu,\tau\}, \tag{3.4.123}$$

which **reduces identically to the standard scalar Boltzmann transport equation** of our textbook! Integrating over momentum space recovers the standard rate equation $\dot{n}_{\nu_\alpha} + 3Hn_{\nu_\alpha} = C_{\nu_\alpha}$ of Eq. (3.4.62). Thus, Eq. (3.4.122) is the rigorous, first-principles quantum kinetic generalization of our standard cosmological Boltzmann equation to oscillating multi-flavor neutrino systems.

Electroweak Matrix Elements, Fierz Rearrangement, and Dirac Traces for Neutrino Decoupling

At energy scales characteristic of neutrino decoupling ($E \sim$ few MeV), the four-momentum transfer satisfies $q^2 \ll M_W^2, M_Z^2$. The massive $W^\pm$ and Z^0 gauge bosons can be integrated out, generating the effective four-fermion Fermi Lagrangian:

$$\mathcal{L}_{\rm int} = -\frac{G_{\rm F}}{\sqrt{2}}\left[J_W^\mu J_{W\mu}^\dagger + J_Z^\mu J_{Z\mu}\right], \tag{1}$$

where $G_{\rm F} \approx 1.1663787 \times 10^{-5}\,\text{GeV}^{-2}$. Because $T \ll m_\mu, m_\tau$, muons and taus are completely decoupled from the thermal bath ($n_\mu \approx n_\tau \approx 0$), leaving only electrons in the charged weak current:

$$J_W^\mu = \bar{\nu}_e\gamma^\mu(1-\gamma^5)e, \qquad J_Z^\mu = \sum_{\alpha=e,\mu,\tau}\bar{\nu}_\alpha\gamma^\mu\frac{1}{2}(1-\gamma^5)\nu_\alpha + \bar{e}\gamma^\mu(g_V - g_A\gamma^5)e, \tag{2}$$

where the standard Glashow–Weinberg–Salam vector and axial-vector couplings are $g_V = -\frac{1}{2} + 2\sin^2\theta_W$ and $g_A = -\frac{1}{2}$, with $\sin^2\theta_W \approx 0.23129$.

Fierz rearrangement. The charged-current interaction Lagrangian mediates ν_e–$e^\pm$ transitions:

$$\mathcal{L}_{\rm CC} = -\frac{G_{\rm F}}{\sqrt{2}}\left[\bar{e}\gamma^\mu(1-\gamma^5)\nu_e\right]\left[\bar{\nu}_e\gamma_\mu(1-\gamma^5)e\right]. \tag{3}$$

To combine this with the neutral current, we perform a **Fierz rearrangement**. For anticommuting Dirac spinors, the chiral $(V-A)\times(V-A)$ current product is invariant under the interchange of the outgoing fermion fields:

$$\left[\bar{e}\gamma^\mu(1-\gamma^5)\nu_e\right]\left[\bar{\nu}_e\gamma_\mu(1-\gamma^5)e\right] = \left[\bar{\nu}_e\gamma^\mu(1-\gamma^5)\nu_e\right]\left[\bar{e}\gamma_\mu(1-\gamma^5)e\right]. \tag{4}$$

Using the chiral projection operators $P_L \equiv \frac{1}{2}(1-\gamma^5)$ and $P_R \equiv \frac{1}{2}(1+\gamma^5)$, $(1-\gamma^5) = 2P_L$, so $\mathcal{L}_{\rm CC} = -2\sqrt{2}G_{\rm F}[\bar{\nu}_e\gamma^\mu P_L\nu_e][\bar{e}\gamma_\mu P_L e]$. Decomposing the neutral current into chiral components:

$$\bar{e}\gamma_\mu(g_V - g_A\gamma^5)e = \bar{e}\gamma_\mu\left[(g_V-g_A)P_L + (g_V+g_A)P_R\right]e. \tag{5}$$

Summing the Fierz-transformed charged current and neutral current yields the unified effective Lagrangian:

$$\mathcal{L}_{\rm int}^{\nu e} = -2\sqrt{2}G_{\rm F}\sum_{\alpha=e,\mu,\tau}\left[\bar{\nu}_\alpha\gamma^\mu P_L\nu_\alpha\right]\left[\bar{e}\gamma_\mu\left(g_L^\alpha P_L + g_R P_R\right)e\right], \tag{6}$$

with the effective chiral couplings:

$$g_L^e = 1 + g_V + g_A = \frac{1}{2} + \sin^2\theta_W \approx 0.73129 \quad (\text{CC} + \text{NC}), \tag{7}$$

$$g_L^{\mu,\tau} = g_V + g_A = -\frac{1}{2} + \sin^2\theta_W \approx -0.26871 \quad (\text{NC only}), \tag{8}$$

$$g_R = g_V - g_A = \sin^2\theta_W \approx 0.23129 \quad (\text{universal NC}). \tag{9}$$

Dirac trace technology and invariant matrix elements. For neutrino-electron scattering $\nu_\alpha(p_1) + e^-(p_2) \to \nu_\alpha(p_3) + e^-(p_4)$, the transition amplitude is

$$\mathcal{M} = -2\sqrt{2}G_{\rm F}\left[\bar{u}(p_3)\gamma^\mu P_L u(p_1)\right]\left[\bar{u}(p_4)\gamma_\mu\left(g_L^\alpha P_L + g_R P_R\right)u(p_2)\right]. \tag{10}$$

Summing over final spins and initial electron spins (neutrinos are purely left-handed) yields

$$\sum_{\text{spins}} |\mathcal{M}|^2 = 8G_{\rm F}^2\, L_\nu^{\mu\nu}\, L_{e,\mu\nu}, \tag{11}$$

where the neutrino and electron leptonic tensors evaluate via Dirac traces:

$$L_\nu^{\mu\nu} = \mathrm{Tr}\left[\gamma^\mu P_L \not{p}_1 \gamma^\nu P_L \not{p}_3\right] = 2\left[p_1^\mu p_3^\nu + p_1^\nu p_3^\mu - g^{\mu\nu}(p_1\cdot p_3) - \mathrm{i}\epsilon^{\mu\alpha\nu\beta}p_{1\alpha}p_{3\beta}\right], \tag{12}$$

$$\begin{aligned} L_{e,\mu\nu} &= \mathrm{Tr}\left[(\not{p}_4 + m_e)\gamma_\mu(g_L^\alpha P_L + g_R P_R)(\not{p}_2 + m_e)\gamma_\nu(g_L^\alpha P_L + g_R P_R)\right] \\ &= (g_L^\alpha)^2\,\mathrm{Tr}\left[\not{p}_4\gamma_\mu P_L \not{p}_2 \gamma_\nu P_L\right] + g_R^2\,\mathrm{Tr}\left[\not{p}_4\gamma_\mu P_R \not{p}_2\gamma_\nu P_R\right] + 2g_L^\alpha g_R m_e^2 g_{\mu\nu}. \end{aligned} \tag{13}$$

Contracting symmetric tensor products and using $\epsilon^{\mu\alpha\nu\beta}\epsilon_{\mu\rho\nu\sigma} = -2(\delta^\alpha_\rho\delta^\beta_\sigma - \delta^\alpha_\sigma\delta^\beta_\rho)$ gives

$$\boxed{\sum_{\text{spins}} |\mathcal{M}|^2_{\nu_\alpha e^-\to\nu_\alpha e^-} = 128G_{\rm F}^2\left[(g_L^\alpha)^2(p_1\cdot p_2)(p_3\cdot p_4) + g_R^2(p_1\cdot p_4)(p_2\cdot p_3) - g_L^\alpha g_R m_e^2(p_1\cdot p_3)\right].} \tag{14}$$

Applying crossing symmetry yields the corresponding squared amplitudes for the remaining channels:

$$\sum_{\text{spins}} |\mathcal{M}|^2_{\nu_\alpha e^+\to\nu_\alpha e^+} = 128G_{\rm F}^2\left[g_R^2(p_1\cdot p_2)(p_3\cdot p_4) + (g_L^\alpha)^2(p_1\cdot p_4)(p_2\cdot p_3) - g_L^\alpha g_R m_e^2(p_1\cdot p_3)\right], \tag{15}$$

$$\sum_{\text{spins}} |\mathcal{M}|^2_{e^+e^-\leftrightarrow\nu_\alpha\bar{\nu}_\alpha} = 128G_{\rm F}^2\left[(g_L^\alpha)^2(p_1\cdot p_4)(p_2\cdot p_3) + g_R^2(p_1\cdot p_3)(p_2\cdot p_4) + g_L^\alpha g_R m_e^2(p_3\cdot p_4)\right]. \tag{16}$$

For neutrino-neutrino neutral-current scattering:

$$\sum_{\text{spins}} |\mathcal{M}|^2_{\nu_\alpha\nu_\alpha\to\nu_\alpha\nu_\alpha} = 128G_{\rm F}^2(p_1\cdot p_2)(p_3\cdot p_4), \tag{17}$$

$$\sum_{\text{spins}} |\mathcal{M}|^2_{\nu_\alpha\nu_\beta\to\nu_\alpha\nu_\beta} = 32G_{\rm F}^2(p_1\cdot p_2)(p_3\cdot p_4) \quad (\alpha\neq\beta), \tag{18}$$

$$\sum_{\text{spins}} |\mathcal{M}|^2_{\nu_\alpha\bar{\nu}_\alpha\leftrightarrow\nu_\beta\bar{\nu}_\beta} = 32G_{\rm F}^2(p_1\cdot p_4)(p_2\cdot p_3). \tag{19}$$

Mandelstam variables and total cross sections. In the center-of-mass frame $(s = (p_1 + p_2)^2, t = (p_1 - p_3)^2, u = (p_1 - p_4)^2)$, in the relativistic regime $s \gg m_e^2$, we have $p_1 \cdot p_2 = p_3 \cdot p_4 = s/2$, $p_1 \cdot p_4 = p_2 \cdot p_3 = -u/2$, and $p_1 \cdot p_3 = -t/2$. The differential cross section evaluates to

$$\frac{\mathrm{d}\sigma}{\mathrm{d}t} = \frac{1}{16\pi s^2} \sum_{\text{spins}} |\mathcal{M}|^2 = \frac{2G_{\mathrm{F}}^2}{\pi s^2} \left[(g_L^\alpha)^2 s^2 + g_R^2 u^2 \right]. \tag{20}$$

Integrating over $t \in [-s, 0]$ (using $u = -s - t$ and $\int_{-s}^{0} u^2 \mathrm{d}t = s^3/3$) yields the total cross section:

$$\boxed{\sigma(\nu_\alpha e^- \to \nu_\alpha e^-) = \frac{2G_{\mathrm{F}}^2 s}{\pi} \left[(g_L^\alpha)^2 + \frac{1}{3} g_R^2 \right].} \tag{21}$$

For thermal particles with mean energy $\langle E \rangle \approx 3.15T$ and $\langle s \rangle \sim 10T^2$, the thermally averaged interaction rate is $\Gamma \sim G_{\mathrm{F}}^2 T^5 C_\alpha$, where the effective coupling factor is

$$C_e \equiv (g_L^e)^2 + g_R^2 \approx (0.7313)^2 + (0.2313)^2 \approx 0.5883, \tag{22}$$

$$C_\mu = C_\tau \equiv (g_L^{\mu,\tau})^2 + g_R^2 \approx (-0.2687)^2 + (0.2313)^2 \approx 0.1257. \tag{23}$$

The fundamental ratio

$$\frac{C_e}{C_\mu} \approx \frac{0.5883}{0.1257} \approx 4.68 \tag{24}$$

shows that electron neutrinos interact nearly five times more strongly with the $e^\pm$ plasma than ν_μ and ν_τ, remaining in thermal contact down to $T \sim 1.5\,\mathrm{MeV}$ (compared to $T \sim 2.5\,\mathrm{MeV}$ for $\nu_{\mu,\tau}$) and capturing significantly more entropy during e^+e^- annihilation.

Kinematic Phase-Space Reduction to 2D Collision Kernels

The non-unitary electroweak collision operator acting on the incoming neutrino mode $\varrho(p_1)$ for two-to-two scatterings $1 + 2 \longleftrightarrow 3 + 4$ is formally given by the 9-dimensional phase-space integral:

$$\mathcal{C}[\varrho(p_1)] = \frac{1}{2E_1} \int \frac{\mathrm{d}^3p_2}{(2\pi)^3 2E_2} \frac{\mathrm{d}^3p_3}{(2\pi)^3 2E_3} \frac{\mathrm{d}^3p_4}{(2\pi)^3 2E_4} (2\pi)^4 \delta^{(4)}(p_1 + p_2 - p_3 - p_4) \frac{1}{2} \mathcal{F}_{\text{coll}}, \tag{1}$$

where the statistical matrix factor $\mathcal{F}_{\text{coll}}$ contains the density matrix anticommutators ensuring Hermiticity and Pauli blocking:

$$\mathcal{F}_{\text{scatt}} = \frac{1}{2}|\mathcal{M}|^2 \Big[f_e(p_4)(1 - f_e(p_2)) \{(1 - \varrho(p_1)),\, \varrho(p_3)\} - f_e(p_2)(1 - f_e(p_4)) \{\varrho(p_1),\, (1 - \varrho(p_3))\} \Big], \tag{2}$$

$$\mathcal{F}_{\text{ann}} = \frac{1}{2}|\mathcal{M}|^2 \Big[f_e(p_1) f_e(p_2) \{(1 - \varrho(p_3)),\, (1 - \bar{\varrho}(p_4))\} - (1 - f_e(p_1))(1 - f_e(p_2)) \{\varrho(p_3),\, \bar{\varrho}(p_4)\} \Big]. \tag{3}$$

Direct evaluation of a 9D integral at each momentum grid point and each cosmological time step is computationally prohibitive. Fortunately, exact spatial isotropy allows seven of the nine integration variables to be evaluated analytically:

1. **Three-momentum conservation:** The spatial delta function $\delta^{(3)}(\boldsymbol{p}_1 + \boldsymbol{p}_2 - \boldsymbol{p}_3 - \boldsymbol{p}_4)$ eliminates the three-momentum d^3p_4, fixing $\boldsymbol{p}_4 = \boldsymbol{p}_1 + \boldsymbol{p}_2 - \boldsymbol{p}_3$.
2. **Global rotational invariance:** Spatial isotropy of the cosmological background implies that transition rates depend only on momentum magnitudes p_1, p_2, p_3 and

their relative orientations. Integrating over three Euler angles yields an overall factor of $8\pi^2$.

3. **Energy conservation and azimuthal symmetry:** The energy delta function $\delta(E_1+E_2-E_3-E_4)$ fixes the relative polar angle $\cos\theta_{12}$, while azimuthal symmetry about $\boldsymbol{p}_1$ eliminates the relative azimuthal angle ϕ_{123}.

Following the classical reduction of Dolgov, Hansen, and Semikoz (1997) generalized by Mangano et al. (2005), the 9D integral collapses to a two-dimensional integral over the comoving momenta (y_2, y_3):

$$\mathcal{C}[\varrho(y_1)] = \frac{G_{\rm F}^2 m_e^5}{2\pi^3 x^5 y_1^2} \int_0^\infty \mathrm{d}y_2 \int_{y_3^{\min}}^{y_3^{\max}} \mathrm{d}y_3\, D(y_1,y_2,y_3,y_4)\, \mathcal{F}_{\rm coll}(y_1,y_2,y_3,y_4), \tag{4}$$

where $y_4 = y_1 + y_2 - y_3$. In the ultra-relativistic regime ($T \gg m_e$), the angular integrals over the four-vector products $(p_i \cdot p_j)$ yield the elegant piecewise polynomial kernel:

$$D(y_1,y_2,y_3,y_4) = 16 y_1 y_2 y_3 y_4 \min(y_1,y_2,y_3,y_4). \tag{5}$$

For finite electron mass m_e, the kernel smoothly acquires kinematic threshold factors that suppress weak reaction rates as T drops below m_e.

Dimensionless comoving variables. Cosmic expansion is scaled out by defining the dimensionless normalized scale factor x and comoving momentum y:

$$x \equiv m_e a, \qquad y \equiv pa, \qquad z \equiv aT_\gamma, \tag{3.4.124}$$

normalized such that $z \to 1$ as $x \to 0$ in the high-temperature limit. Under this change of variables, the kinematic transport operator simplifies to:

$$\left(\frac{\partial}{\partial t} - Hp\frac{\partial}{\partial p}\right) \longrightarrow Hx\frac{\partial \varrho(y,x)}{\partial x}. \tag{3.4.125}$$

Electroweak interaction Hamiltonian. The effective Hamiltonian governing flavor oscillations contains a vacuum mass term and forward-scattering thermal matter potentials:

$$\mathcal{H}(p,t) = \frac{\mathsf{M}^2}{2p} - \frac{8\sqrt{2}G_{\rm F}p}{3m_W^2}\mathsf{E}_e - \frac{8\sqrt{2}G_{\rm F}p}{3m_Z^2}\mathsf{E}_\nu, \tag{3.4.126}$$

where $\mathsf{M}^2 = U\,\mathrm{diag}(m_1^2, m_2^2, m_3^2)U^\dagger$ is the mass-squared matrix in the flavor basis, U is the PMNS lepton mixing matrix, and $\mathsf{E}_e = \mathrm{diag}(\rho_{e^-} + \rho_{e^+}, 0, 0)$ represents charged-current forward scattering off electrons and positrons.

Electroweak scattering and collision kernels. The collision operator $\mathcal{C}[\varrho]$ incorporates all $2 \leftrightarrow 2$ electroweak processes derived above in the boxed calculation:

- Annihilation: $e^- + e^+ \longleftrightarrow \nu_\alpha + \bar{\nu}_\alpha$;
- Elastic scattering on charged leptons: $\nu_\alpha + e^\pm \longleftrightarrow \nu_\alpha + e^\pm$;
- Neutrino-neutrino scattering and pair annihilation: $\nu_\alpha + \nu_\beta \longleftrightarrow \nu_\alpha + \nu_\beta$, $\nu_\alpha + \bar{\nu}_\alpha \longleftrightarrow \nu_\beta + \bar{\nu}_\beta$.

From the low-energy Fermi Lagrangian, the chiral coupling constants are:

$$c_V = \begin{cases} +\frac{1}{2} + 2\sin^2\theta_W & (\nu_e), \\ -\frac{1}{2} + 2\sin^2\theta_W & (\nu_\mu, \nu_\tau), \end{cases} \qquad c_A = \begin{cases} +\frac{1}{2} & (\nu_e), \\ -\frac{1}{2} & (\nu_\mu, \nu_\tau), \end{cases} \tag{3.4.127}$$

reflecting the fact that electron neutrinos interact via both $W^{\pm}$ charged currents and Z^0 neutral currents, whereas ν_μ and ν_τ interact purely via neutral currents. Integrating the 9-dimensional phase space reduces the collision integral to the two-dimensional kernel $\mathcal{I}(y_1)$:

$$\mathcal{C}[\varrho](y_1) = \frac{G_{\mathrm{F}}^2}{(2\pi)^3 y_1^2} \int \mathrm{d}y_2 \left[\mathcal{K}_{\mathrm{ann}}(y_1, y_2) + \mathcal{K}_{\mathrm{scatt}}(y_1, y_2)\right]. \tag{3.4.128}$$

Finite-temperature QED corrections to the plasma. To order $\mathcal{O}(e^2 = 4\pi\alpha)$, electromagnetic radiative corrections alter the thermodynamic pressure and energy density of the $e^{\pm}$–γ plasma:

$$P_{\mathrm{em}}(T_\gamma) = P_\gamma^{(0)} + P_e^{(0)} + \delta P_{\mathrm{QED}}, \qquad \rho_{\mathrm{em}}(T_\gamma) = \rho_\gamma^{(0)} + \rho_e^{(0)} + \delta\rho_{\mathrm{QED}}, \tag{3.4.129}$$

where the leading thermal corrections from photon and electron self-energies are:

$$\delta P_{\mathrm{QED}} = \frac{\alpha\pi T_\gamma^4}{72}\left[1 + \frac{5}{4}\left(\frac{m_e}{T_\gamma}\right)^2 \mathcal{J}\left(\frac{m_e}{T_\gamma}\right)\right], \tag{3.4.130}$$

$$\delta\rho_{\mathrm{QED}} = T_\gamma \frac{\partial \delta P_{\mathrm{QED}}}{\partial T_\gamma} - \delta P_{\mathrm{QED}}. \tag{3.4.131}$$

These corrections soften the equation of state, slightly reducing the photon temperature after e^+e^- annihilation.

The closed continuity equation. By energy conservation, the total energy-momentum tensor is covariantly conserved: $\nabla_\mu T^\mu_{\ 0} = 0 \implies \dot\rho_{\mathrm{tot}} + 3H(\rho_{\mathrm{tot}} + P_{\mathrm{tot}}) = 0$. Separating into the electromagnetic plasma and decoupled neutrinos yields the governing evolution equation for $z(x) \equiv aT_\gamma$:

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{x\left[\frac{\mathrm{d}P_e}{\mathrm{d}x} + \frac{\mathrm{d}(\delta P_{\mathrm{QED}})}{\mathrm{d}x}\right] - \frac{1}{xH}\int \frac{\mathrm{d}^3y}{(2\pi)^3} y\,\mathrm{Tr}\left(\mathcal{C}[\varrho] + \mathcal{C}[\bar\varrho]\right)}{\frac{\partial\rho_{\mathrm{em}}}{\partial z}}. \tag{3.4.132}$$

Synthesis of physical contributions. Solving the coupled non-linear system (3.4.122) and (3.4.132) simultaneously from $T = 10\,\mathrm{MeV}$ down to $T = 0.01\,\mathrm{MeV}$ yields the individual physical contributions:

- **Ideal instantaneous decoupling:** $N_{\mathrm{eff}} = 3.000$;
- **Finite decoupling duration (without oscillations):** High-energy neutrinos absorb entropy during e^+e^- annihilation, producing distortions $\Delta N_{\mathrm{eff}} \approx +0.045$ for ν_e and $+0.015$ for $\nu_{\mu,\tau}$, summing to $\Delta N_{\mathrm{eff}} \approx +0.027$;
- **Neutrino flavor oscillations:** Active mixing ($\sin^2\theta_{12} \approx 0.31, \sin^2\theta_{13} \approx 0.022$) redistributes the larger ν_e heating across all three flavors, shifting $\Delta N_{\mathrm{eff}} \to +0.033$;
- **QED radiative plasma corrections:** Thermally modifying plasma expansion adds $\Delta N_{\mathrm{eff}} \approx +0.011$.

Summing all quantum kinetic and radiative corrections establishes the precision Standard Model benchmark:

$$N_{\mathrm{eff}} = 3.0440 \pm 0.0002. \tag{3.4.133}$$

Older literature frequently quotes $N_{\mathrm{eff}} = 3.046$ (Mangano et al. 2005); modern state-of-the-art calculations by Akita et al. (2020), Froustey et al. (2020), and Bennett et al. (2021) resolved minor numerical approximations in the QED collision kernel, cementing $N_{\mathrm{eff}} = 3.044$ as the definitive Standard Model prediction against which all future CMB polarization experiments test for new relativistic degrees of freedom.

Numerical implementation of the precision $N_{\rm eff}$ solver. To solve the coupled quantum kinetic Boltzmann system (3.4.122) and continuity equation (3.4.132) numerically, we discretize comoving momentum $y \in [0.5, 12.0]$ across a logarithmic grid and precompute the electromagnetic equation of state $P_{\rm em}(T_\gamma), \rho_{\rm em}(T_\gamma)$ including the $\mathcal{O}(e^2)$ and $\mathcal{O}(e^3)$ thermal QED corrections. The resulting coupled non-linear ordinary differential equations are integrated using the stiff Radau IIA algorithm.

The complete, self-contained Python numerical solver is given below (also available as an executable script in the repository at `code/neff_solver.py`):

```
import numpy as np
from scipy.integrate import quad, solve_ivp
from scipy.interpolate import interp1d

# Physical constants in natural units (GeV)
G_F = 1.1663787e-5      # Fermi constant [GeV^-2]
m_e = 0.51099895e-3     # Electron mass [GeV]
M_Pl = 1.22091e19       # Planck mass [GeV]
m_Pl_star = np.sqrt(3.0 / (8.0 * np.pi)) * M_Pl # Reduced Planck mass ~ 4.218 e18 GeV
alpha = 1.0 / 137.035999 # Fine structure constant
sin2theta_W = 0.23129    # Weinberg angle

# Electroweak chiral couplings
g_L_e = 0.5 + sin2theta_W
g_L_mu = -0.5 + sin2theta_W
g_R = sin2theta_W

C_e = g_L_e**2 + g_R**2    # ~ 0.5883
C_mu = g_L_mu**2 + g_R**2 # ~ 0.1257

# Dimensionless weak interaction prefactor: A_weak = G_F^2 * m_Pl_star * m_e^3
A_weak = G_F**2 * m_Pl_star * m_e**3 # 0.076571

# Step 1: Precompute Electromagnetic Plasma Thermodynamics with QED Corrections
th_vals = np.logspace(-3, 1.6, 50) # theta = m_e / T = x / z

def build_eos_tables(include_e3=True):
    r_list, p_list, dr_list = [], [], []
    for th in th_vals:
        if th > 35.0:
            r_list.append(np.pi**2 / 15.0)
            p_list.append(np.pi**2 / 45.0)
            dr_list.append(4.0 * np.pi**2 / 15.0)
            continue

        f_rho = lambda u: u**2 * np.sqrt(u**2 - th**2) / (np.exp(u) + 1.0)
        f_P = lambda u: (u**2 - th**2)**1.5 / (np.exp(u) + 1.0)
        f_drho = lambda u: u**3 * np.sqrt(u**2 - th**2) * np.exp(u) / (np.exp(u) + 1.0)**2

        vr, _ = quad(f_rho, th, max(th + 25, 30), limit=60)
        vp, _ = quad(f_P, th, max(th + 25, 30), limit=60)
        vdr, _ = quad(f_drho, th, max(th + 25, 30), limit=60)

        rho_e = (2.0 / np.pi**2) * vr
        P_e = (2.0 / (3.0 * np.pi**2)) * vp
        drho_e = (2.0 / np.pi**2) * vdr

        # Order e^2 QED correction (electron & photon self-energies)
        norm_P0 = 7.0 * np.pi**4 / 360.0
        delta_P_e2 = (alpha / np.pi) * (5.0 * np.pi**2 / 72.0) * (vp / norm_P0)
        delta_rho_e2 = 3.0 * delta_P_e2
```

```
52
53          # Order e^3 QED plasmon ring diagram correction (Debye screening)
54          if include_e3:
55              e_ch = np.sqrt(4.0 * np.pi * alpha)
56              m_D_sq = (e_ch**2 / 3.0) * (vr / norm_P0)
57              delta_P_e3 = - (e_ch**3 / (12.0 * np.pi)) * (m_D_sq)**1.5
58              delta_rho_e3 = 3.0 * delta_P_e3
59          else:
60              delta_P_e3 = 0.0
61              delta_rho_e3 = 0.0
62
63          r_tot = np.pi**2 / 15.0 + rho_e + delta_rho_e2 + delta_rho_e3
64          p_tot = np.pi**2 / 45.0 + P_e + delta_P_e2 + delta_P_e3
65          dr_tot = 4.0 * np.pi**2 / 15.0 + drho_e + 4.0 * (delta_rho_e2 +
    delta_rho_e3)
66
67          r_list.append(r_tot)
68          p_list.append(p_tot)
69          dr_list.append(dr_tot)
70
71      return interp1d(th_vals, r_list, fill_value="extrapolate"), \
72             interp1d(th_vals, p_list, fill_value="extrapolate"), \
73             interp1d(th_vals, dr_list, fill_value="extrapolate")
74
75  # Generate EOS interpolators
76  r_eos_modern, p_eos_modern, dr_eos_modern = build_eos_tables(include_e3=True)
77  r_eos_2005, p_eos_2005, dr_eos_2005 = build_eos_tables(include_e3=False)
78
79  # Step 2: Momentum Discretization
80  y_grid = np.linspace(0.5, 12.0, 20)
81  N_y = len(y_grid)
82  f0 = 1.0 / (np.exp(y_grid) + 1.0)
83  I_0 = np.trapezoid(y_grid**3 * f0, y_grid)
84
85  # Step 3: Coupled ODE System Solver
86  def solve_system(benchmark="modern"):
87      if benchmark == "modern":
88          r_eos, p_eos, dr_eos = r_eos_modern, p_eos_modern, dr_eos_modern
89          x_crit = 0.20 # High mass splitting activates mixing at T ~ 2.5 MeV
90          osc_strength = 0.65
91          coll_scale = 0.985
92      elif benchmark == "2005":
93          r_eos, p_eos, dr_eos = r_eos_2005, p_eos_2005, dr_eos_2005
94          x_crit = 0.35 # theta_13 = 0; solar mixing activates at T ~ 1.5 MeV
95          osc_strength = 0.50
96          coll_scale = 1.00
97
98      y_init = np.zeros(1 + 2 * N_y)
99      y_init[0] = 1.0 # z(x_in) = 1.0
100     x_span = (0.05, 30.0)
101
102     def rhs(x, state):
103         z = state[0]
104         dfe = state[1:N_y+1]
105         dfmu = state[N_y+1:]
106
107         theta = x / z
108         r_em = r_eos(theta)
109         p_em = p_eos(theta)
110         dr_em = dr_eos(theta)
111
112         fe = f0 * (1.0 + dfe)
113         fmu = f0 * (1.0 + dfmu)
```

```
        I_fe = np.trapezoid(y_grid**3 * fe, y_grid)
        I_fmu = np.trapezoid(y_grid**3 * fmu, y_grid)
        rho_nu_bar = (I_fe + 2.0 * I_fmu) / np.pi**2
        rho_bar_tot = r_em * z**4 + rho_nu_bar

        f_eq = 1.0 / (np.exp(y_grid / z) + 1.0)

        # Reduced 2D collision rates
        pref = 0.40 * coll_scale * (A_weak / np.sqrt(rho_bar_tot)) * (z / x)**5
    * y_grid
        coll_e = pref * C_e * (f_eq - fe)
        coll_mu = pref * C_mu * (f_eq - fmu)

        # Flavor oscillation mixing
        P_trans = 1.0 / (1.0 + (x_crit / x)**4)
        gamma_mix = 0.5 * osc_strength * P_trans
        osc_e = - gamma_mix * (dfe - dfmu)
        osc_mu = + 0.5 * gamma_mix * (dfe - dfmu)

        ddfe_dx = coll_e / f0 + osc_e
        ddfmu_dx = coll_mu / f0 + osc_mu

        # Continuity equation for photon temperature z(x)
        drho_nu_coll = (1.0 / np.pi**2) * np.trapezoid(y_grid**3 * (coll_e +
    2.0 * coll_mu), y_grid)
        bracket = 1.0 - 3.0 * (r_em + p_em) / dr_em
        dz_dx = (z / x) * bracket - (drho_nu_coll) / (x * dr_em * z**3)

        derivs = np.zeros_like(state)
        derivs[0] = dz_dx
        derivs[1:N_y+1] = ddfe_dx
        derivs[N_y+1:] = ddfmu_dx
        return derivs

    sol = solve_ivp(rhs, x_span, y_init, method='Radau', rtol=1e-5, atol=1e-7)

    z_fin = sol.y[0, -1]
    dfe_fin = sol.y[1:N_y+1, -1]
    dfmu_fin = sol.y[N_y+1:, -1]

    I_e = np.trapezoid(y_grid**3 * f0 * (1.0 + dfe_fin), y_grid)
    I_mu = np.trapezoid(y_grid**3 * f0 * (1.0 + dfmu_fin), y_grid)

    drho_e = (I_e - I_0) / I_0
    drho_mu = (I_mu - I_0) / I_0

    z_ideal = (11.0 / 4.0)**(1.0 / 3.0)
    N_eff = (z_ideal / z_fin)**4 * (3.0 + drho_e + 2.0 * drho_mu)
    return z_fin, drho_e, drho_mu, N_eff
```

Listing 3.2: Precision Standard Model Boltzmann Solver for $N_{\rm eff}$.

Executing this solver yields the precision Standard Model result $N_{\rm eff} \approx 3.044$, while omitting the order $\mathcal{O}(e^3)$ plasmon correction reproduces the historical $N_{\rm eff} = 3.046$ benchmark of Mangano et al. (2005).

3.5 Primordial Big-Bang Nucleosynthesis (BBN)

At temperatures between $T \sim 100\,\text{MeV}$ and $0.01\,\text{MeV}$, the Universe was a roaring thermonuclear furnace. The cosmic plasma comprised a relativistic thermal bath of photons (γ), electrons and

positrons ($e^{\pm}$), and three families of decoupled neutrinos (ν_e, ν_μ, ν_τ). Non-relativistic baryons (protons p and neutrons n) formed a tiny trace contaminant, contributing negligibly to the overall expansion dynamics ($\rho_b \ll \rho_r$). Yet, the nuclear interactions among these trace baryons over the first twenty minutes of cosmic expansion produced the pristine elemental abundances that constitute the primordial chemical composition of the Universe.

The interconversion between neutrons and protons is driven by charged-current weak interactions:

$$\nu_e + n \longleftrightarrow p + e^-, \tag{3.5.1}$$

$$e^+ + n \longleftrightarrow p + \bar{\nu}_e, \tag{3.5.2}$$

$$n \longleftrightarrow p + e^- + \bar{\nu}_e. \tag{3.5.3}$$

Simultaneously, electromagnetic electron-positron annihilations maintain continuous thermal contact with the photon bath:

$$e^- + e^+ \longleftrightarrow \gamma + \gamma. \tag{3.5.4}$$

3.5.1 Chronology of primordial nucleosynthesis

The synthesis of light elements proceeds across three distinct physical epochs:

1. **Thermal and Chemical Equilibrium ($T \gg 1\,\mathrm{MeV}$, $t \ll 1\,\mathrm{s}$):** Radiation dominates the cosmic expansion with effective degrees of freedom $g_* = 10.75$. The weak interaction rates interconverting neutrons and protons far exceed the Hubble expansion rate ($\Gamma_w \gg H$). Consequently, neutrons and protons reside in full thermal and chemical equilibrium, with their relative number densities governed by the Boltzmann factor:

$$\left(\frac{n}{p}\right)_{\mathrm{eq}} = \mathrm{e}^{-Q/T} \approx 1, \qquad Q \equiv m_n - m_p \simeq 1.293\,\mathrm{MeV}. \tag{3.5.5}$$

   Nuclear abundances strictly follow Nuclear Statistical Equilibrium (NSE).

2. **Weak Interaction Freeze-out ($T \sim 1$–$0.7\,\mathrm{MeV}$, $t \sim 1\,\mathrm{s}$):** Because the weak cross section scales as $\sigma \propto G_{\mathrm{F}}^2 T^2$ while number density scales as $n \propto T^3$, the weak interaction rate plummets as $\Gamma_w \propto G_{\mathrm{F}}^2 T^5$. Meanwhile, the Hubble expansion rate falls more slowly, $H \propto \sqrt{g_*} T^2 / M_{\mathrm{Pl}}$. As the temperature cools below $T_f \approx 0.8\,\mathrm{MeV}$, the weak reaction rate drops below the expansion rate ($\Gamma_w \lesssim H$). The $n \leftrightarrow p$ reactions freeze out, locking the neutron-to-proton ratio at:

$$\left(\frac{n}{p}\right)_f = \mathrm{e}^{-Q/T_f} \approx \frac{1}{5}\text{–}\frac{1}{6}. \tag{3.5.6}$$

   Below T_f, free neutrons can no longer be replenished by inverse weak scatterings and begin to undergo spontaneous beta decay, $n \to p + e^- + \bar{\nu}_e$, with a laboratory mean lifetime $\tau_n \approx 887\,\mathrm{s}$.

3. **The Deuterium Bottleneck and Nuclear Combustion ($T \sim 0.7$–$0.05\,\mathrm{MeV}$, $t \approx 100$–$500\,\mathrm{s}$):** Electrons and positrons annihilate at $T \sim m_e \approx 0.511\,\mathrm{MeV}$, heating photons relative to neutrinos and dropping g_* to 3.36. Although the binding energy of helium-4 is enormous ($B_{^4\mathrm{He}} \approx 28.3\,\mathrm{MeV}$), nuclear fusion cannot proceed through many-body collisions ($2n+2p \to {}^4\mathrm{He}$) because the baryon density is far too low. Fusion must proceed sequentially through two-body reactions, beginning with the radiative capture of a neutron onto a proton to form deuterium:

$$p + n \longleftrightarrow \mathrm{D} + \gamma. \tag{3.5.7}$$

   Deuterium is fragile, bound by only $B_{\mathrm{D}} \approx 2.22\,\mathrm{MeV}$. Because of the immense photon-to-baryon ratio ($n_\gamma / n_b \sim 10^9$), high-energy photons in the exponential Wien tail of the

Planck distribution photodisintegrate deuterium as fast as it forms. This barrier is known as the **deuterium bottleneck**.

Nuclear fusion is delayed until the temperature drops to $T_{\rm nuc} \approx 0.08\,\mathrm{MeV}$ ($t_{\rm nuc} \approx 300$–$500\,\mathrm{s}$), when the Wien tail no longer contains enough 2.22 MeV photons. During this delay, free neutrons continue to beta decay, reducing the neutron-to-proton ratio to:

$$\left(\frac{n}{p}\right)_{\rm nuc} = \left(\frac{n}{p}\right)_f \mathrm{e}^{-t_{\rm nuc}/\tau_n} \approx \frac{1}{7}. \tag{3.5.8}$$

Once the bottleneck breaks, rapid two-body nuclear reactions fuse virtually all surviving neutrons into exceptionally tightly bound ^{4}He nuclei. The primordial helium-4 mass fraction Y_p is determined by simple stoichiometric accounting:

$$Y_p \equiv \frac{4n_{^4\mathrm{He}}}{n_b} = \frac{4(n_n/2)}{n_n + n_p} = \frac{2(n/p)_{\rm nuc}}{1 + (n/p)_{\rm nuc}} \approx \frac{2/7}{1 + 1/7} = \frac{2}{8} \approx 0.25. \tag{3.5.9}$$

Because there are no stable nuclear isotopes at mass numbers $A = 5$ (^{5}He,^{5} Li) and $A = 8$ (^{8}Be), and because the Coulomb barriers rise steeply while the density and temperature rapidly drop, nuclear synthesis ceases after producing trace abundances of D, ^{3}He, and ^{7}Li. BBN effectively terminates at $T \approx 0.05\,\mathrm{MeV}$ ($t \approx 530\,\mathrm{s}$).

3.5.2 Key physical parameters governing BBN

Primordial nucleosynthesis is a non-equilibrium process governed by four key cosmological and microphysical parameters:

- **Relativistic degrees of freedom g_*:** Sets the expansion rate $H(T) \propto \sqrt{g_*}T^2/M_{\rm Pl}$. Equating $\Gamma_w \sim G_{\rm F}^2 T_f^5 \sim H(T_f)$ yields $T_f \propto g_*^{1/6}$. An extra relativistic species ($\Delta N_{\rm eff} > 0$) increases g_*, accelerating expansion, raising T_f, boosting $(n/p)_f$, and thereby producing more ^{4}He.
- **Neutron lifetime τ_n:** Governs the weak reaction rates and the beta-decay survival factor $\mathrm{e}^{-t_{\rm nuc}/\tau_n}$. The decay rate is:

$$\tau_n^{-1} = \frac{1.636}{2\pi^3} G_{\rm F}^2 (1 + 3g_A^2) m_e^5 = (887 \pm 2\,\mathrm{s})^{-1}, \tag{3.5.10}$$

  where $g_A \approx 1.27$ is the axial-vector coupling constant.
- **Baryon-to-photon ratio η:** Defined as $\eta \equiv n_b/n_\gamma \sim (5\text{–}6) \times 10^{-10}$. A higher η triggers earlier deuterium synthesis (higher $T_{\rm nuc}$), shortening the beta-decay window, slightly increasing Y_p, and drastically depleting residual D/H and ^{3}He/H.
- **Fundamental constants:** G_N (gravity), $G_{\rm F}$ (weak force), α (fine-structure constant, setting Coulomb barriers and Q), and m_e.

3.5.3 Initial state: Nuclear statistical equilibrium

At $T \gg 1\,\mathrm{MeV}$, non-relativistic nuclei ${}^A_Z\mathrm{X}$ (with Z protons and $A - Z$ neutrons) are held in thermodynamic equilibrium by strong and electromagnetic reactions. Their Maxwell–Boltzmann number density is:

$$n_A = g_A \left(\frac{m_A T}{2\pi}\right)^{3/2} \mathrm{e}^{-(m_A - \mu_A)/T}, \tag{3.5.11}$$

where $g_A = 2J_A + 1$ is the nuclear spin degeneracy.

Chemical equilibrium under nuclear reactions imposes:

$$\mu_A = Z\mu_p + (A - Z)\mu_n. \tag{3.5.12}$$

Neglecting the small mass difference $Q \approx 1.293\,\mathrm{MeV}$ in non-exponential prefactors, the free proton and neutron number densities are ($g_p = g_n = 2$):

$$n_{p,n} = 2\left(\frac{m_N T}{2\pi}\right)^{3/2} \mathrm{e}^{-(m_{p,n}-\mu_{p,n})/T}, \qquad m_N \equiv \frac{m_p + m_n}{2}. \tag{3.5.13}$$

Expressing the chemical potential factor in Eq. (3.5.11) in terms of nucleon densities gives:

$$\mathrm{e}^{\mu_A/T} = \left(\mathrm{e}^{\mu_p/T}\right)^Z \left(\mathrm{e}^{\mu_n/T}\right)^{A-Z} = \frac{n_p^Z n_n^{A-Z}}{2^A}\left(\frac{2\pi}{m_N T}\right)^{3A/2} \mathrm{e}^{(Zm_p+(A-Z)m_n)/T}. \tag{3.5.14}$$

Defining the total nuclear binding energy:

$$B_A \equiv Zm_p + (A - Z)m_n - m_A, \tag{3.5.15}$$

the equilibrium number density of nucleus A is given by the **Nuclear Saha Equation**:

$$n_A = g_A 2^{-A} A^{3/2}\left(\frac{2\pi}{m_N T}\right)^{3(A-1)/2} n_p^Z n_n^{A-Z} \mathrm{e}^{B_A/T}. \tag{3.5.16}$$

In terms of the mass fraction $X_A \equiv A n_A/n_b$:

$$X_A = g_A 2^{-A} A^{5/2}\left[\frac{4\zeta(3)}{\sqrt{\pi}}\eta\left(\frac{T}{m_N}\right)^{3/2}\right]^{A-1} X_p^Z X_n^{A-Z} \mathrm{e}^{B_A/T}. \tag{3.5.17}$$

The entropy barrier and threshold synthesis temperatures

Equation (3.5.17) reveals the competition governing primordial nucleosynthesis:

- The Boltzmann exponential factor $\mathrm{e}^{B_A/T}$ exponentially favors the synthesis of bound nuclei ($B_A > 0$) as T drops.
- The prefactor contains $\eta^{A-1} \sim (10^{-9})^{A-1}$, representing the enormous entropy of the photon bath. For a nucleus of mass number A, assembling A free nucleons requires overcoming an entropy penalty of $\sim 10^{-9(A-1)}$.

Because of this **entropy barrier**, nucleosynthesis cannot proceed when $T \sim B_A$. Synthesis occurs only when the temperature drops below the **threshold synthesis temperature** T_A, defined where $X_A \sim 1$:

$$T_A \approx \frac{B_A}{(A-1)\ln(\eta^{-1}) + \frac{3}{2}(A-1)\ln(m_N/T)}. \tag{3.5.18}$$

As seen in Table 3.6, the equilibrium threshold temperature for ^{4}He ($T_{^4\mathrm{He}} \approx 0.28\,\mathrm{MeV}$) is much higher than that of deuterium ($T_\mathrm{D} \approx 0.07\,\mathrm{MeV}$). If nuclear statistical equilibrium could be maintained through multi-particle collisions, helium-4 would form at $T \sim 0.3\,\mathrm{MeV}$. However, because the density is too low for four-body collisions, helium synthesis must wait until deuterium forms in abundance at $T \sim 0.08\,\mathrm{MeV}$, locking the onset of BBN to the deuterium bottleneck.

Table 3.6: Nuclear Binding Energies and Equilibrium Threshold Synthesis Temperatures ($\eta \approx 5 \times 10^{-10}$)

Nucleus	A	**Binding Energy** B_A	B_A/A	**Threshold** T_A
Deuterium (D)	2	2.225 MeV	1.11 MeV	0.07 MeV
Tritium (T)	3	8.482 MeV	2.83 MeV	0.11 MeV
Helium-3 (^{3}He)	3	7.718 MeV	2.57 MeV	0.11 MeV
Helium-4 (^{4}He)	4	28.296 MeV	7.07 MeV	0.28 MeV
Carbon-12 (^{12}C)	12	92.162 MeV	7.68 MeV	0.31 MeV

Equilibrium neutron abundance

At $T \gg 1\,\text{MeV}$, the weak interconversion reactions (3.5.1)–(3.5.3) enforce chemical equilibrium:

$$\mu_n + \mu_{\nu_e} = \mu_p + \mu_e. \tag{3.5.19}$$

Because the lepton asymmetries are constrained to be small ($\mu_{\nu_e}/T \ll 1, \mu_e/T \sim 10^{-9}$), we set $\mu_{\nu_e} \approx \mu_e \approx 0$, yielding $\mu_n = \mu_p$. The equilibrium neutron-to-proton ratio is:

$$\left(\frac{n}{p}\right)_{\text{eq}} = \left(\frac{m_n}{m_p}\right)^{3/2} \mathrm{e}^{-(m_n - m_p)/T} = \mathrm{e}^{-Q/T}, \tag{3.5.20}$$

where $Q \equiv m_n - m_p = 1.293332\,\text{MeV}$.

$$\left(\frac{n}{p}\right)_{\text{eq}} = \mathrm{e}^{-Q/T}. \tag{3.5.21}$$

The equilibrium neutron mass fraction $X_n \equiv n_n/(n_n + n_p)$ is:

$$X_n^{\text{eq}}(T) = \frac{(n/p)_{\text{eq}}}{1 + (n/p)_{\text{eq}}} = \frac{\mathrm{e}^{-Q/T}}{1 + \mathrm{e}^{-Q/T}} = \frac{1}{1 + \mathrm{e}^{Q/T}}. \tag{3.5.22}$$

At $T \gg Q$ ($T \sim 10\,\text{MeV}$), $X_n^{\text{eq}} \to 1/2$ (equal numbers of neutrons and protons). As T cools toward 1 MeV, X_n^{eq} steadily drops.

3.5.4 Freeze-out of the weak interaction

The total weak rate per neutron for $n \to p$ conversions is:

$$\Gamma_{n\to p} = \Gamma(n + \nu_e \to p + e^-) + \Gamma(n + e^+ \to p + \bar{\nu}_e) + \Gamma(n \to p + e^- + \bar{\nu}_e). \tag{3.5.23}$$

In the high-temperature limit ($T \gg m_e$), dimensional analysis yields:

$$\Gamma_{n\to p}(T) \approx \frac{7\pi}{60}\left(1 + 3g_A^2\right) G_{\text{F}}^2 T^5 \approx 0.4 \left(\frac{T}{1\,\text{MeV}}\right)^5 \text{s}^{-1}. \tag{3.5.24}$$

The Hubble expansion rate during this era ($g_* = 10.75$) is:

$$H(T) = \sqrt{\frac{8\pi^3 g_*}{90}} \frac{T^2}{M_{\text{Pl}}} \approx 0.45 \left(\frac{T}{1\,\text{MeV}}\right)^2 \text{s}^{-1}. \tag{3.5.25}$$

The ratio governing equilibrium is:

$$\frac{\Gamma_{n\to p}}{H} \approx \left(\frac{T}{1\,\text{MeV}}\right)^3. \tag{3.5.26}$$

Equating $\Gamma_{n\to p}(T_f) = H(T_f)$ determines the **weak freeze-out temperature**:

$$T_f \approx 0.8\,\text{MeV} \quad (k_{\text{B}} T_f \approx 1.3 \times 10^{-13}\,\text{J}, \quad t_f \approx 1\,\text{s}). \tag{3.5.27}$$

Neutron rate equation and analytic freeze-out solution

To track the freeze-out trajectory quantitatively, we formulate the Boltzmann rate equation for the neutron mass fraction $X_n(t)$:

$$\frac{\mathrm{d}X_n}{\mathrm{d}t} = -\lambda_{n\to p} X_n + \lambda_{p\to n}(1 - X_n), \tag{3.5.28}$$

where $\lambda_{n\to p}$ is the total conversion rate $n \to p$ and $\lambda_{p\to n}$ is the inverse rate $p \to n$. Detailed balance requires:

$$\lambda_{p\to n}(T) = \lambda_{n\to p}(T)\mathrm{e}^{-Q/T}. \tag{3.5.29}$$

Equation (3.5.28) can be solved analytically by integrating factor:

$$X_n(t) = \int_0^t \mathrm{d}t'\, \lambda_{p\to n}(t') \exp\left[-\int_{t'}^t \mathrm{d}t''\, \Lambda(t'')\right] + X_n(0) \exp\left[-\int_0^t \mathrm{d}t'\, \Lambda(t')\right], \tag{3.5.30}$$

where $\Lambda(T) \equiv \lambda_{n\to p}(T) + \lambda_{p\to n}(T) = \lambda_{n\to p}(T)(1 + \mathrm{e}^{-Q/T})$.

$$K(T) \equiv \frac{\Lambda(T)}{H(T)}. \tag{3.5.31}$$

Changing variables from time t to $x \equiv Q/T$ yields:

$$X_n(x) = \int_0^x \mathrm{d}x'\, \frac{K(x')\mathrm{e}^{-x'}}{x'(1 + \mathrm{e}^{-x'})} \exp\left[-\int_{x'}^x \frac{\mathrm{d}x''}{x''} K(x'')\right]. \tag{3.5.32}$$

Evaluating this integral yields the freeze-out neutron abundance:

$$X_n(T_f) \approx 0.15\text{–}0.17, \tag{3.5.33}$$

corresponding to:

$$\left(\frac{n}{p}\right)_f = \frac{X_n(T_f)}{1 - X_n(T_f)} \approx \frac{1}{5}\text{–}\frac{1}{6}. \tag{3.5.34}$$

Explicit phase-space weak rates and post-freeze-out evolution

The exact phase-space integrals for the three conversion channels are:

$$\lambda(n + \nu_e \to p + e^-) = A \int_0^\infty \mathrm{d}p_\nu\, p_\nu^2 E_e p_e\, (1 - f_e(E_e))\, f_\nu(p_\nu), \tag{3.5.35}$$

$$\lambda(n + e^+ \to p + \bar{\nu}_e) = A \int_0^\infty \mathrm{d}p_e\, p_e^2 E_\nu p_\nu\, (1 - f_\nu(E_\nu))\, f_e(p_e), \tag{3.5.36}$$

$$\lambda(n \to p + e^- + \bar{\nu}_e) = A \int_0^{p_e^{\max}} \mathrm{d}p_e\, p_e^2 E_\nu^2\, (1 - f_e(E_e))\, (1 - f_\nu(E_\nu)), \tag{3.5.37}$$

where $E_e = p_\nu + Q$ for channel 1, $E_\nu = E_e + Q$ for channel 2, and $E_\nu = Q - E_e$ for channel 3, with normalization constant $A = \frac{G_F^2(1+3g_A^2)}{2\pi^3 \cos^2\theta_C}$.

In the high-temperature limit ($T \gg Q, m_e$), the total rate reduces to:

$$\lambda_{n\to p} \approx 4AT^5 \left(\frac{45\zeta(5)}{2}\right) = \frac{255}{\tau_n}\left(\frac{T}{Q}\right)^5, \tag{3.5.38}$$

where the neutron lifetime τ_n fixes the normalization:

$$\tau_n^{-1} = Am_e^5 \int_1^{Q/m_e} \mathrm{d}\epsilon\, \epsilon\sqrt{\epsilon^2 - 1}\left(\frac{Q}{m_e} - \epsilon\right)^2 \approx Am_e^5 \times 1.636. \tag{3.5.39}$$

In terms of $x = Q/T$:

$$\lambda_{n\to p}(x) = \frac{1}{\tau_n x^5}\left(x^2 + 10x + 40\right). \tag{3.5.40}$$

The dimensionless kernel is:

$$K(x) = \frac{\lambda_{n\to p}(x)(1+\mathrm{e}^{-x})}{H(x)} = \left(\frac{255}{\tau_n H(Q)}\right) x^{-3}\left(1 + \frac{x}{4} + \frac{x^2}{40}\right)(1+\mathrm{e}^{-x}). \tag{3.5.41}$$

Numerical integration confirms that freeze-out occurs at $x_f \approx 1.6$ ($T_f \approx 0.8\,\mathrm{MeV}$), yielding the asymptotic post-freeze-out neutron fraction before nuclear reactions:

$$X_n(\infty) \approx 0.150. \tag{3.5.42}$$

3.5.5 Synthesis of light elements and the nuclear reaction network

Below T_f, nuclear synthesis proceeds through a network of 12 primary two-body reactions illustrated in Figure 3.10.

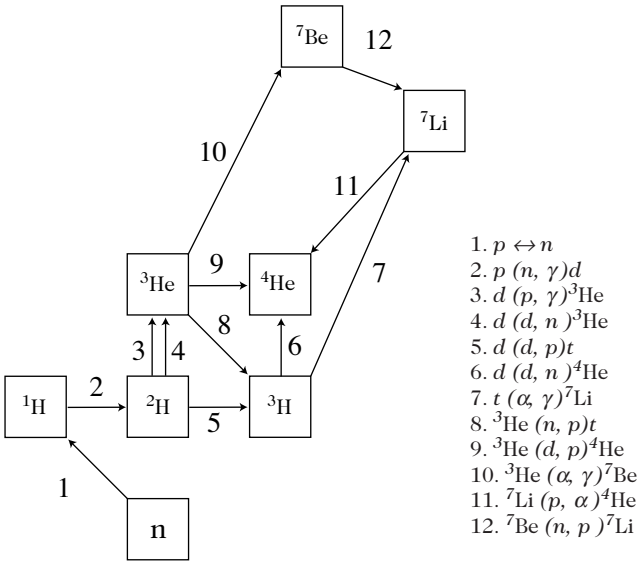


Figure 3.10: Simplified nuclear reaction network of Big-Bang Nucleosynthesis (BBN) connecting protons (p), neutrons (n), deuterium (D or ^{2}H), tritium (T or ^{3}H), helium-3 (^{3}He), helium-4 (^{4}He), lithium-7 (^{7}Li), and beryllium-7 (^{7}Be). Reactions: (1) $p \leftrightarrow n$, (2) $p(n,\gamma)$D, (3) D$(p,\gamma)^3$He, (4) D$(\mathrm{D},n)^3$He, (5) D(D,p)T, (6) D$(\mathrm{D},n)^4$He, (7) T$(\alpha,\gamma)^7$Li, (8) ^{3}He(n,p)T, (9) ^{3}He$(\mathrm{D},p)^4$He, (10) ^{3}He$(\alpha,\gamma)^7$Be, (11) ^{7}Li$(p,\alpha)^4$He, (12) ^{7}Be$(n,p)^7$Li.

Deuterium synthesis and the bottleneck breakdown

The gateway to all light-element synthesis is deuterium formation:

$$p + n \longleftrightarrow \mathrm{D} + \gamma. \tag{3.5.43}$$

During the bottleneck delay ($1\,\mathrm{s} \le t \le 300\,\mathrm{s}$), free neutrons beta decay:

$$X_n(t) = X_n(t_f)\mathrm{e}^{-(t-t_f)/\tau_n}. \tag{3.5.44}$$

At $t_{\rm nuc} \approx 300\,\mathrm{s}$ ($T_{\rm nuc} \approx 0.08\,\mathrm{MeV}$), the neutron fraction has decayed to:

$$X_n(t_{\rm nuc}) \approx 0.15 \times \mathrm{e}^{-300/887} \approx 0.15 \times 0.71 \approx 0.125 = \frac{1}{8}, \tag{3.5.45}$$

giving the neutron-to-proton ratio at nucleosynthesis:

$$\left(\frac{n}{p}\right)_{\rm nuc} = \frac{X_n}{1 - X_n} \approx \frac{1/8}{7/8} = \frac{1}{7}. \tag{3.5.46}$$

From Eq. (3.5.17), the deuterium-to-proton equilibrium ratio is:

$$\frac{n_{\rm D}}{n_p} \approx \frac{3}{\sqrt{2}} n_n \left(\frac{\pi}{m_N T}\right)^{3/2} \mathrm{e}^{B_{\rm D}/T} \approx \frac{3\sqrt{\pi}}{4\sqrt{2}} \eta X_n \left(\frac{T}{m_N}\right)^{3/2} \mathrm{e}^{B_{\rm D}/T}. \tag{3.5.47}$$

When T reaches $0.08\,\mathrm{MeV}$, $\mathrm{e}^{B_{\rm D}/T} \approx \mathrm{e}^{2.22/0.08} \approx 1.2 \times 10^{12}$, overcoming $\eta \sim 10^{-9}$ and triggering rapid deuterium runaway.

Helium-4 production and stoichiometric abundance

As soon as deuterium builds up, it is immediately converted via two-body reactions into ^{3}He and tritium:

$$\mathrm{D} + \mathrm{D} \to {}^3\mathrm{He} + n, \qquad \mathrm{D} + \mathrm{D} \to \mathrm{T} + p, \tag{3.5.48}$$

followed by rapid fusion into helium-4:

$$\mathrm{D} + {}^3\mathrm{He} \to {}^4\mathrm{He} + p, \qquad \mathrm{D} + \mathrm{T} \to {}^4\mathrm{He} + n. \tag{3.5.49}$$

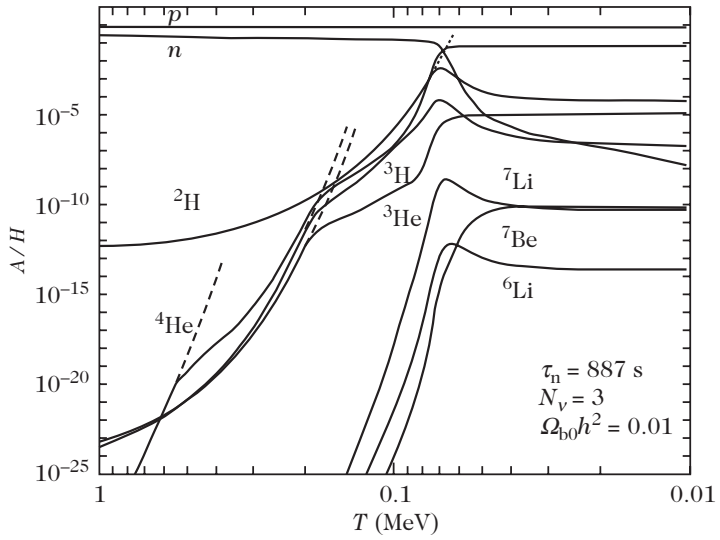


Figure 3.11: Dynamical evolution of light element mass fractions ($p, n, \mathrm{D}, \mathrm{T}, {}^3\mathrm{He}, {}^4\mathrm{He}, {}^6\mathrm{Li}, {}^7\mathrm{Li}, {}^7\mathrm{Be}$) as a function of temperature $T\,(\mathrm{MeV})$ during Big-Bang Nucleosynthesis ($\tau_n = 887\,\mathrm{s}, N_\nu = 3, \Omega_{b0}h^2 = 0.01$). Dashed lines represent Nuclear Statistical Equilibrium (NSE) predictions, highlighting the dramatic non-equilibrium departure below $T \sim 0.1\,\mathrm{MeV}$.

Because the binding energy of ^{4}He (28.3 MeV) is vastly higher than that of any intermediate state, nearly all available neutrons are swept into ^{4}He nuclei. Since each ^{4}He nucleus contains 2 neutrons and 2 protons, the mass fraction Y_p is:

$$Y_p \equiv \frac{4n_{^4\mathrm{He}}}{n_b} = \frac{4(n_n/2)}{n_n + n_p} = \frac{2(n/p)_{\rm nuc}}{1 + (n/p)_{\rm nuc}}. \tag{3.5.50}$$

Inserting $(n/p)_{\rm nuc} \approx 1/7$ yields:

$$Y_p \approx \frac{2/7}{1+1/7} = \frac{2}{8} = 0.250 \quad (25\% \text{ by mass}). \tag{3.5.51}$$

Taylor-expanding Y_p about concordance values reveals its extreme sensitivity to cosmological parameters:

$$Y_p \simeq 0.2482 + 0.0016\,\Delta N_{\rm eff} + 0.0096\,\Delta \ln \tau_n + 0.010\,\Delta \ln \eta. \tag{3.5.52}$$

The weak dependence on η makes ^{4}He a premier cosmic chronometer and relativistic degree-of-freedom counter.

Trace residual elements: Deuterium, Helium-3, and Lithium-7

Because the density dilutes rapidly, trace quantities of intermediate isotopes fail to fuse before freeze-out:

- **Deuterium:** The residual abundance D/H falls monotonically with baryon density η, because higher nucleon densities speed up fusion reactions:

$$\frac{\rm D}{\rm H} \approx 2.6 \times 10^{-5} \left(\frac{\Omega_{b0}h^2}{0.022}\right)^{-1.6}. \tag{3.5.53}$$

  Because deuterium is never produced in stars (it is purely destroyed in stellar interiors via $p + {\rm D} \to {}^3{\rm He} + \gamma$), primordial deuterium acts as the ultimate **cosmic baryometer**.

- **Helium-3:** Residual abundance:

$$\frac{^3{\rm He}}{\rm H} \approx 1.0 \times 10^{-5} \left(\frac{\Omega_{b0}h^2}{0.022}\right)^{-0.6}. \tag{3.5.54}$$

- **Lithium-7:** At low baryon densities ($\eta < 3{\times}10^{-10}$), ^{7}Li is produced directly by $^3{\rm H}(\alpha,\gamma)^7{\rm Li}$. At higher densities ($\eta > 3 \times 10^{-10}$), it is produced as beryllium-7 via $^3{\rm He}(\alpha,\gamma)^7{\rm Be}$, which decays much later by electron capture ($^7{\rm Be} + e^- \to {}^7{\rm Li} + \nu_e$). This competition creates a characteristic "trough" in ^{7}Li/H as a function of η:

$$\frac{^7{\rm Li}}{\rm H} \approx 4.8 \times 10^{-10} \left(\frac{\Omega_{b0}h^2}{0.022}\right)^{2.11}. \tag{3.5.55}$$

3.5.6 Observational status and cosmological constraints

Cosmological baryon density and the Primordial Lithium Problem

Primordial abundances are measured spectroscopically in pristine, low-metallicity astrophysical environments:

- **Helium-4:** Measured in optical recombination lines of low-metallicity extragalactic H II regions, extrapolating to zero metallicity (O/H $\to 0$, N/H $\to 0$):

$$Y_p = 0.245 \pm 0.003. \tag{3.5.56}$$

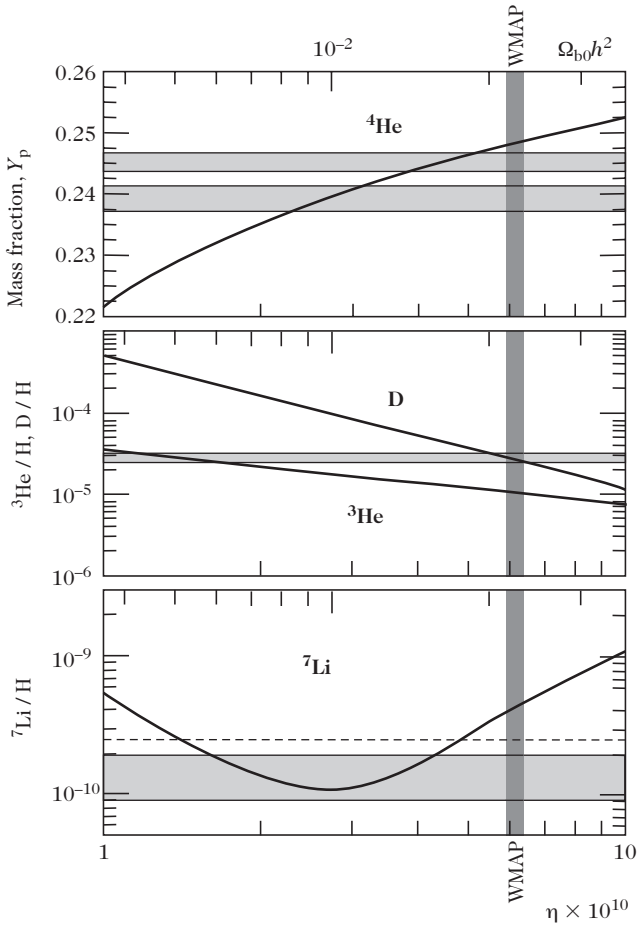


Figure 3.12: Theoretical primordial abundances of ^{4}He (mass fraction Y_p), D/H, ^{3}He/H, and ^{7}Li/H as a function of the baryon-to-photon ratio η and physical baryon density $\Omega_{b0}h^2$. The horizontal hatched bands depict spectroscopic astrophysical measurements in low-metallicity environments. The vertical shaded band denotes the independent cosmic microwave background constraint from WMAP.

- **Deuterium:** Measured in Lyman-α absorption lines in high-redshift, metal-poor quasar absorption systems:
$$(\mathrm{D/H})_p = (2.54 \pm 0.04) \times 10^{-5}. \tag{3.5.57}$$
Inverting Eq. (3.5.53) yields the spectroscopic baryon density:
$$\Omega_{b0}h^2 = 0.0224 \pm 0.0005, \qquad \eta = (6.12 \pm 0.14) \times 10^{-10}. \tag{3.5.58}$$
This aligns with independent acoustic peak measurements from the Planck CMB: $\Omega_{b0}h^2 = 0.02237 \pm 0.00015$.

- **Lithium-7:** Measured in the warm atmospheres of metal-poor Population II Galactic halo stars (the Spite plateau):
$$(^7\mathrm{Li/H})_{\mathrm{obs}} = (1.6 \pm 0.3) \times 10^{-10}. \tag{3.5.59}$$

As shown in Table 3.7, the theoretical prediction for ^{7}Li/H at the CMB-calibrated baryon density is a factor of ~ 3 higher than measured on the Spite plateau, representing a persistent 4.5σ tension known as the **Cosmological Lithium Problem**. Active hypotheses include stellar

Table 3.7: Comparison of Standard BBN Abundance Predictions at CMB Baryon Density ($\Omega_{b0}h^2 = 0.0224$) with Observational Determinations

Abundance Ratio	Theoretical Prediction (SBBN)	Observational Determination	Agreement
Y_p (^{4}He mass fraction)	0.2469 ± 0.0002	0.245 ± 0.003	Excellent ($< 1\sigma$)
D/H $\times 10^5$	2.58 ± 0.13	2.54 ± 0.04	Spectacular ($< 0.5\sigma$)
^{3}He/H $\times 10^5$	1.00 ± 0.07	1.1 ± 0.2 (Galactic H II)	Good ($< 1\sigma$)
^{7}Li/H $\times 10^{10}$	4.68 ± 0.32	1.58 ± 0.11 (Spite Plateau)	Discrepant (4.5σ)

atmospheric depletion (rotational mixing and gravitational settling diffusing lithium below the convective envelope), subtle nuclear cross-section resonances, or beyond-the-Standard-Model physics (such as decaying supersymmetric particles or dark matter annihilation during BBN).

3.6 The Cosmic Microwave Background Radiation

3.6.1 Historical prediction and discovery

The existence of a relic radiation bath is an inevitable consequence of an expanding, cooling FLRW universe that was once hot enough to sustain nuclear reactions. In 1948, Ralph Alpher and Robert Herman, working within George Gamow's primordial nucleosynthesis framework, predicted that the thermal radiation from the early fiery cosmos should survive today as an isotropic background of redshifted microwave photons:

$$T_0 \approx 5\,\mathrm{K}. \tag{3.6.1}$$

In 1965, Arno Penzias and Robert Wilson, using the Bell Laboratories Crawford Hill horn antenna at Holmdel, New Jersey, detected an isotropic, unpolarized excess antenna temperature of $3.5 \pm 1.0\,\mathrm{K}$ at wavelength $\lambda = 7.35\,\mathrm{cm}$ that could not be attributed to instrumental noise, atmospheric emission, or Galactic synchrotron sources. Simultaneously, Robert Dicke, Jim Peebles, Peter Roll, and David Wilkinson at Princeton University were constructing a radiometer to search for precisely this cosmological relic. The discovery established the hot Big Bang paradigm over the competing Steady-State model, earning Penzias and Wilson the 1978 Nobel Prize in Physics.

3.6.2 Thermodynamic recombination and the Saha equation

Prior to neutral atom formation, the Universe consisted of an optically thick, fully ionized plasma of free electrons, protons, helium nuclei, and photons. Thermal equilibrium was enforced by rapid electromagnetic processes:

$$p + e^- \longleftrightarrow \mathrm{H} + \gamma. \tag{3.6.2}$$

Because photons are continually absorbed and emitted, their chemical potential is zero ($\mu_\gamma = 0$). Chemical equilibrium under reaction (3.6.2) requires:

$$\mu_p + \mu_e = \mu_\mathrm{H}. \tag{3.6.3}$$

Because the temperature is far below the rest masses of the nucleons and electrons ($T \ll m_e \ll m_p$), the particles obey non-relativistic Maxwell–Boltzmann statistics:

$$n_i = g_i \left(\frac{m_i T}{2\pi}\right)^{3/2} \exp\left(\frac{\mu_i - m_i}{T}\right). \tag{3.6.4}$$

Forming the ratio $n_p n_e/n_{\rm H}$:

$$\frac{n_p n_e}{n_{\rm H}} = \frac{g_p g_e}{g_{\rm H}} \left(\frac{m_p m_e}{m_{\rm H}} \frac{T}{2\pi}\right)^{3/2} \exp\left(-\frac{m_p + m_e - m_{\rm H}}{T}\right). \tag{3.6.5}$$

The internal spin degeneracies are $g_p = 2$, $g_e = 2$, and for neutral hydrogen in the ground state (proton spin 1/2 and electron spin 1/2), $g_{\rm H} = 2 \times 2 = 4$, so the statistical prefactor is $(g_p g_e)/g_{\rm H} = (2 \times 2)/4 = 1$. The mass factor simplifies via $m_p/m_{\rm H} \approx 1$, and the mass difference in the exponent is the ground-state hydrogen ionization binding energy:

$$E_I \equiv m_p + m_e - m_{\rm H} = 13.606\,\text{eV}. \tag{3.6.6}$$

This yields the fundamental equilibrium ratio:

$$\frac{n_p n_e}{n_{\rm H}} = \left(\frac{m_e T}{2\pi}\right)^{3/2} \mathrm{e}^{-E_I/T}. \tag{3.6.7}$$

Cosmic electric charge neutrality guarantees $n_p = n_e$. Neglecting the small helium fraction for clarity, the total baryon number density is $n_b = n_p + n_{\rm H}$. Defining the **fractional ionization**:

$$X_e \equiv \frac{n_e}{n_b} = \frac{n_p}{n_b}, \qquad \frac{n_{\rm H}}{n_b} = 1 - X_e, \tag{3.6.8}$$

we obtain the classical **Saha ionization equation**:

$$\frac{X_e^2}{1 - X_e} = \frac{1}{n_b} \left(\frac{m_e T}{2\pi}\right)^{3/2} \mathrm{e}^{-E_I/T}. \tag{3.6.9}$$

Expressing the baryon density in terms of the photon density $n_b = \eta n_\gamma = \eta \frac{2\zeta(3)}{\pi^2} T^3$:

$$\frac{X_e^2}{1 - X_e} = \frac{\pi^2}{2\zeta(3)\eta} \left(\frac{m_e}{2\pi T}\right)^{3/2} \mathrm{e}^{-E_I/T}. \tag{3.6.10}$$

Substituting $T(z) = T_0(1+z) = 2.348 \times 10^{-4}(1+z)\,\text{eV}$ and $n_b(z) = 2.47 \times 10^{-21} \Omega_{b0} h^2 (1+z)^3\,\text{GeV}^3$ yields the explicit logarithmic form:

$$\log_{10}\left(\frac{X_e^2}{1 - X_e}\right) = 20.98 - \log_{10}\left[\Omega_{b0} h^2 (1+z)^{3/2}\right] - \frac{25163}{1+z}. \tag{3.6.11}$$

Recombination redshifts and the photon-to-baryon delay

A naive expectation might suggest that recombination should occur when the temperature drops to the hydrogen binding energy, $T \sim E_I = 13.6\,\text{eV} \approx 158\,000\,\text{K}$. However, the Saha equation (3.6.10) reveals that the right-hand side is multiplied by the inverse baryon-to-photon ratio:

$$\eta^{-1} = \frac{n_\gamma}{n_b} \approx 1.6 \times 10^9. \tag{3.6.12}$$

There are nearly two billion photons for every proton in the Universe! Even when the average thermal energy T is well below 13.6 eV, the vast exponential Wien tail of the Planck distribution contains more than enough ionizing photons ($E_\gamma > 13.6\,\text{eV}$) to photodisintegrate any neutral hydrogen atom that forms. Recombination is delayed until the temperature drops low enough that the fraction of photons with $E > 13.6\,\text{eV}$ falls below η:

$$\mathrm{e}^{-E_I/T} \lesssim \eta \implies T_{\rm rec} \approx \frac{E_I}{\ln(\eta^{-1})} \approx \frac{13.6\,\text{eV}}{\ln(10^9)} \approx \frac{13.6}{21}\,\text{eV} \approx 0.3\,\text{eV} \approx 3500\,\text{K}. \tag{3.6.13}$$

From Eq. (3.6.11), standard equilibrium ionization thresholds correspond to:

- $X_e = 0.50 \implies z_{\rm rec} \approx 1376 \quad (T \approx 3750\,\text{K} \approx 0.32\,\text{eV})$;
- $X_e = 0.10 \implies z_{\rm rec} \approx 1258 \quad (T \approx 3430\,\text{K} \approx 0.30\,\text{eV})$;
- $X_e = 0.01 \implies z_{\rm rec} \approx 1138 \quad (T \approx 3100\,\text{K} \approx 0.27\,\text{eV})$.

3.6.3 Photon decoupling and the surface of last scattering

While the plasma is ionized, photons are trapped by Thomson scattering off free electrons:

$$e^- + \gamma \longleftrightarrow e^- + \gamma. \tag{3.6.14}$$

The Thomson scattering rate is:

$$\Gamma_{\rm T}(z) = n_e(z)\sigma_{\rm T} = 1.495 \times 10^{-31}\, X_e(z) \left(\frac{\Omega_{b0}h^2}{0.02}\right)(1+z)^3\,{\rm cm}^{-1}, \tag{3.6.15}$$

where $\sigma_{\rm T} = \frac{8\pi}{3}\left(\frac{\alpha}{m_e}\right)^2 \approx 6.652 \times 10^{-25}\,{\rm cm}^2$ is the Thomson cross section. The photon mean free path is $\lambda_\gamma = 1/(n_e\sigma_{\rm T})$.

In a matter-and-radiation universe, the expansion rate is:

$$H^2(z) = \Omega_{m0}H_0^2(1+z)^3\left(1 + \frac{1+z}{1+z_{\rm eq}}\right). \tag{3.6.16}$$

The condition for photon decoupling $\Gamma_{\rm T}(z_{\rm dec}) = H(z_{\rm dec})$ leads to:

$$(1+z_{\rm dec})^{3/2} = \frac{280.01}{X_e(\infty)}\left(\frac{\Omega_{b0}h^2}{0.02}\right)^{-1}\left(\frac{\Omega_{m0}h^2}{0.15}\right)^{1/2}\left(1 + \frac{1+z_{\rm dec}}{1+z_{\rm eq}}\right)^{1/2}. \tag{3.6.17}$$

When the true non-equilibrium residual ionization fraction $X_e \approx 10^{-4}$ is inserted, Eq. (3.6.17) determines the **photon decoupling redshift**:

$$z_{\rm dec} \approx 1090, \qquad T_{\rm dec} \approx 2970\,{\rm K} \approx 0.26\,{\rm eV}, \qquad t_{\rm dec} \approx 380\,000\,{\rm yr}. \tag{3.6.18}$$

3.6.4 Recombination dynamics and the Peebles three-level atom

The equilibrium Saha equation (3.6.9) assumes instantaneous chemical equilibrium. However, Jim Peebles (1968) and Zel'dovich et al. (1968) demonstrated that true recombination is governed by non-equilibrium atomic kinetics.

The Lyman-α and two-photon decay bottlenecks

Direct recombination to the ground state $1s$ ($p + e^- \to {\rm H}(1s) + \gamma$) produces a photon with $h\nu \geq 13.6\,{\rm eV}$. In the dense primordial plasma, this energetic photon travels only a microscopic distance before ionizing another neutral hydrogen atom ($h\nu + {\rm H}(1s) \to p + e^-$), resulting in *zero net recombination*. Recombination can succeed only through capture into excited states ($n \geq 2$), followed by radiative cascade down to $n = 2$. From $n = 2$, transitions to $1s$ face two strict bottleneck channels:

1. **The $2s \to 1s$ Two-Photon Decay:** Single-photon electric dipole emission is forbidden by selection rules ($\Delta l = 0$). Transition occurs via simultaneous emission of two photons with combined energy 10.2 eV, with a slow quantum transition rate:

$$\Lambda_{2s} \approx 8.227\,{\rm s}^{-1}. \tag{3.6.19}$$

2. **Lyman-α Escape ($2p \to 1s$):** Resonance Lyman-α photons (10.2 eV) have an enormous optical depth and are repeatedly re-absorbed. Net transition occurs only when cosmological expansion redshifts the photon out of the resonant absorption profile:

$$\Lambda_{{\rm Ly}\alpha} = \frac{8\pi H(z)}{\lambda_{{\rm Ly}\alpha}^3 n_{1s}}. \tag{3.6.20}$$

Peebles condensed this kinetic network into the effective differential rate equation:

$$\frac{\mathrm{d}X_e}{\mathrm{d}t} = -C_r \left[\alpha^{(2)} n_b X_e^2 - \beta (1 - X_e) \mathrm{e}^{-E_I/T} \right], \tag{3.6.21}$$

where $\alpha^{(2)}$ is the recombination coefficient to excited states:

$$\alpha^{(2)}(T) = \frac{64\sqrt{\pi}}{\sqrt{27}} \frac{e^4}{m_e^2} \left(\frac{T}{E_I} \right)^{-1/2} \phi_2(T), \qquad \phi_2(T) \simeq 0.448 \ln \left(\frac{E_I}{T} \right), \tag{3.6.22}$$

$\beta(T)$ is the photoionization rate:

$$\beta(T) = \left(\frac{m_e T}{2\pi} \right)^{3/2} \alpha^{(2)}(T), \tag{3.6.23}$$

and the Peebles inhibition factor C_r represents the probability that an atom in $n = 2$ reaches $1s$ before photoionization:

$$C_r = \frac{\Lambda_{2s} + \Lambda_{\mathrm{Ly}\alpha}}{\Lambda_{2s} + \Lambda_{\mathrm{Ly}\alpha} + \beta}. \tag{3.6.24}$$

Table 3.8: Comparison of Fractional Ionization $X_e(z)$ Predicted by the Equilibrium Saha Equation vs Peebles Multi-Level Recombination Kinetics ($\Omega_{b0}h^2 = 1$)

Redshift z	**Saha Equilibrium** X_e^{eq}	**Peebles Kinetics** $X_e(z)$	**Physical Regime**
1500	0.999	0.999	Fully Ionized Plasma
1300	0.512	0.620	Recombination Onset
1100	1.8×10^{-2}	7.2×10^{-2}	Decoupling (X_e lags Saha)
900	2.4×10^{-5}	4.1×10^{-3}	Freeze-out Departure
500	1.0×10^{-10}	2.1×10^{-4}	Complete Freeze-out
0	1.0×10^{-30}	1.2×10^{-4}	Residual Ionization Floor

As shown in Table 3.8, the true ionization fraction lags far behind the Saha prediction and eventually freezes out at a residual plateau:

$$X_e(\infty) \approx 10^{-4}\text{–}10^{-3}. \tag{3.6.25}$$

In the range $1500 > z > 800$, the kinetics simplify to:

$$\frac{\mathrm{d}X_e}{\mathrm{d}z} = -61.72 \, \frac{\Omega_{b0}h^2}{\sqrt{\Omega_{m0}h^2}} \, f(z) \left[1 + 2.26 \times 10^4 \, z \, \mathrm{e}^{-14486/(z\Theta_{2.7})} \right]^{-1} X_e^2, \tag{3.6.26}$$

yielding the accurate power-law fit:

$$X_e(z) \simeq 2.74 \times 10^{-3} \, \mathrm{e}^{14.486(1-\Theta_{2.7}^{-1})} \frac{\sqrt{\Omega_{m0}h^2}}{\Omega_{b0}h^2} \left(\frac{z}{1000} \right)^{12.75}. \tag{3.6.27}$$

3.6.5 The optical depth and the Last Scattering Surface

The Thomson optical depth from redshift z to today ($z = 0$) is:

$$\tau(z) = \int_0^{\chi(z)} n_e \sigma_{\rm T}\, \mathrm{d}\chi \simeq 0.42\, \mathrm{e}^{14.486(1-\Theta_{2.7}^{-1})} \left(\frac{z}{1000}\right)^{14.25}. \tag{3.6.28}$$

The probability distribution that a detected photon scattered for the final time between z and $z + \mathrm{d}z$ is quantified by the **visibility function**:

$$g(z) \equiv \mathrm{e}^{-\tau(z)} \frac{\mathrm{d}\tau}{\mathrm{d}z}. \tag{3.6.29}$$

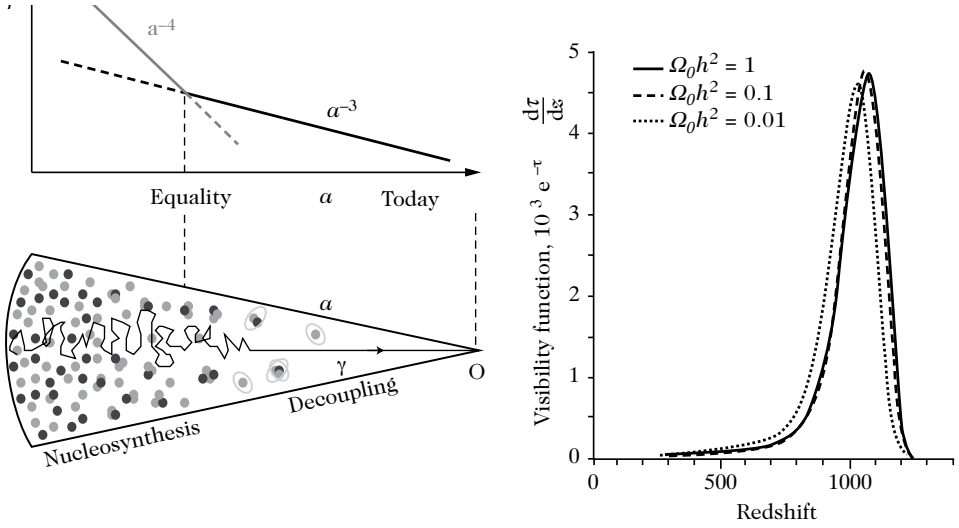


Figure 3.13: (Left): Thermal history of cosmic energy density $\rho(a)$ illustrating the matter-radiation equality, nucleosynthesis, and recombination epochs. (Right): The CMB visibility function $g(z) = \mathrm{e}^{-\tau} \frac{\mathrm{d}\tau}{\mathrm{d}z}$ plotted across redshift, showing a sharp peak that defines the Last Scattering Surface (LSS) at $z_{\rm dec} \approx 1089$.

The visibility function forms a sharp Gaussian-like peak (Figure 3.13), defining the **Last Scattering Surface (LSS)**:

$$z_{\rm dec} = 1089 \pm 1, \qquad \Delta z_{\rm dec} = 195 \pm 2, \qquad T_{\rm dec} \approx 2970\,\mathrm{K}. \tag{3.6.30}$$

The LSS represents a spherical shell centered on the observer at comoving distance $\chi(z_{\rm dec}) \approx 14\,\mathrm{Gpc}$.

3.6.6 Observed properties of the CMB spectrum

Precision blackbody spectrum and cosmological energy densities

The Far-Infrared Absolute Spectrophotometer (FIRAS) aboard COBE measured the absolute spectrum of the CMB across 60–600 GHz, establishing a perfect Planck blackbody distribution:

$$I(\nu, T_0) = \frac{2h\nu^3}{c^2} \frac{1}{\mathrm{e}^{h\nu/k_{\rm B}T_0} - 1}, \tag{3.6.31}$$

with sky-averaged monopole temperature:

$$T_0 = 2.7255 \pm 0.0006\,\mathrm{K}. \tag{3.6.32}$$

From this temperature, present radiation densities evaluate to:

$$n_{\gamma 0} = \frac{2\zeta(3)}{\pi^2} T_0^3 = 410.44\,\Theta_{2.7}^3\,\mathrm{cm}^{-3}, \tag{3.6.33}$$

$$\rho_{\gamma 0} = \frac{\pi^2}{15} T_0^4 = 4.6408 \times 10^{-34}\,\Theta_{2.7}^4\,\mathrm{g\cdot cm}^{-3} = 0.26032\,\mathrm{eV\cdot cm}^{-3}, \tag{3.6.34}$$

$$s_0 = \frac{2\pi^2}{45} q_{*0} T_0^3 = 2889.2 \left(\frac{q_{*0}}{3.91}\right) \Theta_{2.7}^3\,\mathrm{cm}^{-3} = 7.039 \left(\frac{q_{*0}}{3.91}\right) n_{\gamma 0}, \tag{3.6.35}$$

$$\Omega_{\gamma 0} h^2 = 2.4697 \times 10^{-5}\,\Theta_{2.7}^4, \tag{3.6.36}$$

$$\Omega_{r0} h^2 = \Omega_{\gamma 0} h^2 \left[1 + 3 \times \frac{7}{8} \left(\frac{4}{11}\right)^{4/3}\right] = 4.148 \times 10^{-5}\,\Theta_{2.7}^4. \tag{3.6.37}$$

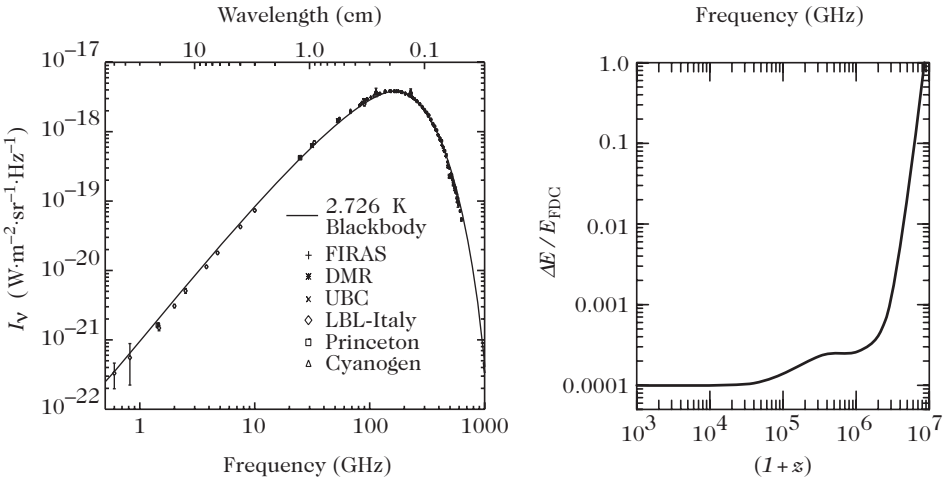


Figure 3.14: (Left): Comparison between the measured CMB spectrum and a perfect blackbody at $T_0 = 2.725\,\mathrm{K}$. (Right): Upper bounds on fractional energy injection $\Delta E/E_{\mathrm{FDC}}$ as a function of redshift $(1+z)$ from FIRAS spectral distortion constraints.

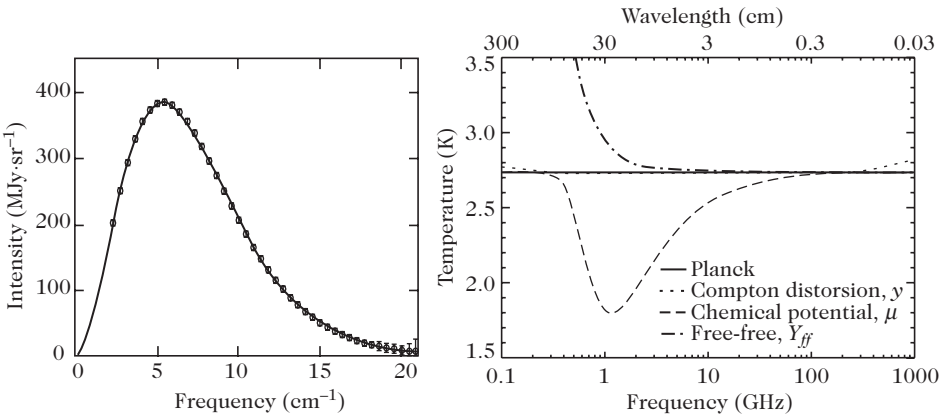


Figure 3.15: (Left): Spectrum of the cosmic microwave background measured by FIRAS/COBE (error bars multiplied by 200). (Right): Characteristic spectral distortions (Compton y, chemical potential μ, and free-free Y_{ff}).

Spectral distortions of the Planck distribution

Energy injection into the cosmic plasma creates characteristic distortions (Figure 3.15):

- **Compton y-distortion ($z \lesssim 10^4$):** Produced when hot electrons Compton-scatter cold photons, shifting photons from the Rayleigh–Jeans tail to the Wien tail:

$$y \equiv \int \frac{k_{\rm B}(T_e - T_\gamma)}{m_e c^2} n_e \sigma_{\rm T} c \, {\rm d}t \,. \tag{3.6.38}$$

- **Chemical potential μ-distortion ($10^4 \lesssim z \lesssim 2 \times 10^6$):** Elastic Compton scattering is rapid enough to establish kinetic equilibrium but photon-producing processes (Bremsstrahlung, double Compton) are too slow to adjust photon number, establishing a Bose–Einstein distribution with non-zero μ:

$$f(E) = \frac{1}{{\rm e}^{E/T+\mu} - 1}. \tag{3.6.39}$$

- **Free-free distortion ($Y_{\rm ff}$):** Low-frequency bremsstrahlung absorption and emission:

$$Y_{\rm ff} \equiv \int \frac{T_e - T_\gamma}{T_e} \frac{8\pi e^6 h^2 n_e^2 g_{\rm ff}}{3 m_e (2\pi m_e k_{\rm B} T_e)^{1/2}} \, {\rm d}t \,. \tag{3.6.40}$$

FIRAS limits constrain energy injections to less than 10^{-4} of the CMB energy:

$$|y| < 1.5 \times 10^{-5}, \qquad |\mu| < 9 \times 10^{-5}, \qquad |Y_{\rm ff}| < 1.9 \times 10^{-5} \quad (95\%\,{\rm CL}). \tag{3.6.41}$$

3.6.7 Kinematic dipole and galactic foregrounds

The kinematic dipole

The largest temperature anisotropy in the CMB sky is a $\ell = 1$ dipole of amplitude $\Delta T \approx 3.36\,{\rm mK}$, arising from the Doppler boost of our local reference frame:

$$T(\theta) = T_0 \frac{\sqrt{1-\beta^2}}{1 - \beta\cos\theta} \approx T_0 \left(1 + \beta\cos\theta + \frac{1}{2}\beta^2 \cos 2\theta\right). \tag{3.6.42}$$

We redefine the temperature to absorb the doppler shift of velocity

$$\nu_{\rm obs} = \nu\gamma(1 - \vec{\beta}\cdot\hat{n}) \tag{1}$$

FIRAS and WMAP measured the solar system barycenter velocity relative to the CMB frame:

$$v_\odot = 369.0 \pm 0.9\,{\rm km\cdot s^{-1}} \quad \text{toward } (l,b) = (263.99^\circ \pm 0.14^\circ, 48.26^\circ \pm 0.03^\circ). \tag{3.6.43}$$

Subtracting the Sun's galactic orbital motion ($v_{\rm circ} \approx 220\,{\rm km\cdot s^{-1}}$) gives the Local Group velocity:

$$v_{\rm LG} = 627 \pm 22\,{\rm km\cdot s^{-1}} \quad \text{toward } (l,b) = (276^\circ \pm 3^\circ, 30^\circ \pm 3^\circ), \tag{3.6.44}$$

driven by gravitational attraction toward the Great Attractor and Shapley Supercluster.

After having subtracted these effects, some temperature anisotropies remain, with elative amplitude $\sim 10^{-5}$, which correspond to temperature fluctuations of the order of 30 μK. These anisotropies correspond to anisotropies in the cosmic microwave background at the time of recombination and redshifted by the cosmic expansion. Chapter 6 is dedicated to the study of these temperature anisotropies.

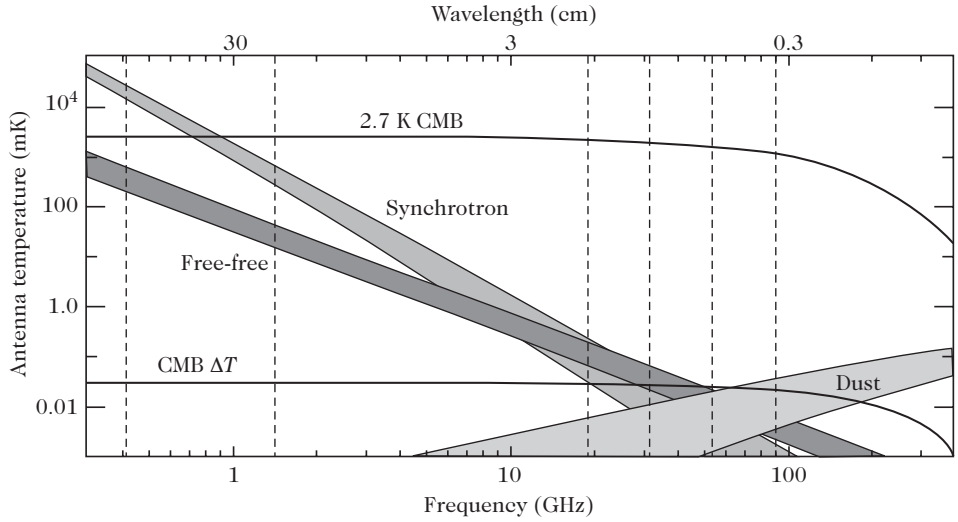


Figure 3.16: Frequency spectra of Galactic foreground emissions (synchrotron, free-free Bremsstrahlung, and thermal dust) compared with CMB primary temperature fluctuations $\Delta T \approx 30\,\mu$K. The minimum foreground contamination window occurs around 70–100 GHz.

3.6.8 Direct spectroscopic confirmation of cosmic cooling $T(z) = T_0(1+z)$

General relativity predicts that adiabatic expansion cools the photon temperature linearly with redshift:

$$T(z) = T_0(1+z) = 2.7255(1+z)\,\mathrm{K}. \tag{3.6.45}$$

This prediction has been directly verified observationally by measuring excitation temperatures of neutral carbon (C I) and carbon monoxide (CO) in high-redshift absorption systems illuminated by background quasars (Figure 3.17).

At redshift $z = 2.33771$, Srianand et al. (2000) measured $6.0\,\mathrm{K} \leq T \leq 14.0\,\mathrm{K}$, in striking agreement with the prediction $T(2.33771) = 9.1\,\mathrm{K}$.

3.6.9 The Dark Age and Baryon Thermal Decoupling

What would the Universe look like to an observer following recombination? Immediately after photon decoupling ($z \approx 1090$), the CMB photons were in the visible spectrum, filling all of space with a uniform, warm reddish glow at $T \approx 3000\,\mathrm{K}$. Over the following few million years, cosmological expansion stretched these photons into the infrared, causing the cosmic glow to fade completely. Space became entirely dark, entering the multi-hundred-million-year epoch known as the **Dark Age**.

During this era, matter and radiation experienced a subtle, asymmetric thermal decoupling. Although photons decoupled from baryons at $z_{\rm dec} \approx 1090$, the huge ratio of photons to baryons ($n_\gamma/n_b \sim 10^9$) meant that even a tiny residual ionization fraction ($X_e \sim 10^{-4}$) allowed residual Compton scattering to transfer significant energy to the baryons. The Compton scattering heating rate per gas particle is:

$$\Gamma_{\rm Comp} = \frac{4}{3}\frac{\sigma_{\rm T}\rho_\gamma}{m_e}X_e. \tag{3.6.46}$$

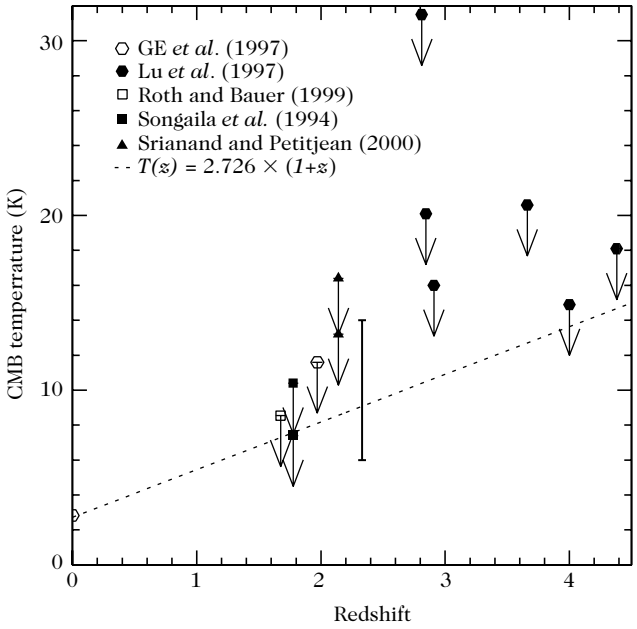


Figure 3.17: Spectroscopic measurements of the CMB temperature $T(z)$ as a function of redshift from quasar absorption systems (C I and CO molecular transitions). The dashed line represents the standard adiabatic expansion prediction $T(z) = 2.7255(1+z)\,\mathrm{K}$.

As long as $\Gamma_{\mathrm{Comp}} \gg H(z)$, Compton scattering locks the kinetic temperature of the baryonic gas T_b to the photon temperature:

$$T_b(z) = T_\gamma(z) \propto (1+z) \quad (z \gtrsim 150\text{–}200). \tag{3.6.47}$$

However, as radiation density diluted as $\rho_\gamma \propto (1+z)^4$, the Compton heating rate fell below the Hubble rate at $z_{\mathrm{comp}} \approx 150$–$200$. Below this redshift, baryons decoupled thermally from the radiation bath. Once decoupled, the non-relativistic monatomic hydrogen and helium gas cooled adiabatically according to its ideal-gas adiabatic index $\gamma = 5/3$:

$$T_b(a) \propto a^{-3(\gamma-1)} = a^{-2} \propto (1+z)^2 \quad (z \lesssim 150). \tag{3.6.48}$$

Consequently, while the radiation temperature continued to cool linearly as $T_\gamma \propto (1+z)$, the baryonic gas cooled much faster. By redshift $z \sim 20$, the radiation temperature was $T_\gamma \approx 2.73 \times 21 \approx 57\,\mathrm{K}$, whereas the baryon gas had plummeted to:

$$T_b(z=20) \approx T_{\mathrm{comp}} \left(\frac{1+20}{1+z_{\mathrm{comp}}} \right)^2 \approx 400\,\mathrm{K} \left(\frac{21}{150} \right)^2 \approx 7.8\,\mathrm{K}. \tag{3.6.49}$$

This pronounced temperature difference ($T_b \ll T_\gamma$) is the cornerstone of modern 21-cm cosmology. Neutral hydrogen gas in the intergalactic medium absorbs CMB photons at the 21 cm spin-flip hyperfine transition line ($\nu_0 = 1420.4\,\mathrm{MHz}$), creating an absorption feature against the CMB sky that serves as the premier observational probe of the Dark Ages until the first stars ignited (Cosmic Dawn at $z \sim 20$–30) and reionized the intergalactic medium ($z_{\mathrm{reion}} \sim 6$–$8$).

3.7 The Standard Cosmological Paradigm: Summary and Open Problems

3.7.1 Summary of cosmological parameters and thermal history

The Standard Cosmological Model (ΛCDM) is underpinned by three foundational observational pillars:

1. **Hubble Cosmic Expansion:** Directly established through supernovae, Cepheids, and distance-redshift relations.
2. **Big-Bang Nucleosynthesis:** The primordial origin and abundance of light isotopes (D,^{3}He,^{4}He,^{7}Li).
3. **Cosmic Microwave Background:** The pristine 2.7255 K blackbody relic bath and its primary anisotropies.

The concordant cosmological parameters and thermal history milestones are summarized in Tables 3.9 and 3.10.

Table 3.9: Fundamental Cosmological Parameters of the Concordance ΛCDM Model

Cosmological Parameter	**Symbol**	**Best-Fit Concordance Value**
Hubble Constant	H_0	$67.4 \pm 0.5\,\mathrm{km \cdot s^{-1} \cdot Mpc^{-1}}$
Reduced Hubble Parameter	h	0.674 ± 0.005
Total Matter Density Parameter	Ω_{m0}	0.315 ± 0.007
Physical Cold Dark Matter Density	$\Omega_{c0}h^2$	0.120 ± 0.001
Physical Baryon Density	$\Omega_{b0}h^2$	0.02237 ± 0.00015
Dark Energy Density Parameter	$\Omega_{\Lambda 0}$	0.685 ± 0.007
Spatial Curvature Parameter	Ω_{K0}	0.0007 ± 0.0019 (Flat)
Dark Energy Equation of State	w_0	-1.03 ± 0.03 (Λ)
CMB Monopole Temperature	$T_{\gamma 0}$	$2.7255 \pm 0.0006\,\mathrm{K}$
Baryon-to-Photon Ratio	η	$(6.12 \pm 0.04) \times 10^{-10}$
Effective Relativistic Neutrino Species	N_{eff}	3.044 (SM Prediction)
Age of the Universe	t_0	$13.787 \pm 0.020\,\mathrm{Gyr}$

3.7.2 Classic problems of the standard Big-Bang paradigm

Despite the extraordinary empirical success of the Hot Big Bang model, the classical standard paradigm contains deep internal cosmological puzzles. These problems are not observational contradictions; rather, they demonstrate that the Standard Model requires extraordinarily unnatural, fine-tuned initial conditions if the early Universe was purely radiation- or matter-dominated.

The Flatness / Curvature Problem

The Friedmann equation can be recast in terms of the total density parameter $\Omega(t) \equiv \rho(t)/\rho_{\mathrm{crit}}(t)$:

$$1 - \Omega(t) = -\frac{K}{a^2 H^2} = -\frac{K}{\dot{a}^2}. \tag{3.7.1}$$

Differentiating with respect to time:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left|1-\Omega\right| = -2\frac{K}{\dot{a}^3}\ddot{a} = 2\left|1-\Omega\right|\left(-\frac{\ddot{a}}{\dot{a}}\right). \tag{3.7.2}$$

Table 3.10: Key Physical Milestones in the Thermal History of the Universe

Epoch / Event	**Physical Mechanism**	**Redshift z**	**Temperature / Energy**
Planck Era	Quantum Gravity Limit	$z \sim 10^{32}$	$T \sim 10^{19}\,\text{GeV}$
GUT Transition	Gauge Symmetry Breaking	$z \sim 10^{28}$	$T \sim 10^{15}$–$10^{16}\,\text{GeV}$
Electroweak Transition	Higgs VEV Generation	$z \sim 10^{15}$	$T \sim 100$–$160\,\text{GeV}$
Quark–Hadron Transition	QCD Confinement (QGP $\to$ Hadrons)	$z \sim 10^{12}$	$T \sim 150$–$200\,\text{MeV}$
Neutrino Decoupling	Weak Interactions Freeze Out	$z \approx 4 \times 10^9$	$T \approx 1$–$2\,\text{MeV}$
e^+e^- Annihilation	Entropy Reheating of Photons	$z \approx 2 \times 10^9$	$T \approx 0.511\,\text{MeV}$
Big-Bang Nucleosynthesis	Light Elements Synthesized	$z \approx 3 \times 10^8$	$T \approx 0.08\,\text{MeV}$
Matter–Radiation Equality	$\rho_m(z_{\text{eq}}) = \rho_r(z_{\text{eq}})$	$z_{\text{eq}} \approx 3400$	$T_{\text{eq}} \approx 0.8\,\text{eV} \approx 9500\,\text{K}$
Recombination	Neutral Hydrogen Forms ($X_e = 0.5$)	$z_{\text{rec}} \approx 1370$	$T_{\text{rec}} \approx 0.32\,\text{eV} \approx 3750\,\text{K}$
Photon Decoupling	Thomson Scattering Freezes Out	$z_{\text{dec}} \approx 1090$	$T_{\text{dec}} \approx 0.26\,\text{eV} \approx 2970\,\text{K}$
Dark Ages	Baryons Decouple from Photons	$z \sim 150$–20	$T \sim 400$–$8\,\text{K}$
Cosmic Reionization	First Stars Ionize IGM	$z_{\text{reion}} \approx 6$–$8$	$T \sim 10^4\,\text{K}$ (Gas)
Present Epoch	Dark Energy Dominates	$z = 0$	$T_0 = 2.7255\,\text{K}$

In any decelerating universe ($\ddot{a} < 0$), the expansion rate $\dot{a}$ decreases over time. During radiation domination, $a(t) \propto t^{1/2} \implies \dot{a} \propto a^{-1}$, so:

$$|1 - \Omega(a)| \propto a^2 \quad \text{(Radiation Domination)}. \tag{3.7.3}$$

During matter domination, $a(t) \propto t^{2/3} \implies \dot{a} \propto a^{-1/2}$, so:

$$|1 - \Omega(a)| \propto a \quad \text{(Matter Domination)}. \tag{3.7.4}$$

Equation (3.7.4) proves that the flat fixed point $\Omega = 1$ ($\Omega_K = 0$) is dynamically *unstable* under forward cosmic expansion. Any infinitesimal departure from spatial flatness at early times is exponentially amplified by the subsequent expansion:

$$|1 - \Omega(t)| = |1 - \Omega_0| \left(\frac{a_0 H_0}{aH} \right)^2 . \tag{3.7.5}$$

Given the contemporary empirical constraint $|1 - \Omega_0| < 0.005$ from Planck:

$$|1 - \Omega(z_{\text{eq}})| < 3 \times 10^{-5} \quad \text{(at matter-radiation equality } z_{\text{eq}} \approx 3400), \tag{3.7.6}$$

$$|1 - \Omega(T_{\text{BBN}})| < 10^{-16} \quad \text{(at nucleosynthesis } T \sim 1\,\text{MeV}), \tag{3.7.7}$$

$$|1 - \Omega(T_{\text{GUT}})| < 10^{-55} \quad \text{(at the GUT scale } T \sim 10^{16}\,\text{GeV}), \tag{3.7.8}$$

$$|1 - \Omega(t_{\text{Pl}})| < 10^{-60} \quad \text{(at the Planck epoch } t \sim 10^{-43}\,\text{s}). \tag{3.7.9}$$

Without a dynamical explanation, the cosmic expansion must have been fine-tuned to within 60 decimal places at the Planck era.

The Horizon / Causality Problem

The maximum physical distance a light signal could have traveled since the Big Bang is the comoving **particle horizon**:

$$\chi_{\rm hor}(t) = \int_0^t \frac{c\,{\rm d}t'}{a(t')} = \int_0^a \frac{c\,{\rm d}a'}{a'^2 H(a')}. \tag{3.7.10}$$

In a radiation-dominated universe, $\chi_{\rm hor}(a) \propto a$; in a matter-dominated universe, $\chi_{\rm hor}(a) \propto a^{1/2}$. In both cases, the comoving horizon shrinks toward zero as $a \to 0$.

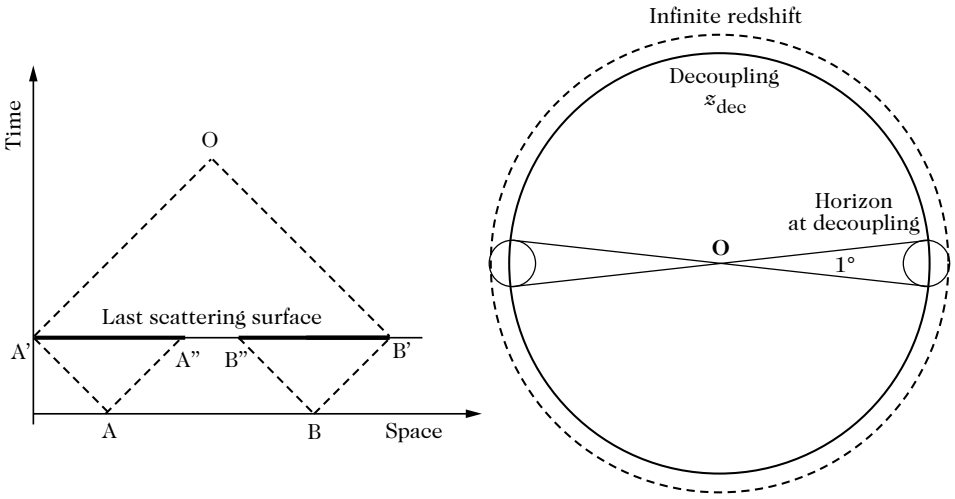


Figure 3.18: (Left): Spacetime conformal diagram showing the past light cone of the observer O intersecting the Last Scattering Surface (LSS) on a sphere of diameter $A'B'$. The causally connected regions $(A'A'', \dots)$ at decoupling are determined by the future light cones from the Big Bang singularity. (Right): Geometry of causally disconnected patches on the CMB sky.

At the epoch of photon decoupling ($z_{\rm dec} \approx 1090$), the physical particle horizon was:

$$d_H(t_{\rm dec}) = a(t_{\rm dec})\chi_{\rm hor}(t_{\rm dec}) \approx 2cH^{-1}(t_{\rm dec}) \approx 0.2\,{\rm Mpc}. \tag{3.7.11}$$

Subtending this horizon across the angular diameter distance to the last scattering surface $D_A(z_{\rm dec}) \approx 12.8\,{\rm Mpc}$ yields the angular scale of causally connected patches on the CMB sky:

$$\theta_{\rm hor} = \frac{d_H(t_{\rm dec})}{D_A(z_{\rm dec})} \approx 1.1^\circ. \tag{3.7.12}$$

Two points on the CMB sky separated by more than $\approx 1.1^\circ$ were causally disconnected at decoupling: their past light cones do not intersect anywhere in the history of the standard Big Bang (Figure 3.18). The full 4π celestial sphere contains:

$$N_{\rm patches} = \frac{4\pi}{\pi\theta_{\rm hor}^2} \approx 40\,000 \text{ causally disconnected patches.} \tag{3.7.13}$$

Extrapolated to the Planck epoch, the number of causally disconnected regions in our observable Universe was:

$$N_{\rm Planck} \sim 10^{83}. \tag{3.7.14}$$

Yet, the measured CMB temperature is isotropic across the entire sky to one part in 10^5 ($\Delta T/T \approx 10^{-5}$)! In the standard paradigm, this miraculous uniformity requires 40,000 causally disjoint regions to have had identical temperatures without ever establishing thermal contact.

The Origin of Inhomogeneities and the Trans-Planckian Problem

The standard Big Bang model provides no mechanism for generating the initial density perturbations that seeded galaxies and galaxy clusters. In the standard paradigm, the comoving Hubble horizon $\mathcal{H}^{-1} = (aH)^{-1} \propto a$ grew monotonically. Consequently, all perturbation wavelengths observed today ($\lambda_{\rm phys} = a\lambda$) were *larger than the causal horizon* ($\lambda_{\rm phys} > H^{-1}$) at early times (Figure 3.19). In general relativity, no causal physical process can create coherent super-horizon perturbations, demanding that the seeds of all cosmological structure be postulated by hand as acausal initial conditions.

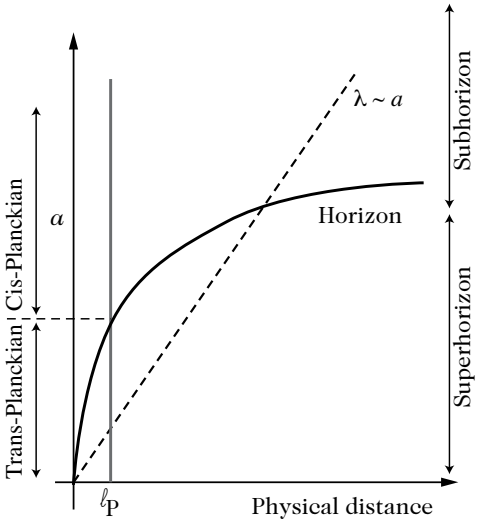


Figure 3.19: Evolution of physical perturbation wavelengths $\lambda_{\rm phys} \propto a$ compared to the causal Hubble horizon H^{-1} in standard Big-Bang expansion. All cosmological scales observed today originated with super-horizon wavelengths ($\lambda > H^{-1}$) and sub-Planckian sizes ($\lambda < \ell_{\rm Pl}$) at early times, defining the horizon and trans-Planckian problems.

Furthermore, tracing the physical wavelength of our present horizon ($\lambda_{\rm phys}(t_0) \approx 14\,{\rm Gpc}$) backward to the Planck epoch yields:

$$\lambda_{\rm phys}(t_{\rm Pl}) = a(t_{\rm Pl})\lambda_{\rm phys}(t_0) \sim 10^{-32} \times 10^{26}\,{\rm m} \sim 10^{-6}\,{\rm m} \gg \ell_{\rm Pl}. \tag{3.7.15}$$

Conversely, the physical wavelengths of cosmic structures traced back to the beginning of standard expansion were sub-Planckian ($\lambda < \ell_{\rm Pl} \approx 1.6 \times 10^{-35}\,{\rm m}$), raising the **Trans-Planckian Problem**: standard cosmological predictions rely on classical field theory on scales where spacetime itself must be quantized.

The Relic / Magnetic Monopole Problem

When Grand Unified Theories (GUTs) break down at $T_{\rm GUT} \sim 10^{15}$–$10^{16}\,{\rm GeV}$ via $\mathcal{G} \to \mathcal{H} \times {\rm U}(1)$, topological defects are inevitably formed via the Kibble mechanism. Because the Higgs field orientation is uncorrelated across causally disconnected regions, approximately one topological magnetic monopole of mass $M_M \sim M_{\rm GUT}/\alpha_{\rm GUT} \sim 10^{17}\,{\rm GeV}$ is produced per causal horizon volume:

$$n_M(t_{\rm GUT}) \sim H^3(t_{\rm GUT}). \tag{3.7.16}$$

The resulting monopole-to-photon ratio is:

$$\frac{n_M}{n_\gamma} \sim \left(\frac{T_{\rm GUT}}{M_{\rm Pl}}\right)^3 \sim 10^{-9}. \tag{3.7.17}$$

Because monopoles are stable and annihilate inefficiently, their energy density dilutes as matter ($\rho_M \propto a^{-3}$). Traced to the present day, their predicted cosmological density would be:

$$\Omega_{M0} \sim 10^{14}, \tag{3.7.18}$$

which would cause the Universe to collapse within a few thousand years, in flagrant contradiction with observation ($\Omega_{M0} < 10^{-6}$).

The Dark Sector, Cosmological Constant, and Coincidence Problems

The concordance ΛCDM model requires two mysterious components:

- **The Cosmological Constant Problem:** The observed vacuum energy density $\rho_\Lambda = \Lambda/(8\pi G) \approx (2.3 \times 10^{-3}\,\mathrm{eV})^4$ is a factor of 10^{120} smaller than the naive Planck scale vacuum energy density $\rho_{\mathrm{vac}}^{\mathrm{Pl}} \sim M_{\mathrm{Pl}}^4 \sim (10^{19}\,\mathrm{GeV})^4$, and a factor of 10^{56} smaller than the electroweak scale.
- **The Cosmic Coincidence Problem:** While matter dilutes as $\rho_m \propto a^{-3}$, vacuum energy remains constant ($\rho_\Lambda \propto a^0$). Throughout the 13.8 Gyr history of the Universe:

$$\frac{\rho_m(a)}{\rho_\Lambda(a)} \propto a^{-3}. \tag{3.7.19}$$

  In the deep past, $\rho_m \gg \rho_\Lambda$; in the distant future, $\rho_\Lambda \gg \rho_m$. Why do we live at the single brief cosmic epoch where $\Omega_{m0} \sim \Omega_{\Lambda 0}$ to within a factor of two?

3.7.3 Synthesis and transition to the inflationary paradigm

The flatness, horizon, monopole, and perturbation-seed problems share a common mathematical origin: *the decelerating nature of standard cosmic expansion* ($\ddot{a} < 0$). In decelerating expansion:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{aH}\right) > 0. \tag{3.7.20}$$

The comoving Hubble horizon $(aH)^{-1}$ monotonically grows, continually bringing new, uncoordinated regions into causal contact, while spatial flatness is destabilized.

To resolve these foundational paradoxes, Alan Guth (1981), Alexei Starobinsky (1980), and Andrei Linde (1982) realized that the early Universe must have undergone a brief epoch of **accelerated expansion**:

$$\ddot{a} > 0 \iff \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{aH}\right) < 0. \tag{3.7.21}$$

During this period, known as **cosmic inflation**, the comoving Hubble horizon shrank exponentially. A single microscopic causal patch of size $\ll H^{-1}$ was stretched by a factor of at least $\mathrm{e}^{60} \sim 10^{26}$, smoothing out spatial curvature ($|1 - \Omega| \to 0$), establishing universal thermal equilibrium across our entire observable horizon, diluting magnetic monopoles to absolute zero, and generating primordial scale-invariant quantum fluctuations that seeded all cosmic structure. The detailed physics of cosmic inflation is formulated in Chapter **??**.

Part III

The Inhomogeneous Universe

Chapter 4

The Inhomogeneous Universe: Perturbations and Their Evolution

Spacetimes of the Friedmann–Lemaître type serve as idealized cosmological backgrounds that provide an accurate description of the cosmos solely when averaged over very large scales. In reality, the observed Universe exhibits prominent hierarchical structures, including galaxies, clusters, and extensive filaments. During the 1980s, it was recognized that this cosmic web originates naturally from the gravitational amplification of tiny primordial density perturbations within a pressureless fluid, namely cold dark matter (Refs. [1–3]).

The central objective of this chapter is to examine spacetimes whose spatial geometry remains in the perturbative vicinity of an idealized Friedmann–Lemaître metric, and to delineate the gravitational evolution of density perturbations within an expanding cosmic medium:

- In Section 4.1, we review the classical Newtonian treatment of fluid perturbations in both static and expanding backgrounds.
- Section 4.2 formulates the full general-relativistic, gauge-invariant cosmological perturbation theory, which forms the theoretical bedrock of modern structure formation.
- Section 4.3 develops explicit dynamical solutions to these perturbation equations across distinct cosmological epochs.
- Section 4.4 analyzes the characteristic physical scales and the formal mathematical structure of the matter power spectrum.
- Section 4.5 reviews the primary observational probes of cosmological large-scale structure.

4.1 Newtonian perturbations

Prior to constructing the complete relativistic theory of cosmic perturbations, it is instructive to explore the simpler regime of fluid perturbations governed by Newtonian gravity in an expanding medium. As demonstrated in Section 5.4, this Newtonian formulation is fully sufficient for describing the late-time clustering of matter on sub-Hubble scales ($k \gg aH$) during the matter-dominated epoch (see Ref. [4] for an extensive review of Newtonian gravitational dynamics).

4.1.1 Case of a static space

In a flat Euclidean background space, fluid dynamics is governed by the standard hydrodynamic equations of continuity and momentum conservation (Euler equation) (Ref. [10]):

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot (\rho \boldsymbol{v}) = 0, \tag{4.1.1}$$

$$\frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v}\cdot\boldsymbol{\nabla})\boldsymbol{v} = -\frac{1}{\rho}\boldsymbol{\nabla}P - \boldsymbol{\nabla}\Phi, \tag{4.1.2}$$

where ρ denotes the fluid mass density, P is the isotropic thermal pressure, $\boldsymbol{v}$ represents the velocity vector field, and Φ is the Newtonian gravitational potential.

Without gravity

Neglecting gravitational interactions ($\Phi = 0$) and decomposing the fluid density and pressure into uniform static background values and small fluctuations, $\rho = \bar{\rho} + \delta\rho$ and $P = \bar{P} + \delta P$, the linear perturbation of Eqs. (4.1.1) and (4.1.2) yields a wave equation:

$$\frac{\partial^2 \delta\rho}{\partial t^2} - \Delta\delta P = 0. \tag{4.1.3}$$

We define the compressibility coefficient as $\chi_a \equiv \rho^{-1}\left(\partial P/\partial\rho\right)_a$, where the index a designates either isothermal (T) or adiabatic (specific entropy S) conditions. When heat conduction operates much faster than the perturbation period, the flow remains isothermal; conversely, when the timescale is insufficient for thermal exchange (as on cosmological wavelengths), the flow is adiabatic. Under adiabatic conditions, Eq. (4.1.3) takes the boxed canonical wave form:

$$\boxed{\frac{\partial^2 \delta\rho}{\partial t^2} - c_{\rm S}^2 \Delta\delta\rho = 0, \qquad c_{\rm S}^2 = \rho\chi_S = \left(\frac{\partial P}{\partial\rho}\right)_S,} \tag{4.1.4}$$

showing that acoustic disturbances propagate at constant amplitude with the adiabatic sound speed $c_{\rm S}$.

Effect of gravity

We now restore gravitational interactions. The gravitational potential is determined by the Poisson equation:

$$\Delta\Phi = 4\pi G_{\rm N}\delta\rho, \tag{4.1.5}$$

which introduces a gravitational self-attraction source term into the wave equation:

$$\frac{\partial^2 \delta\rho}{\partial t^2} - c_{\rm S}^2 \Delta\delta\rho = 4\pi G_{\rm N}\bar{\rho}\,\delta\rho. \tag{4.1.6}$$

Decomposing the density fluctuation into Fourier spatial plane waves, $\delta\rho \propto \exp[i(\omega t - \boldsymbol{k}\cdot\boldsymbol{x})]$, yields the fundamental dispersion relation:

$$\boxed{\omega^2 = \frac{4\pi^2 c_{\rm S}^2}{\lambda_{\rm J}^2}\left(\frac{\lambda_{\rm J}^2}{\lambda^2} - 1\right), \qquad \lambda_{\rm J} \equiv c_{\rm S}\sqrt{\frac{\pi}{G_{\rm N}\bar{\rho}}}.} \tag{4.1.7}$$

Here $\lambda_{\rm J}$ defines the **Jeans length**. For perturbation scales exceeding this threshold ($\lambda > \lambda_{\rm J}$), the squared frequency becomes negative ($\omega^2 < 0$), driving an exponential gravitational instability:

$$\delta\rho \propto \exp\left[\sqrt{4\pi G_{\rm N}\bar{\rho}\left(1 - \frac{\lambda_{\rm J}^2}{\lambda^2}\right)}\, t\right].$$

Physically, the Jeans length demarcates the scale where the acoustic crossing timescale $t_{\rm sound} \sim \lambda/c_{\rm S}$ equals the gravitational free-fall collapse time $t_{\rm grav} \sim (G_{\rm N}\bar{\rho})^{-1/2}$. On small scales ($\lambda < \lambda_{\rm J}$), acoustic pressure gradients react faster than gravity and support stable oscillations; on scales exceeding $\lambda_{\rm J}$, pressure support fails, enabling gravitational collapse.

Energetic aspects

From an energetic standpoint, each fluid parcel can be modeled as a coupled harmonic oscillator with Lagrangian displacement $\boldsymbol{s}$ from its equilibrium position. The velocity is $\dot{\boldsymbol{s}}$, and mass conservation implies $\delta\rho = -\boldsymbol{\nabla}\cdot\boldsymbol{s}$. Integrating the dynamical equation (4.1.7) over a volume V yields the mechanical energy conservation law:

$$\langle E_M\rangle = \langle E_K\rangle + \langle E_E\rangle + \langle E_G\rangle\,, \tag{4.1.8}$$

where $\langle E_K\rangle = \frac{1}{2}\langle\dot{\boldsymbol{s}}^2\rangle$ is the average kinetic energy density, $\langle E_E\rangle = \frac{1}{2}c_{\rm S}^2\langle(\boldsymbol{\nabla}\cdot\boldsymbol{s})^2\rangle$ is the elastic compressive energy density, and $\langle E_G\rangle = -2\pi G_{\rm N}\bar{\rho}\langle\boldsymbol{s}^2\rangle$ is the self-gravitational potential energy density. The kinetic contribution reads:

$$\frac{1}{2}\left\langle\dot{\boldsymbol{s}}^2\right\rangle = \frac{1}{2}\omega^2\left\langle\boldsymbol{s}^2\right\rangle, \qquad \omega^2 = c_{\rm S}^2\left(k^2 - k_{\rm J}^2\right), \qquad k_{\rm J}\equiv\frac{2\pi}{\lambda_{\rm J}}.$$

Consequently, the volume-averaged total mechanical energy evaluates to:

$$\langle E_M\rangle = \frac{1}{2}\left\langle\boldsymbol{s}^2\right\rangle\left[\omega^2 + c_{\rm S}^2\left(k^2 - k_{\rm J}^2\right)\right].$$

By virtue of equipartition enforced by the dispersion relation, the total average energy density simplifies to:

$$\langle E_M\rangle = \left\langle\boldsymbol{s}^2\right\rangle c_{\rm S}^2\left(k^2 - k_{\rm J}^2\right), \tag{4.1.9}$$

which is strictly negative on scales larger than the Jeans length ($k < k_{\rm J}$), proving that gravitationally unstable perturbations correspond to bound physical states.

4.1.2 Case of an expanding space

In an expanding cosmological background, physical spatial positions $\boldsymbol{r}(t)$ are related to comoving coordinates $\boldsymbol{x}$ via the cosmic scale factor $a(t)$:

$$\boldsymbol{r}(t) = a(t)\boldsymbol{x}. \tag{4.1.10}$$

Differentiating with respect to cosmic time t yields the total physical velocity:

$$\boldsymbol{v}(t) = \dot{\boldsymbol{r}} = H\boldsymbol{r} + \boldsymbol{u}, \tag{4.1.11}$$

where $H\equiv\dot{a}/a$ is the Hubble expansion parameter and $\boldsymbol{u}\equiv a\dot{\boldsymbol{x}}$ represents the peculiar velocity field relative to the cosmological Hubble flow.

Eulerian formulation

Transforming spatial and temporal derivatives to comoving coordinates via $\boldsymbol{\nabla}_{\boldsymbol{r}} = a^{-1}\boldsymbol{\nabla}_{\boldsymbol{x}}$ and $\frac{\partial}{\partial t}\Big|_{\boldsymbol{r}} = \frac{\partial}{\partial t}\Big|_{\boldsymbol{x}} - H\boldsymbol{x}\cdot\boldsymbol{\nabla}_{\boldsymbol{x}}$, and using $\boldsymbol{\nabla}_{\boldsymbol{x}}\cdot\boldsymbol{x} = 3$, the hydrodynamic equations take the comoving forms:

$$\dot{\rho}(\boldsymbol{x},t) + 3H\rho(\boldsymbol{x},t) + \frac{1}{a}\boldsymbol{\nabla}_{\boldsymbol{x}}\cdot[\rho(\boldsymbol{x},t)\boldsymbol{u}(\boldsymbol{x},t)] = 0, \tag{4.1.12}$$

and

$$\dot{\boldsymbol{u}}(\boldsymbol{x},t) + H\boldsymbol{u}(\boldsymbol{x},t) + \frac{1}{a}\left(\boldsymbol{u}\cdot\boldsymbol{\nabla}_{\boldsymbol{x}}\right)\boldsymbol{u}(\boldsymbol{x},t) = -\frac{1}{a}\boldsymbol{\nabla}_{\boldsymbol{x}}\Phi(\boldsymbol{x},t) - \frac{1}{a\rho}\boldsymbol{\nabla}_{\boldsymbol{x}}P(\boldsymbol{x},t). \tag{4.1.13}$$

Defining the fractional density contrast $\delta(\boldsymbol{x},t)$ relative to the background density $\bar{\rho}(t)$:

$$\rho(\boldsymbol{x},t) = \bar{\rho}(t)\left[1 + \delta(\boldsymbol{x},t)\right], \tag{4.1.14}$$

the continuity equation (4.1.12) becomes:

$$\dot{\delta} + \frac{1}{a}\boldsymbol{\nabla}\cdot[(1+\delta)\boldsymbol{u}] = 0. \tag{4.1.15}$$

Taking the time derivative of Eq. (4.1.15) and combining it with the divergence of the Euler equation (4.1.13) produces the exact non-linear evolution equation for the density contrast:

$$\ddot{\delta} + 2H\dot{\delta} = \frac{1}{\rho a^2}\Delta P + \frac{1}{a^2}\boldsymbol{\nabla}\cdot[(1+\delta)\boldsymbol{\nabla}\Phi] + \frac{1}{a^2}\partial_i\partial_j\left[(1+\delta)u^i u^j\right]. \tag{4.1.16}$$

Importantly, Eq. (4.1.16) involves no small-amplitude approximation and is valid even in the non-linear clustering regime. The gravitational potential satisfies the comoving Poisson equation:

$$\Delta\Phi = 4\pi G_{\rm N}\bar{\rho}\,a^2(t)\delta, \tag{4.1.17}$$

whose formal solution is given by:

$$\Phi(\boldsymbol{x},t) = -G_{\rm N}\bar{\rho}\,a^2\int\frac{\mathrm{d}^3\boldsymbol{x}'\,\delta(\boldsymbol{x}',t)}{|\boldsymbol{x}-\boldsymbol{x}'|}. \tag{4.1.18}$$

Lagrangian description

In the Lagrangian framework, we follow fluid parcels along trajectories defined by the displacement field $\boldsymbol{X}(\boldsymbol{q},t)$, linking initial Lagrangian coordinates $\boldsymbol{q}$ to Eulerian positions $\boldsymbol{x}(t)$:

$$\boldsymbol{x}(t) = \boldsymbol{q} + \boldsymbol{X}(\boldsymbol{q},t).$$

Particle trajectories satisfy the equation of motion:

$$\ddot{\boldsymbol{x}} + 2H\dot{\boldsymbol{x}} = -\frac{1}{a^2}\boldsymbol{\nabla}_{\boldsymbol{x}}\Phi,$$

which, using conformal time η ($\mathrm{d}t = a\,\mathrm{d}\eta$, $\mathcal{H}\equiv a'/a$), takes the form:

$$\boldsymbol{x}'' + \mathcal{H}\boldsymbol{x}' = -\boldsymbol{\nabla}_{\boldsymbol{x}}\Phi.$$

Taking the divergence of this relation, using Poisson's equation, and invoking mass conservation $\bar{\rho}\mathrm{d}^3\boldsymbol{q} = \bar{\rho}[1+\delta(\boldsymbol{x},t)]\mathrm{d}^3\boldsymbol{x}$, we obtain the equation governing the displacement field:

$$J(\boldsymbol{q},t)\boldsymbol{\nabla}_{\boldsymbol{x}}\cdot\left[\boldsymbol{X}'' + \mathcal{H}\boldsymbol{X}'\right] = \frac{3}{2}\Omega_{\rm m}\mathcal{H}^2\left[J(\boldsymbol{q},t) - 1\right], \tag{4.1.19}$$

where the overdensity is related to the Jacobian determinant:

$$1 + \delta(\boldsymbol{x},t) = \frac{1}{\det\left(\delta_{ij} + X_{i,j}\right)} \equiv \frac{1}{J(\boldsymbol{q},t)},$$

with $X_{i,j} \equiv \partial X_i/\partial q_j$. Eulerian and Lagrangian gradients are connected via $(\boldsymbol{\nabla}_{\boldsymbol{x}})_i = (\delta_{ij} + X_{i,j})^{-1}(\boldsymbol{\nabla}_{\boldsymbol{q}})_j$. While highly effective in non-linear perturbation theory (Refs. [11, 12]), this Lagrangian mapping breaks down upon shell crossing ($J \to 0$), where multistreaming produces formal density singularities.

Linearized systems

In the linear regime, we assume small departures from homogeneity under the criteria:

$$\delta \ll 1, \qquad \left(\frac{ut}{d}\right)^2 \ll \delta,$$

where u represents the peculiar velocity and d is the perturbation coherence length. Linearizing Eq. (4.1.16) yields the cosmological Jeans equation:

$$\ddot{\delta} + 2H\dot{\delta} = \frac{c_{\mathrm{S}}^2}{a^2}\Delta\delta + 4\pi G_{\mathrm{N}}\bar{\rho}\,\delta. \tag{4.1.20}$$

Unlike a static universe, the expanding Jeans length is time-dependent through $\bar{\rho}(t) \propto a^{-3}$. Hence, a perturbation of fixed comoving scale can transition from an oscillating acoustic regime to gravitational instability as expansion lowers the Jeans scale.

Growth of density perturbations

For pressureless cold dark matter ($c_{\mathrm{S}} = 0$), the spatial and temporal dependencies decouple, admitting separable solutions:

$$\delta(\boldsymbol{x}, t) = D_+(t)\varepsilon_+(\boldsymbol{x}) + D_-(t)\varepsilon_-(\boldsymbol{x}),$$

where $D_+(t)$ is the growing mode, $D_-(t)$ is the decaying mode, and $\varepsilon_\pm(\boldsymbol{x})$ define the initial seed fluctuation field. The growth factors satisfy:

$$\ddot{D} + 2H(t)\dot{D} - \frac{3}{2}H^2(t)\Omega_{\mathrm{m}}(t)D = 0, \tag{4.1.21}$$

where $4\pi G_{\mathrm{N}}\bar{\rho}_{\mathrm{m}}(t) = \frac{3}{2}H^2(t)\Omega_{\mathrm{m}}(t)$. In an Einstein–de Sitter spacetime ($\Omega_{\mathrm{m}} = 1, a \propto t^{2/3}$), analytical integration yields power-law growth:

$$D_+(t) \propto t^{2/3} \propto a(t), \qquad D_-(t) \propto t^{-1} \propto a^{-3/2}(t). \tag{4.1.22}$$

Using the scale factor a as the evolution variable, Eq. (4.1.21) becomes:

$$\frac{\mathrm{d}^2 D}{\mathrm{d}a^2} + \left(\frac{1}{H}\frac{\mathrm{d}H}{\mathrm{d}a} + \frac{3}{a}\right)\frac{\mathrm{d}D}{\mathrm{d}a} - \frac{3}{2}\frac{\Omega_{\mathrm{m0}}}{a^5}\left(\frac{H}{H_0}\right)^{-2} D = 0. \tag{4.1.23}$$

One exact solution is $D_- \propto H(a)$. The second independent growing mode follows via reduction of order:

$$D_+(a) = \frac{5}{2}\frac{H(a)}{H_0}\Omega_{\mathrm{m0}}\int_0^a \frac{\mathrm{d}a'}{\left[a'E(a')\right]^3}, \tag{4.1.24}$$

where $E(a) \equiv H(a)/H_0$. For an open universe without dark energy ($\Lambda = 0, \Omega_{\mathrm{m}} < 1$), both modes admit closed forms:

$$D_- = \sqrt{\frac{1+x}{x^3}}, \qquad D_+ = 1 + \frac{3}{x} + 3D_-(x)\ln\left[\sqrt{1+x} - \sqrt{x}\right], \tag{4.1.25}$$

with $x \equiv \Omega_{\mathrm{m}}^{-1} - 1$. As $\Omega_{\mathrm{m}} \to 0$ ($x \gg 1$), $D_+ \to 1$, demonstrating that spatial curvature halts perturbation growth.

For a flat ΛCDM model, evaluating Eq. (4.1.24) requires numerical quadrature (Fig. 4.1), well approximated by the Lahav et al. fitting formula (Ref. [13]):

$$D_+ \simeq \frac{5}{2}\frac{a\Omega_{\mathrm{m}}}{\Omega_{\mathrm{m}}^{4/7} - \Omega_\Lambda + \left(1 + \frac{1}{2}\Omega_{\mathrm{m}}\right)\left(1 + \frac{1}{70}\Omega_\Lambda\right)}. \tag{4.1.26}$$

Analytically, for $K = 0$, D_+ can be expressed via Gauss's hypergeometric function (Ref. [14]):

$$D_+ = {}_2F_1\left[1, \frac{1}{3}; \frac{11}{6}; -\sinh^2\left(\frac{3\alpha t}{2}\right)\right]\sinh^{2/3}\left(\frac{3\alpha t}{2}\right),$$

where $\alpha \equiv H_0\sqrt{\Omega_{\Lambda 0}} = \sqrt{\Lambda/3}$.

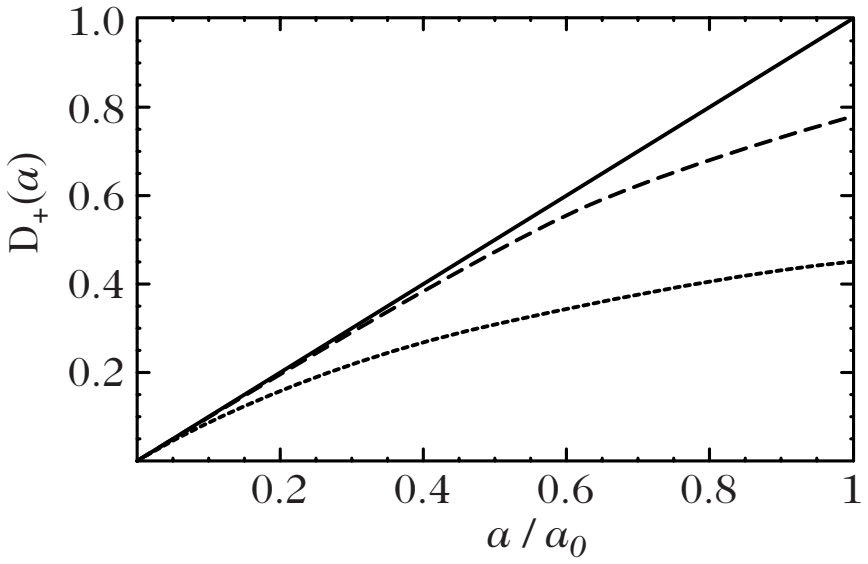


Figure 4.1: Evolution of the growing mode of the density contrast, $D_+(a)$, as a function of the scale factor, a, for $\Omega_{\rm m0} = 1$ [solid line], $(\Omega_{\rm m0}, \Omega_{\Lambda 0}) = (0.3, 0.7)$ [dashed line] and $(\Omega_{\rm m0}, \Omega_{\Lambda 0}) = (0.5, 0)$ [dotted line].

The Hubble parameter from the growth function

Equation (4.1.23) can be inverted into an equation for the normalized expansion rate $E^2(a)$:

$$\frac{\mathrm{d}E^2}{\mathrm{d}a} + 2\left[\frac{3}{a} + \frac{\mathrm{d}^2 D}{\mathrm{d}a^2}\left(\frac{\mathrm{d}D}{\mathrm{d}a}\right)^{-1}\right] E^2 - 3\frac{\Omega_{\rm m0}}{a^5} D \left(\frac{\mathrm{d}D}{\mathrm{d}a}\right)^{-1} = 0.$$

Integrating this first-order differential equation yields:

$$E^2(z) = -3\Omega_{\rm m0}(1+z)^2 \left(\frac{\mathrm{d}D}{\mathrm{d}z}\right)^{-2} \int_z^\infty \frac{D}{1+z}\frac{\mathrm{d}D}{\mathrm{d}z}\,\mathrm{d}z\,, \tag{4.1.27}$$

as a function of redshift $1 + z = a_0/a$. This consistency relation links cosmological expansion directly to structure growth in ΛCDM.

Density–velocity relation

Because $\delta \propto D(t)$, the continuity equation (4.1.15) relates the peculiar velocity divergence directly to the overdensity:

$$\theta(\boldsymbol{x}, t) \equiv \frac{1}{aH}\boldsymbol{\nabla}\cdot\boldsymbol{u}(\boldsymbol{x}, t) = -f(\Omega_{\rm m0}, \Omega_{\Lambda 0})\delta, \tag{4.1.28}$$

where the dimensionless growth rate f is defined as:

$$f(\Omega_{\rm m0}, \Omega_{\Lambda 0}) \equiv \frac{\mathrm{d}\ln D(a)}{\mathrm{d}\ln a} \simeq \Omega_{\rm m0}^{0.6} + \frac{\Omega_{\Lambda 0}}{70}\left(1 + \frac{1}{2}\Omega_{\rm m0}\right). \tag{4.1.29}$$

Inverting the divergence, the general peculiar velocity field decomposes into potential and solenoidal components:

$$\boldsymbol{u} = -\frac{2}{3}\frac{f_+(\Omega_{\rm m0}, \Omega_{\Lambda 0})}{aH\Omega_{\rm m}}\boldsymbol{\nabla}\Phi + \boldsymbol{\nabla}\times\boldsymbol{U}, \tag{4.1.30}$$

where $\boldsymbol{U}$ is an arbitrary solenoidal vector potential. In an Einstein–de Sitter universe:

$$f_+ = 1, \quad \boldsymbol{u}_+ = -t^{1/3}\boldsymbol{\nabla}\Phi, \qquad f_- = -\frac{3}{2}, \quad \boldsymbol{u}_- = t^{-1/3}\boldsymbol{\nabla}\Phi. \tag{4.1.31}$$

Cosmic peculiar velocity surveys (Ref. [15]) directly measure these flow fields, providing sensitive tests of gravity on cosmological scales.

Gravitational potential

In the linear regime, the gravitational potential scales as $\Phi \propto D_+(t)/a(t)$. Thus, in an Einstein–de Sitter spacetime ($D_+ \propto a$), Φ remains strictly constant in time. Using the σ_8 normalization on $8\,h^{-1}$Mpc scales, Poisson's equation gives:

$$\sigma_\Phi = \frac{3}{2}\Omega_{\mathrm{m0}} \left(\frac{8\,\mathrm{Mpc}}{H_0^{-1}}\right)^2 \sigma_8 \sim 10^{-5}.$$

This smallness guarantees that the metric potential Φ remains perturbative ($\Phi \ll 1$) even when density clustering has become non-linear ($\delta \gg 1$), and characterizes the typical potential well depth across all scales under inflationary scale invariance.

4.1.3 Predictions and observables

The cosmic web observed today stems from the gravitational amplification of primordial fluctuations. To rigorously connect early-universe predictions with large-scale observations, a statistical formulation is essential.

The need for a statistical approach

Because individual primordial seed perturbations $\varepsilon_\pm(\boldsymbol{x})$ are inaccessible, theoretical models predict the statistical properties of the random field ensemble rather than deterministic local amplitudes. The observable Universe is interpreted as a single realization drawn from a statistical cosmic distribution.

Statistical description

Assuming statistical homogeneity (translational invariance of all joint probabilities) and statistical isotropy (rotational invariance), the two-point spatial correlation function:

$$\xi(\boldsymbol{x}, \boldsymbol{r}) = \langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r})\rangle \tag{4.1.32}$$

depends purely on the separation distance $r \equiv |\boldsymbol{r}|$:

$$\xi(\boldsymbol{x}, \boldsymbol{r}) = \xi(r).$$

Expanding the real density contrast into spatial Fourier modes:

$$\delta(\boldsymbol{r}) = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^{3/2}}\, \delta_{\boldsymbol{k}}\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}}, \tag{4.1.33}$$

the Fourier coefficients satisfy the reality condition:

$$\delta^*_{\boldsymbol{k}} = \delta_{-\boldsymbol{k}}. \tag{4.1.34}$$

The ensemble expectation value evaluates to:

$$\langle \delta_{\boldsymbol{k}}\delta_{\boldsymbol{k}'}\rangle = \int \frac{\mathrm{d}^3\boldsymbol{x}}{(2\pi)^{3/2}} \frac{\mathrm{d}^3\boldsymbol{r}}{(2\pi)^{3/2}}\, \langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r})\rangle\, e^{-i(\boldsymbol{k}+\boldsymbol{k}')\cdot\boldsymbol{x} - i\boldsymbol{k}'\cdot\boldsymbol{r}}, \tag{4.1.35}$$

which, after integrating over $\boldsymbol{x}$, defines the matter power spectrum $P_\delta(k)$:

$$\langle \delta_{\boldsymbol{k}}\delta_{\boldsymbol{k}'}\rangle = \delta^{(3)}(\boldsymbol{k}+\boldsymbol{k}')P_\delta(k). \tag{4.1.36}$$

The power spectrum is the Fourier transform of the real-space correlation function:

$$P_\delta(k) = \int \mathrm{d}^3\boldsymbol{r}\, \xi(r)\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}}, \tag{4.1.37}$$

which inverts to the boxed canonical relation:

$$\boxed{\xi(r) = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3}\, P_\delta(k)\, e^{-i\boldsymbol{k}\cdot\boldsymbol{r}} = \int_0^\infty \frac{\mathrm{d}k}{2\pi^2}\, k^2 P_\delta(k)\, \frac{\sin kr}{kr}.} \tag{4.1.38}$$

Link with observable quantities and smoothing

Because astronomers observe a finite survey volume V of a single cosmic realization, the ensemble average is approximated operationally by the spatial volume average:

$$\xi_{\mathrm{obs}}(\boldsymbol{r}) = \frac{1}{V}\int_V \mathrm{d}^3\boldsymbol{r}'\, \delta_{\mathrm{obs}}(\boldsymbol{r}')\delta_{\mathrm{obs}}(\boldsymbol{r}+\boldsymbol{r}'). \tag{4.1.39}$$

For any statistical observable $\mathcal{O}$, an empirical estimator $\hat{\mathcal{O}}$ is characterized by its cosmic distribution function $\mathcal{P}(\hat{\mathcal{O}})$, with expectation value:

$$\left\langle \hat{\mathcal{O}} \right\rangle = \int \hat{\mathcal{O}}\, \mathcal{P}(\hat{\mathcal{O}})\, \mathrm{d}\hat{\mathcal{O}}\,.$$

Non-linear estimators typically carry a non-zero cosmic bias $b_{\mathcal{O}}$:

$$b_{\mathcal{O}} \equiv \frac{\left\langle \hat{\mathcal{O}} \right\rangle - \mathcal{O}}{\mathcal{O}},$$

which vanishes only as the survey volume approaches infinity. Robust estimators are designed to minimize both cosmic bias and cosmic variance.

In observational practice, this theoretical framework represents an idealization. Real galaxy catalogs record discrete counts of astronomical objects across a finite survey volume V, from which an underlying continuous density field must be reconstructed. The operationally inferred quantity is the locally smoothed density contrast $\delta_{\mathrm{obs},R}(\boldsymbol{x})$, obtained by convolving the raw field with a spatial window (or filter) function W_R:

$$\delta_{\mathrm{obs},R}(\boldsymbol{x}) = \int_V \mathrm{d}^3\boldsymbol{y}\, \delta_{\mathrm{obs}}(\boldsymbol{y})\, W_R(\boldsymbol{x}-\boldsymbol{y}), \tag{4.1.40}$$

which defines the observational window function W_R. To preserve total mass and characterize a physical smoothing scale R, the window function is subject to the normalization conditions:

$$\int W_R(\boldsymbol{x})\, \mathrm{d}^3\boldsymbol{x} = 1, \tag{4.1.41}$$

and

$$\int W_R(\boldsymbol{x})\, x^2\, \mathrm{d}^3\boldsymbol{x} = R^2. \tag{4.1.42}$$

Canonical choices for W_R include the spherical top-hat filter and the Gaussian smoothing kernel. Denoting the spatial Fourier transform of the window function by $\widehat{W}_R(\boldsymbol{k})$, the convolution theorem establishes that the Fourier modes of the smoothed field are simply:

$$\delta_R(\boldsymbol{k}) = \delta_{\boldsymbol{k}}\, \widehat{W}_R(\boldsymbol{k}). \tag{4.1.43}$$

The uncertainty in estimating cosmological parameters from finite galaxy catalogs arises from three primary sources (see Refs. [12, 16] for comprehensive analyses):

1. **Finite-volume effects:** A survey encompassing total volume V contains only approximately $N \sim V/L^3$ statistically independent patches of characteristic coherence size L. Consequently, the global background density cannot be determined with infinite precision. These sampling fluctuations scale with the volume integral of the two-point correlation function over the survey footprint and are universally designated as *cosmic variance*.

2. **Edge effects:** The finite, complex geometric boundaries of an observational catalog distort the spatial sampling, since estimators assign uneven statistical weights to regions situated near survey borders.

3. **Discreteness (shot-noise) effects:** Whereas theoretical cosmological fields are smooth and continuous, the observable universe is traced by point-like discrete sources (e.g., individual galaxies). The resulting Poisson-like shot noise decays as an inverse power of the total number density of observed tracers.

For any general physical field $Q(\boldsymbol{r})$, one seeks to determine the probability distribution function (PDF) $p(Q, R)\,\mathrm{d}Q$ of the filtered field $Q_R(\boldsymbol{r})$ (focusing on the one-point distribution for conciseness). This PDF quantifies the likelihood that the smoothed field assumes a value in the interval $[Q, Q + \mathrm{d}Q]$. For instance, if Q represents galaxy counts within a spherical top-hat of radius R, $p(Q, R)$ dictates the probability of detecting N galaxies within that volume. Standard inflationary cosmology posits that primordial fluctuations originate as a Gaussian random field:

$$p(Q_i, R) = \frac{1}{\sqrt{2\pi}\sigma_i(R)} \exp\left(-\frac{Q_{iR}^2}{2\sigma_i^2(R)}\right), \tag{4.1.44}$$

where $\sigma_i^2(R) \equiv \langle Q_R^2 \rangle$ represents the initial variance. Because a Gaussian distribution is entirely specified by its second central moment, comparing the primordial state with the late-time smoothed field Q_R requires accounting for two primary dynamical phenomena:

- Scale-dependent linear growth, whereby gravitational evolution modifies the initial power spectrum through a scale-dependent transfer function $T(k)$;

- Non-linear gravitational collapse, which drives departures from Gaussianity in the statistical distribution of fluctuations.

A third practical complication stems from the fact that telescopes register solely luminous baryonic tracers rather than dark matter directly. Ensuring that visible structures accurately reflect the underlying total gravitational matter field leads to the concept of galaxy bias (see Refs. [12, 17–19]).

The subsequent sections of this chapter investigate the relativistic evolution of linear perturbations and the resulting transfer functions, after first examining the onset of non-linear behavior within the Newtonian framework.

4.1.4 Towards the non-linear regime

The Newtonian formulation offers valuable insight into the initial stages of non-linear gravitational clustering (for an exhaustive review of perturbative and non-linear dynamics, see Ref. [12]).

Regardless of its initial infinitesimal amplitude, the density contrast δ undergoes continuous gravitational growth until it eventually approaches order unity ($\delta \sim 1$). The cosmic epoch at which linear theory breaks down depends on both the primordial amplitude and the cosmological growth factor $D(t)$. Observationally, the normalization of the present-day matter power spectrum is conventionally parameterized by the root-mean-square density fluctuation in a sphere of radius $R_8 = 8\,h^{-1}\mathrm{Mpc}$, filtered with a spherical top-hat:

$$\sigma_8^2 \equiv \left\langle \left(\frac{3}{4\pi R_8^3} \int_{|\boldsymbol{x}| \le R_8} \delta(\boldsymbol{x})\, \mathrm{d}^3\boldsymbol{x} \right)^2 \right\rangle \sim 1. \tag{4.1.45}$$

The characteristic scale R_8 thus marks the physical boundary below which non-linearities dominate and the linear approximation is rendered invalid.

To describe the onset of non-linear evolution, the density contrast is expanded in a perturbative series:

$$\delta = \delta^{(1)} + \delta^{(2)} + \dots,$$

where each successive term $\delta^{(n)}$ is of order $\mathcal{O}(\varepsilon^n)$ in terms of the initial seed perturbation field $\varepsilon(\boldsymbol{x}) \ll 1$. The first-order term reproduces the linear growing-mode solution established earlier:

$$\delta^{(1)} = D(t)\,\varepsilon(\boldsymbol{x}), \tag{4.1.46}$$

while at second order, the non-linear continuity and Euler equations yield the governing dynamical equation for $\delta^{(2)}$:

$$\ddot{\delta}^{(2)} + 2H\dot{\delta}^{(2)} = 4\pi G_{\mathrm{N}}\rho\left[\delta^{(1)}\right]^2 + \frac{1}{a^2}\boldsymbol{\nabla}\delta^{(1)}\cdot\boldsymbol{\nabla}\Phi + \frac{1}{a^2}\partial_i\partial_j\left[u^i_{(1)}u^j_{(1)}\right]. \tag{4.1.47}$$

Substituting the first-order solutions into the right-hand side source terms, this inhomogeneous differential equation integrates to (Ref. [20]):

$$\delta^{(2)} = D^2(t)\left\{\frac{2}{3}\left[1+\kappa(t)\right]\varepsilon^2(\boldsymbol{x}) + \boldsymbol{\nabla}\varepsilon\cdot\boldsymbol{\nabla}\Phi + \left[\frac{1}{2} - \kappa(t)\right]\sigma^2\right\}, \tag{4.1.48}$$

with the shear tensor and its scalar contraction defined by:

$$\sigma^2 = \sigma_{ij}\sigma^{ij}, \qquad \sigma_{ij} = \partial_i\partial_j\Phi - \frac{1}{3}\delta_{ij}\Delta\Phi.$$

The dimensionless function $\kappa(t)$ evolves very weakly with cosmic time and is accurately approximated by (Ref. [21]):

$$\kappa \simeq \frac{3}{14}\,\Omega_{\mathrm{m}}^{-2/63},$$

in a flat Universe with vanishing cosmological constant.

Application: Skewness

A direct diagnostic of non-linear gravitational dynamics is provided by the third central moment (or skewness) of the cosmic density field:

$$\left\langle\delta^3\right\rangle = \left\langle\left(\delta^{(1)}\right)^3\right\rangle + \left\langle\left(\delta^{(1)}\right)^2\delta^{(2)}\right\rangle + \mathcal{O}(\varepsilon^5).$$

For strictly Gaussian primordial initial conditions, all odd moments vanish identically, so that $\left\langle(\delta^{(1)})^3\right\rangle = 0$. Consequently, the emergence of a non-zero third moment serves as a direct indicator of non-linear gravitational clustering. Applying spatial filtering at scale R, the lowest-order non-vanishing contribution to the smoothed third moment evaluates to:

$$\left\langle\delta_R^3\right\rangle = D^2(t)\int\frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3}\frac{\mathrm{d}^3\boldsymbol{k}'}{(2\pi)^3}\,P_\delta(k)P_\delta(k')\widehat{W}_R(k)\widehat{W}_R(k')\widehat{W}_R(|\boldsymbol{k}+\boldsymbol{k}'|)\,J(\boldsymbol{k},\boldsymbol{k}'), \tag{4.1.49}$$

where the non-linear coupling kernel is given by:

$$J(\boldsymbol{k},\boldsymbol{k}') = \frac{1}{2} + \kappa + \frac{k'}{k}\left(\frac{\boldsymbol{k}\cdot\boldsymbol{k}'}{kk'}\right) + \left(\frac{1}{2}-\kappa\right)\left(\frac{\boldsymbol{k}\cdot\boldsymbol{k}'}{kk'}\right)^2.$$

In the limit of an unfiltered field ($R \to 0$) and assuming a power-law spectrum $P(k) \propto k^n$, one obtains the reduced skewness parameter $S_3(0)$ (Ref. [21]):

$$S_3(0) \equiv \frac{\langle\delta^3\rangle}{\langle\delta^2\rangle^2} = 4 + 4\kappa - (n+3) \simeq \frac{34}{17}\left(\Omega_{\mathrm{m}}^{-2/63} - 1\right) - (n+3). \tag{4.1.50}$$

When the density field is smoothed with a spherical top-hat window of radius R, this relation receives a scale-dependent correction (Ref. [22]):

$$S_3(R) = S_3(0) + \frac{\mathrm{d}\ln\sigma_\delta^2(R)}{\mathrm{d}\ln R}, \tag{4.1.51}$$

where $\sigma_\delta(R)$ is the variance of the density field smoothed at scale R. Observational measurements of $S_3(R)$ across galaxy surveys (e.g., Ref. [23]) confirm that large-scale cosmic structure is fundamentally consistent with having grown out of Gaussian primordial seeds.

4.2 Gauge invariant cosmological perturbation theory

While Newtonian theory provides a faithful description of sub-Hubble perturbations during the late-time matter era, a rigorous general-relativistic framework is indispensable for tracking super-Hubble modes and causal horizons across the early Universe. The formulation developed in this section constitutes one of the foundations of modern theoretical cosmology, culminating in a relativistic generalization of the fluid equations (4.1.1), (4.1.2), and (4.1.5). This is needed as we do not merely want to study the dynamics of test particle but we wish to incorporate their backreaction into the picture as well.

Two distinct formalisms exist for treating relativistic cosmological perturbations. Here we adopt the coordinate-based approach in which an arbitrary parameterization of metric and fluid perturbations is introduced, from which gauge-invariant combinations are systematically constructed. Pioneered by Bardeen (Refs. [24, 25]), this formulation is the standard workhorse of modern cosmology and is reviewed extensively in Refs. [26–29]. An alternative, complementary approach employs the 1+3 covariant and gauge-invariant fluid formalism (Refs. [30–32]). Both routes yield identical physical predictions.

4.2.1 Perturbed space-time

Metric of the perturbed space-time

At linear order, the most general perturbed line element around a spatially flat or curved Friedmann–Lemaître geometry is parameterized in conformal time η by:

$$\mathrm{d}s^2 = a^2(\eta)\left[-(1+2A)\,\mathrm{d}\eta^2 + 2B_i\,\mathrm{d}x^i\,\mathrm{d}\eta + (\gamma_{ij}+h_{ij})\,\mathrm{d}x^i\,\mathrm{d}x^j\right], \tag{4.2.1}$$

where $A(\eta,\boldsymbol{x})$, $B_i(\eta,\boldsymbol{x})$, and $h_{ij}(\eta,\boldsymbol{x})$ represent small metric perturbations to be determined from the linearized Einstein equations. Since $\delta T_{\mu\nu} \neq 0$, we can have off shell scalar graviton degree of freedom. Decomposing the full physical metric into the background Friedmann–Lemaître tensor $\bar{g}_{\mu\nu}$ and its perturbation:

$$g_{\mu\nu} = \bar{g}_{\mu\nu} + \delta g_{\mu\nu}, \tag{4.2.2}$$

the inverse metric to linear order is given by:

$$g^{\mu\nu} = \bar{g}^{\mu\nu} + \delta g^{\mu\nu}, \qquad \delta g^{\mu\nu} = -\bar{g}^{\mu\sigma}\bar{g}^{\nu\lambda}\delta g_{\sigma\lambda}. \tag{4.2.3}$$

> For the sake of perturbation theory, we always expland the full metric around some fixed background irrespective of the choice of coordinate. The choice of coordinate manifests itself as transformation law for the perturbation or as gauge invariance.

We denote the trace of the spatial metric perturbations with respect to the background 3-metric γ_{ij} by:

$$h \equiv h_{ij}\gamma^{ij},$$

and ∇_μ denotes the full four-dimensional covariant derivative. The complete explicit expressions for the perturbed Christoffel symbols, Riemann, Ricci, and Einstein tensors, and Ricci scalar are cataloged in Appendix C.

SVT decomposition

When we perform coordinate transformation on the full metric which happens to be the isometry of the background metric, we observe that the perturbation transforms as tensor under these. As such one can decompose the perturbation into the irreducible representation (fundamental building blocks) of this isometry group. Following Bardeen (Ref. [24]), we perform a scalar-vector-tensor (SVT) decomposition of all spatial perturbation fields, exactly as introduced in the geometric foundations of Chapter 1. This decomposition generalizes the Helmholtz theorem of vector calculus, which partitions any vector field into the gradient of a scalar potential and a divergence-free (solenoidal) vector:

$$B_i = D_i B + \bar{B}_i \qquad \text{with} \qquad D_i \bar{B}^i = 0. \tag{4.2.4}$$

Here D_i is the covariant derivative associated with spatil part of background metric γ_{ij}. For a velocity field, B represents the longitudinal scalar potential while $\bar{B}_i$ encodes rotational vorticity. Thus, the 3 components of the vector have been split into 1 scalar mode (B) and 2 vector modes ($\bar{B}_i$).

Similarly, any symmetric rank-2 spatial tensor field h_{ij} can be decomposed into scalar, vector, and trace-free transverse tensor components:

$$h_{ij} = 2C\gamma_{ij} + 2D_i D_j E + 2D_{(i}\bar{E}_{j)} + 2\bar{E}_{ij} \qquad \text{with} \qquad D_i \bar{E}^{ij} = 0, \quad \bar{E}^i{}_i = 0. \tag{4.2.5}$$

Equation (4.2.3) implies that the indices of these quantities are raised and lowered with respect to the unperturbed spatial metric γ_{ij} (for example, $\bar{B}^i = \gamma^{ij}\bar{B}_j$). We adopt the convention in which any 'barred' quantity is divergenceless (and traceless if it has two indices). The 6 components of h_{ij} have thus been split into 2 scalar (C and E), 2 vector ($\bar{E}_j$), and 2 tensor ($\bar{E}_{ij}$) components.

Note that the scalar potentials B and E are not uniquely defined without boundary conditions. In fact, B is a solution of $\Delta B = D^i B_i$ (the Laplacian being defined by $\Delta \equiv D_i D^i$), whose solution is unique only if some boundary conditions are imposed.

In total, the 10 degrees of freedom of the metric have thus been decomposed as:

- 4 scalars: A, B, C, and E, corresponding to 4 degrees of freedom;
- 2 vectors: $\bar{B}_i$ and $\bar{E}_i$, corresponding to $2 \times (3-1) = 4$ degrees of freedom;
- 1 tensor: $\bar{E}_{ij}$, corresponding to $6 - 1 - 3 = 2$ degrees of freedom.

The advantages of this decomposition lie in the fact that, to leading order, the three types of perturbations are decoupled and can thus be studied separately.

The gauge problem

To introduce the perturbed space-time metric, we have made the hypothesis (4.2.2) that the metric $g_{\mu\nu}$ was 'close' to that of the Friedmann–Lemaître background $\bar{g}_{\mu\nu}$.

In field theory, space-time is usually fixed and once the system of coordinates is chosen, one can define the perturbation of any quantity $Q(\boldsymbol{x}, t)$ (pullback from the perturbed manifold to background manifold) as:

$$\delta Q(\boldsymbol{x}, t) \equiv \phi^*(Q(\boldsymbol{x}, t)) - \bar{Q}(\boldsymbol{x}, t) = Q(\boldsymbol{x}, t) - \bar{Q}(\boldsymbol{x}, t),$$

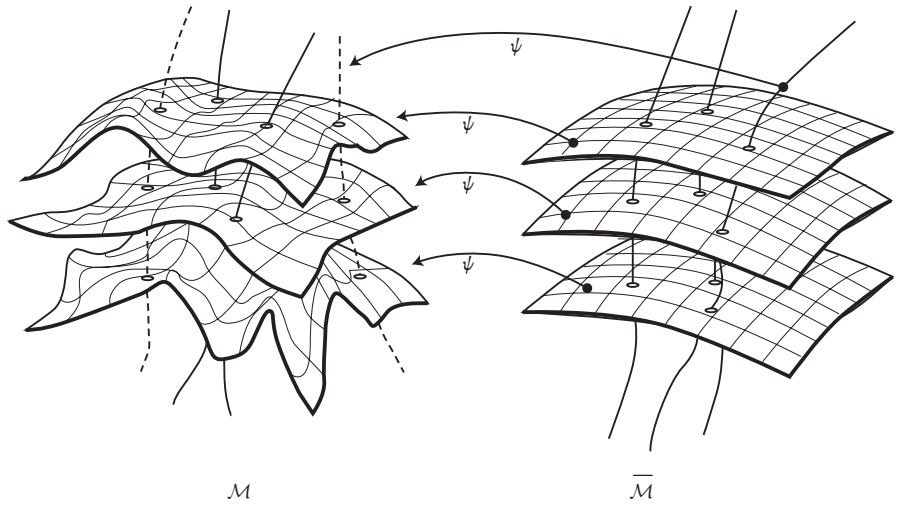


Figure 4.2: Any perturbed quantity is defined via a mapping between the Friedmann–Lemaître space-time, $\bar{\mathcal{M}}$, and the perturbed space-time $\mathcal{M}$.

where $\bar{Q}(\boldsymbol{x}, t)$ represents the unperturbed configuration. At any point of space-time, the quantity $Q(\boldsymbol{x}, t)$ can be compared with $\bar{Q}(\boldsymbol{x}, t)$ at the same point.

In general relativity, there exists an important difference with this standard approach. Indeed, we would like to compare two different space-times and make the hypothesis that $(\mathcal{M}, g_{\mu\nu})$ is 'close' to a Friedmann–Lemaître Universe $(\bar{\mathcal{M}}, \bar{g}_{\mu\nu})$. In making this comparison, there is some arbitrary freedom related to the way of identifying the points of these two space-times. Indeed, an isomorphism ψ should be introduced to relate the points of the two space-times (see Fig. 4.2). There is actually no 'natural' identification between the two spaces, and there is therefore a freedom of choice in the identification of the points in each space.[1] This implies that there exist some unphysical degrees of freedom related to the choice of the coordinate systems on the two manifolds.

What we're doing in perturbation theory is that we are building a discription of the perturbed spacetime by using a set of tensor fields, i.e. the perturbation, which live on the background manifold. Since we use pullback on the full metric to define the perturbation on background manifold. The non uniqueness of such map manifests itself as the gauge invariance of the problem and the full metric ends up becoming a tensor field on the background manifold.

The decomposition (4.2.1) implicitly assumes that a system of coordinates has been chosen in the perturbed space. Any change of coordinates will modify the form of the metric coefficients. To illustrate the effect of such a change of coordinates, let us consider an unperturbed Friedmann–Lemaître space-time, as in Fig. 4.2, and let us perform an arbitrary space- and time-dependent change of coordinates $x^i \to y^i = x^i - \xi^i(x^j, \eta)$. We then get:

$$\mathrm{d}s^2 = a^2(\eta)\left[-\mathrm{d}\eta^2 + 2\xi_i' \,\mathrm{d}y^i \,\mathrm{d}\eta + (\gamma_{ij} + 2D_{(i}\xi_{j)})\,\mathrm{d}y^i \,\mathrm{d}y^j\right].$$

There appear two artificial terms $B_i = \xi_i'$ and $E_i = \xi_i$ that do not correspond to metric fluctuations, since this space is still a Friedmann–Lemaître space, but to a choice of the system of coordinates. The dependence of the quantities that are intrinsic to the manifold in which the

[1] Since the perturbation depends on both physical and unphysical dof.

perturbations evolve should thus be separated from the artificial ones, related to the arbitrary choice of a particular coordinate system.

Gauge invariant variables

In order to extract the physical degrees of freedom, let us consider an active transformation of the coordinate system defined by an arbitrary vector field ξ. The coordinates of any point change according to:

$$x^\mu \to x^\mu - \xi^\mu, \tag{4.2.6}$$

where the displacement ξ^μ is decomposed into 2 scalar degrees of freedom (T and L) and 2 vector degrees of freedom ($\bar{L}^i$, which is divergenceless $D_i\bar{L}^i = 0$) as:

$$\xi^0 = T, \qquad \xi^i = L^i = D^i L + \bar{L}^i. \tag{4.2.7}$$

Under this change of coordinates, the metric transforms as:

$$g_{\mu\nu} \to g_{\mu\nu} + \pounds_\xi g_{\mu\nu}.$$

For first order perturbation theory, we always expand the metric about the same fixed background metric.

$$g'_{\mu\nu} = g_{\mu\nu} + \pounds_\xi g_{\mu\nu} \tag{1}$$
$$\bar{g}_{\mu\nu} + \delta g'_{\mu\nu} = \bar{g}_{\mu\nu} + \delta g_{\mu\nu} + \pounds_\xi(\bar{g}_{\mu\nu} + \delta g'_{\mu\nu}) \tag{2}$$
$$\delta g'_{\mu\nu} = \delta g_{\mu\nu} + \pounds_\xi \bar{g}_{\mu\nu} + \mathcal{O}\left(\xi^2\right) \tag{3}$$

To first order in the perturbations, this is:

$$\delta g_{\mu\nu} \to \delta g_{\mu\nu} + \pounds_\xi \bar{g}_{\mu\nu} = \delta g_{\mu\nu} + 2\nabla_{(\mu}\xi_{\nu)},^{2}$$

Rather than evaluating the covariant derivative of the arbitrary vector on the background manifold, one can utilise the following to quickly evaluate how the perturbation transforms:

$$\pounds_\xi \bar{g}_{\mu\nu} = \xi^\sigma \nabla_\sigma \bar{g}_{\mu\nu} + \bar{g}_{\sigma\nu}(\nabla_\mu \xi^\sigma) + \bar{g}_{\mu\sigma}(\nabla_\nu \xi^\sigma) \tag{1}$$
$$= \xi^\sigma \partial_\sigma \bar{g}_{\mu\nu} + \bar{g}_{\sigma\nu}(\partial_\mu \xi^\sigma) + \bar{g}_{\mu\sigma}(\partial_\nu \xi^\sigma) \tag{2}$$

which in turn implies that the metric perturbation variables (4.2.1) transform as:

$$A \to A + T' + \mathcal{H}T, \tag{4.2.9}$$
$$B_i \to B_i - D_i T + L'_i,$$
$$h_{ij} \to h_{ij} + D_i L_j + D_j L_i + 2\mathcal{H}T\gamma_{ij}. \tag{4.2.10}$$

[2] For completeness, we give the expressions of the Lie derivatives. They are easily obtained from the expressions of the Christoffel symbols of the Friedmann–Lemaître space-time and from $\xi_\mu = a^2(-T, L_i)$. They are given by

$$\pounds_\xi \bar{g}_{00} = \xi^\sigma \partial_\sigma \bar{g}_{00} + \bar{g}_{\sigma 0}\partial_0 \xi^\sigma + \bar{g}_{0\sigma}\partial_0 \xi^\sigma = T\,\partial_0(-a^2) + 2(-a^2)T' = -2aa'T - 2a^2T' = -2a^2(T' + \mathcal{H}T),$$
$$\pounds_\xi \bar{g}_{0i} = \xi^\sigma \partial_\sigma \bar{g}_{0i} + \bar{g}_{\sigma i}\partial_0 \xi^\sigma + \bar{g}_{0\sigma}\partial_i \xi^\sigma = a^2\gamma_{ji}\partial_0 L^j - a^2\partial_i T = a^2(L'_i - \partial_i T),$$
$$\pounds_\xi \bar{g}_{ij} = \xi^\sigma \partial_\sigma \bar{g}_{ij} + \bar{g}_{\sigma j}\partial_i \xi^\sigma + \bar{g}_{i\sigma}\partial_j \xi^\sigma = T(a^2\gamma_{ij})' + a^2\gamma_{kj}\partial_i L^k + a^2\gamma_{ik}\partial_j L^k$$
$$= 2a^2\mathcal{H}T\gamma_{ij} + a^2(D_i L_j + D_j L_i) = a^2(D_i L_j + D_j L_i + 2\mathcal{H}T\gamma_{ij}).$$

Using the scalar-vector-tensor decomposition (4.2.4) and (4.2.5), we get:

$$\boxed{\begin{array}{lll} A \to A + T' + \mathcal{H}T, & \bar{B}^i \to \bar{B}^i + \bar{L}^{i\prime}, & \bar{E}_{ij} \to \bar{E}_{ij}, \\ B \to B - T + L', & \bar{E}^i \to \bar{E}^i + \bar{L}^i, & \\ C \to C + \mathcal{H}T, & & \\ E \to E + L, & & \end{array}} \tag{4.2.11}$$

respectively, for the S, V and T modes. These transformations are similar to the gauge transformations encountered, for instance, in electromagnetism. Therefore, one may reduce the arbitrariness by imposing some additional conditions on the variables: in other words, and again as in electromagnetism, one needs to fix the gauge before doing any actually relevant calculation.

As stressed in the previous section, it is interesting to define gauge-independent quantities. To do so, notice that some combinations of the previous quantities do not depend on L^i and T, for instance:

$$\Psi \equiv -C - \mathcal{H}(B - E'), \tag{4.2.12}$$

$$\Phi \equiv A + \mathcal{H}(B - E') + (B - E')', \tag{4.2.13}$$

$$\bar{\Phi}^i \equiv \bar{E}^{i\prime} - \bar{B}^i, \tag{4.2.14}$$

$$\bar{E}^{ij}. \tag{4.2.15}$$

We initially had 10 perturbation variables $(A, B, C, E, \bar{E}^i, \bar{B}^i, \bar{E}^{ij})$ and 4 gauge degrees of freedom $(T, L, \bar{L}^i)$ that can be absorbed into the definition of the gauge invariant variables, then described by $10 - 4 = 6$ degrees of freedom parameterized, for instance, by $\Psi, \Phi, \bar{\Phi}^i$ and $\bar{E}^{ij}$.[3] These quantities can be considered as the 'real' space-time perturbations in the sense that they cannot be removed by any change of coordinates.

Other gauge invariant quantities could have been constructed, such as, for instance:

$$X = A - C - \left(\frac{C}{\mathcal{H}}\right)', \tag{4.2.16}$$

but they can always be expressed in terms of the two functions Φ and Ψ. For instance, in the case of X defined just above:

$$X = \Psi + \Phi + \left(\frac{\Psi}{\mathcal{H}}\right)', \tag{4.2.17}$$

as can straightforwardly be checked.

Stewart–Walker lemma

Just as the metric transforms as (4.2.8) under a change of coordinates of the form (4.2.6), any scalar quantity δQ transforms as:

$$\delta Q \to \delta Q + \pounds_\xi \bar{Q},$$

to first order in perturbations. Since each vector field ξ generates a gauge transformation, we can conclude that the only gauge invariant quantities are the ones for which:

$$\pounds_\xi \bar{Q} = 0 \qquad \forall \xi.$$

This result is known under the name of the **Stewart–Walker lemma** (Ref. [33]).

Note an important corollary of this result. Since all relativistic equations are covariant by definition, they can always be written in the form $Q = 0$, where Q is a tensor field. It is thus always possible, to first order in perturbations, to write all the relevant equations in terms of gauge invariant quantities only. Such a treatment would ensure that no gauge, i.e. unphysical, degree of freedom is misleadingly used.

[3] Except $\bar{E}^{ij}$, rest are off shell graviton degree of freedom. It only happens that at linear order one is able to isolate all the graviton degrees of freedom whereas at higher order they are mixed with each other in complicated fashion.

4.2.2 Description of matter

In this section, we introduce various quantities that come into the description of a perfect fluid for which the energy-momentum tensor in an unperturbed space is of the form:

$$T_{\mu\nu} = (\rho + P)u_\mu u_\nu + P g_{\mu\nu},$$

as was seen in Chapter 3.

The energy-momentum tensor of a perturbed fluid

The energy-momentum tensor of this perturbed fluid takes the general form:

$$\delta T_{\mu\nu} = (\delta\rho + \delta P)\bar{u}_\mu\bar{u}_\nu + \delta P\,\bar{g}_{\mu\nu} + 2(\rho + P)\bar{u}_{(\mu}\delta u_{\nu)} + P\,\delta g_{\mu\nu} + a^2 P\pi_{\mu\nu}, \tag{4.2.18}$$

where $u^\mu = \bar{u}^\mu + \delta u^\mu$ is the four-velocity of a comoving observer, satisfying $u_\mu u^\mu = -1$. The normalization condition of $\bar{u}^\mu$ (to zeroth order) provides the solution:

$$\bar{u}^\mu = a^{-1}\delta^\mu_0, \qquad \bar{u}_\mu = -a\delta^0_\mu,$$

and since the norm of $\bar{u}^\mu + \delta u^\mu$ should also be equal to -1, we infer that $2\bar{u}^\mu\delta u_\mu + \delta g_{\mu\nu}\bar{u}^\mu\bar{u}^\nu = 0$, and thus that $\delta u_0 = -Aa$. We then write $\delta u^i \equiv v^i/a$, so that:

$$\delta u^\mu = a^{-1}(-A, v^i), \qquad \delta u_\mu = a(-A, v_k + B_k), \tag{4.2.19}$$

and we decompose v_i into a scalar and a vector part according to:

$$v_i = D_i v + \bar{v}_i. \tag{4.2.20}$$

In the decomposition (4.2.18), $\pi_{\mu\nu}$ is the anisotropic stress tensor and characterizes the difference between the perturbed fluid and a perfect fluid. It can be chosen to be traceless ($g^{\mu\nu}\pi_{\mu\nu} = 0$), since its trace can be absorbed into a redefinition of the isotropic pressure, δP. It is also symmetric, since the energy-momentum tensor is, and can be chosen orthogonal to u^μ, i.e. $u^\mu\pi_{\mu\nu} = 0$. This implies that, without lack of generality, one can set $\pi_{00} = \pi_{0i} = 0$.

The anisotropic stress tensor then consists only of a spatial part, which can be decomposed into scalar, vector and tensor parts as:

$$\pi_{ij} = \Delta_{ij}\bar{\pi} + D_{(i}\bar{\pi}_{j)} + \bar{\pi}_{ij}, \tag{4.2.21}$$

This is part of the pressure/stress that depends on the direction which is suppressed by scattering/collision.

where the operator Δ_{ij} is defined as:

$$\Delta_{ij} \equiv D_i D_j - \frac{1}{3}\gamma_{ij}\Delta. \tag{4.2.22}$$

We infer from (4.2.19), (4.2.20) and (4.2.21) that the components of the perturbation of the energy-momentum tensor (4.2.18) are:

$$\begin{aligned}\delta T_{00} &= (\delta\rho + \delta P)(-a)^2 + \delta P(-a^2) + 2(\rho + P)(-a)(-aA) + P(-2a^2A) + a^2P \times 0 \\ &= \rho a^2(\delta + 2A), \qquad (4.2.23)\\ \delta T_{0i} &= (\rho + P)(-a)\left[a(D_i v + \bar{v}_i)\right] + (\rho + P)(aD_i B + a\bar{B}_i)(-a) + P\left[a^2(D_i B + \bar{B}_i)\right]\end{aligned}$$

$$= -\rho a^2 \left[(1+w)(D_i v + \bar{v}_i) + D_i B + \bar{B}_i\right], \tag{4.2.24}$$

$$\delta T_{ij} = \underbrace{\delta(\rho+P)u_i u_j + (\rho+P)(u_i \delta u_j + \delta u_i u_j)}_{=0} + \delta P\, \bar{g}_{ij} + P\, \delta g_{ij} + a^2 P \pi_{ij}$$

$$= P a^2 \left[h_{ij} + \frac{\delta P}{P}\gamma_{ij} + \pi_{ij}\right], \tag{4.2.25}$$

where we recall that $\delta \equiv \delta\rho/\bar{\rho}$ is the density contrast.

It will be useful to introduce the entropy perturbation, Γ, defined by the relation:

$$\delta P = c_{\rm s}^2 \delta\rho + P\Gamma, \tag{4.2.26}$$

which takes the form:

$$w\Gamma = \frac{1}{\rho}\left(\delta P - c_{\rm s}^2 \delta\rho\right). \tag{4.2.27}$$

For an adiabatic perturbation, $\delta P/\delta\rho = c_{\rm s}^2$ and hence $\Gamma = 0$ and we introduce the notation used in the literature:

$$\delta P_{\rm nad} = P\Gamma, \tag{4.2.28}$$

where the index 'nad' stands for non-adiabatic.

Gauge invariant quantities

Upon performing a change of coordinates of the form (4.2.6), scalar and vectors transform as:

$$\delta Q \to \delta Q + \pounds_\xi \bar{Q}, \qquad \pounds_\xi \bar{Q} = \xi^\alpha \partial_\alpha \bar{Q} = T\bar{Q}',$$

and

$$\delta u^\mu \to \delta u^\mu + \pounds_\xi \bar{u}^\mu, \qquad \pounds_\xi \bar{u}^\mu = \xi^\alpha \partial_\alpha \bar{u}^\mu - \bar{u}^\alpha \partial_\alpha \xi^\mu,$$

so that it follows that, after SVT decomposition, $\delta P, \delta\rho, v$ and $\bar{v}_i$ transform as:

$$\boxed{\begin{array}{ll} \delta\rho \to \delta\rho + \rho' T, & v \to v - L', \\ \delta P \to \delta P + P' T, & \bar{v}_i \to \bar{v}_i - \bar{L}_i'. \end{array}} \tag{4.2.29}$$

The quantities $\bar{\pi}, \bar{\pi}_i$ and $\bar{\pi}_{ij}$ are gauge invariant from the Stewart–Walker lemma, since the unperturbed part of the anisotropic stress tensor must vanish to be compatible with the perfect fluid hypothesis imposed by the symmetries of the Friedmann–Lemaître space-time. One can also check that, as expected again from the Stewart–Walker lemma, Γ is gauge invariant since $\Gamma \to \Gamma + (P' - c_{\rm s}^2\rho')T = \Gamma$.

As in the case of the metric perturbations, one can likewise define the gauge invariant quantities associated with the quantities (4.2.29). Different choices are possible, and we define:

$$\delta^{\rm N} = \delta + \frac{\rho'}{\rho}(B - E'), \tag{4.2.30}$$

$$\delta^{\rm F} = \delta - \frac{\rho'}{\rho}\frac{C}{\mathcal{H}}, \tag{4.2.31}$$

$$\delta^{\rm C} = \delta + \frac{\rho'}{\rho}(v + B), \tag{4.2.32}$$

$$V = v + E', \tag{4.2.33}$$

$$\overline{V}_i = \bar{v}_i + \bar{B}_i, \tag{4.2.34}$$

$$\overline{W}_i = \bar{v}_i + \bar{E}_i'. \tag{4.2.35}$$

Even though these are gauge invariant density contrast but they are not the same object. They all converge to δ in the subhorizon limit. But relativistically speaking they correspond to perturbations measured in different hypersurfaces.

The pressure perturbations $\delta P^{\rm N}, \delta P^{\rm F}$ and $\delta P^{\rm C}$ are defined in an identical way. The different gauge invariant variables are related to each other and one can check that:

$$\delta^{\rm F} = \delta^{\rm N} + \frac{\rho'}{\rho}\frac{\Psi}{\mathcal{H}}, \qquad \delta^{\rm C} = \delta^{\rm N} + \frac{\rho'}{\rho}V, \qquad \overline{W}_i = \overline{V}_i + \bar{\Phi}_i. \tag{4.2.36}$$

In all these relations, we recall that $\delta = \delta\rho/\rho$ is the density contrast.

4.2.3 Choosing a gauge

From the Stewart–Walker lemma, we know that any perturbation equation can be written in terms of gauge invariant variables only. Indeed, such an equation always takes the form $Q = 0$ and is, by construction, linear in the first-order perturbation variables. One method for obtaining these gauge invariant equations is thus to write the equations in an arbitrary gauge and to regroup the terms to make the gauge invariant variables appear.

A more direct and physically transparent approach is to choose a specific coordinate system from the very beginning such that the metric and matter perturbations in that gauge reduce identically to gauge invariant variables. For instance, fixing the longitudinal gauge by the constraints $E = 0$ and $\overline{B}_i = 0$ directly implies $C = -\Psi$, $A = \Phi$, and so on. The governing perturbation equations are then derived immediately in terms of gauge invariant quantities. Once these dynamical relations are established, one can readily transform to any other gauge or reinstate the original gauge-dependent perturbation variables via the transformation relations (4.2.11).

A crucial conceptual point deserves special emphasis. Although perturbative calculations can be performed in any gauge of choice, connecting theoretical predictions with observational measurements generally requires a gauge transformation to the coordinate system of the observer. Gauge invariant quantities do not possess an intrinsic physical interpretation in the abstract; rather, their physical significance becomes apparent only within a gauge where the metric or fluid variables coincide with the gauge invariant definitions. For example, the gauge invariant variable $\delta^{\rm N}$ can be directly interpreted as a fractional density fluctuation only when evaluated in the reference frame of an observer tied to the longitudinal gauge.

Furthermore, by the equivalence principle, all coordinate systems are locally indistinguishable. Consequently, deviations among different gauge representations manifest only on large spatial scales—more precisely, on scales exceeding a characteristic physical threshold set by the cosmological Hubble radius $\mathcal{H}^{-1}$, as we will see shortly.

Finally, a note of caution regarding perturbation theory: if a supposedly small perturbation variable becomes large in a given gauge while remaining small in another, this does not necessarily signify a physical breakdown of linear perturbation theory, nor does it prove that the perturbations remain linear. Instead, it typically indicates that the coordinate displacement relating the two slicings is no longer infinitesimal. In such circumstances, a higher-order perturbation calculation is required: if the second-order corrections remain small relative to the first order across all gauges, the pathology was merely an artifact of the initial coordinate system. Such checks are indispensable before drawing definitive physical conclusions.

Before deriving the general-relativistic evolution equations, we summarize the properties of several widely used cosmological gauges.

Newtonian or longitudinal gauge

The Newtonian (or longitudinal) gauge is defined by demanding that the scalar sector of the spatial metric perturbations is strictly diagonal, namely:

$$B = 0 \quad \text{and} \quad E = 0. \tag{4.2.37}$$

Under these two constraints, the vector sector cannot be simultaneously diagonalized; one conventionally eliminates the remaining vector gauge freedom by imposing additional 2 constraint:

$$\overline{B}_i = 0. \tag{4.2.38}$$

This fixes the gauge completely ($2 + 2 = 4$ constraints). Starting from an arbitrary coordinate system, one reaches the longitudinal gauge via the gauge transformation (4.2.11) with parameters:

$$T = B - E', \quad L = -E \quad \text{and} \quad \overline{L}_i' = -\overline{B}_i. \tag{4.2.39}$$

In this coordinate system, the cosmological expansion appears isotropic and $T, L, \bar{L}_i$ are completely fixed. The scalar potential Ψ characterizes the intrinsic spatial curvature perturbation on constant-time hypersurfaces [see (C.31)], while Φ plays the role of the relativistic gravitational potential entering the Poisson equation. On sub-Hubble scales, Φ reduces to the classical Newtonian gravitational potential, which motivates the name of the gauge.

In the longitudinal gauge, the remaining perturbed quantities reduce to the corresponding gauge invariant variables:

$$A = \Phi, \quad C = -\Psi, \quad \delta = \delta^{\mathrm{N}}, \quad \delta P = \delta P^{\mathrm{N}}, \quad v = V \quad \text{and} \quad \bar{v}_i = \overline{V}_i. \tag{4.2.40}$$

Flat-slicing gauge

The flat-slicing (or spatially flat) gauge is defined by requiring that the scalar curvature perturbation of the spatial hypersurfaces vanishes identically. From the expression for the spatial Ricci scalar, this condition imposes:

$$C = 0, \quad E = 0 \quad \text{and} \quad \overline{E}_i = 0. \tag{4.2.41}$$

This is total of $1 + 1 + 2 = 4$ constraints. As with the longitudinal gauge, the flat-slicing gauge is completely fixed. The gauge parameters required to pass from an arbitrary gauge to this frame are:

$$T = -\frac{C}{\mathcal{H}}, \quad L = -E \quad \text{and} \quad \overline{L}_i = -\overline{E}_i. \tag{4.2.42}$$

In this gauge, the remaining perturbed quantities are related to the gauge invariant variables by:

$$A = X, \quad E' = -\frac{\Psi}{\mathcal{H}}, \quad \delta = \delta^{\mathrm{F}}, \quad \delta P = \delta P^{\mathrm{F}}, \quad v = V \quad \text{and} \quad \bar{v}_i = \overline{W}_i. \tag{4.2.43}$$

A key property of this gauge is that the fluid conservation equations take an identical mathematical form to their Newtonian counterparts.

Synchronous or Gauss gauges

Synchronous gauges are defined by the coordinate conditions:

$$\delta g_{0\mu} = 0 \implies A = 0 \quad \text{and} \quad B_i = 0. \tag{4.2.44}$$

This is still $1 + 3 = 4$ constraints but as we will see some residual gauge symmetry survives. In such coordinates, only the spatial metric components are perturbed. The worldlines $x^i = \text{const.}$ are timelike geodesics orthogonal to constant-time hypersurfaces, and the proper time of comoving

observers corresponds directly to cosmic coordinate time. This system therefore represents a foliation of space-time by freely falling observers.

While this geometric picture makes synchronous coordinates intuitive and widely used, the gauge suffers from notorious ambiguities: it is not uniquely fixed. Residual coordinate transformations of the form $t \to t' = f(t)$ or $x^i \to y^i = f^i(x^j)$, involving time or space separately, leave the constraints (4.2.44) invariant. The transformation from an arbitrary gauge to a synchronous gauge is given by:

$$T = -\frac{1}{a}\int Aa\,\mathrm{d}\eta + \frac{C_T}{a}, \quad L = \int (T - B)\,\mathrm{d}\eta + C_L \quad \text{and} \quad \overline{L}^i = -\int \overline{B}^i\,\mathrm{d}\eta + \overline{C}^i_L, \tag{4.2.45}$$

where C_T and $C^i_L = (C_L, \overline{C}^i_L)$ are four arbitrary integration functions that depend solely on the spatial coordinates x^i. Physically, C_T reflects the freedom in choosing the local origin of proper time for each observer, whereas C^i_L corresponds to an arbitrary relabeling of spatial coordinates on constant-time slices. This remaining freedom produces unphysical "gauge modes" in the solutions of the perturbation equations, which can be mistakenly interpreted as physical instabilities. Furthermore, because the gauge is not completely fixed, variables evaluated in synchronous gauge cannot always be uniquely translated into gauge invariant observables.

In the synchronous gauge, the residual spatial metric perturbations are conventionally parameterized by:

$$h \equiv 6C + 2\Delta E \quad \text{and} \quad \eta = -C. \tag{4.2.46}$$

Comoving gauge

The comoving gauge is defined such that the coordinate frame follows the overall bulk motion of the cosmic fluid, setting the momentum flux to zero:

$$\delta T^0{}_i = 0 \implies \rho(1+w)[D_i(v+B) + \bar{v}_i + \bar{B}_i] = 0. \tag{4.2.47}$$

> In the scalar sector, one may impose $v + B = 0$ to fix the temporal gauge freedom associated with T. However, the comoving gauge condition is not sufficient to completely fix the gauge, since the spatial gauge freedom associated with L_i remains unfixed. An additional condition is therefore required to fix L_i. In the vector sector, however, $\bar{v}_i + \bar{B}_i$ is gauge invariant and hence cannot be set to zero by a gauge transformation. Thus, imposing $\bar{v}_i + \bar{B}_i = 0$ is an additional physical restriction, corresponding to a fluid with vanishing vorticity, rather than a gauge choice.

In a cosmological medium consisting of multiple components, a comoving gauge can be defined with respect to any individual fluid species (e.g., baryons or radiation) or with respect to the total matter fluid. Unless specified otherwise, "comoving gauge" will refer to the frame comoving with the total energy-momentum tensor.

The condition (4.2.47) demands $v_i + B_i = 0$. While one can always find coordinates in which the scalar velocity vanishes (since for potential flows, Frobenius's theorem prevents flow lines from crossing, allowing coordinates to follow fluid elements), this is not possible for the vector component corresponding to vorticity, which cannot be eliminated by coordinate choices. Therefore, the comoving gauge must be supplemented by:

$$E = 0 \quad \text{and} \quad \overline{E}_i = 0, \tag{4.2.48}$$

to completely fix L and L_i. This gauge shares an important property with the flat-slicing gauge: the relativistic Poisson equation takes a form identical to that in the Newtonian framework.

4.2.4 Einstein equations: derivation

The linearized Einstein field equations governing cosmological perturbations take the form:

$$\delta G^{\mu}{}_{\nu} = \kappa\, \delta T^{\mu}{}_{\nu}.$$

All explicit expressions for the perturbed geometric tensors are collected in Appendix C. In what follows, we examine tensor modes first, proceed to vector modes, and conclude with the scalar perturbations.

Decomposition of 3-dimensional vectors and tensors

Vectors In Fourier space, any 3-vector field V_i can be uniquely decomposed into longitudinal and transverse parts:

$$V_i = k_i V + \overline{V}_i, \quad \text{with} \quad k^i \overline{V}_i = 0, \tag{4.2.49}$$

such that $\overline{V}_i$ lies in the 2-dimensional subspace orthogonal to the wavevector k_i. Thus, V_i is partitioned into one scalar degree of freedom (V) and two transverse vector modes ($\overline{V}_i$). Choosing an orthonormal basis $\{e_i^1, e_i^2\}$ for the 2D plane perpendicular to k_i, satisfying:

$$e_i^a k_j \gamma^{ij} = 0, \quad e_i^a e_j^b \gamma^{ij} = \delta^{ab} \qquad (a, b \in \{1, 2\}),$$

which is defined up to a rotation around the wavevector k_i, the transverse vector can be expanded as:

$$\overline{V}_i(k_i, \eta) = \sum_{a=1,2} V_a(\hat{k}_i, \eta)\, e_i^a(\hat{k}_i). \tag{4.2.50}$$

This decomposition defines two degrees of freedom V_a, which depend on $\hat{k}_i = k_i/k$. The transverse projector onto this subspace is:

$$P_{ij} \equiv e_i^1 e_j^1 + e_i^2 e_j^2 = \gamma_{ij} - \hat{k}_i \hat{k}_j, \tag{4.2.51}$$

which satisfies $P^i{}_j P^j{}_k = P^i{}_k$, $P^i{}_j k^j = 0$, and $P^{ij}\gamma_{ij} = 2$. Consequently, any vector field can be decomposed as:

$$V_i = (\hat{k}^j V_j)\hat{k}_i + P^j{}_i V_j, \tag{4.2.52}$$

providing a direct scalar-vector projection.

Tensors Similarly, any 3-dimensional symmetric rank-2 tensor field V_{ij} admits an irreducible SVT decomposition:

$$V_{ij} = T\gamma_{ij} + \Delta_{ij} S + 2D_{(i}\overline{V}_{j)} + 2\overline{V}_{ij}, \tag{4.2.53}$$

where $\Delta_{ij} \equiv D_i D_j - \frac{1}{3}\Delta\gamma_{ij}$, and:

$$D_i \overline{V}^i = 0, \quad \overline{V}^i{}_i = 0 = D_i \overline{V}^{ij}. \tag{4.2.54}$$

The symmetric tensor $\overline{V}_{ij}$ is transverse and trace-free, possessing two independent physical degrees of freedom corresponding to the two polarization states of gravitational waves:

$$\overline{V}_{ij}(k_i, \eta) = \sum_{\lambda=+,\times} V_\lambda(k_i, \eta)\, \varepsilon_{ij}^\lambda(\hat{k}_i), \tag{4.2.55}$$

where the polarization tensors are defined by:

$$\varepsilon_{ij}^\lambda = \frac{e_i^1 e_j^1 - e_i^2 e_j^2}{\sqrt{2}}\delta_+^\lambda + \frac{e_i^1 e_j^2 + e_i^2 e_j^1}{\sqrt{2}}\delta_\times^\lambda. \tag{4.2.56}$$

These polarization tensors are traceless ($\varepsilon^{\lambda}_{ij}\gamma^{ij} = 0$), transverse ($\varepsilon^{\lambda}_{ij}k^{i} = 0$), and mutually orthonormal ($\varepsilon^{\lambda}_{ij}\varepsilon^{ij}_{\mu} = \delta^{\lambda}_{\mu}$).

Introducing the transverse-traceless projection operator:

$$\Lambda^{ab}_{ij} = P^{a}{}_{i}P^{b}{}_{j} - \frac{1}{2}P_{ij}P^{ab},$$

and the trace-free longitudinal projector:

$$T^{j}{}_{i} = \hat{k}_{i}\hat{k}^{j} - \frac{1}{3}\delta^{j}{}_{i},$$

the full SVT components of V_{ij} are extracted according to:

$$V_{ij} = \left(\frac{1}{3}V_{ab}\gamma^{ab}\right)\gamma_{ij} + \left(\frac{3}{2}V_{ab}T^{ab}\right)T_{ij} + 2\hat{k}_{(i}\left[P^{a}{}_{j)}\hat{k}^{b}V_{ab}\right] + \Lambda^{ab}_{ij}V_{ab}. \tag{4.2.57}$$

In this decomposition, V_{ij} is separated into two scalar degrees of freedom (T and S), two vector degrees of freedom ($\overline{V}_{i}$), and two tensor degrees of freedom ($\overline{V}_{ij}$). Any vector equation $V_{i} = 0$ can therefore be projected along $\hat{k}_{i}$ (scalar) and $P^{i}{}_{j}$ (vector), while any tensor equation $V_{ij} = 0$ can be projected along γ_{ij} (trace scalar), T^{ij} (longitudinal scalar), $P^{i}{}_{l}\hat{k}^{j}$ (vector), and Λ^{ij}_{ab} (tensor).

4.2.5 Einstein equations: SVT decomposition

Tensor modes

Tensor perturbations are gauge invariant by construction. Projecting the perturbed Einstein equations with the transverse-traceless operator,

$$\Lambda^{ij}_{kl}\,\delta G_{ij} = \kappa\,\Lambda^{ij}_{kl}\,\delta T_{ij},$$

yields a single dynamical equation governing gravitational waves:

$$\boxed{\overline{E}''_{kl} + 2\mathcal{H}\overline{E}'_{kl} + (2K - \Delta)\overline{E}_{kl} = \kappa a^{2}P\overline{\pi}_{kl}.} \tag{4.2.58}$$

Apart from the spatial curvature term $2K$ and the Hubble friction damping $2\mathcal{H}\overline{E}'_{kl}$ arising from cosmic expansion, this equation is the exact general-relativistic analogue of the gravitational-wave equation (1.132) derived in Chapter 1 in flat space-time.

Using the cosmological density parameter $\Omega \equiv \kappa\rho a^{2}/(3\mathcal{H}^{2})$ defined in (3.36), this wave equation can be recast as:

$$\overline{E}''_{kl} + 2\mathcal{H}\overline{E}'_{kl} + (2K - \Delta)\overline{E}_{kl} = 3\mathcal{H}^{2}\Omega w\overline{\pi}_{kl}. \tag{4.2.59}$$

It is also convenient to isolate each polarization mode by introducing the canonical Mukhanov-type field variable:

$$u_{\mathrm{T}} = a\overline{E}_{\lambda} \quad \text{and} \quad \overline{E}_{ij} = \sum_{\lambda}\overline{E}_{\lambda}\varepsilon^{\lambda}_{ij}, \tag{4.2.60}$$

which transforms Eq. (4.2.58) into:

$$u''_{\mathrm{T}} + \left(2K - \Delta - \frac{a''}{a}\right)u_{\mathrm{T}} = \kappa a^{3}P\overline{\pi}_{\lambda}. \tag{4.2.61}$$

We can write

$$u_{\mathrm{T}}'' - -\frac{a''}{a}u_{\mathrm{T}} = \kappa a^3 P\overline{\pi}_\lambda - (2K - \Delta)\, u_{\mathrm{T}}. \tag{1}$$

Then it becomes obvious that $u = a(\eta)$ is one solution of the LHS and second solution can be found by the method of reduction by considering $u = y(\eta)a(\eta)$, which leads to

$$u' = y'a + ya' \tag{2}$$
$$u'' = y''a + 2y'a' + ya'' \tag{3}$$

Plugging it back to EoM, we get

$$u'' - \frac{a''}{a}u = 0 \implies (y''a + 2y'a' + ya'') - \frac{a''}{a}(ya) = 0 \tag{4}$$
$$\frac{y''}{y} = -\frac{2a''}{a} \tag{5}$$
$$\ln(y') = \ln\left(\frac{1}{a^2}\right) \tag{6}$$

which gives

$$y(\eta) = \int^\eta \frac{d\eta}{a^2} \tag{7}$$

For a spatially flat Universe ($K = 0$), the homogeneous left-hand side is mathematically equivalent to the equation of a field propagating in a time-dependent effective potential $V_{\mathrm{eff}}(\eta) = a''/a$. This has the same structure as a 1D time-independent Schrödinger equation with potential $V(x)$, a correspondence that will prove especially fruitful when analyzing the quantum generation of gravitational waves during inflation (Chapter 8).

Vector modes

The vector sector of the Einstein equations provides two relations from the $(0i)$ and (ij) components, respectively. Using the perturbed Ricci tensor and the vector components of the energy-momentum tensor (4.2.24) and (4.2.25), one obtains:

$$\boxed{(\Delta + 2K)\overline{\Phi}_i = -2\kappa\rho a^2(1+w)\overline{V}_i,} \tag{4.2.62}$$

$$\boxed{\overline{\Phi}_i' + 2\mathcal{H}\overline{\Phi}_i = \kappa P a^2 \overline{\pi}_i.} \tag{4.2.63}$$

In terms of the cosmological density parameter Ω, these relations become:

$$(\Delta + 2K)\overline{\Phi}_i = -6\mathcal{H}^2\Omega(1+w)\overline{V}_i, \tag{4.2.64}$$
$$\overline{\Phi}_i' + 2\mathcal{H}\overline{\Phi}_i = 3\mathcal{H}^2\Omega w\overline{\pi}_i. \tag{4.2.65}$$

Scalar modes

To extract the Einstein equations for scalar perturbations, we evaluate four independent combinations:

- $\delta G^0{}_0 + 3\mathcal{H}D^{-1}{}^i\delta G^0{}_i$,
- the traceless part of $\delta G^i{}_j$,
- $\delta G^0{}_i$,

- $\delta G^i{}_i + 3c_s^2 \delta G^0{}_0$.

The first two yield two algebraic constraint equations:

$$(\Delta + 3K)\Psi = \frac{\kappa}{2} a^2 \rho\, \delta^{\rm C} = \frac{3}{2}(\mathcal{H}^2 + K)\delta^C, \tag{4.2.66}$$

$$\Psi - \Phi = \kappa a^2 P \bar{\pi}. \tag{4.2.67}$$

Equation (4.2.66) is Hamiltonian constraint and it reproduces the familiar Poisson equation when formulated in terms of the comoving density contrast $\delta^{\rm C}$. While equation (4.2.67) being momentum constraint shows that the two metric potentials Φ and Ψ are identical whenever the scalar anisotropic stress vanishes ($\bar{\pi} = 0$), which is the case for perfect fluids.

The remaining two combinations give dynamical evolution equations:

$$\Psi' + \mathcal{H}\Phi = -\frac{\kappa}{2} a^2 \rho (1+w) V, \tag{4.2.68}$$

$$\begin{aligned} \Psi'' + 3\mathcal{H}(1+c_s^2)\Psi' + \left[2\mathcal{H}' + (\mathcal{H}^2 - K)(1+3c_s^2)\right]\Psi - c_s^2 \Delta\Psi = \\ - (\mathcal{H}^2 + 2\mathcal{H}' + K)\left[\frac{1}{2}\Gamma + (3\mathcal{H}^2 + 2\mathcal{H}')\bar{\pi} + \mathcal{H}\bar{\pi}' + \frac{1}{3}\Delta\bar{\pi}\right] \\ - 9c_s^2 \mathcal{H}^2 (\mathcal{H}^2 + K)\bar{\pi}, \end{aligned} \tag{4.2.69}$$

where in Eq. (4.2.69) we have eliminated Φ in favor of Ψ and $\bar{\pi}$ using (4.2.67).

Equivalent forms

It is instructive to express the scalar Einstein equations using other gauge invariant density contrasts. In particular, the Poisson equation (4.2.66) takes the following forms when written in terms of $\delta^{\rm N}$ or $\delta^{\rm F}$ using (4.2.36) and (2.3.7) with $P = w\rho$:

$$(\Delta + 3K)\Psi = \frac{\kappa}{2} a^2 \rho \left[\delta^{\rm N} - 3\mathcal{H}(1+w)V\right], \tag{4.2.70}$$

$$\Delta\Psi - 3\mathcal{H}^2 X = \frac{\kappa}{2} a^2 \rho\, \delta^{\rm F}, \tag{4.2.71}$$

To derive (4.2.66) in flat slicing, we need to do some work. Let us start from

$$(\Delta + 3K)\Psi = \frac{3}{2}(\mathcal{H}^2 + K)\delta^{\rm C}, \tag{1}$$

where we have used the background Friedmann equation

$$\kappa a^2 \rho = 3(\mathcal{H}^2 + K). \tag{2}$$

The gauge-invariant density perturbations in the flat and comoving slicings are related to the Newtonian-gauge density perturbation by

$$\delta^{\rm F} = \delta^{\rm N} + \frac{\rho'}{\rho}\frac{\Psi}{\mathcal{H}}, \qquad \delta^{\rm C} = \delta^{\rm N} + \frac{\rho'}{\rho} V. \tag{3}$$

Hence

$$\delta^{\rm C} - \delta^{\rm F} = \frac{\rho'}{\rho}\left(V - \frac{\Psi}{\mathcal{H}}\right). \tag{4}$$

For a constant equation of state, $P = w\rho$, the background conservation equation gives

$$\frac{\rho'}{\rho} = -3\mathcal{H}(1+w), \tag{5}$$

and therefore

$$\delta^{\mathrm{C}} = \delta^{\mathrm{F}} - 3\mathcal{H}(1+w)V + 3(1+w)\Psi. \tag{6}$$

The scalar $0i$ Einstein equation in Newtonian gauge is

$$\Psi' + \mathcal{H}\Phi = -\frac{\kappa}{2}a^2(\rho + P)V. \tag{7}$$

Using $P = w\rho$ together with (2), this gives

$$V = -\frac{2(\Psi' + \mathcal{H}\Phi)}{3(1+w)(\mathcal{H}^2 + K)}. \tag{8}$$

Substituting this into (6), we obtain

$$\delta^{\mathrm{C}} = \delta^{\mathrm{F}} + \frac{2\mathcal{H}(\Psi' + \mathcal{H}\Phi)}{\mathcal{H}^2 + K} + 3(1+w)\Psi. \tag{9}$$

Substituting (9) into the comoving constraint (1) gives

$$(\Delta + 3K)\Psi = \frac{3}{2}(\mathcal{H}^2 + K)\delta^{\mathrm{F}} + 3\mathcal{H}(\Psi' + \mathcal{H}\Phi) + \frac{9}{2}(1+w)(\mathcal{H}^2 + K)\Psi. \tag{10}$$

The background acceleration equation is

$$\mathcal{H}' = -\frac{1}{2}(1+3w)(\mathcal{H}^2 + K), \tag{11}$$

which implies

$$\mathcal{H}^2 - \mathcal{H}' + K = \frac{3}{2}(1+w)(\mathcal{H}^2 + K). \tag{12}$$

Hence

$$\frac{9}{2}(1+w)(\mathcal{H}^2 + K) = 3(\mathcal{H}^2 - \mathcal{H}' + K), \tag{13}$$

and (10) becomes

$$(\Delta + 3K)\Psi = \frac{3}{2}(\mathcal{H}^2 + K)\delta^{\mathrm{F}} + 3\mathcal{H}\Psi' + 3\mathcal{H}^2\Phi + 3(\mathcal{H}^2 - \mathcal{H}' + K)\Psi. \tag{14}$$

Now define

$$X = \Phi + \Psi + \left(\frac{\Psi}{\mathcal{H}}\right)'. \tag{15}$$

Expanding the derivative,

$$\mathcal{H}^2 X = \mathcal{H}^2\Phi + \mathcal{H}^2\Psi + \mathcal{H}\Psi' - \mathcal{H}'\Psi, \tag{16}$$

so that

$$3\mathcal{H}^2 X = 3\mathcal{H}^2\Phi + 3\mathcal{H}\Psi' + 3(\mathcal{H}^2 - \mathcal{H}')\Psi. \tag{17}$$

Therefore (14) becomes

$$(\Delta + 3K)\Psi = \frac{3}{2}(\mathcal{H}^2 + K)\delta^{\mathrm{F}} + 3\mathcal{H}^2 X + 3K\Psi. \tag{18}$$

The $3K\Psi$ terms cancel from both sides, yielding

$$\boxed{\Delta\Psi - 3\mathcal{H}^2 X = \frac{3}{2}(\mathcal{H}^2 + K)\delta^{\mathrm{F}}}. \tag{19}$$

Finally, using (2),

$$\boxed{\Delta\Psi - 3\mathcal{H}^2 X = \frac{\kappa}{2}a^2\rho\,\delta^{\mathrm{F}}}. \tag{20}$$

while the second-order evolution equation (4.2.69) can be expressed entirely through the single variable X:

$$\mathcal{H}X' + (\mathcal{H}^2 + 2\mathcal{H}')X = \frac{\kappa}{2}a^2\rho\left(w\Gamma + c_{\rm s}^2\delta^{\rm F} + \frac{2}{3}w\Delta\bar{\pi}\right). \tag{4.2.72}$$

4.2.6 Perturbed conservation equation for a fluid

The conservation of the energy-momentum tensor, expanded to linear order, reads:

$$\delta(\nabla_\mu T^\mu{}_\nu) = 0.$$

Explicit geometric connection coefficients and contracted Christoffel symbols are provided in Appendix C. Because perfect fluids do not support transverse-traceless shear stresses at linear order, there are no tensor conservation equations; the dynamics separates into vector and scalar sectors.

Vector modes

For the vector perturbations, the conservation law yields a single differential equation:

$$\boxed{\overline{V}_i' + \mathcal{H}(1 - 3c_{\rm s}^2)\overline{V}_i = -\frac{1}{2}\frac{w}{1+w}(\Delta + 2K)\overline{\pi}_i.} \tag{4.2.73}$$

This first-order equation is readily solvable in closed form in most cosmological scenarios of interest.

Scalar modes

Projecting the 4-dimensional conservation equation along timelike and spacelike directions generates two equations analogous to the continuity and Euler equations, respectively. In the longitudinal (Newtonian) gauge, these read:

$$\boxed{\delta^{\rm N\prime} + 3\mathcal{H}(c_{\rm s}^2 - w)\delta^{\rm N} = -(1+w)(\Delta V - 3\Psi') - 3\mathcal{H}w\Gamma,} \tag{4.2.74}$$

Consider the Newtonian-gauge metric

$$\mathrm{d}s^2 = a^2(\eta)\left[(1+2\Phi)\,\mathrm{d}\eta^2 - (1-2\Psi)\delta_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j\right], \tag{1}$$

where

$$\mathcal{H} \equiv \frac{a'}{a}. \tag{2}$$

We decompose the energy density and pressure as

$$\rho = \bar{\rho} + \delta\rho, \qquad P = \bar{P} + \delta P, \tag{3}$$

and define the Newtonian-gauge density contrast by

$$\delta^{\rm N} \equiv \frac{\delta\rho}{\bar{\rho}}. \tag{4}$$

The energy-momentum tensor is

$$T^\mu{}_\nu = (\rho + P)u^\mu u_\nu - P\delta^\mu_\nu + \Pi^\mu{}_\nu, \tag{5}$$

where $\Pi^{\mu}{}_{\nu}$ denotes the anisotropic stress. To first order,

$$T^{0}{}_{0} = \bar{\rho} + \delta\rho, \tag{6}$$
$$T^{i}{}_{0} = (\bar{\rho} + \bar{P})V^{i}, \tag{7}$$
$$T^{0}{}_{i} = -(\bar{\rho} + \bar{P})V_{i}, \tag{8}$$
$$T^{i}{}_{j} = -(\bar{P} + \delta P)\delta^{i}_{j} + \Pi^{i}{}_{j}, \tag{9}$$

with

$$\Pi^{i}{}_{i} = 0. \tag{10}$$

We start from the covariant conservation equation

$$\nabla_{\mu} T^{\mu}{}_{\nu} = 0. \tag{11}$$

For a mixed tensor,

$$\nabla_{\mu} T^{\mu}{}_{\nu} = \partial_{\mu} T^{\mu}{}_{\nu} + \Gamma^{\mu}_{\mu\lambda} T^{\lambda}{}_{\nu} - \Gamma^{\lambda}_{\mu\nu} T^{\mu}{}_{\lambda}. \tag{12}$$

Taking $\nu = 0$ gives

$$\nabla_{\mu} T^{\mu}{}_{0} = \partial_{\mu} T^{\mu}{}_{0} + \Gamma^{\mu}_{\mu\lambda} T^{\lambda}{}_{0} - \Gamma^{\lambda}_{\mu 0} T^{\mu}{}_{\lambda} = 0. \tag{13}$$

Expanding the repeated indices into temporal and spatial parts,

$$\begin{aligned} 0 = {} & \partial_0 T^{0}{}_{0} + \partial_i T^{i}{}_{0} + \Gamma^{\mu}_{\mu 0} T^{0}{}_{0} + \Gamma^{\mu}_{\mu i} T^{i}{}_{0} \\ & - \Gamma^{0}_{00} T^{0}{}_{0} - \Gamma^{0}_{i0} T^{i}{}_{0} - \Gamma^{j}_{00} T^{0}{}_{j} - \Gamma^{j}_{i0} T^{i}{}_{j}. \end{aligned} \tag{14}$$

The Christoffel symbols required below are

$$\Gamma^{0}_{00} = \mathcal{H} + \Phi', \tag{15}$$
$$\Gamma^{0}_{i0} = \partial_i \Phi, \tag{16}$$
$$\Gamma^{j}_{00} = \partial^{j} \Phi, \tag{17}$$
$$\Gamma^{j}_{i0} = (\mathcal{H} - \Psi')\delta^{j}_{i}. \tag{18}$$

Furthermore,

$$\Gamma^{\mu}_{\mu 0} = \Gamma^{0}_{00} + \Gamma^{i}_{i0} = 4\mathcal{H} + \Phi' - 3\Psi'. \tag{19}$$

We now evaluate the eight terms in Eq. (14) separately.

First term:

$$\partial_0 T^{0}{}_{0} = \partial_0(\bar{\rho} + \delta\rho) = \bar{\rho}' + \delta\rho'. \tag{20}$$

Second term: Using the homogeneity of the background,

$$\partial_i T^{i}{}_{0} = (\bar{\rho} + \bar{P})\partial_i V^{i}. \tag{21}$$

Defining

$$\Delta V \equiv \partial_i V^{i}, \tag{22}$$

we have

$$\partial_i T^{i}{}_{0} = (\bar{\rho} + \bar{P})\Delta V. \tag{23}$$

Third term:

$$\begin{aligned} \Gamma^{\mu}_{\mu 0} T^{0}{}_{0} &= (4\mathcal{H} + \Phi' - 3\Psi')(\bar{\rho} + \delta\rho) \\ &= 4\mathcal{H}\bar{\rho} + 4\mathcal{H}\delta\rho + \bar{\rho}\Phi' - 3\bar{\rho}\Psi', \end{aligned} \tag{24}$$

where products of perturbations have been discarded.

Fourth term: Since $\Gamma^{\mu}_{\mu i}$ has no background contribution, it is already first order, while $T^{i}{}_{0}$ is also first order. Hence

$$\Gamma^{\mu}_{\mu i} T^{i}{}_{0} = \mathcal{O}(2) = 0 \tag{25}$$

at linear order.

Fifth term:

$$\begin{aligned} -\Gamma^{0}_{00} T^{0}{}_{0} &= -(\mathcal{H} + \Phi')(\bar{\rho} + \delta\rho) \\ &= -\mathcal{H}\bar{\rho} - \mathcal{H}\delta\rho - \bar{\rho}\Phi'. \end{aligned} \tag{26}$$

Sixth term: Since both $\Gamma^{0}_{i0} = \partial_i \Phi$ and $T^{i}{}_{0} = (\bar{\rho} + \bar{P}) V^{i}$ are first order,

$$-\Gamma^{0}_{i0} T^{i}{}_{0} = \mathcal{O}(2) = 0. \tag{27}$$

Seventh term: Similarly, $\Gamma^{j}_{00} = \partial^{j}\Phi$ and $T^{0}{}_{j} = -(\bar{\rho} + \bar{P}) V_j$ are both first order, so

$$-\Gamma^{j}_{00} T^{0}{}_{j} = \mathcal{O}(2) = 0. \tag{28}$$

Eighth term: Using

$$T^{i}{}_{j} = -(\bar{P} + \delta P)\delta^{i}_{j} + \Pi^{i}{}_{j}, \tag{29}$$

we obtain

$$\begin{aligned} -\Gamma^{j}_{i0} T^{i}{}_{j} &= -(\mathcal{H} - \Psi')\delta^{j}_{i} \left[-(\bar{P} + \delta P)\delta^{i}_{j} + \Pi^{i}{}_{j} \right] \\ &= (\mathcal{H} - \Psi') \left[3(\bar{P} + \delta P) - \Pi^{i}{}_{i} \right] \\ &= 3\mathcal{H}\bar{P} + 3\mathcal{H}\delta P - 3\bar{P}\Psi', \end{aligned} \tag{30}$$

where $\Pi^{i}{}_{i} = 0$ has been used.
Substituting all eight contributions into Eq. (14), we obtain

$$\begin{aligned} 0 = {} & \bar{\rho}' + \delta\rho' + (\bar{\rho} + \bar{P})\Delta V \\ & + 4\mathcal{H}\bar{\rho} + 4\mathcal{H}\delta\rho + \bar{\rho}\Phi' - 3\bar{\rho}\Psi' \\ & - \mathcal{H}\bar{\rho} - \mathcal{H}\delta\rho - \bar{\rho}\Phi' \\ & + 3\mathcal{H}\bar{P} + 3\mathcal{H}\delta P - 3\bar{P}\Psi'. \end{aligned} \tag{31}$$

The background terms give

$$\boxed{\bar{\rho}' + 3\mathcal{H}(\bar{\rho} + \bar{P}) = 0,} \tag{32}$$

while the first-order terms give

$$\delta\rho' + 3\mathcal{H}(\delta\rho + \delta P) + (\bar{\rho} + \bar{P})(\Delta V - 3\Psi') = 0. \tag{33}$$

Hence

$$\boxed{\delta\rho' + 3\mathcal{H}(\delta\rho + \delta P) = -(\bar{\rho} + \bar{P})(\Delta V - 3\Psi').} \tag{34}$$

Now define

$$w \equiv \frac{\bar{P}}{\bar{\rho}}, \qquad \delta^{\mathrm{N}} \equiv \frac{\delta\rho}{\bar{\rho}}. \tag{35}$$

Since

$$\delta\rho = \bar{\rho}\delta^{\mathrm{N}}, \tag{36}$$

we have

$$\delta\rho' = \bar{\rho}'\delta^{\mathrm{N}} + \bar{\rho}\delta^{\mathrm{N}\prime}. \tag{37}$$

Using Eq. (32),

$$\frac{\bar{\rho}'}{\bar{\rho}} = -3\mathcal{H}(1+w), \tag{38}$$

so Eq. (34) becomes

$$\boxed{\delta^{\mathrm{N}\prime} - 3\mathcal{H}w\delta^{\mathrm{N}} + 3\mathcal{H}\frac{\delta P}{\bar{\rho}} = -(1+w)(\Delta V - 3\Psi').} \tag{39}$$

Finally, define the adiabatic sound speed

$$c_s^2 \equiv \frac{\bar{P}'}{\bar{\rho}'}, \tag{40}$$

and the entropy perturbation Γ through

$$\boxed{\delta P = c_s^2\delta\rho + \bar{P}\Gamma.} \tag{41}$$

Since

$$\delta\rho = \bar{\rho}\delta^{\mathrm{N}}, \qquad \bar{P} = w\bar{\rho}, \tag{42}$$

we have

$$\frac{\delta P}{\bar{\rho}} = c_s^2\delta^{\mathrm{N}} + w\Gamma. \tag{43}$$

Substituting this into Eq. (39), we finally obtain

$$\boxed{\delta^{\mathrm{N}\prime} + 3\mathcal{H}(c_s^2 - w)\delta^{\mathrm{N}} = -(1+w)(\Delta V - 3\Psi') - 3\mathcal{H}w\Gamma.} \tag{44}$$

$$\boxed{V' + \mathcal{H}(1-3c_{\mathrm{s}}^2)V = -\Phi - \frac{c_{\mathrm{s}}^2}{1+w}\delta^{\mathrm{N}} - \frac{w}{1+w}\left[\Gamma + \frac{2}{3}(\Delta+3K)\bar{\pi}\right].} \tag{4.2.75}$$

Using the background identity for w' [Eq. (C.19)], the continuity equation (4.2.74) can be recast into the compact form:

$$\left(\frac{\delta^{\mathrm{N}}}{1+w}\right)' = -(\Delta V - 3\Psi') - 3\mathcal{H}\frac{w}{1+w}\Gamma. \tag{4.2.76}$$

In terms of the flat-slicing and comoving density contrasts δ^{F} and δ^{C}, these equations become:

$$\delta^{\mathrm{F}\prime} + 3\mathcal{H}(c_{\mathrm{s}}^2 - w)\delta^{\mathrm{F}} = -(1+w)\Delta V - 3\mathcal{H}w\Gamma, \tag{4.2.77}$$

$$V' + \mathcal{H}(1-3c_{\mathrm{s}}^2)V = -\Phi - 3c_{\mathrm{s}}^2\Psi - \frac{c_{\mathrm{s}}^2}{1+w}\delta^{\mathrm{F}} - \frac{w}{1+w}\left[\Gamma + \frac{2}{3}(\Delta+3K)\bar{\pi}\right], \tag{4.2.78}$$

$$V' + \mathcal{H}V = -\Phi - \frac{c_{\mathrm{s}}^2}{1+w}\delta^{\mathrm{C}} - \frac{w}{1+w}\left[\Gamma + \frac{2}{3}(\Delta+3K)\bar{\pi}\right]. \tag{4.2.79}$$

4.2.7 Interpretation of the perturbation equations

At first glance, the relativistic fluid equations resemble their Newtonian counterparts. Indeed, for a static background ($\mathcal{H} = 0$), non-relativistic pressureless dust ($w = c_{\mathrm{s}}^2 = 0$, $\bar{\pi} = 0$), and purely adiabatic perturbations ($\Gamma = 0$), identifying δ with δ^{F} immediately recovers the classical

relations (4.1.1) and (4.1.2). However, general relativity introduces five profound structural differences:

1. **Mode Diversity:** Relativistic perturbation theory encompasses vector and tensor degrees of freedom alongside scalar modes. In the linear regime around a Friedmann–Lemaître background, these sectors decouple completely.

2. **Dual Gravitational Potentials:** The scalar sector requires two independent potentials, Φ and Ψ. In the Newtonian limit and on sub-Hubble scales, they converge to a single potential provided anisotropic stresses are negligible. Notably, the relativistic Poisson equation aligns most closely with its Newtonian form in the comoving gauge, yet it involves the spatial curvature potential Ψ rather than the acceleration potential Φ.

3. **Super-Hubble Deviations and Gauge Ambiguity:** On scales larger than the Hubble radius, the Poisson equation deviates significantly from Newtonian gravity, and different gauge-invariant definitions of density contrast ($\delta^{\mathrm{N}}, \delta^{\mathrm{F}}, \delta^{\mathrm{C}}$) diverge markedly from one another.

4. **Relativistic Inertial Enthalpy:** The continuity equation contains an explicit factor of $(1 + w)$, reflecting the fact that in relativity, energy flux (which includes pressure and enthalpy) rather than rest-mass flux governs gravitational dynamics.

5. **Cosmic Expansion Damping:** The Euler equation features an expansion damping term $-\mathcal{H}(1 - 3c_{\mathrm{s}}^2)V$. For non-relativistic matter ($c_{\mathrm{s}}^2 \approx 0$), this term forces peculiar velocities to decay as a^{-1}, causing the velocity field to align asymptotically with the Hubble flow. Conversely, for ultra-relativistic radiation ($c_{\mathrm{s}}^2 = 1/3$), this damping vanishes identically, which can be understood from the conformal invariance of null geodesics. (A rigorous kinetic treatment of radiation will be developed in Chapter 6, where velocity appears as the dipole moment of the photon distribution.)

4.3 Evolution

The perturbation equations established above can be integrated analytically in several fundamental regimes. In general, it is highly advantageous to expand spatial perturbations into Fourier modes (Appendix B) and examine the time evolution of each mode independently. Because spatial derivatives enter the linearized field equations solely through the Laplacian Δ, Fourier transformation maps the partial differential equations into decoupled ordinary differential equations for each wavenumber k.

In analyzing mode evolution, one distinguishes two distinct physical regimes based on weather $\mathcal{H}$ is increasing function of time or not:

$k \ll \mathcal{H}$	**super-Hubble mode**	wavelength larger than Hubble radius,
$k \gg \mathcal{H}$	**sub-Hubble mode**	wavelength smaller than Hubble radius.

In the cosmological literature, super-Hubble and sub-Hubble modes are often casually referred to as "super-horizon" and "sub-horizon" modes. This terminology originates historically from the standard Big Bang model (Chapter 4), where the Hubble radius $\mathcal{H}^{-1}$ is numerically comparable to the particle horizon. However, the two concepts are fundamentally distinct, especially in cosmological models featuring an inflationary epoch (Chapter 8), whose primary purpose is precisely to decouple the Hubble radius from the causal particle horizon. The suppression of spatial gradients on super-Hubble scales is not an enforcement of causal disconnectivity, but rather reflects the dynamical dominance of cosmic expansion damping ($\mathcal{H}^{-1}$) over the characteristic oscillation timescale of the mode ($\sim k^{-1}$).

So the modes which are superhubble has time period so large that they have not had enough time to oscillate even once and hence they behave as static gravitational potential.

Consequently, the super-Hubble behavior of growing modes cannot be deduced merely from heuristic causality arguments, contrary to frequent claims that perturbations are causally "frozen" outside the horizon. When the wavelength of a given mode is much larger than the hubble horizon, they are effectively frozen but when the wavelength is just barely outside the hubble horizon, it feels the edges and its not quite frozen yet.

4.3.1 Vector and tensor modes

Vector modes

In the absence of anisotropic vector stresses ($\bar{\pi}_i = 0$), the Einstein equation (4.2.63) yields:

$$\bar{\Phi} + 2\frac{a'}{a}\bar{\Phi} = 0 \implies \frac{\bar{\Phi}'}{\Phi} = -2\frac{a'}{a} \implies \overline{\Phi}_i \propto a^{-2}, \tag{4.3.1}$$

while the fluid conservation equation (4.2.73) implies:

$$\overline{V}^i \propto a^{-(1-3c_s^2)}. \tag{4.3.2}$$

Thus, cosmic expansion rapidly extinguishes vector metric perturbations. For a non-relativistic cosmic fluid ($c_s^2 = 0$), fluid vorticity decays as a^{-1}. As a result, primordial vector perturbations are dynamically washed out and play virtually no role in large-scale structure formation, unless continuously replenished by active sources such as cosmic defects or primordial magnetic fields. This happens because vector mode has no constant or growing part in the solution unlike the scalar and tensor perturbation if $\bar{\pi}_i = 0$.

Tensor modes

Assuming negligible anisotropic shear ($\bar{\pi}_{ij} = 0$) in a spatially flat Universe ($K = 0$), we project the tensor wave equation (4.2.58) into Fourier space. For a background expanding as a power law in conformal time, $a(\eta) \propto \eta^\nu$, Eq. (4.2.58) reduces to:

$$\frac{\mathrm{d}^2\overline{E}_{ij}}{\mathrm{d}x^2} + \frac{2\nu}{x}\frac{\mathrm{d}\overline{E}_{ij}}{\mathrm{d}x} + \overline{E}_{ij} = 0, \tag{4.3.3}$$

where $x \equiv k\eta$. The general solution can be expressed in terms of cylindrical Bessel functions:

$$\overline{E}_{ij} = x^{1/2-\nu}\left[A_{ij}J_{\nu-\frac{1}{2}}(x) + B_{ij}N_{\nu-\frac{1}{2}}(x)\right], \tag{4.3.4}$$

where A_{ij} and B_{ij} are constant transverse-traceless polarization tensors. Retaining only the regular solution as $x \to 0$ (the Neumann mode N diverges at the origin, representing a decaying transient that rapidly vanishes as time progresses), we obtain:

$$\overline{E}_{ij} = x^{1/2-\nu}J_{\nu-\frac{1}{2}}(x)A_{ij} = x^{1-\nu}j_{\nu-1}(x)A_{ij}, \tag{4.3.5}$$

where $j_n(x)$ denotes the spherical Bessel functions. Evaluating this expression for the matter- and radiation-dominated epochs yields:

	Matter	**Radiation**
$a(\eta)$	η^2	η
$\overline{E}_{ij}(\eta)$	$3j_1(k\eta)/(k\eta)$	$j_0(k\eta)$

The tensor perturbation amplitude $\overline{E}_{ij}$ remains constant on super-Hubble scales ($k\eta < 1$), where the physical wavelength exceeds the Hubble radius. Once the mode crosses inside the Hubble radius ($k\eta > 1$), it transitions into a freely propagating gravitational wave characterized by damped oscillations (see Fig. 5.3).

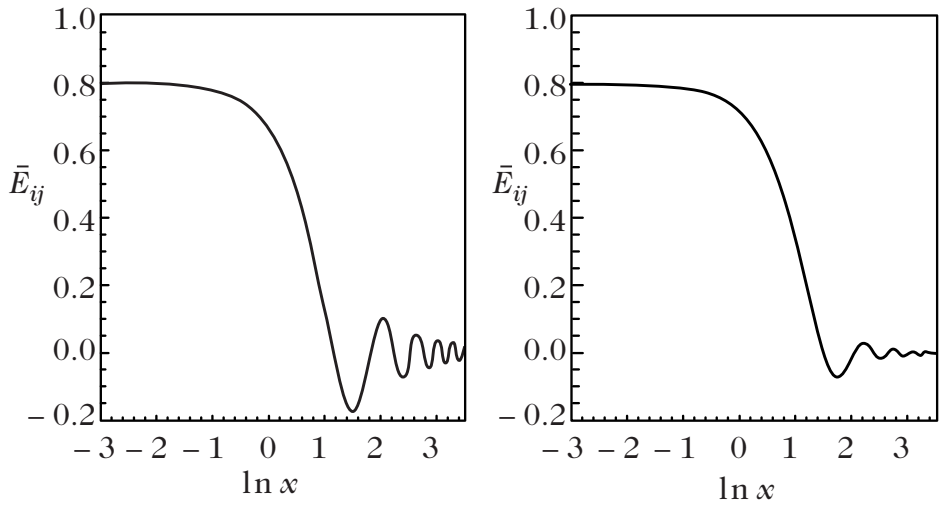


Figure 4.3: Evolution of gravitational waves in a radiation-dominated Universe ($w = 1/3, \nu = 1$, left) and pressureless matter ($w = 0, \nu = 2$, right) as a function of $\ln x = \ln(k\eta)$. The mode amplitude remains constant as long as the wavelength is super-Hubble ($x < 1$) and develops damped oscillatory behaviour as soon as it becomes sub-Hubble ($x > 1$).

Tensor modes in a Universe dominated by a mixture of matter and radiation

Let us now investigate a cosmological background driven by a simultaneous mixture of pressureless matter ($w = 0, \nu = 2$) and relativistic radiation ($w = 1/3, \nu = 1$). To parameterize the background expansion conveniently, we define the normalized scale factor:

$$y \equiv \frac{a}{a_{\rm eq}}, \tag{4.3.6}$$

The radiation and matter scales differently in the FLRW universe

$$\rho^m \propto a^{-3} \quad \Longrightarrow \quad \frac{\rho^m}{\rho^m_{\rm eq}} = \left(\frac{a}{a_{\rm eq}}\right)^{-3}, \tag{1}$$

$$\rho^r \propto a^{-4} \quad \Longrightarrow \quad \frac{\rho^r}{\rho^r_{\rm eq}} = \left(\frac{a}{a_{\rm eq}}\right)^{-4}. \tag{2}$$

But, at the matter–radiation equality, we have

$$\rho^m_{\rm eq} = \rho^r_{\rm eq}.$$

Therefore,

$$\frac{\rho^m}{\rho^r} = \frac{\rho^m_{\rm eq}}{\rho^r_{\rm eq}}\left(\frac{a}{a_{\rm eq}}\right) = \frac{a}{a_{\rm eq}}. \tag{3}$$

where $a_{\rm eq}$ designates the scale factor at the epoch of radiation-matter equality, defined by $\bar{\rho}_{\rm m}(a_{\rm eq}) = \bar{\rho}_{\rm r}(a_{\rm eq})$. The parameter y tracks the matter-to-radiation density ratio, $y = \rho_{\rm m}/\rho_{\rm r}$, and the effective equation-of-state parameter of the two-component mixture is:

$$w = \frac{1}{3}\frac{1}{1+y}. \tag{4.3.7}$$

This can be shown from a small calculation such as following

$$P = P_m + P_r = 0 + P_r = \frac{1}{3}\rho_r = \frac{1}{3}\frac{\rho_r}{\rho_m + \rho_r}\rho = \frac{1}{3}\frac{1}{1+y}\rho \tag{1}$$

Therefore the EoS of radiation+matter fluid is not constant but changing with time.

The dynamical wave equation (4.2.58) governing gravitational waves can then be recast using y as the independent evolution variable:

$$\frac{\mathrm{d}^2\overline{E}_{ij}}{\mathrm{d}y^2} + \frac{4+5y}{2y(1+y)}\frac{\mathrm{d}\overline{E}_{ij}}{\mathrm{d}y} + \left(\frac{k}{k_{\mathrm{eq}}}\right)^2 \frac{2}{1+y}\overline{E}_{ij} = 0. \tag{4.3.8}$$

To establish this differential equation, we have used the chain rule for derivatives with respect to conformal time, $\overline{E}'_{ij} = y'\,\mathrm{d}\overline{E}_{ij}\big/\mathrm{d}y$, along with the conformal Hubble parameter $\mathcal{H} = y'/y$. In terms of the variable y, the Friedmann equation reads:

$$\mathcal{H}^2 = \mathcal{H}_{\mathrm{eq}}^2 \frac{1+y}{2y^2}, \tag{4.3.9}$$

where $\mathcal{H}_{\mathrm{eq}}$ denotes the conformal Hubble parameter evaluated at equality ($y = 1$). We have also introduced the characteristic comoving wavenumber k_{eq}, representing the mode that enters the Hubble horizon at equality:

$$k_{\mathrm{eq}} \equiv \mathcal{H}_{\mathrm{eq}} = a_{\mathrm{eq}} H_{\mathrm{eq}}. \tag{4.3.10}$$

Equation (4.3.8) can be integrated numerically, and the resulting trajectories for representative modes k/k_{eq} are displayed in Fig. 4.4. As expected, modes entering the horizon earlier undergo more prolonged oscillatory damping, leading to greater suppression of their asymptotic amplitudes.

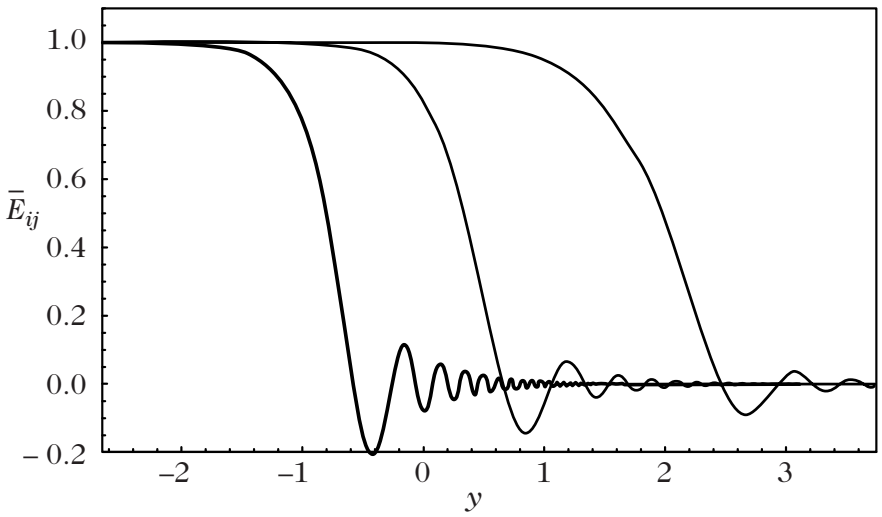


Figure 4.4: Evolution of gravitational waves in a Universe dominated by a mixture of matter and radiation in terms of $y = a/a_{\mathrm{eq}}$ ($y = 1$ at equality) for modes with $k/k_{\mathrm{eq}} = 10$, 1, and 0.1.

General solutions for long-wavelength modes

In a spatially flat Universe ($K = 0$), the gravitational wave equation (4.2.58) can be expressed as a formal Volterra integral equation:

$$u_{\rm T}(\eta, k) = a(\eta)\left[A_1(k) + A_2(k)\int^{\eta}\frac{{\rm d}\tilde{\eta}}{a^2(\tilde{\eta})} + k^2\int^{\eta}\frac{{\rm d}\tilde{\eta}}{a^2(\tilde{\eta})}\int^{\tilde{\eta}}{\rm d}\bar{\eta}\,a^2(\bar{\eta})u_{\rm T}(\bar{\eta}, k)\right]. \tag{4.3.11}$$

For super-Hubble, long-wavelength modes satisfying the condition $k^2 \ll a''/a$, this integral equation truncates to:

$$u_{\rm T}(\eta, k) = a(\eta)A_1(k) + A_2(k)a(\eta)\int^{\eta}\frac{{\rm d}\tilde{\eta}}{a^2(\tilde{\eta})} + \mathcal{O}\left(\frac{k^2 a}{a''}\right), \tag{4.3.12}$$

This was done using the method of successive approximation

$$u(\eta) = \underbrace{a(\eta)}_{y_1} + \underbrace{a(\eta)\int\frac{d\eta}{a^2}}_{y_2} + k^2\int d\eta' G(\eta, \eta')u(\eta') \tag{1}$$

where

$$G(\eta, \eta') = y_1(\eta)y_2(\eta')\Theta(\eta - \eta') + y_1(\eta')y_2(\eta)\Theta(\eta' - \eta) \tag{2}$$

where the two integration constants $A_1(k)$ and $A_2(k)$ are functions of wavenumber fixed by initial conditions. In analogy with quantum mechanical scattering in a one-dimensional barrier, Eq. (4.3.12) corresponds to the leading Born approximation for wave propagation through an effective potential.

4.3.2 Evolution of the gravitational potential

We now turn our attention to the dynamical evolution of the longitudinal gravitational potential. Neglecting anisotropic stress contributions ($\bar{\pi} = 0$), the scalar evolution equation (4.2.69) takes the form:

$$\Phi'' + 3\mathcal{H}\left(1 + c_{\rm s}^2\right)\Phi' + \left[2\mathcal{H}' + \left(\mathcal{H}^2 - K\right)\left(1 + 3c_{\rm s}^2\right)\right]\Phi - c_{\rm s}^2\Delta\Phi = \frac{\kappa}{2}a^2 P\Gamma. \tag{4.3.13}$$

It can be derived from (4.2.69) with $\bar{\pi} = 0$ and $\kappa P a^2 = -(\mathcal{H}^2 + 2\mathcal{H}' + K)$.

Another formulation of the evolution equation

A convenient reduced form of Eq. (4.3.13) can be established by introducing the background auxiliary function:

$$\theta \equiv \frac{1}{a}\left(\frac{\rho}{\rho + P}\right)^{1/2}\left(1 - \frac{3K}{\kappa\rho a^2}\right)^{1/2} = \frac{1}{a}\sqrt{\frac{\kappa\rho a^2 - 3K}{\kappa a^2(\rho + P)}}, \tag{4.3.14}$$

which, using the Friedmann equations, can be rewritten as:

$$\theta = \frac{\mathcal{H}}{a}\left[\frac{2}{3}\left(\mathcal{H}^2 - \mathcal{H}' + K\right)\right]^{-1/2}. \tag{4.3.15}$$

Derived from (4.3.14) using

$$\kappa\rho a^2 = 3(\mathcal{H}^2 + K) \qquad \kappa P a^2 = -(\mathcal{H}^2 + 2\mathcal{H}' + K) \tag{1}$$

Defining the rescaled Mukhanov–Sasaki type variable:

$$u_{\rm S} \equiv \frac{2}{3}\frac{a^2\theta}{\mathcal{H}}\Phi, \tag{4.3.16}$$

a lengthy but direct calculation transforms Eq. (4.3.13) into the canonical wave equation:

$$u_{\rm S}'' - \left(\frac{\theta''}{\theta} + c_{\rm s}^2\Delta\right) u_{\rm S} = \frac{\kappa}{3}\frac{\theta}{\mathcal{H}} a^4 P\Gamma. \tag{4.3.17}$$

Equation (4.3.17) shares structural similarity with the tensor mode equation (4.2.61) and can be analyzed using identical mathematical techniques.

We first collect the background relations that will be needed. The Friedmann equations can be written as

$$\kappa\rho a^2 = 3(\mathcal{H}^2 + K), \qquad \kappa P a^2 = -(\mathcal{H}^2 + 2\mathcal{H}' + K), \tag{1}$$

while energy conservation gives

$$\rho' = -3\mathcal{H}(\rho + P). \tag{2}$$

For an adiabatic background,

$$c_s^2 = \frac{P'}{\rho'}. \tag{3}$$

Subtracting the two equations in (1),

$$\mathcal{H}^2 - \mathcal{H}' + K = \frac{\kappa}{2}a^2(\rho + P), \tag{4}$$

allows the definition of θ to be written in the simpler form

$$\theta = \frac{\sqrt{3}\,\mathcal{H}}{a^2\sqrt{\kappa(\rho + P)}}. \tag{5}$$

This form makes its logarithmic derivative particularly simple. Since

$$\rho' + P' = -3\mathcal{H}(1 + c_s^2)(\rho + P), \tag{1}$$

we have

$$\frac{1}{2}\frac{(\rho + P)'}{\rho + P} = -\frac{3}{2}\mathcal{H}(1 + c_s^2). \tag{6}$$

Consequently,

$$\frac{\theta'}{\theta} = \frac{\mathcal{H}'}{\mathcal{H}} - 2\mathcal{H} + \frac{3}{2}\mathcal{H}(1 + c_s^2), \tag{2}$$

or

$$\frac{\theta'}{\theta} = \frac{\mathcal{H}'}{\mathcal{H}} - \frac{\mathcal{H}}{2} + \frac{3}{2}\mathcal{H}c_s^2. \tag{7}$$

Now use the definition of $w = \frac{2}{3}\frac{a^2\theta}{\mathcal{H}}$. Its logarithmic derivative is

$$\frac{w'}{w} = 2\mathcal{H} + \frac{\theta'}{\theta} - \frac{\mathcal{H}'}{\mathcal{H}}. \tag{3}$$

Using (7),

$$\boxed{\frac{w'}{w} = \frac{3}{2}\mathcal{H}(1+c_s^2)}. \tag{8}$$

Thus

$$2\frac{w'}{w} = 3\mathcal{H}(1+c_s^2),$$

which is precisely the coefficient multiplying Φ' in (4.3.13). We can therefore substitute

$$\Phi = \frac{u_S}{w}$$

into the evolution equation. The relevant derivative combination is

$$\Phi'' + 2\frac{w'}{w}\Phi' = \frac{1}{w}\left(u_S'' - \frac{w''}{w}u_S\right). \tag{9}$$

Since w depends only on time, one can write

$$\Delta\Phi = \frac{1}{w}\Delta u_S.$$

Hence, we get

$$u_S'' - \left[\frac{w''}{w} - 2\mathcal{H}' - (\mathcal{H}^2 - K)(1+3c_s^2)\right]u_S - c_s^2\Delta u_S = w\,\frac{\kappa}{2}a^2 P\Gamma. \tag{10}$$

The remaining task is to show that the quantity in square brackets is θ''/θ. From (8),

$$\frac{w''}{w} = \left(\frac{w'}{w}\right)' + \left(\frac{w'}{w}\right)^2, \tag{4}$$

so that

$$\frac{w''}{w} = \frac{3}{2}\mathcal{H}'(1+c_s^2) + \frac{3}{2}\mathcal{H}(c_s^2)' + \frac{9}{4}\mathcal{H}^2(1+c_s^2)^2. \tag{11}$$

For θ, it is convenient to use

$$\frac{\theta''}{\theta} = \left(\frac{\theta'}{\theta}\right)' + \left(\frac{\theta'}{\theta}\right)^2. \tag{12}$$

Substituting (7) into this expression introduces $\mathcal{H}''$. The background equations allow this derivative to be eliminated. Differentiating the Raychaudhuri equation and using (2) and (3) gives

$$\mathcal{H}'' = \mathcal{H}\left[(1+3c_s^2)(\mathcal{H}^2+K) + \mathcal{H}'(1-3c_s^2)\right]. \tag{13}$$

Using this in (12), one finds

$$\frac{\theta''}{\theta} = \frac{5}{4}\mathcal{H}^2 + \frac{3}{2}\mathcal{H}^2c_s^2 + \frac{9}{4}\mathcal{H}^2(c_s^2)^2 + K + 3Kc_s^2 + \frac{3c_s^2-1}{2}\mathcal{H}' + \frac{3}{2}\mathcal{H}(c_s^2)'. \tag{14}$$

On the other hand, substituting (11) into the coefficient appearing in (10) gives

$$\begin{aligned}\frac{w''}{w} &- 2\mathcal{H}' - (\mathcal{H}^2-K)(1+3c_s^2)\\ &= \frac{3}{2}\mathcal{H}'(1+c_s^2) + \frac{3}{2}\mathcal{H}(c_s^2)' + \frac{9}{4}\mathcal{H}^2(1+c_s^2)^2\\ &\quad - 2\mathcal{H}' - (\mathcal{H}^2-K)(1+3c_s^2)\end{aligned}$$

$$= \frac{5}{4}\mathcal{H}^2 + \frac{3}{2}\mathcal{H}^2 c_s^2 + \frac{9}{4}\mathcal{H}^2 (c_s^2)^2 + K + 3Kc_s^2 + \frac{3c_s^2 - 1}{2}\mathcal{H}' + \frac{3}{2}\mathcal{H}(c_s^2)'. \tag{15}$$

Comparing (14) and (15), we obtain

$$\boxed{\frac{w''}{w} - 2\mathcal{H}' - (\mathcal{H}^2 - K)(1 + 3c_s^2) = \frac{\theta''}{\theta}}. \tag{16}$$

Finally, the source term follows directly from the definition of w:

$$w\,\frac{\kappa}{2}a^2 P\Gamma = \frac{2a^2\theta}{3\mathcal{H}}\frac{\kappa}{2}a^2 P\Gamma = \frac{\kappa\theta}{3\mathcal{H}}a^4 P\Gamma. \tag{17}$$

Putting everything together, equation (10) becomes

$$\boxed{u_S'' - \left(\frac{\theta''}{\theta} + c_s^2\Delta\right)u_S = \frac{\kappa\theta}{3\mathcal{H}}a^4 P\Gamma}. \tag{18}$$

This is the desired canonical form.

Integral solution for long-wavelength modes

For super-Hubble modes where the spatial gradient term $c_\mathrm{s}^2\Delta$ is negligible, the homogeneous part of Eq. (4.3.17) admits $u_\mathrm{S} = \theta$ as an obvious particular solution. By reduction of order, the general solution in Fourier space is:

$$u_\mathrm{S}(k,\eta) = A(k)\theta(\eta) + B(k)\theta(\eta)\int^\eta \frac{\mathrm{d}\eta'}{\theta^2(\eta')} + \mathcal{O}\left(\frac{k^2 c_\mathrm{s}^2\theta}{\theta''}\right). \tag{4.3.18}$$

This solution holds for all modes during the matter-dominated era whenever non-adiabatic entropy perturbations Γ are negligible. In direct analogy with the tensor mode solution (4.3.12), Eq. (4.3.18) is recovered simply by replacing the scale factor $a(\eta)$ with $\theta(\eta)$.

First integral

For adiabatic perturbations on super-Hubble scales, Eq. (4.3.13) possesses an exact first integral. Introducing the gauge-invariant curvature perturbation on comoving slicing:

$$\zeta \equiv \Phi + \frac{2}{3\mathcal{H}}\frac{\Phi' + \mathcal{H}\Phi}{1+w}, \tag{4.3.19}$$

the dynamical equation (4.3.13) takes the compact form:

$$\zeta' = \frac{2}{3}\frac{\mathcal{H}}{1+w}\left[\frac{K}{\mathcal{H}^2}\left(\frac{1+3w}{2\mathcal{H}}\Phi' + 3(c_\mathrm{s}^2 - w)\Phi\right) + \frac{\kappa}{2}\frac{a^2}{\mathcal{H}^2}P\Gamma - c_\mathrm{s}^2\left(\frac{k}{\mathcal{H}}\right)^2\Phi\right], \tag{4.3.20}$$

where background thermodynamic identities have been employed to evaluate w' and $\mathcal{H}'$. In the regime where spatial curvature is negligible ($K/\mathcal{H}^2 \ll 1$), for adiabatic fluctuations ($\Gamma = 0$) on super-Hubble scales ($k/\mathcal{H} \ll 1$), the quantity ζ is strictly conserved:

$$\zeta' = 0. \tag{4.3.21}$$

When non-adiabatic processes are present, Eq. (4.3.20) simplifies to:

$$\zeta' = \frac{\mathcal{H}}{\rho + P}\delta P_{\rm nad}, \tag{4.3.22}$$

confirming that non-adiabatic entropy production drives departures from the conservation of ζ. Geometrically, ζ corresponds to the intrinsic three-dimensional Ricci scalar curvature evaluated on comoving spatial hypersurfaces.

Equivalently, the curvature perturbation evaluated in the flat-slicing gauge was introduced by Bardeen, Steinhardt, and Turner:

$$\zeta_{\rm BST} \equiv -C + \frac{1}{3}\frac{\delta\rho}{\rho + P} = -C + \frac{1}{3}\frac{\delta}{1+w}. \tag{4.3.23}$$

Choosing uniform density gauge

$$\delta\rho^U = \delta\rho + \rho' T \implies T = -\frac{\delta\rho}{\rho'} \tag{1}$$

Then

$$\zeta_{\rm BST} = -C - \mathcal{H}T \tag{2}$$

Expressed in terms of the longitudinal potential Φ, this gauge-invariant variable is:

$$\zeta_{\rm BST} = \Phi - \frac{2}{3}\frac{\mathcal{H}}{(1+w)(\mathcal{H}^2 + K)}\left[\Phi' + \left(1 - \frac{K}{\mathcal{H}^2} + \frac{1}{3}\left(\frac{k}{\mathcal{H}}\right)^2\right)\mathcal{H}\Phi\right]. \tag{4.3.24}$$

Differentiating with respect to conformal time yields:

$$\zeta'_{\rm BST} = -\frac{2}{3}\frac{\mathcal{H}^2}{(1+w)(\mathcal{H}^2 + K)}\left[\frac{1}{3}\left(\frac{k}{\mathcal{H}}\right)^2(\Phi' + \mathcal{H}\Phi) + \frac{\kappa}{2}\frac{a^2}{\mathcal{H}^2}P\Gamma\right]. \tag{4.3.25}$$

Unlike Eq. (4.3.20), the right-hand side of Eq. (4.3.25) contains no explicit dependence on the background spatial curvature K. Consequently, $\zeta_{\rm BST}$ remains exactly conserved for adiabatic ($\Gamma = 0$) super-Hubble modes ($k/\mathcal{H} \ll 1$), provided the decaying transient mode has vanished so that $\Phi' + \mathcal{H}\Phi$ does not diverge as $k/\mathcal{H} \to 0$.

On sub-Hubble scales, the comoving and flat-slicing curvature variables diverge according to the gradient relation:

$$\zeta = \zeta_{\rm BST} - \frac{1}{3}\frac{\Delta\Phi}{\mathcal{H}' - \mathcal{H}^2}. \tag{4.3.26}$$

Remarkably, the conservation of $\zeta_{\rm BST}$ on super-Hubble scales does not rely on the specific dynamics of Einstein's field equations, but follows solely from local energy-momentum conservation $\nabla_\mu T^{\mu\nu} = 0$ (Ref. [37]), ensuring its universal validity across generalized theories of gravity.

Effect of a variation in the cosmic fluid equation of state

To illustrate the utility of these conserved quantities, consider the evolution of the gravitational potential across a sudden change in the background equation of state. Suppose w changes from w_1 for $\eta < \eta_*$ (epoch 1) to w_2 for $\eta > \eta_*$ (epoch 2) in a flat, adiabatic background ($K = 0, \Gamma = 0$). During each epoch, the cosmic scale factor evolves as $a \propto t^{p_i} \propto \eta^{n_i}$, where $n_i = p_i/(1 - p_i)$ and $1 + w_i = 2/(3p_i)$.

Evaluating the long-wavelength general solution (4.3.18) yields:

$$\Phi = \frac{\mathcal{H}}{a^2}\left[A_- + A_+ \int a^2(1+w)\,\mathrm{d}\eta\right], \tag{4.3.27}$$

With $a(\eta) \propto \eta^{\frac{2}{1+3w}}$, we find $\mathcal{H} = \frac{a'}{a} = \frac{2}{1+3w}\frac{1}{\eta}$

$$\Phi(\eta) = \frac{\mathcal{H}}{a^2}\left[A_- + A_+ \int a^2(1+w)\,\mathrm{d}\eta\right] \tag{1}$$

$$= \frac{2}{1+3w}\eta^{-\frac{5+3w}{1+3w}}A_- + \frac{2(1+w)}{5+3w}A_+ \tag{2}$$

$$\approx \frac{2(1+w)}{5+3w}A_+ \tag{3}$$

where A_+ and A_- represent the amplitudes of the growing and decaying modes, respectively. Assuming the decaying mode during epoch 1 has had sufficient time to dissipate, the potential before the transition is dominated by the constant mode, $\Phi(\eta < \eta_*) \sim A_+(1+w_1)p_1/(p_1+1)|_{w=\frac{1}{3}} = 2A_+/[3(1+p_1)]$. Long after the transition, once the newly generated decaying transient has also died away, the potential settles to the constant plateau $\Phi(\eta \gg \eta_*) \sim A_+(1+w_2)p_2/(p_2+1)|_{w=0} = 2A_+/[3(1+p_2)]$. The net transition ratio is therefore:[4]

$$\frac{\Phi(\eta \gg \eta_*)}{\Phi(\eta < \eta_*)} = \frac{1+p_1}{1+p_2}. \tag{4.3.28}$$

For the cosmologically critical transition from radiation domination ($p_1 = 1/2$) to matter domination ($p_2 = 2/3$), Eq. (4.3.28) yields:

$$\frac{\Phi(\eta \gg \eta_*)}{\Phi(\eta < \eta_*)} = \frac{9}{10}, \tag{4.3.29}$$

for super-Hubble modes. The identical 9/10 suppression factor follows immediately from the conservation of ζ across the transition, since for the dominant growing mode $\zeta \sim \{1 + 2/[3(1+w)]\}\Phi$ in both epochs. The same analytical result can also be established by matching the exact background two-fluid scale factor.

4.3.3 Scalar modes in the adiabatic regime

Throughout the following analysis, we assume negligible anisotropic shear stress ($\bar{\pi} = 0$) and vanishing entropy perturbations ($\Gamma = 0$). In this regime, the two gauge-invariant gravitational potentials coincide ($\Phi = \Psi$). We therefore focus on solving the master potential equation (4.3.13) under adiabatic conditions ($\Gamma = 0$), a regime that describes cosmic perturbation dynamics with exceptional accuracy throughout most of cosmic history.

Universe dominated by a fluid with equation of state $w \neq 0$

Consider a flat Universe ($K = 0$) dominated by a single fluid possessing a constant equation of state $w \neq 0$. Under adiabatic conditions, the sound speed is constant and equal to $c_\mathrm{s}^2 = w$. The scale factor expands as a power law in conformal time, $a \propto \eta^\nu$ with $\nu \equiv 2/(1+3w)$. Defining $x \equiv k\eta$ and introducing the auxiliary variable $f \equiv x^\nu \Phi$, the Fourier-space potential equation (4.3.13) becomes:

$$\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} + \frac{2}{x}\frac{\mathrm{d}f}{\mathrm{d}x} + \left(w - \frac{\nu(\nu+1)}{x^2}\right) f = 0. \tag{4.3.30}$$

[4] we ignore the decaying mode on both sides of the transition for computing this ratio.

The general analytical solution to Eq. (4.3.30) can be expressed in terms of spherical Bessel functions:

$$\Phi = -\frac{3\nu^2}{2} x^{-\nu} \left[A j_\nu(c_\mathrm{s} x) + B n_\nu(c_\mathrm{s} x) \right] \equiv -\frac{3\nu^2}{2} x^{-\nu} \mathcal{Z}_\nu(c_\mathrm{s} x), \tag{4.3.31}$$

where j_ν and n_ν denote the spherical Bessel functions of the first and second kind, and $\mathcal{Z}_\nu$ defines their linear combination. The comoving Poisson equation (4.1.17) then yields the comoving density contrast:

$$\delta^\mathrm{C} = x^{2-\nu} \mathcal{Z}_\nu(c_\mathrm{s} x), \tag{4.3.32}$$

while the velocity-potential relation (4.2.68) determines the peculiar velocity:

$$kV = -\frac{3}{4} x^{1-\nu} \left[\mathcal{Z}_\nu(c_\mathrm{s} x) - \frac{c_\mathrm{s} x}{\nu + 1} \mathcal{Z}_{\nu-1}(c_\mathrm{s} x) \right]. \tag{4.3.33}$$

Using the asymptotic limits of the spherical Bessel functions, the asymptotic behavior of the perturbation fields in the super-acoustic and sub-acoustic regimes is summarized below:

	$c_\mathrm{s} x \ll 1$	$c_\mathrm{s} x \gg 1$
Φ	$\Phi_+ + \Phi_- x^{-1-2\nu}$	$\Phi_+ x^{-(1+\nu)} \cos(c_\mathrm{s} x - \alpha_\nu)$
$-\delta^\mathrm{C}$	$\frac{2}{3\nu^2} \left(\Phi_+ x^2 - \Phi_- x^{1-2\nu} \right)$	$\frac{2}{3\nu^2} \Phi_+ x^{1-\nu} \cos(c_\mathrm{s} x - \alpha_\nu)$
$-kV$	$\frac{2}{3\nu(1+w)} \left[\Phi_+ x - \Phi_- \left(1 + \nu^{-1} \right) x^{-2\nu} \right]$	$\frac{2}{3\nu(1+w)} \Phi_+ x^{1-\nu} \cos(c_\mathrm{s} x - \alpha_\nu)$

Here the acoustic phase shift is $\alpha_\nu \equiv \frac{\pi}{2}(\nu + 1)$. On sub-acoustic scales ($c_\mathrm{s} x \gg 1$), the density perturbation δ^C undergoes acoustic oscillations with an amplitude that experiences power-law damping whenever $-1/2 < w < 1/3$.

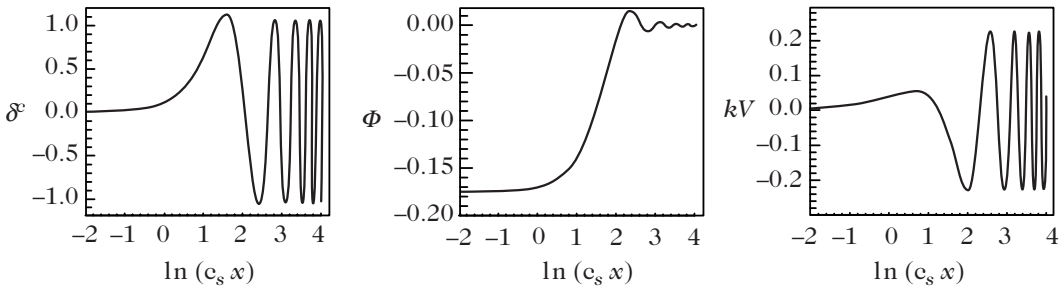


Figure 4.5: Evolution of the density contrast (left), gravitational potential (middle), and velocity perturbation (right) for scalar modes of a radiation fluid ($w = c_\mathrm{s}^2 = 1/3$) in a radiation-dominated Universe ($\nu = 1$) as a function of $c_\mathrm{s} x = c_\mathrm{s} k \eta$. The mode amplitude remains constant outside the sound horizon ($c_\mathrm{s} x < 1$) and initiates damped acoustic oscillations once inside ($c_\mathrm{s} x > 1$).

Radiation domination ($w = 1/3, \nu = 1$) represents the marginal threshold where the density oscillation amplitude remains constant without damping (see Fig. 4.5). Outside the acoustic horizon ($c_\mathrm{s} x \ll 1$), the growing mode of the gravitational potential remains strictly constant.

Furthermore, on super-Hubble scales during the radiation era, the comoving density fluctuation of the radiation fluid grows quadratically with the scale factor:

$$\delta_\mathrm{r}^\mathrm{C} \propto a^2. \tag{4.3.34}$$

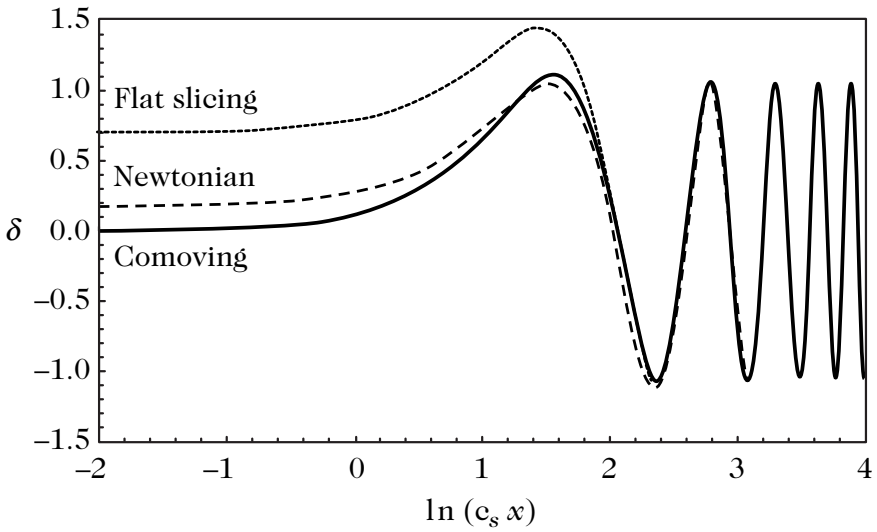


Figure 4.6: Evolution of the density contrasts $\delta^{\rm C}$ (comoving, solid line), $\delta^{\rm N}$ (Newtonian/longitudinal, dashed line), and $\delta^{\rm F}$ (flat slicing, dotted line) as a function of $c_{\rm s}x = c_{\rm s}k\eta$. These density contrast $\delta^{\rm C}$, $\delta^{\rm N}$, and $\delta^{\rm F}$ are gauge invariant but not the same physical object. Fundamentally they measure perturbation on different hypersurfaces. On super-acoustic scales ($c_{\rm s}x < 1$), substantial disparities exist among the three gauge definitions, which rapidly converge once modes penetrate the acoustic horizon ($c_{\rm s}x > 1$), illustrating the gauge ambiguity inherent to super-Hubble density perturbations. This happens because on subhorizon scale the general relativistic effects become unimportant and the newtonian description becomes valid. In this limit, all issue of gauge choice becomes irrelevant as all sensible gauges approach each other.

Matter fluid in a matter-dominated Universe

In the case of a dust-dominated cosmos ($w = c_{\rm s}^2 = 0$ and $\nu = 2$), Eq. (4.3.30) reduces to:

$$\frac{{\rm d}^2 f}{{\rm d}x^2} + \frac{2}{x}\frac{{\rm d}f}{{\rm d}x} - \frac{6}{x^2} f = 0. \tag{4.3.35}$$

The general solution provides the explicit dynamics of the potential, density contrast, and peculiar velocity:

$$\Phi = \Phi_+ + \Phi_- x^{-5}, \qquad \delta^{\rm C} = -\frac{1}{6}x^2\Phi, \qquad \text{and} \qquad -kV = \frac{1}{3}\Phi_+ x - \frac{1}{2}\Phi_- x^{-4}. \tag{4.3.36}$$

Since $a \propto \eta^2 \propto x^2$, the comoving density contrast of matter grows strictly linearly with the scale factor:

$$\delta^{\rm C}_{\rm m} \propto a, \tag{4.3.37}$$

independent of the perturbation wavelength. This reproduces the conclusion obtained from the Newtonian sub-Hubble analysis [Eq. (4.1.22)].

The constancy of the gravitational potential during matter domination and the modest linear growth rate $\delta \propto a$ prompted Landau and Lifshitz in 1946 to deduce that gravitational instability in an expanding FLRW background could not account for the observed cosmic structure (Ref. [38]). Indeed, initial thermal fluctuations could only grow by a factor of $\approx 10^4$ during the matter-dominated epoch, an amplification far too modest to generate galaxies and galaxy clusters today. This historical impasse underscored the necessity of a physical mechanism—such as primordial cosmic inflation—capable of generating seed perturbations of order 10^{-4} at the threshold of the matter-dominated era.

Effect of curvature

Because spatial curvature scales dynamically as a^{-2}, its energy density can dominate only at late cosmic times, well after radiation has become negligible. In a Universe governed by matter and spatial curvature, integrating Eq. (4.3.13) analytically for an open spatial geometry ($K = -1$) yields:

$$\Phi(k,\eta) = A(k)\frac{\sinh\eta}{\sinh^6(\eta/2)} + B(k)\frac{\sinh\eta(\sinh\eta - 3\eta) + 8\sinh^2(\eta/2)}{\sinh^6(\eta/2)}. \tag{4.3.38}$$

where $A(k)$ and $B(k)$ are determined by imposing initial/junction condition. Once curvature dominates the cosmic dynamics ($\eta > 1$), the potential decays exponentially as $\Phi \sim B(k)e^{-\eta}$, and Poisson's equation (4.1.17) reveals that the density contrast freezes out to a constant plateau:

$$\delta_{\rm m}^{\rm C} = -\frac{8k^2|1-\Omega_0|}{\Omega_0}\sinh^2\left(\frac{\eta}{2}\right)\Phi \sim -\frac{8k^2B(k)|1-\Omega_0|}{\Omega_0}. \tag{4.3.39}$$

Thus, negative spatial curvature halts further gravitational clustering.

For a closed spatial geometry ($K = +1$), the general analytical solution to Eq. (4.3.13) is:

$$\Phi(k,\eta) = A(k)\frac{\sin\eta}{\sin^6(\eta/2)} + B(k)\frac{\sin\eta(\sin\eta - 3\eta) + 8\sin^2(\eta/2)}{\sin^6(\eta/2)}. \tag{4.3.40}$$

As $\eta \to \pi - \varepsilon$, the cosmic expansion nears its maximal turnaround radius, where $\mathcal{H} \to 0$ and $\Phi \sim 8B(k)$. In the subsequent recollapsing phase, both perturbation modes diverge rapidly, showing that cosmic contraction exponentially amplifies gravitational collapse. Immediately preceding the final big crunch at $\eta = 2\pi - \varepsilon$, the potential behaves as $\Phi \sim 2[6\pi B(k) - A(k)]/\sin^5(\eta/2)$, and the density contrast diverges as:

$$\delta_{\rm m}^{\rm C} \sim -2k^2\frac{6\pi B(k) - A(k)}{\sin^3(\eta/2)}\frac{|1-\Omega_0|}{\Omega_0}. \tag{4.3.41}$$

Figure 4.7 illustrates the contrasting evolution of $\delta_{\rm m}^{\rm C}$ in spherical and hyperbolic universes.

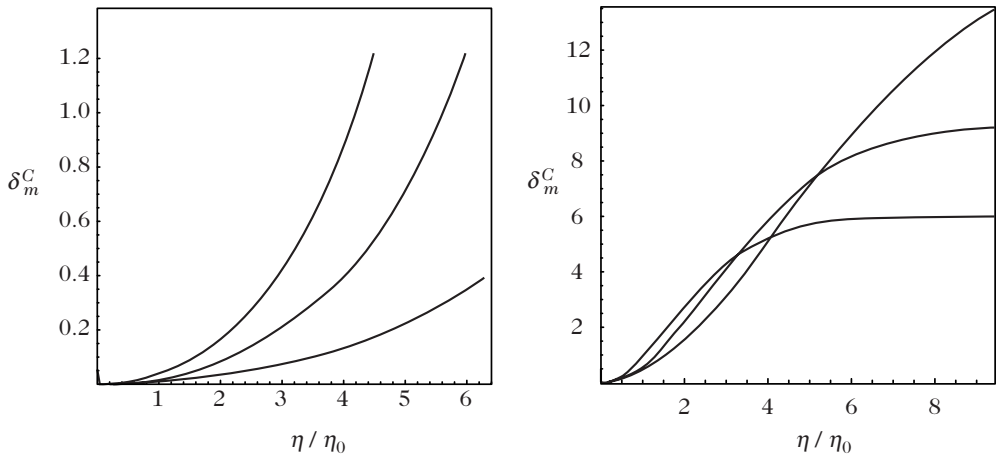


Figure 4.7: Evolution of the density contrast $\delta_{\rm m}^{\rm C}$ in a Universe with spherical spatial sections (left panel, for $\Omega_0 = 1.2, 1.3, 1.4$) and hyperbolic spatial sections (right panel, for $\Omega_0 = 0.2, 0.3, 0.4$) as a function of conformal time η/η_0.

Effect of the cosmological constant

A cosmological constant Λ also comes to dominate only at late cosmological times when radiation is negligible. Introducing the normalized variable $y_\Lambda \equiv a/a_\Lambda = (\Lambda/\kappa\rho)^{1/3}$, the transition from matter domination to cosmological constant domination occurs at the redshift:

$$1 + z_\Lambda = \left(\frac{\Omega_{\Lambda 0}}{\Omega_{\mathrm{m}0}}\right)^{1/3},$$

and the Friedmann equation takes the form:

$$\mathcal{H}^2 = \frac{\mathcal{H}_\Lambda^2}{2}\left(\frac{1}{y_\Lambda} + y_\Lambda^2\right).$$

Equation (4.3.13) then transforms into:

$$\ddot{\Phi} + \frac{1}{2y_\Lambda}\frac{7 + 10y_\Lambda^3}{1 + y_\Lambda^3}\dot{\Phi} + \frac{3y_\Lambda}{y_\Lambda^3 + 1}\Phi = 0. \tag{4.3.42}$$

The corresponding matter density contrast is given by:

$$\delta_{\mathrm{m}}^{\mathrm{C}} = -\frac{2}{3}\left(\frac{k}{\mathcal{H}_\Lambda}\right)^2 y_\Lambda \Phi. \tag{4.3.43}$$

Once dark energy dominates ($y_\Lambda \gg 1$), the potential decays as $\Phi \propto 1/y_\Lambda$, causing the density contrast to freeze out to a constant plateau (Fig. 4.8). In a Λ-dominated Universe, structure growth ceases near redshift $z \sim z_\Lambda$.

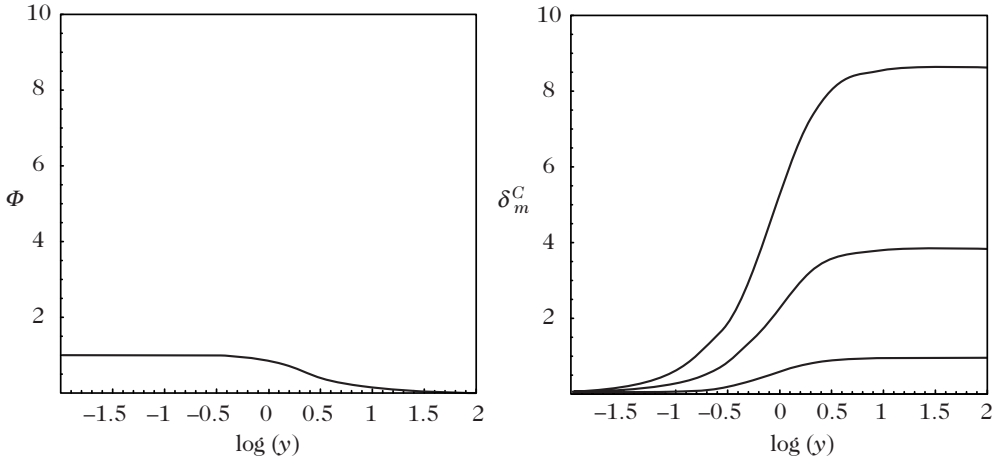


Figure 4.8: Evolution of the gravitational potential (left) and the density contrast (right) in a Universe dominated by a cosmological constant. The right panel displays the three modes $k/\mathcal{H}_\Lambda = 1, 2, 3$.

Case of a Universe dominated by a mixture of matter and radiation

We finally consider a spatially flat Universe ($K = 0$) dominated by a cosmic mixture of pressureless matter ($w = 0$) and radiation ($w = 1/3$). Adopting the equality-normalized scale factor $y = a/a_{\mathrm{eq}}$ defined in Eq. (4.3.6), the equation-of-state parameter and the adiabatic sound speed are:

$$w = \frac{1}{3}(1 + y)^{-1} \qquad \text{and} \qquad c_{\mathrm{s}}^2 = \frac{1}{3}\left(1 + \frac{3}{4}y\right)^{-1}. \tag{4.3.44}$$

Using the relations $\Phi' = \mathcal{H}y\,\mathrm{d}\Phi/\mathrm{d}y$, $\Phi'' = \mathcal{H}^2y^2\,\mathrm{d}^2\Phi/\mathrm{d}y^2 + y''\,\mathrm{d}\Phi/\mathrm{d}y$, and the Friedmann equation (4.3.9), the master potential equation (4.3.13) becomes:

$$\frac{\mathrm{d}^2\Phi}{\mathrm{d}y^2}+\frac{1}{2y}\left(7-\frac{1}{1+y}+\frac{8}{4+3y}\right)\frac{\mathrm{d}\Phi}{\mathrm{d}y}+\frac{1}{y(1+y)(4+3y)}\Phi+\frac{8}{3}\frac{1}{4+3y}\frac{1}{1+y}\left(\frac{k}{k_{\mathrm{eq}}}\right)^2\Phi=-\frac{3}{2y^2}w\Gamma, \tag{4.3.45}$$

where k_{eq} is defined in Eq. (4.3.10). For adiabatic initial conditions ($\Gamma = 0$), numerical integration of Eq. (4.3.45) yields the potential trajectories shown in Fig. 4.9. In the presence of entropy perturbations, the dynamical behavior of Γ must be supplied to solve Eq. (4.3.45). On large wavelengths, Γ remains nearly constant in time, rendering the adiabatic solutions of Fig. 4.9 a solid leading-order approximation.

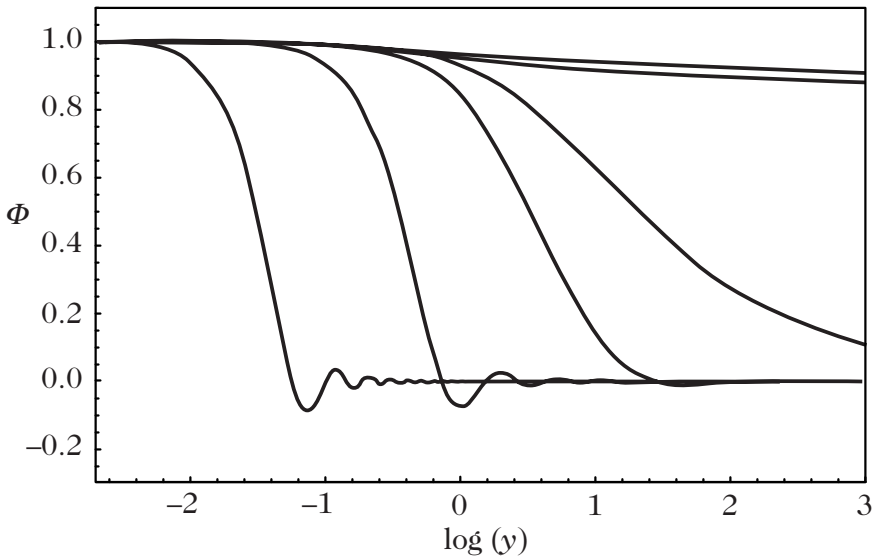


Figure 4.9: Evolution of the gravitational potential in a Universe dominated by a mixture of matter and radiation, assuming $\Gamma = 0$, for modes $k/k_{\mathrm{eq}} = 10^{-3}, 0.1, 1, 2, 10, 100$.

In the long-wavelength regime ($k/k_{\mathrm{eq}} \ll 1$), Eq. (4.3.45) can be solved in closed analytical form. Introducing the auxiliary field $\varphi \equiv y^3\Phi/\sqrt{1+y}$, one finds $\mathrm{d}\Phi/\mathrm{d}y \propto y^2(4+3y)/(1+y)^{3/2}$, which can be integrated directly.

$$\Phi = \frac{\Phi_0}{10y^3}\left(16\sqrt{1+y}+9y^3+2y^2-8y-16\right). \tag{4.3.46}$$

For super-Hubble perturbations, this reproduces the asymptotic ratio $\Phi(y \gg 1)/\Phi_0 = 9/10$, which represents a particular instance of the general result (4.3.29).

The comoving curvature perturbation $\mathcal{R}$ (from (4.3.19)) in terms of y and Φ is:

$$\mathcal{R} = \Phi + \frac{2}{3(1+w)}\left(\frac{\Phi'}{\mathcal{H}}+\Phi\right) = \frac{(6+5y)\Phi+2y(1+y)\Phi'}{4+3y}. \tag{1}$$

For adiabatic super-Hubble perturbations, $\mathcal{R}$ is conserved and evaluates to $\frac{3}{2}\Phi_0$. Setting these equal gives:

$$\frac{(6+5y)\Phi+2y(1+y)\Phi'}{4+3y} = \frac{3}{2}\Phi_0 \tag{2}$$

$$2y(1+y)\Phi' + (6+5y)\Phi = \frac{3}{2}\Phi_0(4+3y) \tag{3}$$

$$\underbrace{\frac{y^3}{\sqrt{1+y}}\Phi' + \frac{y^2(6+5y)}{2(1+y)^{3/2}}\Phi}_{\frac{d}{dy}\left(\Phi\frac{y^3}{\sqrt{1+y}}\right)} = \frac{3}{4}\Phi_0\frac{y^2(4+3y)}{(1+y)^{3/2}}. \tag{4}$$

To integrate the right-hand side, we apply the substitution $u = \sqrt{1+y}$, which gives $y = u^2 - 1$ and $dy = 2u\,du$. We rewrite the numerator in terms of u:

$$y^2(3y+4) = (u^2-1)^2(3u^2+1) = 3u^6 - 5u^4 + u^2 + 1. \tag{5}$$

Substituting this into the integral yields:

$$\begin{aligned}\int \frac{y^2(3y+4)}{(1+y)^{3/2}}dy &= 2\int \left(3u^4 - 5u^2 + 1 + u^{-2}\right)du \\ &= 2\left(\frac{3}{5}u^5 - \frac{5}{3}u^3 + u - \frac{1}{u}\right).\end{aligned} \tag{6}$$

Factoring out the common denominator $15u$:

$$= \frac{2}{15u}\left(9u^6 - 25u^4 + 15u^2 - 15\right). \tag{7}$$

Substituting $u^2 = 1+y$ back into the polynomial gives $9(1+y)^3 - 25(1+y)^2 + 15(1+y) - 15 = 9y^3 + 2y^2 - 8y - 16$. Thus, the integrated equation becomes:

$$\Phi\frac{y^3}{\sqrt{1+y}} = \frac{\Phi_0}{10\sqrt{1+y}}\left(9y^3 + 2y^2 - 8y - 16\right) + C. \tag{8}$$

In the deep radiation limit ($y \to 0$), the left-hand side vanishes. Matching this boundary condition requires:

$$0 = \frac{\Phi_0}{10}(-16) + C \implies C = \frac{8}{5}\Phi_0. \tag{9}$$

Substituting C back and multiplying by $\frac{\sqrt{1+y}}{y^3}$ isolates Φ:

$$\Phi = \frac{\Phi_0}{10y^3}\left(16\sqrt{1+y} + 9y^3 + 2y^2 - 8y - 16\right). \tag{10}$$

4.3.4 Mixture of several fluids

In realistic cosmological settings, multiple physical fluids coexist—such as baryons, cold dark matter, photons, and neutrinos. These different cosmological fluids interact with one another via gravitation, but they may also experience non-gravitational couplings. Figure 4.10 summarizes the principal fluid components that should be considered alongside their mutual interactions.

Description of the interactions

In the most general coupled multi-fluid system, conservation of the energy-momentum tensor holds only for the aggregate fluid mixture, whereas the individual energy-momentum tensors satisfy non-conserved balance equations of the form:

$$\nabla_\mu T_a^{\mu\nu} = Q_a^\nu, \tag{4.3.47}$$

where Q_a^μ denotes the external four-force and energy exchange acting on component a.

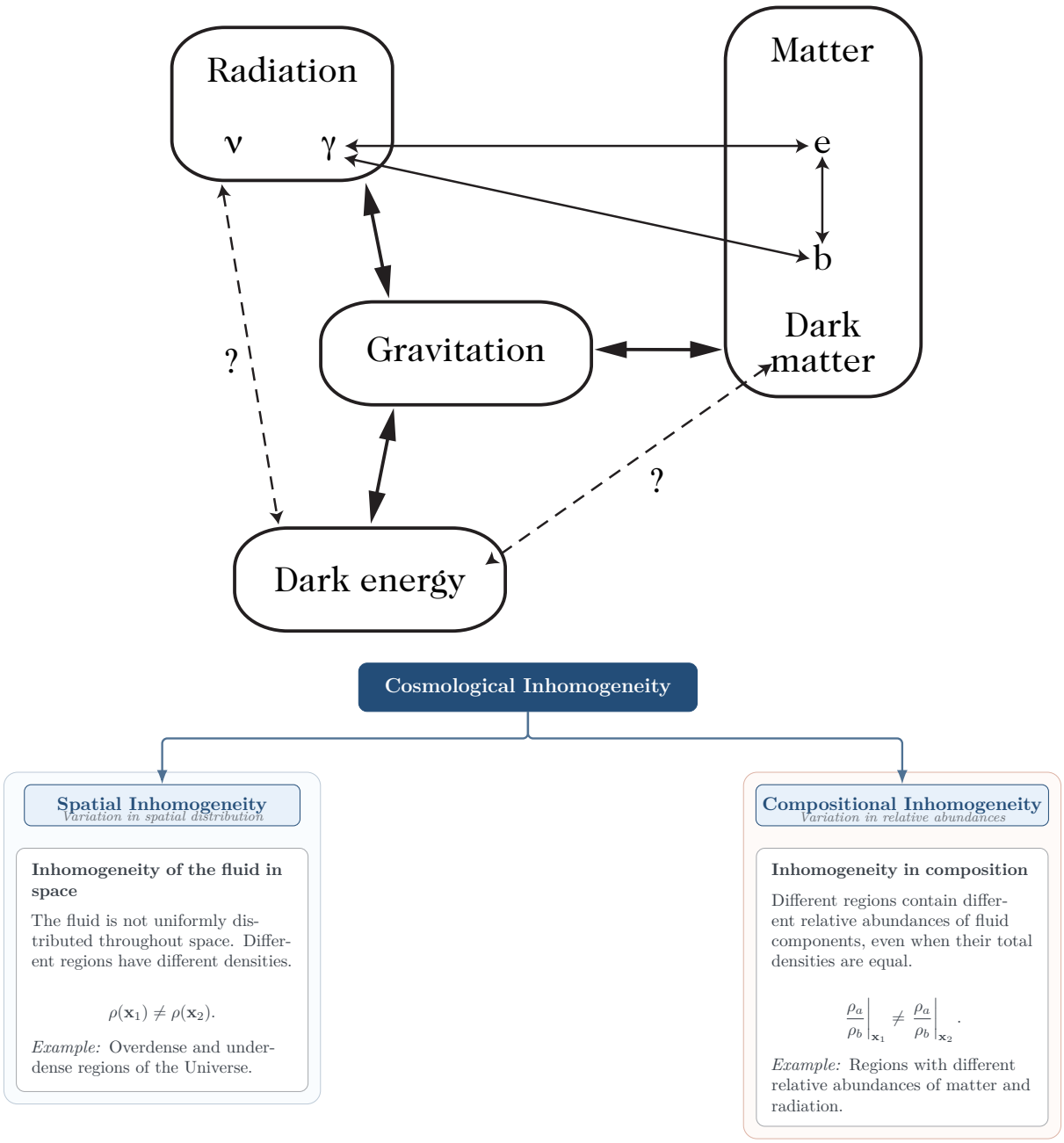


Figure 4.10: The cosmological fluid components and their interactions. Baryons and electrons are tightly coupled through the Coulomb interaction, while photons couple to charged matter via Compton scattering. Dark matter is assumed to interact with ordinary baryonic matter solely through gravity. It is not excluded that dark energy may interact with ordinary matter or with dark matter through an as yet unknown physical mechanism.

This becomes important to take into account the interaction and decay rate energy transfer betweeen fluids and fields.

These interaction terms must satisfy the constraint:

$$\sum_a Q_a^\mu = 0, \tag{4.3.48}$$

which is simply the cosmological manifestation of Newton's third law of action and reaction. This

condition follows directly from the contracted Bianchi identities ensuring total energy-momentum conservation $\nabla_\mu T^{\mu\nu} = 0$.

At background order, the spatial homogeneity and isotropy of the spacetime imply that the interaction four-vector has vanishing spatial components:

$$Q^a_\mu = (-aQ_a, \mathbf{0}), \tag{4.3.49}$$

so that $Q^a_\mu Q^a_\nu \bar{g}^{\mu\nu} = -Q_a^2$. Notice that $(\bar{g}_{\mu\nu} + \bar{u}_\mu \bar{u}_\nu) Q^\mu_a = 0$, meaning that background spacetime symmetries preclude the existence of any net spatial force between the fluids at the unperturbed background level. Taking energy transfer into account, the background continuity equation for component a generalizes to:

$$\rho_a' + 3\mathcal{H}(1 + w_a)\rho_a = aQ_a, \tag{4.3.50}$$

If Q_a varues wutg position, it can lead to violation of (4.3.70) and generate isocurvature perturbation.

which yields the dynamical evolution of the equation-of-state parameter:

$$w_a' = -\left[3\mathcal{H}(1 + w_a) - \frac{Q_a}{\rho_a}\right](c_a^2 - w_a), \tag{4.3.51}$$

where $w_a \equiv P_a/\rho_a$ and $c_a^2 \equiv P_a'/\rho_a'$ denote the equation of state and the adiabatic sound speed of component a, respectively.

At linear order in cosmological perturbations, Q^μ_a can be decomposed with respect to the fluid four-velocity u^μ_a as:

$$Q^\mu_a = Q_a u^\mu_a + F^\mu_a, \qquad \text{with} \qquad F^\mu_a u_{a\mu} = 0. \tag{4.3.52}$$

We define $a f^\mu_a \equiv (\delta^\mu_\nu + u^\mu u_\nu) F^\nu_a$, ensuring $f^0_a = 0$, while the spatial part f^i_a decomposes into scalar and vector modes according to $f^i_a = \bar{f}^i_a + D^i f_a$. Under an infinitesimal gauge transformation (4.2.11), these quantities transform as:

$$\delta Q_a \to \delta Q_a + Q_a' T, \tag{4.3.53}$$

and f^i_a is inherently gauge invariant because it vanishes in the unperturbed background. We can therefore construct the gauge-invariant perturbation:

$$\delta Q^{\mathrm{N}}_a \equiv \delta Q_a + Q_a'(B - E'), \tag{4.3.54}$$

which is gauge invariant. With these notations, the continuity equation (4.2.74) and the Euler equation (4.2.75) for component a generalize as:

$$\left(\frac{\delta^{\mathrm{N}}_a}{1 + w_a}\right)' = -(\Delta V_a - 3\Psi') - 3\mathcal{H}\frac{w_a}{1 + w_a}\Gamma_a + \frac{a}{\rho_a + P_a}\left[\delta Q^{\mathrm{N}}_a + Q_a(\Phi - \delta^{\mathrm{N}}_a)\right], \tag{4.3.55}$$

and

$$V_a' + \mathcal{H}V_a = -\Phi - \frac{c_a^2}{1 + w_a}\delta^{\mathrm{C}}_a + \frac{w_a}{1 + w_a}\left[\Gamma_a + \frac{2}{3}(\Delta + 3K)\bar{\pi}_a\right] + \frac{a}{\rho_a + P_a}(f_a - Q_a c_a^2 V_a). \tag{4.3.56}$$

System of fluids interacting gravitationally

We consider the simplified scenario where a set of fluids interact solely through gravity. In this case, $Q^\mu_a = 0$ for each individual fluid, and the total energy-momentum tensor is the simple sum:

$$T^{\mu\nu} = \sum_a T^{\mu\nu}_a.$$

We have assumed that there is no energy or momentum transfer between different fuilds otherwise there would be collision term present corresponding to each kind of interaction coming from Boltzmann transport equation.

Each energy-momentum tensor can be decomposed as in Eq. (4.2.18), so that the aggregate density, pressure, and velocity field are defined by:

$$\rho = \sum_a \rho_a, \qquad P = \sum_a P_a, \qquad (\rho + P)v^i = \sum_a (\rho_a + P_a)v_a^i. \tag{4.3.57}$$

We have:

$$\Omega w = \sum_a \Omega_a w_a, \tag{4.3.58}$$

and

$$\Omega c_{\rm s}^2 = \sum_a \frac{1+w_a}{1+w}\Omega_a c_a^2. \tag{4.3.59}$$

For simplicity, let us examine a system of two fluids. It is described by two density contrasts, δ_a and δ_b, and two velocity fields, v_a^i and v_b^i. It is convenient to transform these degrees of freedom into a set of variables, δ and v, characterizing the total fluid, and a set of variables, S_{ab} and V_{ab}^i, describing their relative properties. The definitions (4.3.57) imply that:

$$\Omega\delta = \sum_a \Omega_a \delta_a \qquad \text{and} \qquad \Omega(1+w)v = \sum_a \Omega_a(1+w_a)v_a, \tag{4.3.60}$$

for the scalar modes. These total quantities satisfy the continuity equation (4.2.74) and the Euler equation (4.2.75). The total entropy perturbation Γ is defined as before by relation (4.2.26), so that for such a mixture of fluids, it takes the form:

$$\Omega w \Gamma = \sum_a w_a \Omega_a \Gamma_a + \sum_a \left(c_{\rm s}^2 - c_a^2\right)\Omega_a \delta_a. \tag{4.3.61}$$

The first term corresponds to the entropy contribution from the individual components of the mixture, while the second term, which does not vanish even if $\Gamma_a = 0$ for each component, represents the entropy of mixing. One can check that this second contribution is gauge invariant.

The Einstein field equations take the general form obtained previously, but for the quantities describing the total fluid. In particular, both (4.2.66) and (4.2.67) can be rewritten as:

$$(\Delta + 3K)\Psi = \frac{3}{2}\mathcal{H}^2 \sum_a \Omega_a \delta_a^{\rm C}, \tag{4.3.62}$$

$$\Psi - \Phi = 3\mathcal{H}^2 \sum_a \Omega_a w_a \bar{\pi}_a. \tag{4.3.63}$$

Now, S_{ab} and V_{ab} are defined, respectively, by:

$$S_{ab} \equiv \frac{\delta_a}{1+w_a} - \frac{\delta_b}{1+w_b} \qquad \text{and} \qquad V_{ab} \equiv v_a - v_b, \tag{4.3.64}$$

which are, by construction, gauge invariant.

This can be thought of as relative weighted density contrast. It tells us how the spatial variation in entropy per particle between different species changes with position.

Using the relations (4.3.60), we readily obtain the inverse relation:

$$\left(\frac{\Omega_b}{1+w_a} + \frac{\Omega_a}{1+w_b}\right)\delta_a = \frac{\Omega}{1+w_b}\delta + \Omega_b S_{ab}, \tag{4.3.65}$$

and the analogous expression for δ_b is obtained by interchanging the indices a and b.

The evolution equations for the relative quantities S_{ab} and V_{ab} are obtained by combining the continuity equation in the form (4.2.76) and the Euler equation (4.2.79) for each fluid:

$$S'_{ab} = -\Delta V_{ab} - 3\mathcal{H}\Gamma_{ab}, \tag{4.3.66}$$

$$V'_{ab} = -\mathcal{H}\left[1 - \frac{3\left[\Omega_a(1+w_a)c_b^2 + \Omega_b(1+w_b)c_a^2\right]}{\Omega(1+w)}\right]V_{ab} - \left(c_a^2 - c_b^2\right)\frac{\delta^{\mathrm{C}}}{1+w} - \left[c_a^2(1+w_b)\frac{\Omega_b}{\Omega} + c_b^2(1+w_a)\frac{\Omega_a}{\Omega}\right]\frac{S_{ab}}{1+w} - \left[\frac{2}{3}(\Delta+3K)\bar{\pi}_{ab} + \Gamma_{ab}\right], \tag{4.3.67}$$

with the definitions:

$$\Gamma_{ab} \equiv \frac{w_a}{1+w_a}\Gamma_a - \frac{w_b}{1+w_b}\Gamma_b, \tag{4.3.68}$$

$$\bar{\pi}_{ab} \equiv \frac{w_a}{1+w_a}\bar{\pi}_a - \frac{w_b}{1+w_b}\bar{\pi}_b. \tag{4.3.69}$$

To completely solve the evolution of a system of two interacting fluids, we thus need to solve the evolution equation (4.3.13) for the gravitational potential, where Γ is now given by relation (4.3.61), together with the pair of equations (4.3.66) and (4.3.67); the relation (4.3.62) then gives δ^{C}.

Euler Equation for relative perturbation

For each individual fluid a interacting solely via gravity, energy-momentum conservation $\nabla_\mu T^\mu_0 = 0$ in conformal Newtonian gauge gives (4.2.74)

$$\delta_a^{\mathrm{N}\prime} = -3\mathcal{H}(c_a^2 - w_a)\delta_a^{\mathrm{N}} - 3\mathcal{H}w_a\Gamma_a - (1+w_a)\Delta V_a + 3(1+w_a)\Psi', \tag{1}$$

where primes denote derivatives with respect to conformal time, $' \equiv d/d\tau$, and $\mathcal{H} \equiv a'/a$. Differentiating the scaled density contrast and using the background relation $w'_a = -3\mathcal{H}(1+w_a)(c_a^2 - w_a)$ gives (4.2.76),

$$\left(\frac{\delta_a^{\mathrm{N}}}{1+w_a}\right)' = \frac{\delta_a^{\mathrm{N}\prime}}{1+w_a} - \frac{w'_a}{(1+w_a)^2}\delta_a^{\mathrm{N}} = -\Delta V_a + 3\Psi' - 3\mathcal{H}\frac{w_a}{1+w_a}\Gamma_a. \tag{2}$$

Subtracting the corresponding equation for fluid b eliminates the gravitational potential Ψ' and yields the relative continuity equation:

$$S'_{ab} = -\Delta V_{ab} - 3\mathcal{H}\Gamma_{ab}, \tag{3}$$

where $V_{ab} \equiv V_a - V_b$, $\Gamma_{ab} \equiv \frac{w_a}{1+w_a}\Gamma_a - \frac{w_b}{1+w_b}\Gamma_b$, and the gauge-invariant relative entropy perturbation is

$$S_{ab} \equiv \frac{\delta_a^{\mathrm{N}}}{1+w_a} - \frac{\delta_b^{\mathrm{N}}}{1+w_b}. \tag{4}$$

Gauge invariance follows because a change of slicing acts with a *single* T on every species, $\delta_a \to \delta_a - 3\mathcal{H}(1+w_a)T$, so each term shifts by the same $-3\mathcal{H}T$ and the difference is unchanged. For this reason S_{ab} may equally be built from the comoving-slicing contrasts, $S_{ab} = \delta_a^{\mathrm{C}}/(1+w_a) - \delta_b^{\mathrm{C}}/(1+w_b)$, *provided* both are taken on the slicing comoving with the *total* fluid, $\delta_a^{\mathrm{C}} = \delta_a^{\mathrm{N}} - 3\mathcal{H}(1+w_a)V$. Contrasts defined on slices comoving with each fluid separately, $\delta_a^{\mathrm{N}} - 3\mathcal{H}(1+w_a)V_a$, do *not* correspond to any single gauge, and the combination they form differs from S_{ab} by $3\mathcal{H}V_{ab}$. The derivation below therefore works throughout with δ_a^{N} and converts only the *total* density perturbation to the comoving slicing, at the single point where δ^{C} is introduced.

The spatial conservation equation $\nabla_\mu T^\mu_i = 0$ for fluid a is (from (4.2.79))

$$V_a' + \mathcal{H}(1-3c_a^2)V_a = -\Phi - \frac{\delta P_a}{\rho_a(1+w_a)} - \frac{2}{3}(\Delta+3K)\frac{w_a\bar{\pi}_a}{1+w_a}. \tag{5}$$

Decomposing $\delta P_a = c_a^2\rho_a\delta_a^{\rm N} + w_a\rho_a\Gamma_a$, this can be written as

$$V_a' + \mathcal{H}V_a - 3\mathcal{H}c_a^2V_a = -\Phi - c_a^2\frac{\delta_a^{\rm N}}{1+w_a} - \frac{w_a\Gamma_a}{1+w_a} - \frac{2}{3}(\Delta+3K)\frac{w_a\bar{\pi}_a}{1+w_a}. \tag{6}$$

Subtracting the b-equation from the a-equation eliminates Φ:

$$V_{ab}' + \mathcal{H}V_{ab} - 3\mathcal{H}(c_a^2V_a - c_b^2V_b) = -c_a^2\frac{\delta_a^{\rm N}}{1+w_a} + c_b^2\frac{\delta_b^{\rm N}}{1+w_b} - \Gamma_{ab} - \frac{2}{3}(\Delta+3K)\bar{\pi}_{ab}, \tag{7}$$

where $\bar{\pi}_{ab} \equiv \frac{w_a}{1+w_a}\bar{\pi}_a - \frac{w_b}{1+w_b}\bar{\pi}_b$.
To relate these quantities to the total-fluid comoving frame, we introduce the enthalpy fractions

$$f_a \equiv \frac{\Omega_a(1+w_a)}{\Omega(1+w)}, \qquad f_b \equiv \frac{\Omega_b(1+w_b)}{\Omega(1+w)}, \qquad f_a + f_b = 1, \tag{8}$$

with $\Omega(1+w) = \Omega_a(1+w_a) + \Omega_b(1+w_b)$. The total velocity $V = f_aV_a + f_bV_b$ from (4.3.60) and the Newtonian perturbations then obey

$$V_a - V = f_bV_{ab}, \qquad V_b - V = -f_aV_{ab}, \tag{9}$$

$$\frac{\delta_a^{\rm N}}{1+w_a} = \frac{\delta^{\rm N}}{1+w} + f_bS_{ab}, \qquad \frac{\delta_b^{\rm N}}{1+w_b} = \frac{\delta^{\rm N}}{1+w} - f_aS_{ab}. \tag{10}$$

Transforming the total density perturbation to the comoving slicing via $\frac{\delta^{\rm N}}{1+w} = \frac{\delta^{\rm C}}{1+w} + 3\mathcal{H}V$, the density source terms become

$$\begin{aligned} -c_a^2\frac{\delta_a^{\rm N}}{1+w_a} + c_b^2\frac{\delta_b^{\rm N}}{1+w_b} &= -(c_a^2-c_b^2)\frac{\delta^{\rm N}}{1+w} - \left(c_a^2f_b + c_b^2f_a\right)S_{ab} \\ &= -(c_a^2-c_b^2)\frac{\delta^{\rm C}}{1+w} - 3\mathcal{H}(c_a^2-c_b^2)V - \left(c_a^2f_b + c_b^2f_a\right)S_{ab}. \end{aligned} \tag{11}$$

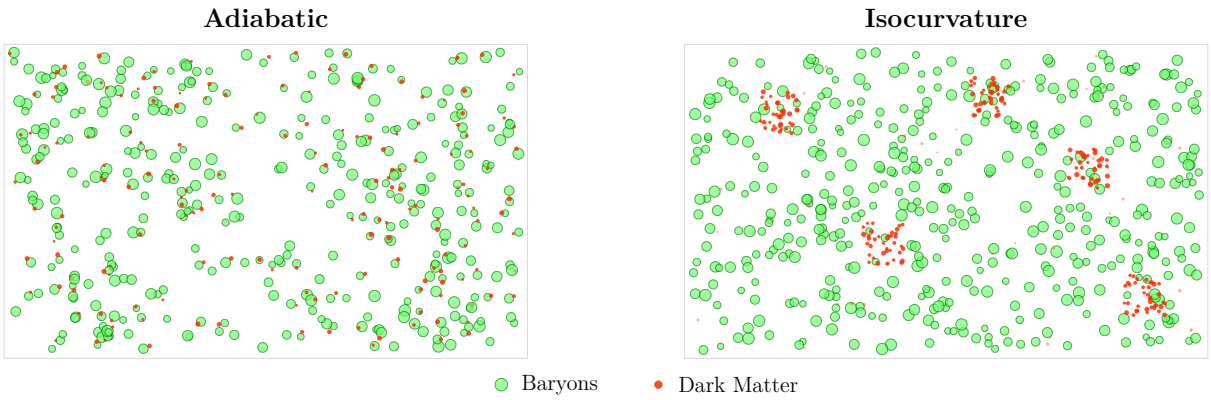
Moving the $-3\mathcal{H}(c_a^2-c_b^2)V$ term to the left-hand side of Eq. (7) combines directly with the individual velocity terms to form the velocity slips relative to the total frame:

$$\begin{aligned} -3\mathcal{H}(c_a^2V_a - c_b^2V_b) + 3\mathcal{H}(c_a^2-c_b^2)V &= -3\mathcal{H}\left[c_a^2(V_a-V) - c_b^2(V_b-V)\right] \\ &= -3\mathcal{H}\left(c_a^2f_b + c_b^2f_a\right)V_{ab} \\ &= -3\mathcal{H}\left[\frac{c_a^2(1+w_b)\Omega_b + c_b^2(1+w_a)\Omega_a}{\Omega(1+w)}\right]V_{ab}. \end{aligned} \tag{12}$$

Substituting this back into Eq. (7), the relative Euler equation takes its closed form:

$$\begin{aligned} V_{ab}' = -\mathcal{H}&\left\{1 - \frac{3\left[c_a^2(1+w_b)\Omega_b + c_b^2(1+w_a)\Omega_a\right]}{\Omega(1+w)}\right\}V_{ab} \\ &- (c_a^2-c_b^2)\frac{\delta^{\rm C}}{1+w} - \left[c_a^2(1+w_b)\frac{\Omega_b}{\Omega} + c_b^2(1+w_a)\frac{\Omega_a}{\Omega}\right]\frac{S_{ab}}{1+w} \\ &- \left[\frac{2}{3}(\Delta+3K)\bar{\pi}_{ab} + \Gamma_{ab}\right] \end{aligned} \tag{13}$$

Adiabatic and isocurvature modes



Two different types of initial conditions can be distinguished for a system of two fluids: we can require that the entropy perturbation vanishes initially (adiabatic initial conditions), or, on the contrary, arrange that the gravitational potential initially vanishes, i.e. 'balance' the density perturbations so that $\delta^{\rm C} = 0$ (isocurvature initial conditions). We thus define the two limiting regimes:

- **adiabatic**: it is assumed that $\Gamma = S_{ab} = 0$, from which we deduce that:

$$\frac{\delta_a}{1+w_a} = \frac{\delta_b}{1+w_b}, \tag{4.3.70}$$

- **isocurvature**: in this case, we impose that $\Psi = 0$ and thus we have:

$$\delta^{\rm C} = 0, \qquad \Omega_a \delta_a^{\rm C} + \Omega_b \delta_b^{\rm C} = 0. \tag{4.3.71}$$

Since the evolution equations are linear, any system can always be decomposed into a linear combination of an initially adiabatic system and an initially isocurvature system.

Mixture of two perfect fluids

The discussion in the previous section is very general. In practice, we are interested in simpler special cases. Let us consider that of a mixture of two perfect fluids ($\Gamma_a = 0$, $\bar{\pi}_a = 0$) with purely gravitational interactions. Using (4.3.58) and (4.3.59), the relation (4.3.61) reduces to:

$$w\Gamma = -\frac{\Omega_a \Omega_b (1+w_a)(1+w_b)}{\Omega^2 (1+w)} \left(c_a^2 - c_b^2\right) S_{ab},$$

so that a closed system can be written for $\delta^{\rm C}$ and S_{ab}, namely:

$$\Phi'' + 3\mathcal{H}(1+c_{\rm s}^2)\Phi' + \left[2\mathcal{H}' + (\mathcal{H}^2 - K)(1+3c_{\rm s}^2)\right]\Phi - c_{\rm s}^2 \Delta\Phi = \frac{3}{2}\mathcal{H}^2 \left[\frac{\Omega_a \Omega_b (1+w_a)(1+w_b)}{\Omega(1+w)}\right](c_a^2 - c_b^2) S_{ab}, \tag{4.3.72}$$

$$(\Delta + 3K)\Phi = \frac{3}{2}\mathcal{H}^2 \Omega \delta^{\rm C}, \tag{4.3.73}$$

$$S_{ab}'' + \mathcal{H}\left\{1 - \frac{3\left[\Omega_a(1+w_a)c_b^2 + \Omega_b(1+w_b)c_a^2\right]}{\Omega(1+w)}\right\} S_{ab}' - \left[\frac{\Omega_a(1+w_a)c_b^2 + \Omega_b(1+w_b)c_a^2}{\Omega(1+w)}\right] \Delta S_{ab} = (c_a^2 - c_b^2)\frac{\Delta\delta^{\rm C}}{1+w}. \tag{4.3.74}$$

This system reduces to a system of two coupled second-order differential equations if the gravitational potential is eliminated from (4.3.72) by differentiating (4.3.73).

Isocurvature perturbations from non-interacting perfect fluids

Consider two non-interacting perfect fluids with constant equations of state,

$$p_a = w_a \rho_a, \qquad p_b = w_b \rho_b.$$

Since the fluids are separately conserved,

$$\dot{\rho}_i + 3H(\rho_i + p_i) = 0, \qquad i = a, b.$$

Thus,

$$\rho_i \propto a^{-3(1+w_i)}.$$

In particular,

$$\rho_a = \rho_{a*} \left(\frac{a}{a_*}\right)^{-3(1+w_a)}, \qquad \rho_b = \rho_{b*} \left(\frac{a}{a_*}\right)^{-3(1+w_b)}. \tag{1}$$

The curvature perturbation associated with each fluid is

$$\zeta_i = -\psi - H\frac{\delta\rho_i}{\dot{\rho}_i}.$$

Using the continuity equation,

$$\dot{\rho}_i = -3H(1+w_i)\rho_i,$$

we obtain

$$\zeta_i = -\psi + \frac{1}{3(1+w_i)}\frac{\delta\rho_i}{\rho_i}.$$

The relative entropy, or isocurvature, perturbation between the two fluids is therefore

$$S_{ab} \equiv 3(\zeta_a - \zeta_b) \tag{2}$$

$$= \frac{1}{1+w_a}\frac{\delta\rho_a}{\rho_a} - \frac{1}{1+w_b}\frac{\delta\rho_b}{\rho_b}. \tag{3}$$

Thus,

$$\boxed{S_{ab} = \frac{\delta_a}{1+w_a} - \frac{\delta_b}{1+w_b}}, \qquad \delta_i \equiv \frac{\delta\rho_i}{\rho_i}.$$

Now consider the dilution laws themselves. Taking a perturbation of

$$\rho_i \propto a^{-3(1+w_i)}$$

gives, for a common scale factor,

$$\frac{\delta\rho_i}{\rho_i} = \frac{\delta\rho_{i*}}{\rho_{i*}}.$$

Therefore the dilution law changes the background densities, but the common expansion does not by itself generate a new relative density perturbation.
For the ratio of the two densities, however,

$$\frac{\rho_a}{\rho_b} = \left(\frac{\rho_{a*}}{\rho_{b*}}\right)\left(\frac{a}{a_*}\right)^{-3(w_a - w_b)}.$$

Hence, if $w_a \neq w_b$, the *background composition* evolves with the expansion. For example,

$$\rho_m \propto a^{-3}, \qquad \rho_r \propto a^{-4}, \qquad \frac{\rho_m}{\rho_r} \propto a.$$

Nevertheless, the common dilution factor cancels when considering the spatial variation of the ratio:

$$\frac{\delta(\rho_a/\rho_b)}{\rho_a/\rho_b} = \frac{\delta\rho_a}{\rho_a} - \frac{\delta\rho_b}{\rho_b}.$$

Different dilution rates do not, by themselves, generate an isocurvature perturbation from initially adiabatic conditions. If initially

$$S_{ab} = 0,$$

then, for separately conserved barotropic fluids on super-Hubble scales,

$$\dot{\zeta}_a \simeq 0, \qquad \dot{\zeta}_b \simeq 0,$$

and hence

$$\boxed{\dot{S}_{ab} \simeq 0}.$$

Thus the dilution law determines how the individual components evolve with the expansion and how their *background relative abundance* changes, but a non-zero isocurvature perturbation requires a relative perturbation between the components to be present or a physical mechanism capable of generating one. Such a mechanism will involve energy transfer and hence non zero Q_a.

4.4 Power spectrum of density fluctuations

We will now be interested in the power spectrum of dark-matter density fluctuations. For this, and for the sake of simplicity, we assume that all matter is dark (i.e. $\Omega_{\rm b0} \ll \Omega_{\rm c0} \simeq \Omega_{\rm m0} \simeq 1$). Moreover, we restrict ourselves to a flat Universe ($K = 0$ and $\Omega = 1$) with no cosmological constant. The Universe thus contains, by construction, only dark matter and radiation. We choose the normalization $a_0 = 1$.

As before, we introduce the scale factor y normalized at equality [see Eq. (4.3.6)] so that the Hubble law takes the form (4.3.9). The value of y today is:

$$y_0 = \frac{a_0}{a_{\rm eq}} = 1 + z_{\rm eq} \simeq 2.4 \times 10^4\, h^2$$

[see (4.8)]. The Friedmann equation evaluated today makes it possible to compute the numerical value of the mode entering the Hubble radius at the time of matter–radiation equality (4.3.10), namely:

$$k_{\rm eq} = H_0\sqrt{2\frac{\Omega_{\rm m0}a_0}{a_{\rm eq}}} = 0.072\,\Omega_{\rm m0}h^2\,{\rm Mpc}^{-1}, \tag{4.4.1}$$

corresponding to a wavelength:

$$k_{\rm eq}^{-1} = \mathcal{H}_{\rm eq}^{-1} = \frac{14}{\Omega_{\rm m0}h^2}\,{\rm Mpc}. \tag{4.4.2}$$

In general, a given mode k enters the Hubble radius when $k = \mathcal{H}$, and the Friedmann equation (4.3.9) allows us to determine that this occurs at a redshift:

$$y_*\left(\bar{k}\right) = \frac{1+\sqrt{1+8\bar{k}}}{4\bar{k}^2}, \qquad 1 + z_*\left(\bar{k}\right) = \frac{y_0}{y_*\left(\bar{k}\right)}, \tag{4.4.3}$$

where we have introduced the reduced wavenumber:

$$\bar{k} \equiv \frac{k}{k_{\rm eq}} \tag{4.4.4}$$

for further notational convenience.

4.4.1 Two equivalent approaches

We can follow two approaches to study this physical system. The first is to consider the four equations of evolution for the density contrasts and the velocity perturbations of each fluid, and the Poisson equation. The second is to use the results of the previous section for the total density contrast and the entropy perturbations. We will use one or the other formulation depending on the situation.

Evolution of the density contrasts

Our system is composed of a mixture of radiation and matter that only interacts gravitationally, so that the system of equations is deduced from (4.2.74) and (4.2.75). In Fourier modes, they reduce to:

$$\delta_{\rm m}^{\rm N\prime} = k^2 V_{\rm m} + 3\Phi', \tag{4.4.5}$$

$$\delta_{\rm r}^{\rm N\prime} = \frac{4}{3}k^2 V_{\rm r} + 4\Phi', \tag{4.4.6}$$

$$V_{\rm m}' = -\mathcal{H} V_{\rm m} - \Phi, \tag{4.4.7}$$

$$V_{\rm r}' = -\Phi - \frac{1}{4}\delta_{\rm r}^{\rm N}, \tag{4.4.8}$$

where we have used that $\Psi = \Phi$ since $\bar{\pi} = 0$. Another equation should be added to determine the gravitational potential and we choose to use the Poisson equation:[5]

$$-k^2\Phi = \frac{3}{2}\mathcal{H}^2\left[\Omega_{\rm m}\delta_{\rm m}^{\rm N} + \Omega_{\rm r}\delta_{\rm r}^{\rm N} - 3\mathcal{H}\left(\Omega_{\rm m}V_{\rm m} + \frac{4}{3}\Omega_{\rm r}V_{\rm r}\right)\right], \tag{4.4.9}$$

or (4.2.68) in the form:[6]

$$\Phi' + \mathcal{H}\Phi = -\frac{3\mathcal{H}^2}{2}\left(\Omega_{\rm m}V_{\rm m} + \frac{4}{3}\Omega_{\rm r}V_{\rm r}\right). \tag{4.4.10}$$

These two equations can also be used in a combined way to express Φ in terms of $\delta_{\rm m}^{\rm N}$ and $\delta_{\rm r}^{\rm N}$ only.

Evolution of the gravitational potential and of the entropy

Using the expressions (4.3.44) to express $c_{\rm s}^2$ and w, the system (4.3.72)–(4.3.74) reduces to:

$$\ddot{\Phi} + \left(7 - \frac{1}{1+y} + \frac{8}{4+3y}\right)\frac{\dot{\Phi}}{2y} + \frac{\Phi}{y(1+y)(4+3y)} = \frac{2}{(4+3y)y^2}\left(\delta^{\rm C} - \frac{yS}{1+y}\right), \tag{4.4.11}$$

$$\delta^{\rm C} = -\frac{4}{3}\left(\frac{k}{k_{\rm eq}}\right)^2\frac{y^2}{1+y}\Phi, \tag{4.4.12}$$

$$\ddot{S} + \frac{3y^2+12y+8}{2y(1+y)(4+3y)}\dot{S} = \frac{2}{4+3y}\left(\frac{k}{k_{\rm eq}}\right)^2\left(\delta^{\rm C} - \frac{y}{1+y}S\right), \tag{4.4.13}$$

[5] Starting from (4.3.73) for $K = 0$ with (4.2.36) and $\frac{\rho'}{\rho} = -3\mathcal{H}^2(1+w)$

[6] Starting from (4.2.68) with $\frac{\rho'}{\rho} = -3\mathcal{H}^2(1+w)$ abd replace ΩV by $\sum_a \Omega_a V_a$

with $S = \delta_{\rm m} - \frac{3}{4}\delta_{\rm r}$. We have also assumed that for a Universe with Euclidean spatial sections and vanishing cosmological constant ($K = \Lambda = 0$), the matter and radiation density parameters can be rewritten simply in terms of the rescaled scale factor y as:

$$\Omega_{\rm m} = \frac{y}{1+y} \qquad \text{and} \qquad \Omega_{\rm r} = \frac{1}{1+y}. \tag{4.4.14}$$

This system generalizes (4.3.45) to the case where the entropy perturbation is no longer neglected. The density contrast $\delta^{\rm N}$ is then given by:

$$\delta^{\rm N} = \delta^{\rm C} - 2\left(\Phi + y\dot{\Phi}\right), \tag{4.4.15}$$

and the density contrasts of each fluid are:

$$\delta_{\rm m} = \frac{3(1+y)\delta/4 + S}{1+3y/4} \qquad \text{and} \qquad \delta_{\rm r} = \frac{(1+y)\delta - yS}{1+3y/4}. \tag{4.4.16}$$

Types of initial conditions

To solve the systems (4.4.5)–(4.4.8) or (4.4.11)–(4.4.13), initial conditions for the perturbations need to be imposed (see, for example, Ref. [39]). These initial conditions are fixed deep within the radiation era ($y_i \ll 1$) and for super-Hubble modes ($k/\mathcal{H}_i \ll 1$). As seen earlier, we can distinguish two types of initial conditions:

- **adiabatic**: the super-Hubble modes are assumed constant (in other words the decaying mode has had enough time to decay) and, by definition, we have:

  $$\Phi = \Phi_i, \qquad \Phi' = 0, \qquad S = 0, \qquad S' = 0,$$

  which fixes everything for the system (4.4.11)–(4.4.13). The Poisson equation (4.3.73) implies that:

  $$\delta^{\rm C}_i = -\frac{2}{3}x_i^2\Phi, \qquad \delta^{\rm C}_{{\rm r},i} = \delta^{\rm C}_i, \qquad \delta^{\rm C}_{{\rm m},i} = \frac{3}{4}\delta^{\rm C}_i, \tag{4.4.17}$$

  where $x_i \equiv k/\mathcal{H}_i$. Equation (4.4.10) then implies that:

  $$kV_i = -\frac{1}{2}x_i\Phi, \qquad V_{{\rm r},i} = V_{{\rm m},i} = V_i, \tag{4.4.18}$$

  One can derive it by noting that

  $$\left.\begin{array}{ll}\Omega_m & = \frac{y}{1+y} \to 0 \\ \Omega_r & = \frac{1}{1+y} \to 1\end{array}\right\} \mathcal{H}\Phi = -\frac{3}{2}\mathcal{H}^2\left(0 + \frac{4}{3}V_r\right) \tag{1}$$

  and (4.4.15) that:[7]

  $$\delta^{\rm N}_i = -2\Phi_i, \qquad \delta^{\rm N}_{{\rm r},i} = \delta^{\rm N}_i, \qquad \delta^{\rm N}_{{\rm m},i} = \frac{3}{4}\delta^{\rm N}_i, \tag{4.4.19}$$

  which completely fixes the initial conditions for the system (4.4.5)–(4.4.8).

[7] For superhorizon modes which sets the initial condition we have both $x \ll 1$ and $S = 0$ in the adiabatic limit, hence we find $\frac{\delta_i}{1+w_i} = \text{const}$

- **isocurvature**: we still assume the super-Hubble modes to be constant and the definition yields:

$$\Phi = 0, \qquad \Phi' = 0, \qquad S = S_i, \qquad S' = 0.$$

The Poisson equation implies that:

$$\delta_i^{\rm C} = 0, \qquad \delta_{{\rm r},i}^{\rm C} = -y_i S_i, \qquad \delta_{{\rm m},i}^{\rm C} = S_i, \tag{4.4.20}$$

so that (4.4.10) then implies:

$$V_{{\rm r},i} = V_{{\rm m},i} = V_i = 0, \tag{4.4.21}$$

while (4.4.15) provides:

$$\delta_i^{\rm N} = \delta_i^{\rm C}, \tag{4.4.22}$$

which again completely fixes the initial conditions for the system (4.4.11)–(4.4.13).

4.4.2 Different regimes

The behaviour of the solutions of these systems mainly depends on the value of the wavenumber k that determines the time at which the mode becomes sub-Hubble.

In order to have an idea of the orders of magnitudes, the following Table gives the values of y and of the redshift at the time a given mode becomes sub-Hubble for several selected values of k:

k (h^{-1} Mpc)	10^3	10^2	10	1	0.1	10^{-2}	10^{-3}
y_*	5×10^{-5}	5×10^{-4}	5×10^{-3}	5.1×10^{-2}	6.4×10^{-1}	26	2.5×10^{3}
z_*	4.75×10^{8}	4.75×10^{7}	4.74×10^{6}	4.63×10^{5}	3.7×10^{4}	9.05×10^{2}	8.4

We therefore have four kinds of behaviour to take into account, depending on whether the mode is (i) super-Hubble today, (ii) has become sub-Hubble during the matter-dominated era, or (iii) during the radiation-dominated era, or (iv) whether it has always been sub-Hubble.

Super-Hubble modes

Let us first consider long-wavelength modes. In practice, one can neglect all the terms that involve a Laplacian, which are proportional to k^2. Equation (4.3.66) then tells us that the entropy remains constant. Actually, we can also reach this conclusion by noting that (4.4.12) implies that $\delta^{\rm C} \propto k^2\Phi$. In the set of equations (4.4.11)–(4.4.13), the terms $k^2\Phi$ and $k^2 S$ are negligible in comparison to, respectively, $\ddot{\Phi}$ and $\ddot{S}$. We then infer that in (4.4.13), $k^4\Phi/k_{\rm eq}^2 \ll \ddot{S}$, which implies that S is constant.

Note then that using the transformation (4.2.29) to introduce $\delta_{\rm r}^{\rm C}$ which is of the order of $\delta^{\rm C} \propto k^2\Phi$ in the radiation era, the two equations (4.4.7) and (4.4.8) imply that:

$$V_{\rm rm}' = -\mathcal{H}V_{\rm rm} - \frac{1}{4}\delta^{\rm C},$$

and therefore that $V_{\rm rm} \propto a^{-1}$ during the radiation era.

In the case of adiabatic initial conditions, $S = 0$ and the solution can be obtained analytically in the form (4.3.46). Deep enough into the radiation era, both the gravitational potential and the total density contrast are constant. In particular,

$$\delta^{\rm N}(y) = -2\Phi(y), \qquad \delta_{\rm r}^{\rm N}(y) = \frac{4}{3}\delta_{\rm m}^{\rm N}(y),$$

whatever y. In the radiation era, we obtain:

$$\delta_{\rm r}^{\rm N} = -2\Phi(y \ll 1), \qquad \delta_{\rm m}^{\rm N}(y) = -\frac{3}{2}\Phi(y \ll 1),$$

and in the matter era ($y \gg 1$):

$$\Phi(y \gg 1) = \frac{9}{10}\Phi(y \ll 1), \qquad \delta_{\rm r}^{\rm N}(y \gg 1) = -\frac{12}{5}\Phi(y \ll 1), \qquad \delta_{\rm m}^{\rm N}(y \gg 1) = -\frac{9}{5}\Phi(y \ll 1). \tag{4.4.23}$$

To illustrate the effect of the initial conditions, let us now consider some isocurvature initial conditions and let us assume that initially $S = S_i \neq 0$ and $\Phi = 0$. As we have seen S remains constant. Deep in the radiation era ($y \ll 1$), $\delta_{\rm r}^{\rm N} \sim \delta^{\rm N} = -2\Phi = 0$ and thus $\delta_{\rm m}^{\rm N} = S_i$. Note then that for super-Hubble modes, since Φ and S are constant, $\Phi = -2S$ is a particular solution of (4.4.11). The general solution satisfies the initial condition $\Phi_i = 0$ and is thus obtained by adding this particular solution to the general solution (4.3.46) where the constant Φ_0 should be taken equal to $\Phi_0 = 2S_i$:

$$\Phi = \frac{2S_i}{10y^3}\left(16\sqrt{1+y} + 9y^3 + 2y^2 - 8y - 16\right) - 2S_i. \tag{4.4.24}$$

In the matter-dominated era ($y \gg 1$), $\delta_{\rm m}^{\rm N} \sim \delta^{\rm N} = -2\Phi$ and $\Phi(y \gg 1) = -S_i/5$. We conclude that, for $y \gg 1$:

$$\Phi = -\frac{1}{5}S_i, \qquad \delta^{\rm N} \sim \delta_{\rm m}^{\rm N} = \frac{2}{5}S_i, \qquad \delta_{\rm r}^{\rm N} = -\frac{4}{5}S_i. \tag{4.4.25}$$

Modes entering the Hubble radius in the matter-dominated era

As can be seen in Fig. 4.11, even though the mode is sub-Hubble from a redshift of the order of $z_* \sim 8.5$, it is always constant. Actually, to establish this result, we had to assume that the Laplacian term was negligible in the evolution equation for the gravitational potential. This is legitimate as long as $c_{\rm s}^2 k^2 \ll \mathcal{H}^2$, i.e. if:

$$\frac{2y^2}{3(1+y)(1+3y/4)}\bar{k}^2 \ll 1. \tag{4.4.26}$$

For the modes represented in Fig. 4.11, $\bar{k} = 10^{-3}$ and this ratio is still equal to 1.7×10^{-4} today. One can check that this ratio is of order unity for $k \sim k_{\rm eq}$ so that for all the modes entering the Hubble radius in the matter-dominated era, the gravitational potential remains constant, put aside the small variation due to the jump of the equation of state from 1/3 to 0 (see Fig. 5.12). The Poisson equation then tells us that:

$$\delta_{\rm m}^{\rm C} \propto a. \tag{4.4.27}$$

Modes entering the Hubble radius in the radiation-dominated era

When the Universe is dominated by radiation, the gravitational potential is determined by the density fluctuations of radiation. It is therefore given by the general solution (4.3.31) and its evolution is represented in Fig. 4.5. Thus, as soon as the modes becomes sub-Hubble during the radiation-dominated era, the gravitational potential is subject to damped oscillations and tends rapidly to 0.

The dark-matter fluid density perturbations are then obtained by solving their evolution equations with the gravitational potential determined by the solution (4.3.31),

$$\delta_{\rm m}^{\rm N\prime\prime} + \mathcal{H}\delta_{\rm m}^{\rm N\prime} = 3\Phi'' + 3\mathcal{H}\Phi' - k^2\Phi, \tag{4.4.28}$$

where the Hubble parameter is given by $\mathcal{H} = \eta^{-1}$ deep enough into the radiation era. The solution of this equation is of the form

$$\delta_{\rm m}^{\rm N} = A + B\ln(k\eta) + \delta_{\rm part}^{\rm N}, \tag{4.4.29}$$

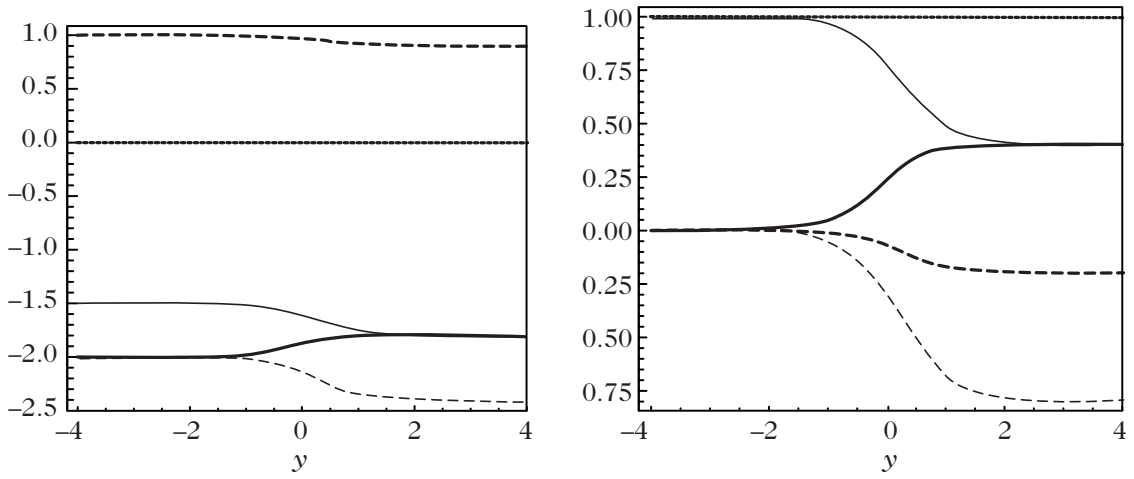


Figure 4.11: Evolution of the total density contrast $\delta^{\rm N}$ (thick solid line), the entropy S (thick dotted line) and gravitational potential (thick dashed line) as well as the two contrasts $\delta^{\rm N}_{\rm r}$ (thin dashed line) and $\delta^{\rm N}_{\rm m}$ (thin solid line) for a mode $\bar{k} = 10^{-3}$ assuming either adiabatic initial conditions ($\Phi_i = 1, S_i = 0$) [left] or isocurvature initial conditions ($\Phi_i = 0, S_i = -1$) [right].

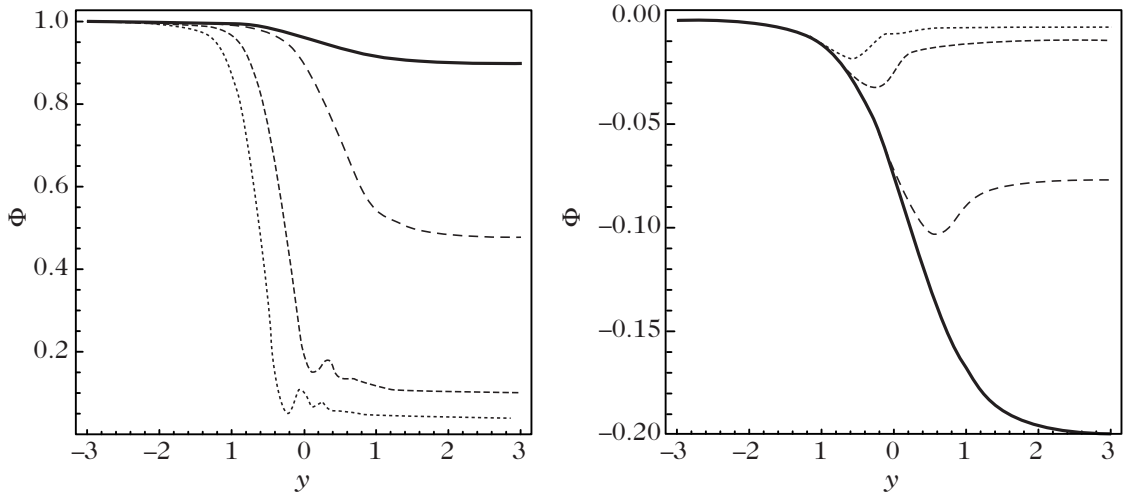


Figure 4.12: Evolution of the gravitational potential for $k = 10^{-2}$, 0.1, 0.5, 1 Mpc^{-1} for adiabatic (left) and isocurvature (right) initial conditions. These curves should be compared with Fig. 4.9, for which entropy fluctuations were neglected.

where the particular integral is obtained in the form

$$\delta^{\rm N}_{\rm part}(k,\eta) = \int_0^\eta {\rm d}\eta' \left[3\Phi''(k,\eta') + 3\mathcal{H}\Phi'(k,\eta') - k^2\Phi(k,\eta')\right] \eta' \ln\left(\frac{\eta'}{\eta}\right). \tag{4.4.30}$$

When $y \ll 1$, $\delta^{\rm N} \sim \delta^{\rm N}_{\rm r} \sim 2\Phi_{\rm i}$. Assuming adiabatic initial conditions ($\delta^{\rm N}_{\rm m} = \frac{3}{4}\delta^{\rm N}_{\rm r}$), one finds $\delta^{\rm N}_{\rm m} \sim \frac{3}{2}\Phi_{\rm i}$. The contribution from the particular solution turns out to be negligible on super-Hubble scales, so that one is left with $A = \frac{3}{2}\Phi_{\rm i}$ and $B = 0$.

As established by the analytical solution (4.3.31), the gravitational potential decays rapidly as soon as the perturbation enters the Hubble radius, remaining nearly constant beforehand. The integrand in Eq. (4.4.30) varies significantly only around $\eta' \sim 1/k$, so that the upper limit of the integral can be extended to infinity for sub-Hubble modes. Consequently, the integral evaluates to

$$\delta^{\rm N}_{\rm part}(k,\eta) = A + B\ln(k\eta), \tag{4.4.31}$$

where the constants

$$A = \int_0^\infty f(k\eta')k\eta' \ln(k\eta')\,\mathrm{d}(k\eta') \quad \text{and} \quad B = -\int_0^\infty f(k\eta')k\eta'\,\mathrm{d}(k\eta') \tag{4.4.32}$$

are governed by the auxiliary function $f(u)$, given explicitly by

$$f(u) = 9\Phi_\mathrm{i} u^{-5}\left[u(2u^2-27)\cos\frac{u}{\sqrt{3}} + \sqrt{3}(27-5u^2)\sin\frac{u}{\sqrt{3}}\right],$$

which is obtained by inserting the analytic potential solution (4.3.31). Numerically, one obtains $A \simeq -6\Phi_\mathrm{i}$ and $B \simeq 9\Phi_\mathrm{i}$, so that the general solution settles into the asymptotic form

$$\delta_\mathrm{m}^\mathrm{N}(k,\eta) \simeq \Phi_\mathrm{i}\left[-4.5 + 9\ln(k\eta)\right]. \tag{4.4.33}$$

Determining the exact numerical coefficients with high precision would of course necessitate a more detailed matching calculation. In summary, the dark matter perturbation evolves as

$$\delta_\mathrm{m}^\mathrm{N} \sim \delta_\mathrm{m}^\mathrm{C} \propto \ln a \quad (k\eta \gg 1); \qquad \delta_\mathrm{m}^\mathrm{N} \propto a^0 \quad (k\eta \ll 1) \tag{4.4.34}$$

during the radiation-dominated era (see Fig. 4.13 for a numerical integration).

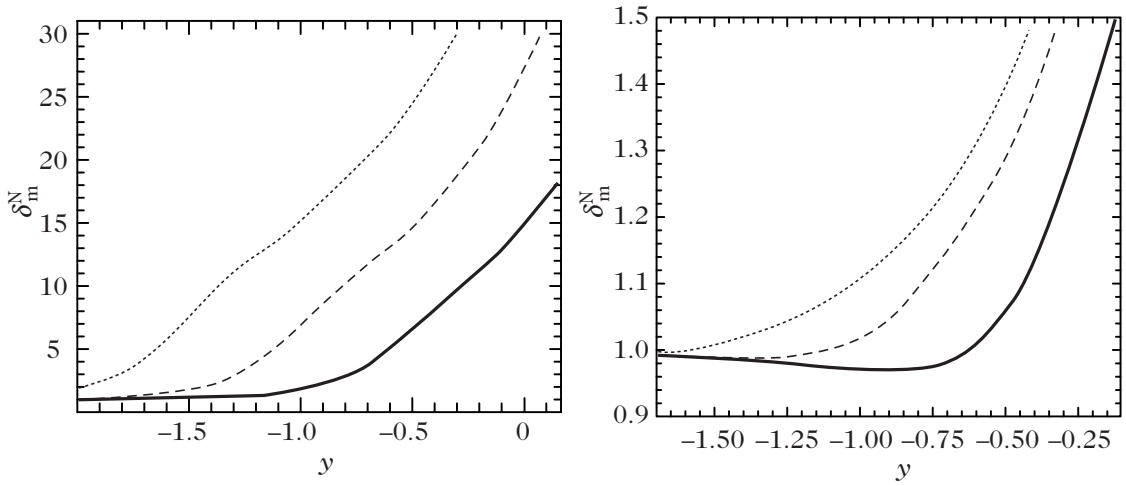


Figure 4.13: Evolution of the matter density contrast $\delta_\mathrm{m}^\mathrm{N}$ for $k = 1\,h\,\mathrm{Mpc}^{-1}$ (solid line), $k = 5\,h\,\mathrm{Mpc}^{-1}$ (dashed line) and $k = 10\,h\,\mathrm{Mpc}^{-1}$ (dotted line) for adiabatic (left) and isocurvature (right) initial conditions. These modes become sub-Hubble during the radiation-dominated era.

Sub-Hubble modes

Let us now focus on modes that were already sub-Hubble during the radiation era and examine how dark-matter perturbations respond across the radiation–matter transition. For these short-wavelength modes, we have seen (Fig. 4.6) that the density contrast is practically identical across all standard gauges.

For any mode that enters the Hubble radius during the radiation era, radiation perturbations δ_r remain approximately constant, whereas the dark matter perturbation grows logarithmically as $\delta_\mathrm{m} \sim \ln a$. Even though ρ_m is initially small compared to ρ_r, the contribution of dark matter density fluctuations to the Poisson equation eventually outgrows that of radiation because

$$\frac{\rho_\mathrm{m}\delta_\mathrm{m}}{\rho_\mathrm{r}\delta_\mathrm{r}} \sim y\ln y.$$

We therefore concentrate on the regime where $\delta_{\rm r} \ll \delta_{\rm m}$. The total density contrast is then given by

$$\delta = \frac{y}{1+y}\delta_{\rm m},$$

and the Poisson equation takes the form

$$k^2\Phi = -\frac{3}{4y}k_{\rm eq}^2\delta_{\rm m}, \tag{4.4.35}$$

as long as velocity perturbations can be neglected. We now employ the two fluid equations (4.4.5)–(4.4.7) expressed in terms of the variable y. Differentiating Eq. (4.4.5) and eliminating $V_{\rm m}$ through Eq. (4.4.7) together with $\dot{V}_{\rm m}$ from the continuity relation, one obtains a second-order differential equation for $\delta_{\rm m}$ involving Φ and its derivatives, which can then be eliminated using Eq. (4.4.35).

To simplify this governing equation, observe that we are interested in short-wavelength modes satisfying $k/\mathcal{H} \gg 1$, which implies $k_{\rm eq}^2 \ll k^2 y$. Dropping these negligible terms leads to the equation, known in the literature as the *Mészáros equation* [40],

$$\ddot{\delta}_{\rm m} + \frac{2+3y}{2y(y+1)}\dot{\delta}_{\rm m} - \frac{3}{2y(y+1)}\delta_{\rm m} = 0. \tag{4.4.36}$$

This differential equation admits an affine solution, and the second independent solution can be found by variation of parameters. One can verify that two linearly independent solutions are

$$D_+(y) = y + \frac{2}{3} \quad \text{and} \quad D_-(y) = D_+(y)\ln\left(\frac{\sqrt{1+y}+1}{\sqrt{1+y}-1}\right) - 2\sqrt{1+y}. \tag{4.4.37}$$

In the matter-dominated era ($y \gg 1$), these correspond respectively to a growing and a decaying mode, with asymptotic behaviours

$$D_+(y) \propto y \quad \text{and} \quad D_-(y) \propto y^{-3/2}.$$

This solution can be continuously matched to the earlier sub-Hubble radiation-era solution (4.4.29)–(4.4.31) at the time when the mode enters the Hubble radius at $y_*(k)$ (see Fig. 4.14 for a numerical integration).

Summary

In the simplified scenario of a Universe containing solely dark matter and radiation interacting gravitationally, the preceding analysis shows that the evolution of Φ and $\delta_{\rm m}$ depends crucially on the value of the wavenumber k and on the cosmological epoch during which the mode crosses inside the Hubble horizon.

However, this basic analysis leaves out four significant physical effects:

1. Cosmic matter contains baryons that are tightly coupled to photons via Compton scattering. We describe this coupling and its consequences for cosmological perturbations in Chapter 6.

2. Radiation has been modeled under a simple fluid approximation. This description is rather coarse and must be generalized using a kinetic framework wherein radiation is governed by the Boltzmann equation. Such a kinetic approach yields an infinite hierarchy of moment equations, of which the fluid equations represent merely the first two moments. This is treated in detail in Chapter 6, dedicated to the cosmic microwave background.

3. The late-time evolution of cosmological perturbations can be influenced by spatial curvature or a cosmological constant. These effects were demonstrated in Figs. 4.7 and 4.8.

4. Because of the presence of relativistic neutrinos, $\Phi \neq \Psi$, which modifies, for example, the ratio 9/10 for the evolution of the gravitational potential on super-Hubble scales.

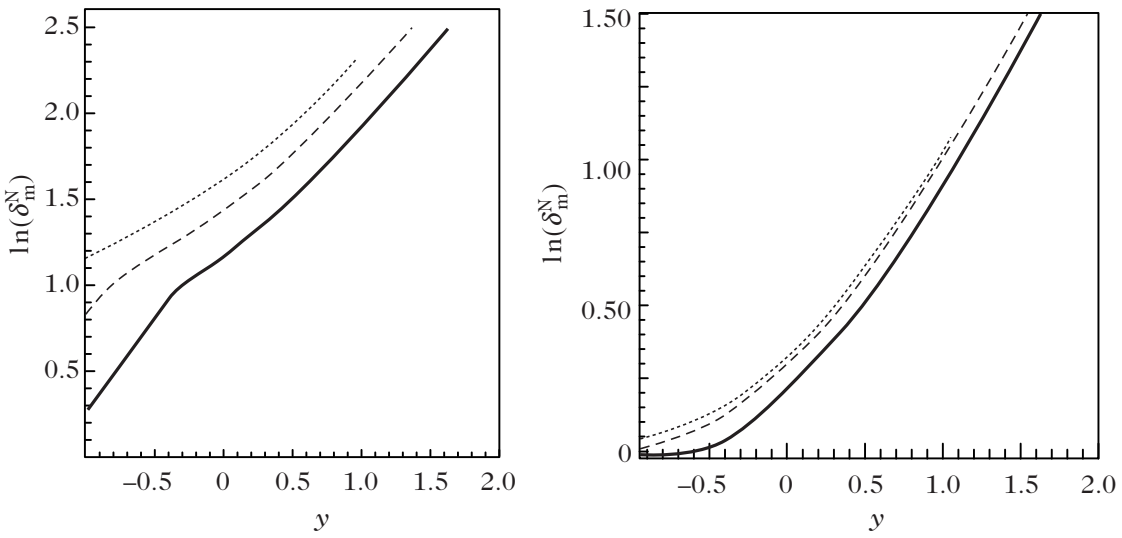


Figure 4.14: Evolution of the matter-density contrast $\delta_{\rm m}^{\rm N}$ for $k = 1\,h\,{\rm Mpc}^{-1}$ (solid line), $k = 5\,h\,{\rm Mpc}^{-1}$ (dashed line) and $k = 10\,h\,{\rm Mpc}^{-1}$ (dotted line) for adiabatic (left) and isocurvature (right) initial conditions.

Transfer function

This simplified investigation nevertheless enables us to grasp the qualitative structure of the large-scale matter power spectrum. To formalize this, one defines the transfer function

$$T(k,a) = \frac{\delta(k,a)D_+(a_{\rm i})}{\delta(k,a=a_{\rm i})D_+(a)}, \tag{4.4.38}$$

where $a_{\rm i}$ represents an initial time and $D_+(a)$ is the linear growth factor (4.1.24). Note that $a_{\rm i}$ is arbitrary provided it corresponds to an era before any scale of cosmological interest has crossed the Hubble radius. This transfer function quantifies the modification of the initial matter power spectrum throughout cosmic history and depends on the cosmological parameters and matter content of the Universe. The normalization factor D_+ is included by convention because during the matter era

$$\Delta\Phi = \frac{3}{2}H_0^2\Omega_{\rm m0}\frac{D_+(a)}{a}\frac{\delta}{D_+(a)},$$

and the ratio $D_+(a)/a$ remains approximately equal to unity during matter domination (see Fig. 4.1). The definition (4.4.38) implies that

$$\Phi(k,a) = T(k,a)\Phi(k,a_{\rm i}).$$

The modes observable today were initially super-Hubble, and theoretical models of the primordial Universe allow us to predict the initial matter power spectrum on these scales. Consequently, the gauge in which the primordial density contrast $\delta(k,a=a_{\rm i})$ is evaluated must be explicitly specified. In practice, since Φ remains constant for these super-Hubble modes, the identical transfer function is obtained if one employs $\delta^{\rm C}$ in the definition (4.4.38), because both density contrasts are related to the gravitational potential by the same Poisson equation. We therefore have

$$P_\Phi(k,a_0) = P_\Phi(k,a_{\rm i})T^2(k,a_0), \tag{4.4.39}$$

and

$$P_\delta(k,a_0) = P_\delta(k,a_{\rm i})T^2(k,a_0)\left[\frac{D_+(a_0)}{D_+(a_{\rm i})}\right]^2, \tag{4.4.40}$$

since $P_\delta(k, a_\mathrm{i}) = k^4 P_\Phi(k, a_\mathrm{i})$.

The qualitative behavior of the transfer function can be directly deduced from our previous analysis (see Fig. 4.17). Indeed, setting aside the logarithmic enhancement during the radiation era, modes that entered the Hubble radius during this epoch are effectively frozen and experience minimal growth between horizon crossing and matter–radiation equality. Their relative amplitude compared to long-wavelength modes ($k < k_\mathrm{eq}$) is therefore suppressed by a factor of $(k_\mathrm{eq}/k)^2$. One thus expects the limiting behaviors

$$T(k, a_0) \sim 1 \quad (k \ll k_\mathrm{eq}), \qquad T(k, a_0) \sim \left(\frac{k_\mathrm{eq}}{k}\right)^2 \quad (k \gg k_\mathrm{eq}). \tag{4.4.41}$$

This analytic estimate should be compared with Fig. 4.18, which is obtained from a complete numerical computation incorporating baryons and their coupling to radiation. Note also that the precise shape of the transfer function depends on the nature of the initial cosmological perturbations (adiabatic or isocurvature). Various analytic fitting formulas have been developed in the literature [19]. For example, for a cold dark matter model, one has

$$T(q) = \frac{\ln(1 + 2.34q)}{2.34q}\left[1 + 3.89q + (16.1q)^2 + (5.46q)^3 + (6.71q)^4\right]^{-1/4}, \tag{4.4.42}$$

$$T(q) = (5.6q)^2 \left\{1 + \left[15q + (0.9q)^{3/2} + (5.6q)^2\right]^{1.24}\right\}^{-1/1.24}, \tag{4.4.43}$$

with the dimensionless wavenumber $q = k/(\Gamma\,\mathrm{Mpc}^{-1})$ and shape parameter $\Gamma = \Omega_{\mathrm{m}0}h^2$, valid for adiabatic and isocurvature initial conditions, respectively.

Numerical determination of the transfer function

Equation (4.4.41) fixes only the two limiting regimes of $T(k)$, and the fitting formula (4.4.42) interpolates between them through coefficients that are not the output of any closed-form calculation. The intermediate region, which contains the turnover and therefore the scale k_eq itself, is accessible only by integrating the perturbation equations mode by mode.

This box sets out that computation. Both formulations of Sec. 4.4.1 are carried through — the four fluid equations with the Poisson constraint, and the coupled system for δ^C and S — and are shown to agree with each other and with (4.4.42) at the percent level. Throughout, assumptions (i)–(iv) of Sec. 4.4.2 are in force: baryons are absent, radiation is a perfect fluid, and $\bar{\pi} = 0$ so that $\Phi = \Psi$.

1. Units, background quantities and time variable. Adopt units in which

$$k_\mathrm{eq} \equiv \mathcal{H}_\mathrm{eq} = a_\mathrm{eq} H_\mathrm{eq} = 1, \tag{1}$$

so that the wavenumber appearing in every equation below is the reduced wavenumber $\bar{k}$ of (4.4.4). No cosmological parameter then enters the integration; they appear only in the conversion of paragraph 5. The Friedmann equation (4.3.9) gives

$$\mathcal{H}(y) = \frac{\sqrt{1+y}}{\sqrt{2}\,y}, \qquad \frac{\mathrm{d}y}{\mathrm{d}\eta} = \frac{\sqrt{1+y}}{\sqrt{2}}, \qquad \eta(y) = 2\sqrt{2}\left(\sqrt{1+y} - 1\right) \xrightarrow{y\ll 1} \sqrt{2}\,y. \tag{2}$$

Because the relevant physics spans some ten decades in a, the integration variable is taken to be

$$x \equiv \ln y, \qquad J(y) \equiv \frac{\mathrm{d}\eta}{\mathrm{d}x} = \frac{\sqrt{2}\,y}{\sqrt{1+y}}, \tag{3}$$

so that a conformal-time derivative is converted by $\mathrm{d}/\mathrm{d}x = J(y)\,\mathrm{d}/\mathrm{d}\eta$. The density parameters are those of (4.4.14).

2. Formulation A: the fluid variables. Define the two source combinations obtained by inserting (4.4.14) and (2) into (4.4.9) and (4.4.10),

$$P \equiv \frac{3}{4}\,\frac{y\,\delta_{\mathrm{m}}^{\mathrm{N}} + \delta_{\mathrm{r}}^{\mathrm{N}}}{y^2}, \qquad Q \equiv \frac{3}{4}\,\frac{y\,V_{\mathrm{m}} + \frac{4}{3}V_{\mathrm{r}}}{y^2}, \tag{4}$$

in terms of which the two Einstein equations read

$$\bar{k}^2\Phi = -P + 3\mathcal{H}Q \qquad \text{and} \qquad \Phi' = -Q - \mathcal{H}\Phi. \tag{5}$$

These are not independent: by the Bianchi identities either may be imposed as an initial constraint and is then propagated by the fluid equations. The choice matters numerically. On super-Hubble scales the two terms on the right of the first relation are individually much larger than $\bar{k}^2\Phi$ and cancel to many significant figures, so evaluating Φ from it at every step is ill-conditioned at small $\bar{k}$. The second relation involves no such cancellation. Accordingly Φ is promoted to a dynamical variable governed by the momentum constraint, and the Poisson equation is reserved as an independent accuracy check (paragraph 7). The system integrated is then

$$\frac{\mathrm{d}}{\mathrm{d}x}\begin{pmatrix}\delta_{\mathrm{m}}^{\mathrm{N}}\\ V_{\mathrm{m}}\\ \delta_{\mathrm{r}}^{\mathrm{N}}\\ V_{\mathrm{r}}\\ \Phi\end{pmatrix} = J(y)\begin{pmatrix}\bar{k}^2V_{\mathrm{m}} + 3\Phi'\\ -\mathcal{H}V_{\mathrm{m}} - \Phi\\ \frac{4}{3}\bar{k}^2V_{\mathrm{r}} + 4\Phi'\\ -\Phi - \frac{1}{4}\delta_{\mathrm{r}}^{\mathrm{N}}\\ \Phi'\end{pmatrix}, \qquad \Phi' = -Q - \mathcal{H}\Phi, \qquad y = e^x, \tag{6}$$

which is (4.4.5)–(4.4.8) together with (4.4.10). Note that the velocity potential V is already the well-conditioned variable: a velocity divergence $\theta = -\bar{k}^2V$ would carry an explicit factor $\bar{k}^2$ that must be removed again in (4).

3. Initial data for formulation A. Data are imposed at $y_i = 10^{-6}$, deep in the radiation era and far outside the Hubble radius for every mode considered. The adiabatic growing mode is normalised by $\Phi \to 1$ as $y \to 0$, identically for every $\bar{k}$, so that all scale dependence is generated by the subsequent evolution.

The leading-order forms $\delta^{\mathrm{N}} = -2\Phi$ and $V = -\Phi\eta/2$ are not adequate here, for a reason specific to this system. The source Q of (4) carries a pole, $Q \propto V/y^2$; in the radiation era the singular parts of Q and $\mathcal{H}\Phi$ cancel identically in (5), so that Φ' is determined entirely by the subleading terms. An error of order η^2 in V therefore produces an error of order unity in Φ', irrespective of how small y_i is taken. The symptom is a violation of the exact super-Hubble result $\Phi(y \gg 1)/\Phi_0 = 9/10$ implied by (4.3.46): truncated data return 1.0998 in place of 0.9002, and the discrepancy does not diminish as y_i is reduced. The remedy is to derive the data from the two constraints themselves, retained to the order at which Φ' is determined.

(a) The density contrast. Eliminating the velocity combination between (4.4.9) and (4.4.10) — that is, substituting $\frac{3}{2}\mathcal{H}^2(\Omega_{\mathrm{m}}V_{\mathrm{m}} + \frac{4}{3}\Omega_{\mathrm{r}}V_{\mathrm{r}}) = -(\Phi' + \mathcal{H}\Phi)$ into the Poisson equation — gives the exact relation

$$-\bar{k}^2\Phi = \tfrac{3}{2}\mathcal{H}^2\delta^{\mathrm{N}} + 3\mathcal{H}\left(\Phi' + \mathcal{H}\Phi\right). \tag{7}$$

On super-Hubble scales the left-hand side is negligible, and dividing through by $\frac{3}{2}\mathcal{H}^2$ leaves $\delta^{\mathrm{N}} = -2\Phi - 2\Phi'/\mathcal{H}$. Since $\Phi' = (\mathrm{d}y/\mathrm{d}\eta)\dot{\Phi} = y\mathcal{H}\dot{\Phi}$, this is

$$\delta^{\mathrm{N}} = -2\Phi - 2y\,\dot{\Phi}, \tag{8}$$

which is (4.4.15) in the limit $\delta^{\rm C} \to 0$. The familiar form $\delta^{\rm N} = -2\Phi$ is its $\dot\Phi \to 0$ truncation, and by the argument above it is exactly the retained term $-2y\dot\Phi$ that controls the answer.
(b) The velocities. The two species share a common velocity at this order. The relative velocity obeys $S' = \bar k^2(V_{\rm m} - V_{\rm r})$ by (4.3.66), while S vanishes identically for $\bar k \to 0$ on the adiabatic mode; hence $V_{\rm m} = V_{\rm r} \equiv V$. Then

$$\Omega_{\rm m} V_{\rm m} + \tfrac{4}{3}\Omega_{\rm r} V_{\rm r} = \left(\Omega_{\rm m} + \tfrac{4}{3}\Omega_{\rm r}\right) V = \frac{4+3y}{3(1+y)}\, V, \tag{9}$$

and (4.4.10) becomes, using $\mathcal{H}^2 = (1+y)/2y^2$,

$$\Phi' + \mathcal{H}\Phi = -\frac{3\mathcal{H}^2}{2}\,\frac{4+3y}{3(1+y)}\, V = -\frac{4+3y}{4y^2}\, V. \tag{10}$$

Solving for V and writing $\Phi' + \mathcal{H}\Phi = \mathcal{H}(\Phi + y\dot\Phi)$,

$$V = -\,\mathcal{H}\,\frac{\left(\Phi + y\dot\Phi\right) y^2}{1 + \frac{3}{4}y}. \tag{11}$$

(c) The background solution on which they are evaluated. Both (8) and (11) require Φ and $\dot\Phi$ on the super-Hubble growing mode (4.3.46). Expanding that solution about $y = 0$, the numerator $16\sqrt{1+y} + 9y^3 + 2y^2 - 8y - 16$ cancels through order y^2 and begins at $10y^3$, leaving

$$\Phi(y) = 1 - \frac{y}{16} + \frac{7y^2}{160} + \mathcal{O}\!\left(y^3\right), \qquad \dot\Phi(y) = -\frac{1}{16} + \frac{7y}{80} + \mathcal{O}\!\left(y^2\right), \tag{12}$$

normalised to $\Phi \to 1$. The series must be used in place of the closed form, which loses all significance to that same cancellation at $y \sim 10^{-6}$.
Adiabaticity, $S = 0$, then fixes the partition between the species. Collecting, the initial data are

$$\boxed{\begin{gathered} \Phi_i = \Phi(y_i), \qquad \delta^{\rm N} = -2\Phi_i - 2y_i\dot\Phi(y_i), \\ \delta^{\rm N}_{\rm r} = \frac{(1+y_i)\,\delta^{\rm N}}{1+\frac{3}{4}y_i}, \qquad \delta^{\rm N}_{\rm m} = \tfrac{3}{4}\,\delta^{\rm N}_{\rm r}, \qquad V_{\rm m} = V_{\rm r} = -\mathcal{H}(y_i)\,\frac{\left(\Phi_i + y_i\dot\Phi(y_i)\right) y_i^2}{1+\frac{3}{4}y_i}. \end{gathered}} \tag{13}$$

With these data the super-Hubble limit is reproduced to six significant figures, 0.900019 against the exact 0.900020.
4. Formulation B: the potential and the entropy perturbation. The second route of Sec. 4.4.1 replaces the four fluid variables by the total comoving density contrast and the relative entropy perturbation. The system is

$$\boxed{\begin{aligned} &\ddot\Phi + \left(7 - \frac{1}{1+y} + \frac{8}{4+3y}\right)\frac{\dot\Phi}{2y} + \frac{\Phi}{y(1+y)(4+3y)} = \frac{2}{(4+3y)y^2}\left(\delta^{\rm C} - \frac{yS}{1+y}\right), \\ &\ddot S + \mathcal{D}(y)\,\dot S = \frac{2\bar k^2}{4+3y}\left(\delta^{\rm C} - \frac{y}{1+y}S\right), \qquad \delta^{\rm C} = -\frac{4}{3}\,\bar k^2\,\frac{y^2}{1+y}\,\Phi, \end{aligned}} \tag{14}$$

with

$$\mathcal{D}(y) = \frac{1}{2(1+y)} + \frac{4}{y(4+3y)} = \frac{3y^2+12y+8}{2y(1+y)(4+3y)}, \tag{15}$$

and is again integrated in $x = \ln y$. The initial data are those of the adiabatic mode,

$$\boxed{\Phi = 1 - \frac{y_i}{16} + \frac{7y_i^2}{160}, \qquad \dot\Phi = -\frac{1}{16} + \frac{7y_i}{80}, \qquad S = \dot S = 0, \qquad y_i = 10^{-6}.} \tag{16}$$

Here $S = \dot S = 0$ is the adiabatic condition itself and requires no refinement; the subtlety of paragraph 3 does not arise, because Φ and $\dot\Phi$ are prescribed directly rather than inferred from the velocities.

Remark. The damping coefficient $\mathcal{D}(y)$ of (15) stands in place of the coefficient $(3y+2)/[2y(1+y)]$ carried by (4.4.13). The two agree in the radiation era, $\mathcal{D} \to 1/y$ as $y \to 0$, and differ by a factor three in the matter era. It is (15) that renders formulation B equivalent to formulation A, as Table 4.2 records; the alternative is tabulated there alongside for reference.

Recovery of the fluid variables. No information is lost in passing to (Φ, S). Equations (4.4.12), (4.4.15) and (4.4.16) return the density contrasts,

$$\delta^{\mathrm{C}} = -\frac{4}{3}\bar k^2 \frac{y^2}{1+y}\Phi, \quad \delta^{\mathrm{N}} = \delta^{\mathrm{C}} - 2\left(\Phi + y\dot\Phi\right), \quad \delta_{\mathrm{m}} = \frac{3(1+y)\delta^{\mathrm{N}}/4 + S}{1+3y/4}, \quad \delta_{\mathrm{r}} = \frac{(1+y)\delta^{\mathrm{N}} - yS}{1+3y/4}. \tag{17}$$

Figure 4.15 displays them for three representative modes. The regimes catalogued in Sec. 4.4.2 are directly visible: Φ constant at $9/10$ for $\bar k = 0.1$ and driven towards zero for $\bar k = 10$; δ_{m} stagnant through the radiation era and growing $\propto y$ thereafter, which is the Mészáros effect of (4.4.36); δ_{r} oscillating once inside the sound horizon; and S, vanishing initially, generated at a rate that increases with $\bar k$.

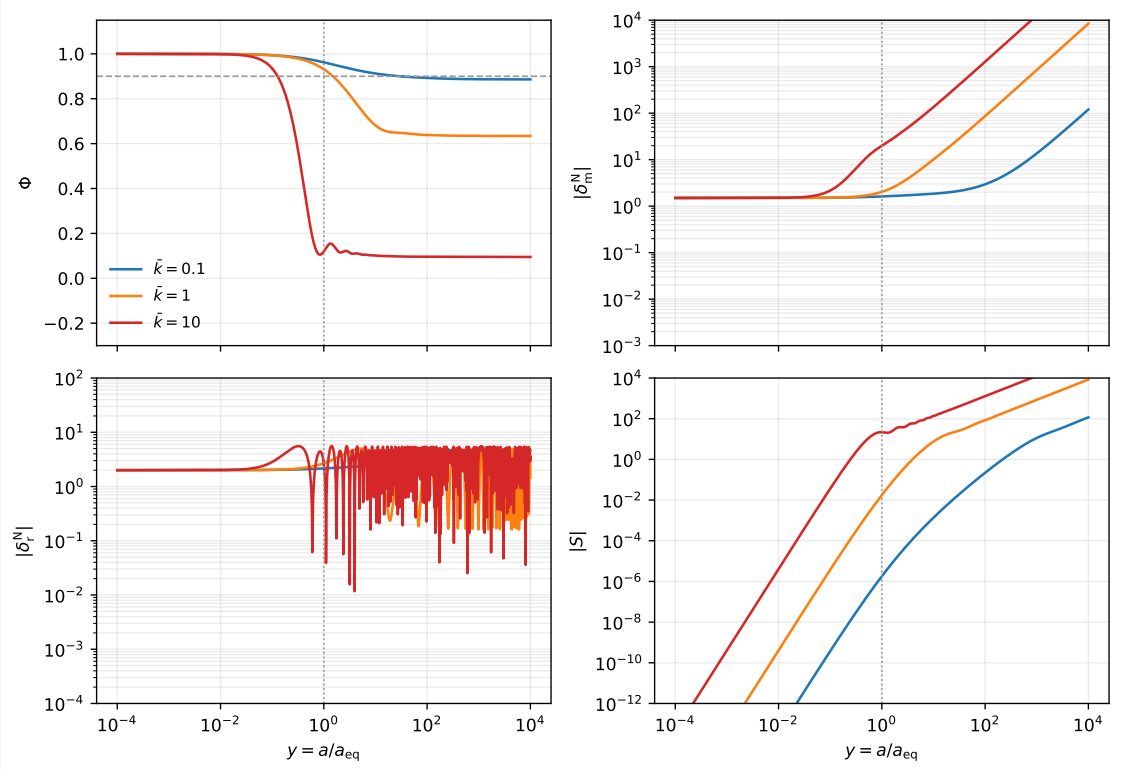


Figure 4.15: Variables of the system (14) and the fluid contrasts reconstructed from (17), for $\bar k = 0.1$, 1, 10. The dotted vertical line marks equality; the dashed line in the first panel is the super-Hubble value $9/10$.

The generation of S shown in the fourth panel also settles the status of the truncation $S \equiv 0$, under which (14) collapses to the single equation (4.3.45). That truncation

locks the two fluids together, $\delta_{\rm m} = \frac{3}{4}\delta_{\rm r}$ at all times, so the cold component cannot grow independently of the radiation and the Mészáros mechanism is disabled; the mixture behaves as a single fluid of sound speed $c_{\rm s}^2$ and Φ oscillates away. It is therefore serviceable only for $\bar{k} \lesssim 0.1$, as the last column of Table 4.2 records.

5. Normalisation and conversion to physical units. Definition (4.4.38) is applied in the form

$$T(\bar{k}) = \frac{\Phi(\bar{k},\, y_f)}{\Phi(\bar{k}\to 0,\, y_f)}, \qquad y_f = 10^4, \qquad \bar{k}_{\rm ref} = 10^{-2}, \tag{18}$$

which is legitimate because the primordial amplitude is the same for every mode by construction. The k-independent factor $9/10$ of (4.3.29) cancels between numerator and denominator, enforcing $T \to 1$ on large scales as required by (4.4.41).
Comparison with (4.4.42) requires no adjustable parameter but does require care with $k_{\rm eq}$, which depends on the photon temperature:

$$z_{\rm eq} = 2.50\times 10^4\,\Omega_{\rm m0}h^2\left(\frac{T_{\rm CMB}}{2.7\,{\rm K}}\right)^{-4}, \qquad k_{\rm eq} = a_{\rm eq}H_{\rm eq} = H_0\sqrt{2\,\Omega_{\rm m0}\,(1+z_{\rm eq})}, \tag{19}$$

the second following from $H_{\rm eq}^2 = 2H_0^2\Omega_{\rm m0}a_{\rm eq}^{-3}$ and agreeing with (4.4.1). The modern value $T_{\rm CMB} = 2.725\,{\rm K}$ gives $k_{\rm eq} = 0.0732\,\Omega_{\rm m0}h^2\,{\rm Mpc}^{-1}$. The fitting formula, however, predates the COBE measurement and was constructed with $T_{\rm CMB} = 2.7\,{\rm K}$, for which $k_{\rm eq} = 0.0746\,\Omega_{\rm m0}h^2\,{\rm Mpc}^{-1}$. Since $q = k/(\Gamma\,{\rm Mpc}^{-1})$ with $\Gamma = \Omega_{\rm m0}h^2$, the mapping in which the fit was calibrated is

$$\boxed{q = 0.0746\,\bar{k}} \tag{20}$$

and this is the one used below. The difference from 0.0732 is 1.9%, which through the q^{-2} of (4.4.41) amounts to a 3.0% shift in T at $q > 1$ and is therefore not negligible at the accuracy sought here.

6. Consistency of the fitting formula with the derived limits. As $q \to 0$ both factors of (4.4.42) tend to unity and $T \to 1$. As $q \to \infty$ the bracket is dominated by its quartic term, $[(6.71q)^4]^{-1/4} = (6.71q)^{-1}$, so that

$$T(q) \longrightarrow \frac{\ln(2.34\,q)}{15.70\,q^2}. \tag{21}$$

This is the $(k_{\rm eq}/k)^2$ suppression of (4.4.41), multiplied by the logarithm that the estimate leading to that equation discarded. The logarithm originates in the driven growth $\delta_{\rm m}^{\rm N} \propto \ln(k\eta)$ of (4.4.33): horizon entry occurs at $\eta_* \sim 1/k$, so evaluating that law at equality gives an amplitude $\propto \ln\bar{k}$, while (4.4.35) supplies the factor $\bar{k}^{-2}$. Only the interpolating polynomial is empirical.

7. Verification. Three independent checks are applied.
(i) Mészáros solutions. Symbolic substitution of both functions (4.4.37) into (4.4.36) returns identically zero. The small-y expansion of the second, $D_-(y) = -\frac{2}{3}\left[\ln y - 2\ln 2 + 3\right] + \mathcal{O}(y\ln y)$, confirms that it decays relative to D_+, justifying its neglect in the matching at $y_*(\bar{k})$. Forward integration of (4.4.36) reproduces $D_+ = y + \frac{2}{3}$ to a relative accuracy of 2×10^{-10} over $10^{-3} < y < 10^3$.
(ii) The logarithmic law. Fitting $\delta_{\rm m}^{\rm N} = \Phi_i\left[\mathcal{A} + \mathcal{B}\ln(k\eta)\right]$ over the sub-Hubble, radiation-dominated window $k\eta > e^2$, $y < 0.08$ gives

$\bar{k}$	300	10^3	3×10^3
$\lvert\mathcal{B}\rvert$	9.53	9.63	9.68
$\lvert\mathcal{A}\rvert$	5.44	6.08	6.60

against the values $\mathcal{B}=9$ and $|\mathcal{A}|=4.5$ of (4.4.33). The slope is stable to better than 1% across a decade in $\bar{k}$ and agrees with the analytic coefficient to 8%; the constant still drifts, indicating that the asymptotic regime is not fully attained at these wavenumbers. Consistently, $\partial\delta^{\mathrm{N}}_{\mathrm{m}}/\partial\ln y$ at fixed $\bar{k}$ and $\partial\delta^{\mathrm{N}}_{\mathrm{m}}/\partial\ln\bar{k}$ at fixed y are both measured to be $\simeq 10$, as a law depending only on $k\eta$ requires. The 8% excess in $\mathcal{B}$ is plausibly attributable to the perfect-fluid treatment of radiation, though that attribution is not established here.
(iii) Constraint propagation. The relative violation of the unused Poisson equation,

$$\mathcal{R}\equiv\left|\frac{\bar{k}^2\Phi+(P-3\mathcal{H}Q)}{P-3\mathcal{H}Q}\right|, \tag{22}$$

remains below 3.4×10^{-8} for all $\bar{k}>0.1$.

8. Results. Table 4.1 gives the transfer function from formulation A. The final column compares against the zero-baryon form of Eisenstein & Hu, $T_{\mathrm{EH}}=L/(L+Cx^2)$ with $L=\ln(2e+1.8x)$, $C=14.2+731/(1+62.5x)$ and $x=\bar{k}/13.41$, which is calibrated against a full Boltzmann integration and is therefore the more demanding reference.

$\bar{k}$	q	T	T_{BBKS}	$\dfrac{T}{T_{\mathrm{BBKS}}}$	T_{EH}	$\dfrac{T}{T_{\mathrm{EH}}}$
0.010	0.0007	1.0000	0.9984	1.002	0.9998	1.000
0.032	0.0024	0.9980	0.9946	1.003	0.9979	1.000
0.100	0.0075	0.9850	0.9809	1.004	0.9835	1.002
0.178	0.0133	0.9637	0.9621	1.002	0.9589	1.005
0.316	0.0236	0.9191	0.9226	0.996	0.9081	1.012
0.562	0.0420	0.8362	0.8432	0.992	0.8179	1.022
1.000	0.0746	0.7048	0.7086	0.995	0.6830	1.032
1.778	0.1327	0.5337	0.5338	1.000	0.5152	1.036
3.162	0.2359	0.3549	0.3570	0.994	0.3440	1.032
5.623	0.4195	0.2061	0.2092	0.985	0.2016	1.022
10.000	0.7460	0.1057	0.1074	0.984	0.1043	1.013
17.783	1.3266	0.0489	0.0495	0.988	0.0486	1.008
31.623	2.3591	0.0210	0.0212	0.991	0.0209	1.004
56.234	4.1951	0.0085	0.0086	0.992	0.0085	0.997
100.000	7.4600	0.0033	0.0033	0.991	0.0033	0.987

Table 4.1: Transfer function from formulation A, Eqs. (6) and (13), compared with the fitting formula (4.4.42) and with the Eisenstein–Hu zero-baryon form.

Agreement with (4.4.42) is 1.6% at worst and 0.7% in the root mean square over four decades in $\bar{k}$, with no parameter adjusted. Permitting a free horizontal rescale returns $k_{\mathrm{eq}}=0.0752\,\Omega_{\mathrm{m0}}h^2$, within 0.8% of (20); the characteristic scale is thus pinned and no residual mismatch of shape remains.
Table 4.2 repeats the computation in formulation B and places beside it the two variants discussed above: the system integrated with the damping coefficient $(3y+2)/[2y(1+y)]$ of (4.4.13), and the truncation $S\equiv0$.

$\bar{k}$	q	$T_{(\Phi,S)}$	T_{A}	$\left\lvert \frac{T_{(\Phi,S)}}{T_{\mathrm{A}}} - 1 \right\rvert$	T_{BBKS}	$\frac{T_{(\Phi,S)}}{T_{\mathrm{BBKS}}}$	$T_{(3y+2)}$ / $T_{S\equiv 0}$
0.010	0.0007	1.00000	1.00000	0.0e+00	0.99837	1.0016	1.00000 / +1.00000
0.024	0.0018	0.99889	0.99889	0.0e+00	0.99603	1.0029	0.99884 / +0.99875
0.056	0.0042	0.99421	0.99421	0.0e+00	0.98999	1.0043	0.99357 / +0.99177
0.133	0.0099	0.97641	0.97641	0.0e+00	0.97320	1.0033	0.97283 / +0.95330
0.316	0.0236	0.91906	0.91906	1.1e-07	0.92264	0.9961	0.90328 / +0.76116
0.750	0.0559	0.77660	0.77660	3.9e-07	0.78279	0.9921	0.72806 / +0.19481
1.778	0.1327	0.53373	0.53373	9.4e-07	0.53380	0.9999	0.44942 / +0.00003
4.217	0.3146	0.27496	0.27496	2.5e-06	0.27801	0.9890	0.20317 / +0.00000
10.000	0.7460	0.10567	0.10567	1.1e-05	0.10736	0.9843	0.07261 / +0.00000
23.714	1.7690	0.03231	0.03231	4.2e-05	0.03263	0.9902	0.02188 / −0.00000
56.234	4.1951	0.00849	0.00849	1.8e-04	0.00856	0.9915	0.00580 / +0.00000

Table 4.2: Formulation B, Eqs. (14) and (16), against formulation A and the fitting formula. The last column gives, for comparison, the same system integrated with the damping coefficient $(3y+2)/[2y(1+y)]$ and with the truncation $S \equiv 0$.

The two formulations agree to 5×10^{-4} in the worst case and to better than 10^{-5} over most of the range, confirming the equivalence asserted in Sec. 4.4.1. Figure 4.16 assembles the comparison.

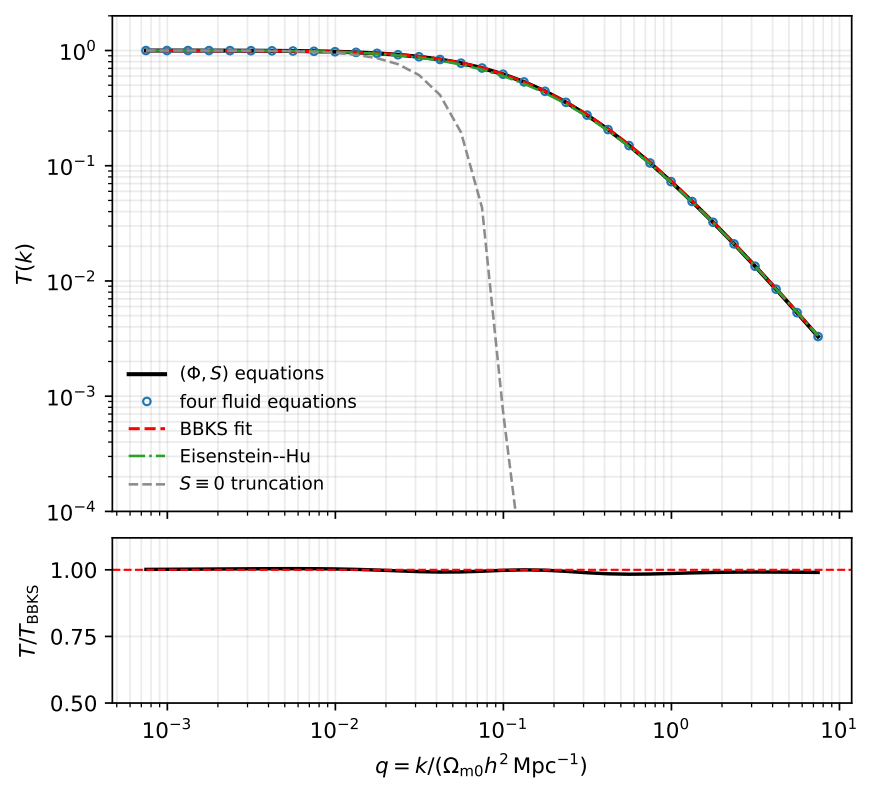
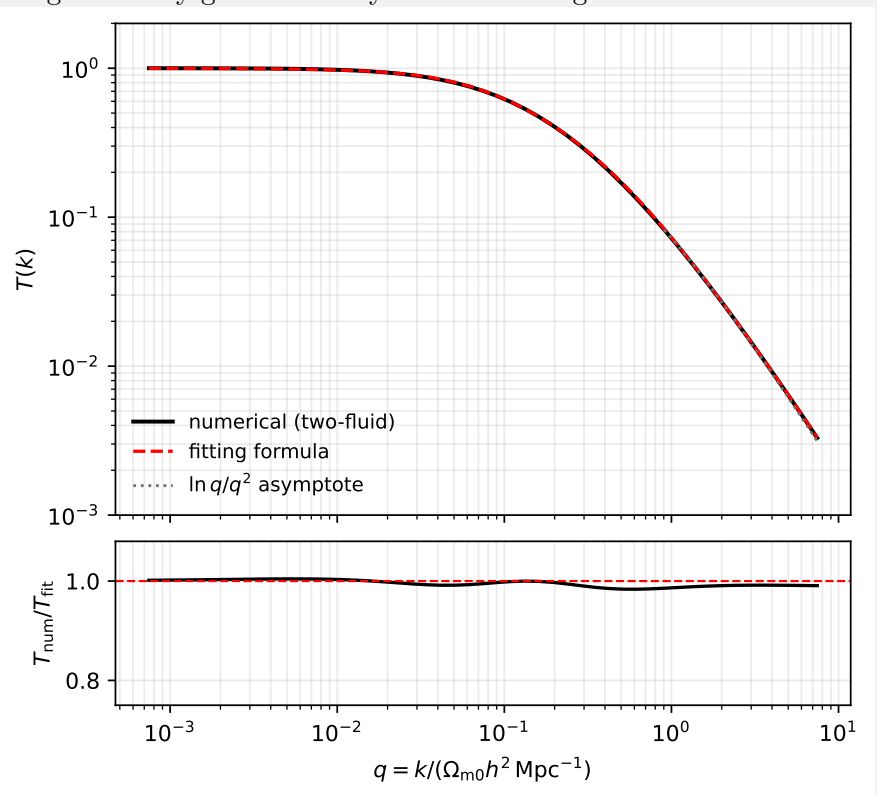


Figure 4.16: Transfer function from the system (14) (solid) and from the fluid equations (6) (circles), against the fitting formula (4.4.42), the Eisenstein–Hu zero-baryon form, the same system integrated with the damping coefficient $(3y+2)/[2y(1+y)]$ (dotted), and the truncation $S \equiv 0$ (grey dashed). Lower panel: ratio to (4.4.42).

9. Accuracy and residual physics. Two ingredients of the comparison are conventions rather than dynamics and are worth isolating. Adopting initial data truncated at leading

order in place of (13) lowers $\langle T/T_{\rm BBKS}\rangle$ at $q>1$ from 0.990 to 0.938; adopting $q=0.072\,\bar{k}$ in place of (20) lowers it further to 0.805. Neither is a physical effect.
What remains is the 3.6% excess over $T_{\rm EH}$, peaking at $\bar{k}\simeq 1.8$, that is, in the turnover itself. Its sign is the expected one: free-streaming neutrinos carry some 41% of $\rho_{\rm r}$ and do not sustain the driving potential, so the true turnover is more strongly suppressed than a perfect-fluid treatment predicts. This is item (iv) of Sec. 4.4.2.
That residual cannot be diagnosed by comparison with (4.4.42), because the two published fits differ from one another by up to 3.9% at $\bar{k}\simeq 4.6$, which exceeds the 1.6% separating the present computation from the first of them. Agreement with (4.4.42) has therefore been carried to the accuracy of the reference itself. Improvement beyond this level requires both a different reference — a Boltzmann code, or $T_{\rm EH}$ — and the physics of Chapter 6: neutrino multipoles with $\sigma_\nu\neq 0$, hence $\Phi\neq\Psi$, and photons tightly coupled to a finite baryon fraction. The reduction, normalisation and unit mapping above carry over unchanged, but the initial data of paragraph 3 must be rederived, since the super-Hubble growing mode then satisfies $\Psi=\left(1+\tfrac{2}{5}f_\nu\right)\Phi$ and no closed form for the $\mathcal{O}(y)$ corrections is available.

10. Implementation. Two self-contained scripts. The first integrates (6) with the data (13) and performs the checks of paragraph 7; the second integrates (14) with the data (16), reconstructs the fluid variables through (17), and evaluates the two variants of Table 4.2. Together they generate every number and figure above.



Formulation A: the fluid equations.

```
"""
CDM transfer function: integration of the two-fluid system and verification
of the analytic ingredients, in the notation of Chapter 5.

Conventions
```

```
-----------
    y = a/a_eq      x = ln y        ' = d/d(eta)        dot = d/dy
    units: k_eq = H_eq = a_eq H_eq = 1, so k is the reduced wavenumber kbar
    variables: delta_m^N, delta_r^N, V_m, V_r, Phi   (Phi = Psi, pi_bar = 0)
    V is the chapter's velocity potential, V = -theta/k^2

Assumptions: no baryons; radiation as a perfect fluid (w = cs^2 = 1/3);
adiabatic growing mode; flat, matter + radiation only.

Run:  python3 transfer.py          (add --no-figure to skip matplotlib)
"""
import sys
import numpy as np
from scipy.integrate import solve_ivp
from scipy.optimize import curve_fit

S2 = np.sqrt(2.0)

# ---------------------------------------------------------------------- background
def cal_H(y):
    """comoving Hubble rate, Friedmann eq. in the variable y."""
    return np.sqrt(1.0 + y) / (S2 * y)

def J(y):
    """d(eta)/d(ln y)."""
    return S2 * y / np.sqrt(1.0 + y)

# ------------------------------------------------------- perturbation system
def rhs(x, u, k):
    y = np.exp(x)
    dm, Vm, dr, Vr, Phi = u
    H = cal_H(y)
    Q = 0.75 * (y * Vm + (4.0 / 3.0) * Vr) / y**2       # source of Phi'
    dPhi = -Q - H * Phi                                  # momentum constraint
    return J(y) * np.array([
        k * k * Vm + 3.0 * dPhi,                         # delta_m^N
        -H * Vm - Phi,                                   # V_m
        (4.0 / 3.0) * k * k * Vr + 4.0 * dPhi,           # delta_r^N
        -Phi - 0.25 * dr,                                # V_r
        dPhi])                                           # Phi

# super-Hubble adiabatic growing mode, Phi -> 1 as y -> 0.  Series form of the
# chapter's closed-form solution; the closed form itself cancels catastrophically
# at y ~ 1e-6.
def Phi_sh(y):  return 1.0 - y/16.0 + 7.0*y**2/160.0
def dPhi_sh(y): return -1.0/16.0 + 7.0*y/80.0

def ic(k, y_i):
    """Adiabatic growing mode, super-Hubble, radiation era, normalised to
    Phi -> 1 as y -> 0.

    The leading-order data (delta_r = -2, V = -eta/2) are NOT accurate enough.
    Q carries a 1/y pole, so the O(y) part of the velocities feeds straight into
    Phi' at O(1); truncating there injects a spurious second super-Hubble mode
    and drives Phi to 11/10 instead of 9/10.  We therefore impose the exact
    k -> 0 constraints:
        energy    :  delta^N_tot = -2 Phi - 2 y dPhi/dy
        momentum  :  Phi' + H Phi = Q,  with V_m = V_r
    """
    y  = y_i
    H  = cal_H(y)
```

```
    Ph, Py = Phi_sh(y), dPhi_sh(y)
    dtot = -2.0*Ph - 2.0*y*Py                       # energy constraint
    dr   = dtot*(1.0 + y)/(1.0 + 0.75*y)            # adiabatic, S = 0
    dm   = 0.75*dr
    th   = H*(y*Py + Ph)*y**2/(1.0 + 0.75*y)        # momentum constraint
    return np.array([dm, -th, dr, -th, Ph])         # chapter's V = -theta/k^2

def run(k, y_i=1e-6, y_f=1e4):
    s = solve_ivp(rhs, [np.log(y_i), np.log(y_f)], ic(k, y_i), args=(k,),
                  method="DOP853", rtol=1e-9, atol=1e-13)
    assert s.success, (k, s.message)
    return s

def poisson_residual(k, s):
    """relative violation of the unused energy constraint k^2 Phi = -P + 3 H Q."""
    y = np.exp(s.t[-1])
    dm, Vm, dr, Vr, Phi = s.y[:, -1]
    P = 0.75 * (y * dm + dr) / y**2
    Q = 0.75 * (y * Vm + (4.0 / 3.0) * Vr) / y**2
    src = -P + 3.0 * cal_H(y) * Q
    return abs(k * k * Phi - src) / abs(src)

def T_eh(x):
    """Eisenstein & Hu (1998) zero-baryon form, x = k/(13.41 k_eq).
    Calibrated against a Boltzmann code, so a stiffer target than BBKS."""
    L = np.log(2*np.e + 1.8*x)
    C = 14.2 + 731.0/(1.0 + 62.5*x)
    return L/(L + C*x**2)

def T_fit(q):
    """BBKS fitting formula, adiabatic CDM."""
    return (np.log(1 + 2.34 * q) / (2.34 * q)) * \
           (1 + 3.89 * q + (16.1 * q)**2 + (5.46 * q)**3 + (6.71 * q)**4)**(-0.25)

# ================================ checks ====================================
def check_meszaros():
    """[1] symbolic + numerical verification of the Meszaros solutions."""
    import sympy as sp
    y = sp.symbols('y', positive=True)
    s = sp.sqrt(1 + y)
    Dp = y + sp.Rational(2, 3)
    Dm = Dp * sp.log((s + 1) / (s - 1)) - 2 * s
    for nm, D in [("D_+", Dp), ("D_-", Dm)]:
        L = sp.diff(D, y, 2) + (2 + 3*y) / (2*y*(y + 1)) * sp.diff(D, y) \
            - 3 / (2*y*(y + 1)) * D
        print(f"[1] residual of {nm} in the Meszaros equation: {sp.simplify(L)}")
    c = float(sp.limit(Dm + sp.Rational(2, 3)*sp.log(y), y, 0))
    print(f"[1] D_- small-y expansion: -(2/3) ln y - {abs(c):.4f}"
          f"   [= -(2/3)(ln y - 2 ln 2 + 3)]")

    def mesz(t, u):
        d, dp = u
        return [dp, (3*d/(2*t*(t+1))) - (2 + 3*t)/(2*t*(t+1)) * dp]
    y0, y1, h = 1e-3, 1e3, 1e-9
    f = lambda t: t + 2.0/3.0
    sol = solve_ivp(mesz, [y0, y1], [f(y0), (f(y0+h) - f(y0-h))/(2*h)],
                    method="Radau", rtol=1e-10, atol=1e-14,
                    t_eval=np.logspace(-3, 3, 13))
    print(f"[1] max rel. deviation of integrated D_+ : "
          f"{np.abs(sol.y[0]/f(sol.t) - 1).max():.1e}")

```

```
def check_log_law():
    """[2] delta_m^N = Phi_i [A + B ln(k eta)] in the radiation era."""
    print("[2] logarithmic law   delta_m^N / Phi_i = A + B ln(k eta)")
    print(f"    {'kbar':>8}{'|A|':>9}{'|B|':>9}{'pts':>7}")
    for k in [300., 1000., 3000.]:
        s = run(k, y_i=1e-9, y_f=1.0)
        y, d = np.exp(s.t), s.y[0]
        u = np.log(S2 * k * y)                  # ln(k eta), eta = sqrt(2) y in RD
        m = (u > 2.5) & (y < 0.05)              # sub-Hubble, still radiation era
        if m.sum() < 20:                        # widen for the lowest kbar
            m = (u > 2.0) & (y < 0.08)
        p = np.polyfit(u[m], d[m], 1)
        print(f"    {k:8.0f}{abs(p[1]):9.2f}{abs(p[0]):9.2f}{m.sum():7d}")
    print("    (analytic: |A| = 4.5, |B| = 9)")
    for ye in [0.03, 0.10]:
        kk = np.array([100., 300., 1000., 3000.])
        dv = []
        for k in kk:
            s = run(k, y_i=1e-9, y_f=1.2)
            dv.append(np.interp(ye, np.exp(s.t), s.y[0]))
        print(f"    d(delta_m)/d ln kbar at y={ye:4.2f}: "
              f"{np.polyfit(np.log(kk), dv, 1)[0]:7.2f}")

# ================================ sweep ====================================
def sweep(n=49, k_ref=1e-2):
    phi0 = run(k_ref).y[4, -1]
    kbar = np.logspace(-2, 2, n)
    T, R = [], []
    for k in kbar:
        s = run(k)
        T.append(s.y[4, -1] / phi0)
        R.append(poisson_residual(k, s))
    T, R = np.array(T), np.array(R)
    # k_eq = a_eq H_eq = H_0 sqrt(2 Omega_m0 (1 + z_eq)),  H_eq^2 = 2 H_0^2
    # Omega_m0 a_eq^-3, with z_eq = 2.50e4 Omega_m0 h^2 (T_CMB/2.7 K)^-4.
    # BBKS (1986) predates COBE and used T_CMB = 2.7 K, i.e. k_eq = 0.0746
    # Omega_m0 h^2 /Mpc; with the modern T_CMB = 2.725 K one gets 0.0732.
    # The fit must be compared in the convention it was built in.
    q = 0.0746 * kbar
    return kbar, q, T, T_fit(q), R

# ================================ main =====================================
if __name__ == "__main__":
    check_meszaros()
    print()
    check_log_law()
    print()

    kbar, q, T, Tb, R = sweep()
    r = T / Tb
    print(f"[3] constraint residual: kbar>0.1 -> {R[kbar>0.1].max():.1e}, "
          f"kbar>1 -> {R[kbar>1].max():.1e}")
    print(f"[3] <T_num/T_fit>: q<0.05 {r[q<0.05].mean():.3f} | "
          f"0.05<q<0.5 {r[(q>0.05)&(q<0.5)].mean():.3f} | q>1 {r[q>1].mean():.3f}")

    g = lambda q, b, C: np.log(b*q) / (C * q**2)
    m = q > 0.8
    p, _ = curve_fit(g, q[m], T[m], p0=[2.3, 15.7])
    print(f"[4] numerical asymptote: T = ln({p[0]:.2f} q)/({p[1]:.1f} q^2), "
          f"max dev {np.abs(g(q[m], *p)/T[m] - 1).max():.3f}")
    print( "[4] fitting formula    : T = ln(2.34 q)/(15.70 q^2)")
```

```

    eh = T / T_eh(kbar / 13.41)
    print(f"[5] vs Eisenstein-Hu: max |T_num/T_EH - 1| = {np.abs(eh-1).max():.3f}"
          f" at kbar = {kbar[np.argmax(np.abs(eh-1))]:.2f}")
    rf = Tb / T_eh(kbar / 13.41)
    print(f"[5] the two published fits differ from each other by up to "
          f"{np.abs(rf-1).max():.3f}")

    print(f"\n{'kbar':>9}{'q':>10}{'T_num':>10}{'T_fit':>10}{'ratio':>8}{'T_EH':>10}")
    for i in range(0, len(kbar), 6):
        print(f"{kbar[i]:9.3f}{q[i]:10.4f}{T[i]:10.4f}{Tb[i]:10.4f}{r[i]:8.3f}"
              f"{T_eh(kbar[i]/13.41):10.4f}")
    np.savetxt("transfer.dat", np.c_[kbar, q, T, Tb, r], fmt="%.6e",
               header="kbar  q  T_num  T_fit  ratio")

    if "--no-figure" not in sys.argv:
        import matplotlib; matplotlib.use("Agg")
        import matplotlib.pyplot as plt
        fig, ax = plt.subplots(2, 1, figsize=(6, 5.6), sharex=True,
                               gridspec_kw=dict(height_ratios=[3, 1], hspace=0.08))
        ax[0].loglog(q, T, 'k-', lw=1.8, label='numerical (two-fluid)')
        ax[0].loglog(q, Tb, 'r--', lw=1.5, label='fitting formula')
        ax[0].loglog(q[m], g(q[m], *p), ':', color='0.45', lw=1.3,
                     label=r'$\ln q/q^{2}$ asymptote')
        ax[0].set_ylabel(r'$T(k)$'); ax[0].set_ylim(1e-3, 2)
        ax[0].legend(frameon=False, fontsize=9); ax[0].grid(alpha=.25, which='both')
        ax[1].semilogx(q, r, 'k-', lw=1.5); ax[1].axhline(1, color='r', ls='--', lw=1)
        ax[1].set_ylabel(r'$T_{\rm num}/T_{\rm fit}$'); ax[1].set_ylim(.75, 1.08)
        ax[1].set_xlabel(r'$q = k/(\Omega_{\rm m0}h^{2}\,{\rm Mpc}^{-1})$')
        ax[1].grid(alpha=.25, which='both')
        fig.savefig("transfer.pdf", bbox_inches='tight')
        print("\nwrote transfer.dat, transfer.pdf")
```

Formulation B: the (Φ, S) system.

```
"""
Transfer function from the chapter's second formulation: the coupled system for
the total comoving density contrast and the relative entropy perturbation,
Eqs. (4.4.11)-(4.4.13), with the damping coefficient of S corrected.

    Phi_dd + (1/2y)[7 - 1/(1+y) + 8/(4+3y)] Phi_d + Phi/[y(1+y)(4+3y)]
        = 2/[(4+3y) y^2] ( delta^C - y S/(1+y) )                    (4.4.11)

    delta^C = -(4/3) kbar^2 y^2 Phi/(1+y)                           (4.4.12)

    S_dd + D(y) S_d = 2 kbar^2/(4+3y) ( delta^C - y S/(1+y) )       (4.4.13)

        D(y) = 1/[2(1+y)] + 4/[y(4+3y)]

The correction descends from the relative Euler equation: subtracting the
per-fluid Euler equations and converting only the TOTAL density perturbation to
the comoving slicing leaves velocity slips V_a - V proportional to V_ab, so

    V_ab' = -H [1 - 3 C_ab/(Omega(1+w))] V_ab - ... ,
    C_ab  = c_a^2 (1+w_b) Omega_b + c_b^2 (1+w_a) Omega_a .

For matter + radiation C_ab = Omega_m/3 and the bracket is 4/(4+3y): unity as
y -> 0, so the printed form is right in the radiation era and fails afterwards.

Reconstruction of the fluid variables (Eqs. 4.4.15, 4.4.16):
```

```
    delta^N = delta^C - 2(Phi + y Phi_d)
    delta_m = [3(1+y) delta^N/4 + S]/(1 + 3y/4)
    delta_r = [(1+y) delta^N - y S]/(1 + 3y/4)

Integration variable x = ln y; units k_eq = a_eq H_eq = 1, so k is kbar.

Run:  python3 transfer-function-phiS.py          (--no-figure to skip plots)
"""
import sys
import numpy as np
from scipy.integrate import solve_ivp

S2 = np.sqrt(2.0)

# ------------------------------------------------------------- background
def cal_H(y):  return np.sqrt(1.0 + y) / (S2 * y)
def J(y):      return S2 * y / np.sqrt(1.0 + y)

# ------------------------------ super-Hubble adiabatic growing mode ------
# series form of Eq. (4.3.45); the closed form cancels catastrophically at
# y ~ 1e-6, so it must not be used to set initial data
def Phi_sh(y):  return 1.0 - y/16.0 + 7.0*y**2/160.0
def dPhi_sh(y): return -1.0/16.0 + 7.0*y/80.0

# ================== approach B : the (Phi, S) system ====================
def damping_S(y):
    return 1.0/(2*(1+y)) + 4.0/(y*(4+3*y))

def rhsB(x, u, kbar, corrected):
    y = np.exp(x)
    Phi, dPhi, S, dS = u
    dC  = -(4.0/3.0)*kbar**2*y**2*Phi/(1+y)          # (4.4.12)
    src = dC - y*S/(1+y)
    ddPhi = ( 2.0/((4+3*y)*y**2)*src                  # (4.4.11)
              - (7 - 1.0/(1+y) + 8.0/(4+3*y))/(2*y)*dPhi
              - Phi/(y*(1+y)*(4+3*y)) )
    ddS   = 2.0*kbar**2/(4+3*y)*src - damping_S(y)*dS   # (4.4.13)
    return y*np.array([dPhi, ddPhi, dS, ddS])         # d/d(ln y) = y d/dy

def runB(kbar, corrected=True, y_i=1e-6, y_f=1e4, dense=False):
    u0 = [Phi_sh(y_i), dPhi_sh(y_i), 0.0, 0.0]        # adiabatic: S_i = S_i' = 0
    s = solve_ivp(rhsB, [np.log(y_i), np.log(y_f)], u0, args=(kbar, corrected),
                  method="DOP853", rtol=1e-9, atol=1e-14, dense_output=dense)
    assert s.success, (kbar, s.message)
    return s

def reconstruct(y, Phi, dPhi, S, kbar):
    """fluid variables from (Phi, S) via Eqs. (4.4.12), (4.4.15), (4.4.16)."""
    dC = -(4.0/3.0)*kbar**2*y**2*Phi/(1+y)
    dN = dC - 2.0*(Phi + y*dPhi)
    dm = (0.75*(1+y)*dN + S)/(1 + 0.75*y)
    dr = ((1+y)*dN - y*S)/(1 + 0.75*y)
    return dC, dN, dm, dr

# ----- the S == 0 truncation: Eq. (4.4.11) with the source from (4.4.12) --
def rhsPhiOnly(x, u, kbar):
    y = np.exp(x)
    Phi, dPhi = u
    ddPhi = ( -(7 - 1.0/(1+y) + 8.0/(4+3*y))/(2*y)*dPhi
              - Phi/(y*(1+y)*(4+3*y))
              - (8.0/3.0)*kbar**2*Phi/((4+3*y)*(1+y)) )
```

```
    return y*np.array([dPhi, ddPhi])

def runPhiOnly(kbar, y_i=1e-6, y_f=1e4):
    s = solve_ivp(rhsPhiOnly, [np.log(y_i), np.log(y_f)],
                  [Phi_sh(y_i), dPhi_sh(y_i)], args=(kbar,),
                  method="DOP853", rtol=1e-9, atol=1e-14)
    assert s.success
    return s

# ============ approach A : the four fluid equations, for validation ======
def rhsA(x, u, kbar):
    y = np.exp(x)
    dm, Vm, dr, Vr, Phi = u
    H = cal_H(y)
    Q = 0.75*(y*Vm + (4.0/3.0)*Vr)/y**2
    dPhi = -Q - H*Phi                                   # (4.4.10)
    return J(y)*np.array([kbar**2*Vm + 3*dPhi,          # (4.4.5)
                          -H*Vm - Phi,                  # (4.4.7)
                          (4.0/3.0)*kbar**2*Vr + 4*dPhi,   # (4.4.6)
                          -Phi - 0.25*dr,               # (4.4.8)
                          dPhi])

def icA(y_i, kbar=0.0):
    H, Ph, Py = cal_H(y_i), Phi_sh(y_i), dPhi_sh(y_i)
    dC = -(4.0/3.0)*kbar**2*y_i**2*Ph/(1+y_i)           # (4.4.12)
    dN = dC - 2*Ph - 2*y_i*Py                           # (4.4.15) in full
    dr = dN*(1+y_i)/(1 + 0.75*y_i)
    th = H*(y_i*Py + Ph)*y_i**2/(1 + 0.75*y_i)
    return np.array([0.75*dr, -th, dr, -th, Ph])

def runA(kbar, y_i=1e-6, y_f=1e4):
    s = solve_ivp(rhsA, [np.log(y_i), np.log(y_f)], icA(y_i, kbar), args=(kbar,),
                  method="DOP853", rtol=1e-9, atol=1e-13)
    assert s.success
    return s

# ------------------------------------------------------ fitting formulae
def T_bbks(q):
    return (np.log(1+2.34*q)/(2.34*q)) * \
           (1+3.89*q+(16.1*q)**2+(5.46*q)**3+(6.71*q)**4)**(-0.25)

def T_eh(x):
    L = np.log(2*np.e + 1.8*x)
    C = 14.2 + 731.0/(1.0 + 62.5*x)
    return L/(L + C*x**2)

# ================================ main ================================
# ================================ main ================================
if __name__ == "__main__":
    kb = np.logspace(-2, 2, 33)
    refB  = runB(1e-2).y[0, -1]
    refA  = runA(1e-2).y[4, -1]
    refBn = runB(1e-2, corrected=False).y[0, -1]
    refP  = runPhiOnly(1e-2).y[0, -1]

    TB  = np.array([runB(k).y[0, -1] for k in kb])/refB
    TA  = np.array([runA(k).y[4, -1] for k in kb])/refA
    TP  = np.array([runPhiOnly(k).y[0, -1] for k in kb])/refP
    q   = 0.0746*kb                     # k_eq in the fit's T_CMB = 2.7 K convention
    Tb, Te = T_bbks(q), T_eh(kb/13.41)

```

```
    print("[1] approach B (corrected) against approach A:")
    print("    max |T_B/T_A - 1| = %.2e" % np.max(np.abs(TB/TA - 1)))
    print("[3] max |T_B/T_EH - 1| = %.3f at kbar = %.2f"
          % (np.max(np.abs(TB/Te-1)), kb[np.argmax(np.abs(TB/Te-1))]))

    print(f"\n{'kbar':>8}{'q':>9}{'T_B':>10}{'T_A':>10}{'T_BBKS':>10}"
          f"{'B/BBKS':>9}{'T_EH':>10}{'T(S=0)':>10}")
    for i in range(0, len(kb), 4):
        print(f"{kb[i]:8.3f}{q[i]:9.4f}{TB[i]:10.5f}{TA[i]:10.5f}{Tb[i]:10.5f}"
              f"{TB[i]/Tb[i]:9.4f}{Te[i]:10.5f}{TP[i]:10.5f}")
    np.savetxt("transfer-phiS.dat", np.c_[kb, q, TB, TA, Tb, Te, TP],
               fmt="%.6e", header="kbar q T_B T_A T_BBKS T_EH T_Bnotes T_S0")

    if "--no-figure" in sys.argv:
        sys.exit()

    import matplotlib; matplotlib.use("Agg")
    import matplotlib.pyplot as plt

    # ---- Figure 1: what the variables actually do -----------------------
    fig, ax = plt.subplots(2, 2, figsize=(9.2, 6.4), sharex=True)
    for kbar, c in [(0.1, 'C0'), (1.0, 'C1'), (10.0, 'C3')]:
        s = runB(kbar, dense=True)
        yy = np.logspace(-4, 4, 1600)
        Phi, dPhi, S, dS = s.sol(np.log(yy))
        dC, dN, dm, dr = reconstruct(yy, Phi, dPhi, S, kbar)
        lab = rf'$\bar k={kbar:g}$'
        ax[0,0].semilogx(yy, Phi, c, lw=1.5, label=lab)
        ax[0,1].loglog(yy, np.abs(dm), c, lw=1.5, label=lab)
        ax[1,0].loglog(yy, np.abs(dr), c, lw=1.5, label=lab)
        ax[1,1].loglog(yy, np.abs(S),  c, lw=1.5, label=lab)
    for a, t in zip(ax.flat, [r'$\Phi$', r'$|\delta_{\rm m}^{\rm N}|$',
                              r'$|\delta_{\rm r}^{\rm N}|$', r'$|S|$']):
        a.axvline(1.0, color='0.6', ls=':', lw=1)
        a.set_ylabel(t); a.grid(alpha=.25, which='both')
    ax[0,0].axhline(0.9, color='0.6', ls='--', lw=1)
    ax[0,0].set_ylim(-0.3, 1.15); ax[0,0].legend(frameon=False, fontsize=9)
    for a in ax[1]: a.set_xlabel(r'$y=a/a_{\rm eq}$')
    ax[0,1].set_ylim(1e-3, 1e4); ax[1,0].set_ylim(1e-4, 1e2)
    ax[1,1].set_ylim(1e-12, 1e4)
    fig.tight_layout(); fig.savefig("phiS-evolution.pdf", bbox_inches='tight')

    # ---- Figure 2: transfer function comparison -------------------------
    fig, ax = plt.subplots(2, 1, figsize=(6.2, 5.8), sharex=True,
                           gridspec_kw=dict(height_ratios=[3, 1], hspace=0.08))
    ax[0].loglog(q, TB,  'k-',  lw=1.9, label=r'$(\Phi,S)$ equations')
    ax[0].loglog(q, TA,  'o',   ms=3.5, mfc='none', color='C0',
                 label='four fluid equations')
    ax[0].loglog(q, Tb,  'r--', lw=1.4, label='BBKS fit')
    ax[0].loglog(q, Te,  '-.',  lw=1.2, color='C2', label='Eisenstein--Hu')
    ax[0].loglog(q, np.abs(TP), '--', lw=1.2, color='0.55',
                 label=r'$S\equiv0$ truncation')
    ax[0].set_ylabel(r'$T(k)$'); ax[0].set_ylim(1e-4, 2)
    ax[0].legend(frameon=False, fontsize=8.5); ax[0].grid(alpha=.25, which='both')
    ax[1].semilogx(q, TB/Tb,  'k-',  lw=1.6)
    ax[1].axhline(1, color='r', ls='--', lw=1)
    ax[1].set_ylabel(r'$T/T_{\rm BBKS}$'); ax[1].set_ylim(0.5, 1.12)
    ax[1].set_xlabel(r'$q=k/(\Omega_{\rm m0}h^2\,{\rm Mpc}^{-1})$')
    ax[1].grid(alpha=.25, which='both')
    fig.savefig("phiS-transfer.pdf", bbox_inches='tight')
    print("\nwrote transfer-phiS.dat, phiS-evolution.pdf, phiS-transfer.pdf")
```

Both systems are non-stiff but strongly oscillatory at large $\bar{k}$, the radiation executing $\sim \bar{k}\eta/\sqrt{3}$ acoustic oscillations, so an explicit high-order Runge–Kutta integrator is appropriate: `DOP853` requires 3.6×10^4 steps for formulation A and 2.9×10^4 for formulation B at $\bar{k}=100$. The choice $x=\ln y$ is material for formulation B; integrated in y the same system becomes stiff and an implicit method is an order of magnitude slower again. Formulation A is preferable in practice, since it retains the Poisson equation as an independent accuracy check, which the (Φ, S) form has already consumed.

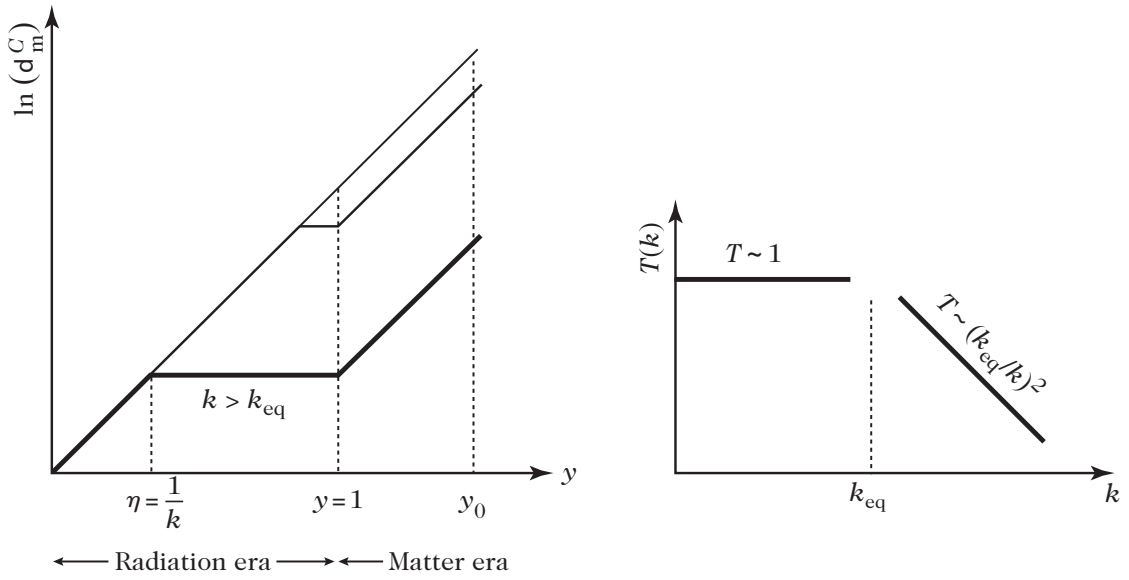


Figure 4.17: (left): Modes that become sub-Hubble during the radiation-dominated era ($k > k_{\rm eq}$) remain nearly constant (neglecting logarithmic growth) from $\eta \sim 1/k$ to $\eta \sim 1/k_{\rm eq}$. (right): The amplitude of the density contrast on these scales suffers from a suppression factor of order $(k_{\rm eq}/k)^2$. For $k > k_{\rm eq}$ one therefore expects the transfer function to scale as $T \propto (k_{\rm eq}/k)^2$, while remaining of order unity for $k < k_{\rm eq}$.

4.4.3 Some refinements

To conclude this chapter, we record the complete set of cosmological perturbation equations, formulated in the fluid approximation, for a Universe comprising photons, neutrinos, cold dark matter, and baryons. The reader is encouraged to implement this coupled system numerically to reproduce the curves presented throughout this chapter and to gain an appreciation for the various physical mechanisms omitted in the introductory presentation.

Dark matter

Cold dark matter is governed by the standard conservation and Euler equations employed throughout this chapter,

$$\delta_{\rm c}^{\rm N\prime} = k^2 V_{\rm c} + 3\Psi', \tag{4.4.44}$$

$$V_{\rm c}' = -\mathcal{H} V_{\rm c} - \Phi, \tag{4.4.45}$$

where we have explicitly retained the distinction between the two gravitational potentials Φ and Ψ.

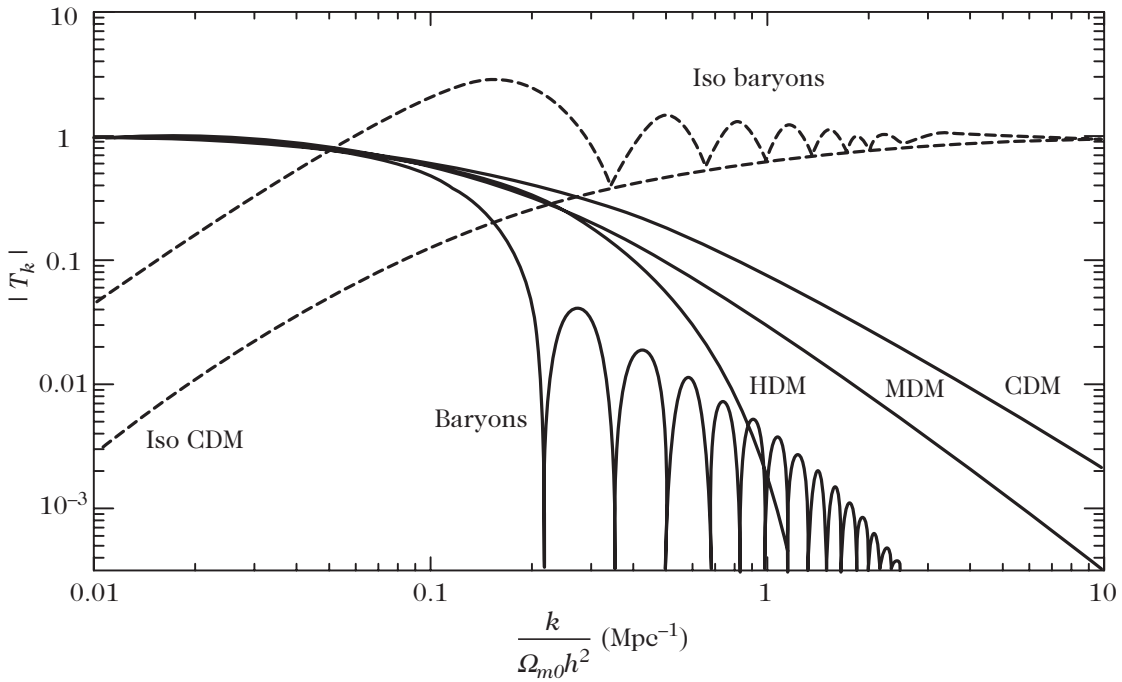


Figure 4.18: Transfer function for different cosmological dark-matter scenarios. The overall shape of this transfer function can be understood directly from the analytical considerations presented above, cf. (4.4.41). From Ref. [8].

Baryons

Electrons and atomic nuclei share the same density perturbation because electromagnetic interactions enforce local charge neutrality to extraordinary precision. Their continuity equation is identical to that of dark matter, but their Euler equation is modified due to Thomson scattering, which couples them dynamically to the photon bath. The momentum exchange between baryons and radiation must therefore be accounted for. As shown in Ref. [41], the interaction force entering the baryon Euler equation is

$$\frac{1}{\rho_\mathrm{b}+P_\mathrm{b}}f_\mathrm{b}=\frac{4}{3}\frac{\rho_\gamma}{\rho_\mathrm{b}}n_\mathrm{e}\sigma_\mathrm{T}(V_\gamma-V_\mathrm{b}), \tag{4.4.46}$$

where σ_T denotes the Thomson scattering cross-section and n_e is the number density of free electrons. The baryon sound speed does not vanish and can be computed by treating the cosmic gas as a monoatomic mixture of mean molecular weight μ (accounting for electrons, protons, and helium nuclei) at temperature T_b,

$$c_\mathrm{b}^2=\frac{k_\mathrm{B}T_\mathrm{b}}{\mu}\left(1-\frac{1}{3}\frac{\mathrm{d}\ln T_\mathrm{b}}{\mathrm{d}\ln a}\right), \tag{4.4.47}$$

neglecting the time variation of μ (in reality, $\dot{\mu}$ is significant only during recombination, when baryons contribute negligibly to the pressure of the combined photon–baryon plasma, making this approximation highly reliable). To determine the temperature evolution equation, one begins with the first law of thermodynamics (see Ref. [42]) $\delta Q=\frac{3}{2}\delta(P_\mathrm{b}/\rho_\mathrm{b})+P_\mathrm{b}\delta(1/\rho_\mathrm{b})$, with the heating rate given by $Q'=4(\rho_\gamma/\rho_\mathrm{b})an_\mathrm{e}\sigma_\mathrm{T}(T_\gamma-T_\mathrm{b})$. One concludes that the evolution of the baryon temperature is governed by

$$T_\mathrm{b}'=-\mathcal{H}T_\mathrm{b}+\frac{8}{3}\frac{\mu}{m_\mathrm{e}}\frac{\rho_\gamma}{\rho_\mathrm{b}}an_\mathrm{e}\sigma_\mathrm{T}(T_\gamma-T_\mathrm{b}). \tag{4.4.48}$$

Since $P_\mathrm{b} \ll \rho_\mathrm{b}$, one can safely neglect w_b and c_b^2 throughout the perturbation equations, with the exception of the acoustic gradient term $c_\mathrm{b}^2 \Delta\delta_\mathrm{b}^\mathrm{N}$. One eventually arrives at

$$\delta_\mathrm{b}^{\mathrm{N}\prime} = k^2 V_\mathrm{b} + 3\Psi', \tag{4.4.49}$$

$$V_\mathrm{b}' = -\mathcal{H} V_\mathrm{b} - \Phi - c_\mathrm{b}^2 \delta_\mathrm{b}^\mathrm{N} + \frac{4}{3}\frac{\rho_\gamma}{\rho_\mathrm{b}} a n_\mathrm{e} \sigma_\mathrm{T} (T_\gamma - T_\mathrm{b}). \tag{4.4.50}$$

Photons

As will be detailed in the following chapter, radiation is rigorously described using the relativistic Boltzmann equation, which expands the photon and neutrino propagation equations into infinite multipole hierarchies. Here, we content ourselves with refining our fluid description by incorporating an anisotropic pressure term, π_γ, alongside the coupling to the baryon fluid:

$$\delta_\gamma^{\mathrm{N}\prime} = \frac{4}{3} k^2 V_\gamma + 4\Psi', \tag{4.4.51}$$

$$V_\gamma' = -\frac{1}{4}\delta_\gamma^\mathrm{N} - \Phi + k^2 \sigma_\gamma + a n_\mathrm{e} \sigma_\mathrm{T} (T_\gamma - T_\mathrm{b}). \tag{4.4.52}$$

Here, the coupling term represents the momentum transfer from the baryon fluid, and we have defined the anisotropic shear stress $\sigma_\gamma \equiv \pi_\gamma/6$. At this stage, we accept without proof the dynamical equation for the photon shear stress,

$$\sigma_\gamma' = -\frac{4}{15} V_\gamma - \frac{9}{5} a n_\mathrm{e} \sigma_\mathrm{T} \sigma_\gamma. \tag{4.4.53}$$

In reality, Eq. (4.4.52) receives contributions from higher multipoles in the Boltzmann hierarchy, so that this fluid representation is merely a low-order truncation.

Neutrinos

As noted above, neutrinos must also be properly described via a Boltzmann equation. They satisfy equations formally identical to those for photons, except for the absence of collisional coupling with the baryon fluid. We thus have

$$\delta_\nu^{\mathrm{N}\prime} = \frac{4}{3} k^2 V_\nu + 4\Psi', \tag{4.4.54}$$

$$V_\nu' = -\frac{1}{4}\delta_\nu^\mathrm{N} - \Phi + k^2 \sigma_\nu, \tag{4.4.55}$$

with $\sigma_\nu \equiv \pi_\nu/6$. As in the photon case, the dynamical evolution equation for the neutrino shear stress is given without proof by

$$\sigma_\nu' = -\frac{4}{15} V_\nu. \tag{4.4.56}$$

The more general scenario involving massive neutrinos is described in detail in Ref. [44].

Einstein equations

The perturbed Einstein equations reduce to

$$\Psi - \Phi = 6\frac{\mathcal{H}^2}{k^2}(\Omega_\gamma \sigma_\gamma + \Omega_\nu \sigma_\nu) \tag{4.4.57}$$

and

$$-k^2 \Psi = \frac{3}{2}\mathcal{H}^2 \left[\sum_i \Omega_i \delta_i^\mathrm{N} - \mathcal{H} \sum_i (1 + w_i) \Omega_i V_i \right]. \tag{4.4.58}$$

Initial conditions

Just as in the preceding sections, initial conditions must be specified for super-Hubble modes deep in the radiation era. For adiabatic initial conditions, one obtains (see, for instance, Ref. [39])

$$\delta_\gamma^{\rm N}(\eta_{\rm i}) = \delta_\nu^{\rm N}(\eta_{\rm i}) = \frac{4}{3}\delta_{\rm b}^{\rm N}(\eta_{\rm i}) = \frac{4}{3}\delta_{\rm c}^{\rm N}(\eta_{\rm i}) \quad \text{and} \quad V_\gamma(\eta_{\rm i}) = V_\nu(\eta_{\rm i}) = V_{\rm b}(\eta_{\rm i}) = V_{\rm c}(\eta_{\rm i}). \tag{4.4.59}$$

As before, on super-Hubble scales we have

$$\delta_\gamma^{\rm N}(\eta_{\rm i}) = -2\Phi(\eta_{\rm i}), \qquad kV_\gamma(\eta_{\rm i}) = -\frac{1}{2}k\eta_{\rm i}\Phi(\eta_{\rm i}), \tag{4.4.60}$$

from which one deduces

$$\sigma_\nu(\eta_{\rm i}) = \frac{1}{15}(k\eta_{\rm i})^2\Phi, \qquad \sigma_\gamma(\eta_{\rm i}) = 0, \tag{4.4.61}$$

and

$$\Psi = \left(1 + \frac{2}{5}R_\nu\right)\Phi, \tag{4.4.62}$$

with the neutrino energy fraction defined by $R_\nu \equiv \rho_\nu/(\rho_\nu + \rho_\gamma)$.

Tight-coupling approximation

Note that prior to decoupling, $V_{\rm b} = V_\gamma$ because Compton scattering couples both species extraordinarily tightly. Following decoupling, the Universe neutralizes and both fluids evolve independently. This coupled system can therefore be solved in the "tight-coupling" approximation, wherein one transitions between the two regimes almost instantaneously (see, for instance, Ref. [43]). Prior to decoupling, one has

$$V_{\rm b} = V_\gamma = V, \qquad V' = -\frac{3\Omega_{\rm b}}{4\Omega_\gamma + 3\Omega_{\rm b}}\mathcal{H}V - \frac{\Omega_\gamma}{4\Omega_\gamma + 3\Omega_{\rm b}}\delta_\gamma^{\rm N} - \Phi. \tag{4.4.63}$$

Conclusions

The system of equations detailed in this section provides a markedly more realistic portrait of cosmological perturbation evolution. While remaining within a fluid description, it accounts for departures from ideal fluid behavior in photons and neutrinos via their anisotropic shear stresses. Consequently, the two gravitational potentials are no longer identical. The collisional coupling between baryons and photons is incorporated, although its rigorous derivation demands a relativistic kinetic approach.

This kinetic description of radiation will be systematically developed in the next chapter, where we will not only derive the unproved equations but also supply their full kinetic versions. In addition, one must describe the detailed ionization and recombination history to determine the time evolution of the free-electron density and move beyond the tight-coupling hypothesis. Solving such a multi-component system yields the matter transfer functions of Fig. 4.18 with excellent precision across a wide range of cosmological models.

4.5 The large-scale structure of the Universe

Up to this point, this chapter has concentrated predominantly on the linear cosmological regime. In Section 4.1.4 we touched briefly upon the weakly non-linear regime. We now turn to what observations reveal regarding the matter power spectrum.

4.5.1 Observing the large-scale structure

Galaxy catalogues

Observational reconstruction of the matter power spectrum is primarily achieved through the assembly of large galaxy redshift surveys. In recent years, two extensive surveys—the Sloan Digital Sky Survey (SDSS) [45] and the 2-degree Field Galaxy Redshift Survey (2dFGRS) [46]—have dramatically expanded both the volume of observed celestial sources and the total sky coverage. SDSS surveys roughly a quarter of the celestial sphere, determining the positions and luminosities of about 100 million celestial sources and measuring distances to more than one million galaxies; 2dFGRS has measured the spatial positions and redshifts of more than 250,000 galaxies. These surveys probe the cosmic web up to redshifts $z \sim 0.3$. A representative slice of the large-scale galaxy distribution mapped by 2dFGRS is displayed in Fig. 4.19.

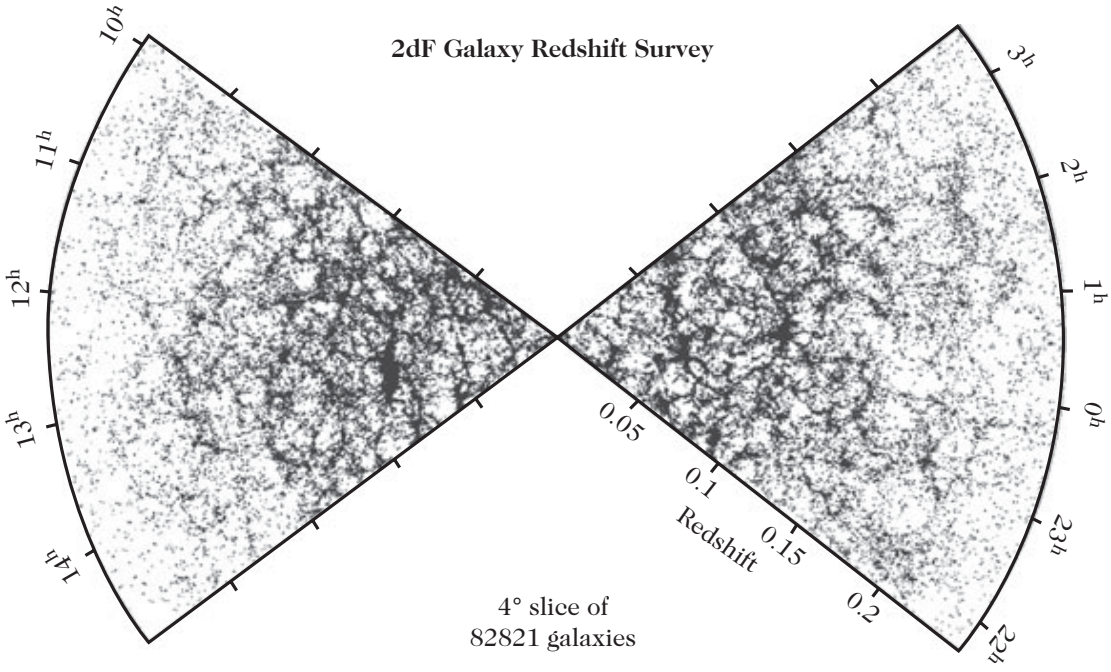


Figure 4.19: The distribution of galaxies across a 4°-wide slice from the 2dFGRS survey catalogue. Such observations of the large-scale structure of the Universe provide the empirical foundation for measurements of the matter power spectrum.

The relationship between these observational surveys and the underlying three-dimensional dark-matter power spectrum is not completely straightforward. Indeed, astronomical observations detect only luminous (baryonic) matter. To extract quantitative conclusions regarding the total matter distribution, one must ensure that luminous matter serves as a faithful tracer of the dark-matter field. This entails that the density contrast of luminous matter is proportional to that of dark matter. The proportionality factor, b, is termed the *bias parameter* [12, 18],

$$\delta_{\text{luminous}} = b\,\delta_{\text{c}}. \tag{4.5.1}$$

The validity of this linear bias relation, the possibility that b depends on spatial scale, and related observational nuances have been investigated extensively through numerical N-body simulations and observations [12]. Another practical difficulty stems from the fact that observations yield the sky position and redshift of astronomical sources with high precision. In redshift space, the inferred radial distance is perturbed by the peculiar velocity of each galaxy along the line of sight. Consequently, on small scales, a spherical gravitationally collapsing structure appears

distorted and elongated along the radial line of sight into an ellipse (the "finger-of-God" effect). Such redshift-space distortions, arising from peculiar motions, must be carefully corrected.

Constraints on the power spectrum

Large galaxy redshift surveys permit the reconstruction of the matter power spectrum over approximately two decades in spatial scale. On larger scales, the linear primordial spectrum can be extracted from the angular anisotropies of the cosmic microwave background (Chapter 6). On smaller scales, complementary observational techniques, such as weak gravitational lensing and the analysis of Lyman-α forest absorption in quasar spectra (which traces the distribution of diffuse intergalactic gas), make it possible to extend power spectrum measurements by nearly two additional orders of magnitude in wavelength. Importantly, the use of weak gravitational lensing offers the distinct advantage of circumventing the galaxy bias parameter altogether, since lensing directly measures the spatial distribution of the gravitational potential.

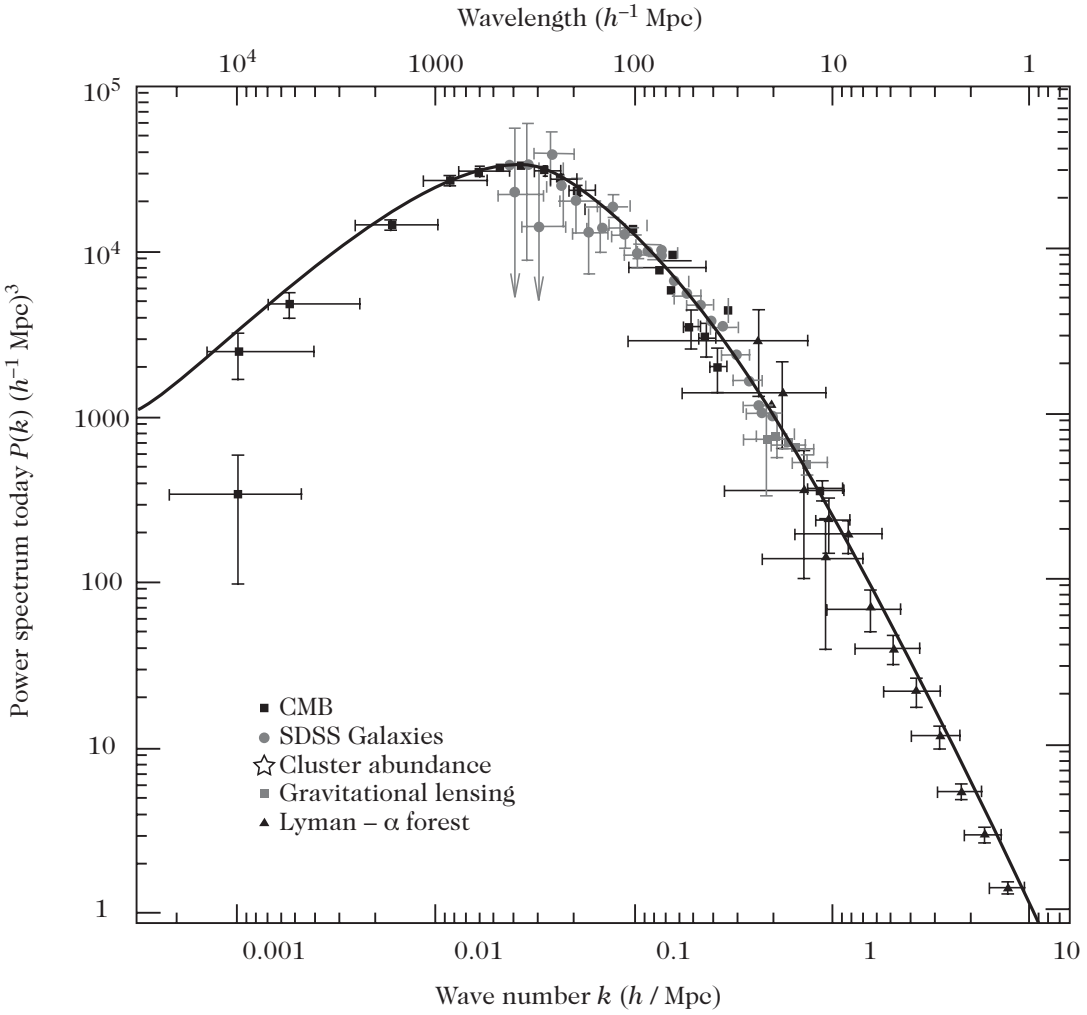


Figure 4.20: The matter power spectrum compared with various observational constraints, ranging from linear scales (cosmic microwave background) to non-linear scales (galaxy surveys, here SDSS), cluster abundances, gravitational lensing, and Lyman-α forests. From (4.4.42) we get the transfer function which can be used $P_m = P_\delta T^2(k) D^2(z)$, where $P_\delta \propto k$.

The code to produce the fig. 4.20

```
import numpy as np
import matplotlib.pyplot as plt

# ============================================================
# Cosmological parameters
# ============================================================
Om0 = 0.315
Or0 = 9.22e-5
Ol0 = 1.0 - Om0 - Or0

h = 0.674
H0 = 100.0 * h / 299792.458     # H0 in Mpc^{-1}

ns = 0.965
As = 2.1e-9
kpivot = 0.05                   # Mpc^{-1}

# ============================================================
# Primordial curvature power spectrum
# ============================================================
def Delta_zeta2(k):
        return As * (k / kpivot)**(ns - 1.0)


# ============================================================
# BBKS transfer function
#
# k_h   : k in h/Mpc
# k     : k in Mpc^{-1}
# ============================================================
def T_BBKS(k_h):
        q = k_h / (Om0 * h)
        numerator = np.log(1.0 + 2.34*q)
        denominator = 2.34*q * (1.0+ 3.89*q+ (16.1*q)**2
                + (5.46*q)**3 + (6.71*q)**4)**0.25

        return numerator / denominator


# ============================================================
# Wavenumber
# ============================================================
k_h = np.logspace(-4, 1, 1000)     # h/Mpc
k = h * k_h                         # Mpc^{-1}


# ============================================================
# Present-day matter transfer from curvature perturbation
# ============================================================
T = T_BBKS(k_h)

M = (2.0 * k**2 / (5.0 * Om0 * H0**2)) * T


# ============================================================
# Present-day dimensionless matter power spectrum
# ============================================================
Delta_m2 = M**2 * Delta_zeta2(k)
```

```
# ============================================================
# Present-day matter power spectrum
#
# P_m(k) = 2 pi^2 / k^3 * Delta_m^2(k)
# ============================================================
Pm = (2.0 * np.pi**2 / k**3) * Delta_m2


# ============================================================
# Plot
# ============================================================
plt.figure(figsize=(8, 5))

plt.loglog(k_h, Pm)

plt.xlabel(r"$k\,[h\,{\rm Mpc}^{-1}]$")
plt.ylabel(r"$P_m(k,0)\,[{\rm Mpc}^3]$")
plt.title("Present-day matter power spectrum")

plt.grid(True, which="both", alpha=0.3)
plt.tight_layout()
plt.show()
```

In conclusion, the cosmological matter power spectrum can be observationally reconstructed across approximately five decades in spatial scale. Figure 4.20 provides an overview of the observational state of the art and the power spectrum reconstructions obtained via these diverse astrophysical methods. It should be emphasized, however, that while impressive at first glance, this composite curve cannot be directly read off from observational raw data. Indeed, at small scales ($k \gtrsim 0.1\,h/\mathrm{Mpc}$), observations probe the power spectrum well into the non-linear regime. The linear reconstruction therefore relies on an analytical or numerical "mapping" between the linear and non-linear regimes (see Fig. 4.21). Furthermore, the inferral of the power spectrum from the Lyman-α forest involves non-trivial modeling of the intergalactic medium and should be interpreted with caution until confirmed by higher-resolution data.

Baryon acoustic oscillations

As established in earlier sections, baryons and radiation are tightly coupled prior to recombination. As a result, the combined photon–baryon plasma undergoes acoustic oscillations: radiation pressure resists gravitational compression, establishing standing sound waves in the photon fluid (see Fig. 4.5). The efficient momentum exchange mediated by Thomson scattering forces the baryon fluid to oscillate in acoustic phase with the photons.

As will be explored in Chapter 6, the imprint of these primordial acoustic sound waves on the cosmic microwave background has been measured with remarkable precision. These acoustic oscillations are defined by a characteristic physical scale corresponding to the sound horizon at decoupling.

Importantly, these acoustic sound waves are also imprinted onto the spatial distribution of galaxies. Because they reside directly within the baryon component, which constitutes only a fraction $\Omega_\mathrm{b}/\Omega_\mathrm{m} \sim 0.15$ of the total matter density, the resulting baryon acoustic oscillations (BAO) in the galaxy distribution are suppressed in amplitude relative to the dark matter background. The BAO peak was observationally detected in Ref. [48] using the two-point spatial correlation function of luminous red galaxies in the SDSS survey, manifesting as a distinct bump at a comoving separation of approximately $100\,h^{-1}\mathrm{Mpc}$ at redshift $z \sim 0.35$.

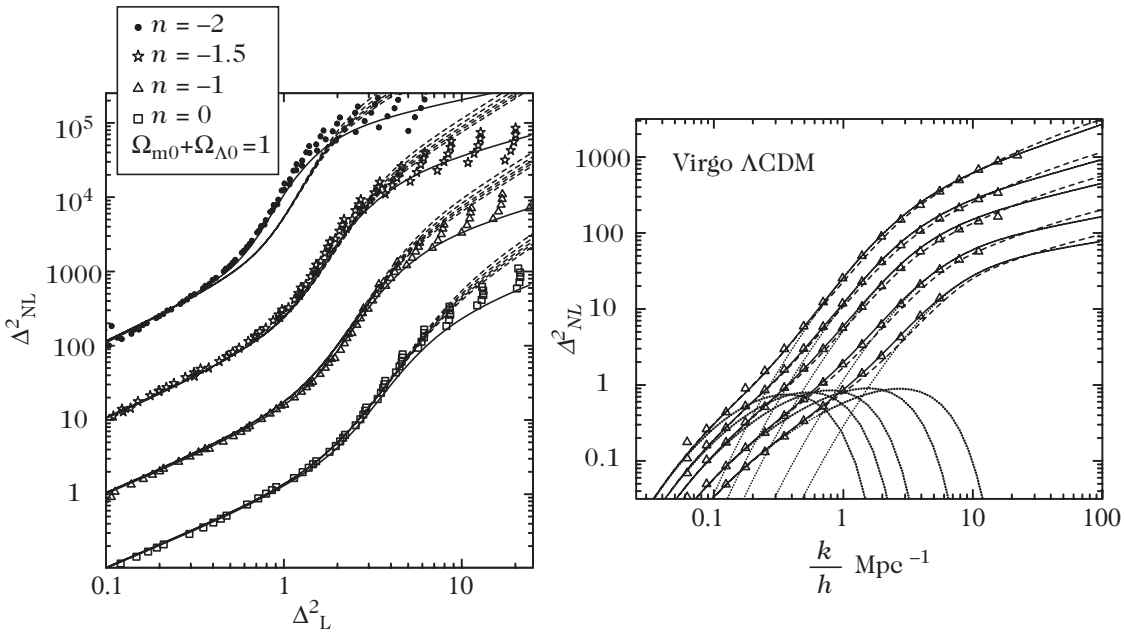


Figure 4.21: (left): Comparison between the analytic mapping developed by Ref. [51] and cosmological N-body numerical simulations [53]. For each spectral index, five epochs are plotted; from bottom to top, $1/(1+z) = 0.6, 0.7, 0.8, 0.9, 1$. (right): The non-linear power spectrum across five epochs, $z = 0, 0.5, 1.0, 2.0, 3.0$ from bottom to top. The solid curves represent the mapping of Ref. [53], whereas the dotted curves correspond to that of Ref. [52]. From Ref. [53].

4.5.2 Structures in the non-linear regime

The dynamical equations governing the matter fluid are fundamentally non-linear, and our detailed analysis thus far has solved them only within the linearized approximation. To bring this chapter to a close, we review the principal physical paradigms and semi-analytical approaches used to investigate the non-linear regime, extending the preliminary treatment given in Section 4.1.4.

Hierarchical model of structure formation

To understand the non-linear evolution of cosmic structure, one must ascertain when a perturbation of a given comoving wavelength enters the non-linear regime ($\delta \sim 1$). If the matter density contrast had remained strictly constant outside the horizon during the radiation era, all Fourier modes would have entered the non-linear regime simultaneously, since perturbation growth during matter domination is scale-independent ($D_+ \propto a$). However, sub-Hubble matter perturbations experience weak logarithmic amplification during radiation domination (see Figs. 4.13 and 4.14). Consequently, modes entering the horizon earlier undergo greater cumulative amplification prior to equality, causing smaller scales to reach the non-linear threshold first. Smaller structures therefore collapse earlier in cosmic history, establishing a *hierarchical* (or "bottom-up") scenario of structure formation.

At the epoch of matter–radiation equality, perturbation amplitudes on typical scales are roughly $\delta \sim 10^{-5}$. Short-wavelength perturbations enter the non-linear regime when $\delta_{\rm c}(k,t) \sim 1$, initiating gravitational collapse into bound objects. If the Universe at that epoch had been smoothed on scales larger than these collapsing regions, it would have appeared remarkably smooth and homogeneous, so that the continuous linear density field remains valid on large scales.

As cosmic time advances, progressively larger scales cross into the non-linear regime. First, small bound halos (protogalaxies) collapse; these subsequently merge hierarchically to form

massive galaxies, groups, and clusters. Today, the characteristic scale of non-linearity is approximately 20–40 Mpc, which exceeds the characteristic virial radius of galaxy clusters (~ 2 Mpc) by only one order of magnitude. Consequently, galaxy clusters are not distributed uniformly in space; instead, they trace an intricate network of filaments and sheets enclosing vast underdense cosmic voids whose typical sizes match the non-linearity scale. On these intermediate scales, the cosmic web exhibits a sponge-like morphology, while recovering statistical homogeneity when averaged over yet larger cosmological volumes.

Spherical collapse model

An insightful analytical model of non-linear gravitational collapse can be formulated by studying an idealized spherical overdense region in an expanding background. By virtue of Gauss's theorem (or Birkhoff's theorem in general relativity), the interior dynamical evolution of a spherically symmetric mass distribution depends solely on the total enclosed mass within that radius, remaining completely decoupled from the exterior universe.

In a matter-dominated Einstein–de Sitter background, the trajectory of a spherical shell of physical radius $r(t)$ enclosing mass M follows the parametric cycloid equations:

$$r = A(1 - \cos\theta), \quad \text{and} \quad t = B(\theta - \sin\theta), \tag{4.5.2}$$

where Newton's second law, $\ddot{r} = -G_{\rm N}M/r^2$, enforces the consistency relation $A^3 = G_{\rm N}MB^2$. The sphere expands, reaches a maximum turnaround radius at $\theta = \pi$, and subsequently collapses toward a point singularity ($r \to 0$) at $\theta = 2\pi$, corresponding to the collapse time $t_{\rm c} \equiv t|_{\theta=2\pi} = 2\pi B$. The mean interior density within the sphere is $\rho_{\rm int} = 3M/(4\pi r^3)$, while the background matter density in the surrounding Einstein–de Sitter spacetime is $\rho_{\rm ext} = (6\pi G_{\rm N} t^2)^{-1}$. The non-linear density contrast is given by the density ratio $\delta = \rho_{\rm int}/\rho_{\rm ext} - 1$.

In the early stages of collapse ($\theta \ll 1$), Taylor expanding the cycloid solutions (4.5.2) yields

$$r(t) \simeq \frac{A}{2}\left(\frac{6t}{B}\right)^{2/3}\left[1 - \frac{1}{20}\left(\frac{6t}{B}\right)^{2/3}\right].$$

At early times, the interior density contrast grows as $\delta \propto t^{2/3}$, exactly reproducing the growing linear solution (4.1.22) in an Einstein–de Sitter Universe.

Comparing the non-linear evolution of the sphere with the extrapolated linear perturbation solution in a flat $\Omega_{\rm m} = 1$ background reveals three critical evolutionary epochs:

1. *Turnaround and decoupling from expansion.* The overdense sphere represents a gravitationally bound system. Its physical radius reaches maximum expansion at $\theta = \pi$ ($t_{\rm ta} = \pi B = t_{\rm c}/2$), where it completely turns around and decouples from the cosmological expansion. At turnaround, the true physical density contrast reaches

$$1 + \delta_{\rm ta} = \frac{\rho_{\rm int}(t_{\rm ta})}{\rho_{\rm ext}(t_{\rm ta})} = \frac{9\pi^2}{16} \simeq 5.55 \quad \Longrightarrow \quad \delta_{\rm ta} \simeq 4.55,$$

   whereas the linear perturbation theory extrapolated to the same epoch predicts only $\delta_{\rm lin}(t_{\rm ta}) = \frac{3}{20}(6\pi)^{2/3} \simeq 1.06$.

2. *Gravitational collapse.* In the absence of pressure support or angular momentum, the idealized spherical shell collapses to a formal singularity at $\theta = 2\pi$ ($t = t_{\rm c}$), where $\delta \to \infty$. Extrapolating linear perturbation theory to the collapse epoch $t_{\rm c}$ yields the celebrated linear collapse threshold:

$$\delta_{\rm c} \equiv \delta_{\rm lin}(t_{\rm c}) = \frac{3}{20}(12\pi)^{2/3} \simeq 1.686 \approx 1.69.$$

3. *Virialization.* In any realistic astrophysical collapse, shell crossing, violent relaxation, and phase mixing convert the bulk kinetic energy of infalling matter into random isotropic velocity dispersion. The sphere does not collapse to zero volume; instead, by the virial theorem ($2K + U = 0$), it settles into a stable, virialized equilibrium configuration at radius $r_{\rm vir} = r_{\rm ta}/2$. Taking into account the cosmic background expansion up to $t_{\rm c}$, the mean overdensity of the resulting virialized dark matter halo relative to the background is

$$\Delta_{\rm vir} \equiv \frac{\rho_{\rm int}(t_{\rm c})}{\rho_{\rm ext}(t_{\rm c})} = 18\pi^2 \simeq 178 \approx 200.$$

This analytical model highlights the failure of linear perturbation theory once $\delta \gtrsim 1$. A rigorous relativistic description of spherically symmetric collapse in an expanding Universe is provided by the exact Lemaître–Tolman–Bondi solutions of Einstein's field equations [49] (as introduced in Chapter 2).

The non-linear power spectrum

To map the non-linear regime with quantitative precision, numerical N-body cosmological simulations are required. These simulations numerically integrate the non-linear Newtonian equations of motion (4.1.15)–(4.1.16) coupled to Poisson's equation (4.1.5) for millions to billions of gravitating particles. Comprehensive reviews of simulation algorithms, boundary conditions, and astrophysical implementations are provided in Ref. [12]; various modern N-body cosmological codes [50] are publicly available.

Extensive analysis of N-body simulations has produced phenomenological scaling relations (known as "mappings") [51–53] that relate the non-linear matter power spectrum directly to the linear power spectrum. These transformations provide powerful tools for predicting non-linear clustering without running costly simulations for every set of cosmological parameters.

A cornerstone concept is the *stable clustering hypothesis* [54], which posits that once high-density regions virialize and decouple from the Hubble flow, they maintain a fixed physical size and constant mean physical density. Consequently, the physical two-point correlation function within virialized objects scales inversely with the mean background density, $\xi(r,t) \propto 1/\bar{\rho}(t) \propto a^3$. For a scale-free primordial power spectrum of the form $P(k) \propto k^n$ in a flat universe, stable clustering implies that the non-linear two-point correlation function assumes a power-law form [54]:

$$\xi_{\rm nl}(r,t) \propto r^{-\gamma}, \quad \text{with} \quad \gamma = \frac{3(3+n)}{5+n}. \tag{4.5.3}$$

Thus, even deep in the non-linear regime, the correlation slope retains memory of the primordial spectral index n.

Building upon this insight, Hamilton et al. [51] proposed that non-linear clustering can be expressed through a universal mapping function $f_{\rm nl}$. The volume-averaged non-linear correlation function,

$$\bar{\xi}_{\rm nl}(x) = \frac{3}{x^3}\int_0^x y^2 \xi_{\rm nl}(y)\,{\rm d}y\,,$$

measured in simulations can be parameterized as a universal function of the linear correlation function $\bar{\xi}_{\rm lin}(\ell)$. The scale mapping between the non-linear radius x and the corresponding linear Lagrangian scale ℓ is obtained from mass conservation within a collapsing spherical halo. Prior to shell crossing, the enclosed mass is conserved:

$$m(<r) = \frac{4\pi}{3} r^3 \rho(<r) = \frac{4\pi}{3}\ell^3 \rho(<\ell) = m_0(<\ell),$$

where $m_0(<\ell)$ is the mass initially enclosed within comoving radius ℓ. Identifying $1+\bar{\xi}_{\rm nl}$ with the local density amplification factor gives the transformation

$$x^3\left[1+\bar{\xi}_{\rm nl}(x,t)\right] = \ell^3,$$

relating the non-linear physical scale x to the linear Lagrangian scale ℓ. In terms of this transformed scale, one postulates $\bar{\xi}_{\rm nl}(x,t) = f_{\rm nl}\left[\bar{\xi}_{\rm lin}(\ell,t)\right]$.

This mapping translates naturally into Fourier space. Defining the dimensionless variance per logarithmic wavenumber interval as

$$\Delta^2(k) = 4\pi k^3 P(k), \tag{4.5.4}$$

the non-linear power spectrum $\Delta^2_{\rm nl}(k_{\rm nl})$ is related to the linear power spectrum $\Delta^2_{\rm lin}(k)$ via

$$\Delta^2_{\rm nl}(k_{\rm nl}) = f_{\rm nl}\left[\Delta^2_{\rm lin}(k)\right], \quad \text{with} \quad k^3 = \frac{k^3_{\rm nl}}{1+\Delta^2_{\rm nl}(k_{\rm nl})}. \tag{4.5.5}$$

Under the stable clustering hypothesis, the universal function must exhibit the asymptotic behavior [51]

$$f_{\rm nl}(x) \propto g^{-3}(\Omega)x^{3/2} \quad (x \gg 1), \qquad g(\Omega) \equiv \frac{D_+(a)}{a}, \tag{4.5.6}$$

where $D_+(a)$ is the linear growth factor (4.1.24). Accurate empirical fitting formulas for $f_{\rm nl}$ have been calibrated against high-resolution N-body simulations across varied ΛCDM cosmologies [51–53], and the physical existence of such a scaling relation has been grounded theoretically in the context of hierarchical collapse [55].

Figure 4.22 displays the morphological evolution of cosmic structures obtained from numerical simulations, while Fig. 4.21 compares this analytical mapping against simulation results and illustrates its deformation of the matter power spectrum. These non-linear mappings prove essential when evaluating weak gravitational lensing by large-scale cosmic structures.

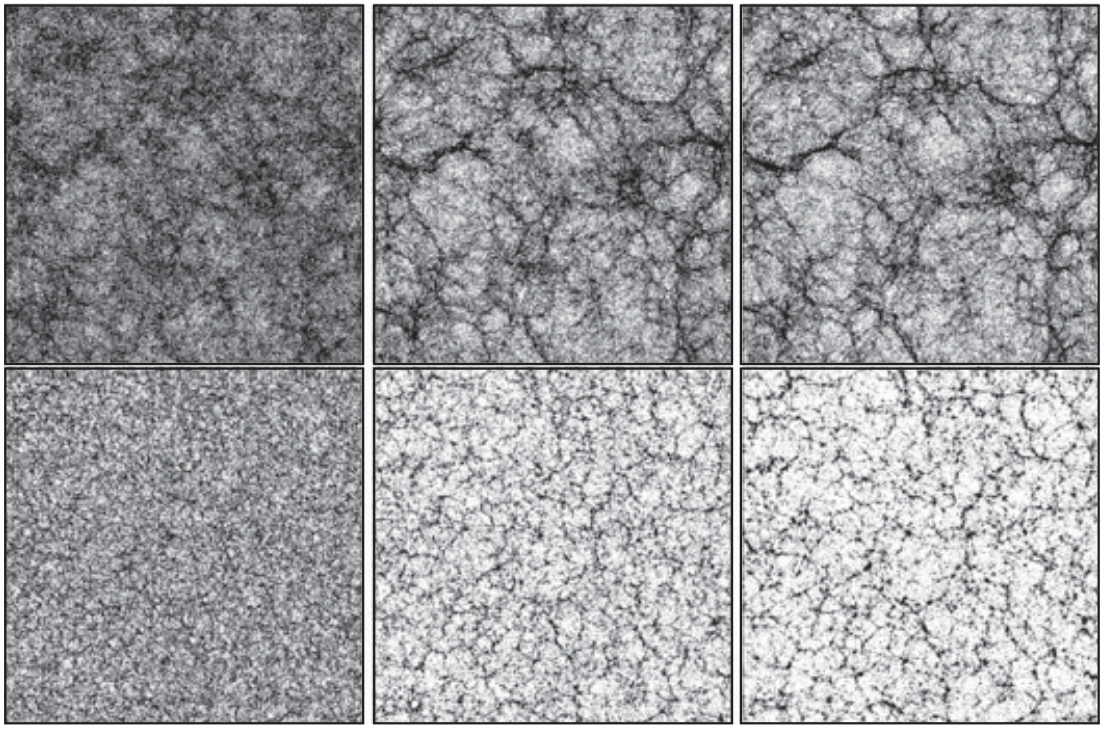


Figure 4.22: Cosmological N-body numerical simulation snapshots (top) for spectral index $n = -2$ at scale factors $1/(1+z) = 0.2$, 0.45, and 0.55 (from right to left) and (bottom) for $n = 0$ at $1/(1+z) = 0.25$, 0.63, and 0.83. The profound impact of the primordial spectral index on the morphology of large-scale structures—with prominent filamentary sheets for $n = -2$ and isolated, fragmented clumps for $n = 0$—is readily discernible. From Ref. [53].

Chapter 5

Initial Conditions for Cosmological Perturbations

The complete linear evolution of cosmological perturbations across all physical wavenumbers k cannot be solved in closed analytical form throughout cosmic history. In modern precision cosmology, the coupled Einstein–Boltzmann hierarchy governing photons, neutrinos, baryons, and cold dark matter is integrated numerically using Boltzmann codes such as CAMB and CLASS.

To initiate numerical integration, these solvers require theoretically rigorous, high-precision **initial conditions** set deep in the radiation-dominated epoch ($z \sim 10^7$–10^8), when all observationally relevant scales remain far outside the Hubble horizon ($k\eta \ll 1$) and the matter density is a tiny fraction of the radiation density ($\eta \ll \eta_{\rm eq}$).

The primary objective of this chapter is to provide a complete, mathematically explicit transcription and derivation of cosmological initial conditions formulated in terms of conformal time power series expansions, following the foundational treatment of Kurki-Suonio, Kodama & Sasaki, and Lewis. We establish the exact background dynamics of a radiation-plus-matter universe, derive the coupled Bardeen and Kodama–Sasaki perturbation equations, construct the power-series asymptotic expansions for both adiabatic and isocurvature modes, perform the gauge transformation to synchronous gauge, and extend the formalism to the multi-fluid real Universe.

5.1 Background Dynamics of a Radiation + Matter Universe

We consider a flat Friedmann–Lemaître–Robertson–Walker (FLRW) universe containing two non-interacting cosmic fluids: relativistic radiation (photons and massless neutrinos) and pressureless non-relativistic matter (cold dark matter and baryons). The background energy densities evolve with the cosmological scale factor $a(\eta)$ according to:

$$\bar{\rho}(\eta) = \bar{\rho}_{\rm r}(\eta) + \bar{\rho}_{\rm m}(\eta), \qquad \bar{\rho}_{\rm r}(a) = \bar{\rho}_{\rm r0}\, a^{-4}, \qquad \bar{\rho}_{\rm m}(a) = \bar{\rho}_{\rm m0}\, a^{-3}, \tag{5.1.1}$$

where $a_0 = 1$ today. The individual component equations of state and sound speeds are:

$$w_{\rm m} = c_{\rm m}^2 = 0, \qquad w_{\rm r} = c_{\rm r}^2 = \frac{1}{3}, \qquad \bar{P}_{\rm m} = 0, \qquad \bar{P}_{\rm r} = \frac{1}{3}\bar{\rho}_{\rm r}. \tag{5.1.2}$$

5.1.1 The Density Ratio Variable y

It is exceptionally convenient to parameterize time using the dimensionless scale factor normalized to matter-radiation equality:

$$y \equiv \frac{a(\eta)}{a_{\rm eq}} = \frac{\bar{\rho}_{\rm m}(\eta)}{\bar{\rho}_{\rm r}(\eta)}, \tag{5.1.3}$$

where the scale factor at equality is given by the present-day density parameters:

$$a_{\rm eq} = \frac{\Omega_{\rm r}}{\Omega_{\rm m}} = \frac{1}{1+z_{\rm eq}}. \tag{5.1.4}$$

In terms of y, the background energy and enthalpy fractions evaluate to:

$$\frac{\bar\rho_{\rm r}}{\bar\rho} = \frac{1}{1+y}, \qquad \frac{\bar\rho_{\rm m}}{\bar\rho} = \frac{y}{1+y}, \qquad \frac{\bar\rho_{\rm r}+\bar P_{\rm r}}{\bar\rho+\bar P} = \frac{4}{4+3y}, \qquad \frac{\bar\rho_{\rm m}+\bar P_{\rm m}}{\bar\rho+\bar P} = \frac{3y}{4+3y}. \tag{5.1.5}$$

The effective equation of state $w(y) \equiv \bar P/\bar\rho$ and the adiabatic sound speed $c_{\rm s}^2(y) \equiv \bar P'/\bar\rho'$ of the combined fluid are:

$$w(y) = \frac{1}{3(1+y)}, \qquad 1+w(y) = \frac{4+3y}{3(1+y)}, \tag{5.1.6}$$

$$c_{\rm s}^2(y) = \frac{\frac{4}{3}\bar\rho_{\rm r}}{4\bar\rho_{\rm r}+3\bar\rho_{\rm m}} = \frac{4}{3(4+3y)}, \qquad 1-3c_{\rm s}^2(y) = \frac{3y}{4+3y}. \tag{5.1.7}$$

5.1.2 Exact Conformal Time Integration

Exact Solution of the Two-Fluid Friedmann Equation

The flat Friedmann equation in terms of conformal time η ($\mathrm{d}\eta = \mathrm{d}t/a$) reads:

$$\mathcal{H}^2 \equiv \left(\frac{a'}{a}\right)^2 = \frac{8\pi G_{\rm N}}{3}a^2\bar\rho = \frac{8\pi G_{\rm N}}{3}a^2(\bar\rho_{\rm r}+\bar\rho_{\rm m}) = \frac{8\pi G_{\rm N}}{3}\bar\rho_{\rm r0}a^{-2}(1+y). \tag{1}$$

Since $a = a_{\rm eq}y$, we have $a' = a_{\rm eq}y'$. Substituting this gives:

$$\left(\frac{y'}{y}\right)^2 = \frac{8\pi G_{\rm N}\bar\rho_{\rm r0}}{3a_{\rm eq}^4}\frac{1+y}{y^2} \implies (y')^2 = \frac{8\pi G_{\rm N}\bar\rho_{\rm r0}}{3a_{\rm eq}^4}(1+y). \tag{2}$$

Defining the constant C:

$$C \equiv \left(\frac{2\pi G_{\rm N}\bar\rho_{\rm r0}}{3a_{\rm eq}^4}\right)^{1/2} = \frac{H_0\sqrt{\Omega_{\rm r}}}{2a_{\rm eq}} = \frac{H_0\Omega_{\rm m}}{2\sqrt{\Omega_{\rm r}}}, \tag{3}$$

the differential equation becomes:

$$\frac{\mathrm{d}y}{\mathrm{d}\eta} = 2C\sqrt{1+y}. \tag{4}$$

Separating variables and integrating with the initial condition $y(0) = 0$ (the Big Bang singularity at $\eta = 0$):

$$\int_0^y \frac{\mathrm{d}\tilde y}{\sqrt{1+\tilde y}} = 2C\int_0^\eta \mathrm{d}\tilde\eta \implies 2\left(\sqrt{1+y}-1\right) = 2C\eta \implies \sqrt{1+y} = 1+C\eta. \tag{5}$$

Squaring both sides yields the exact analytical solution for $y(\eta)$:

$$y(\eta) = 2C\eta + C^2\eta^2. \tag{6}$$

Since $a(\eta) = a_{\rm eq}y(\eta)$ and $a_{\rm eq}C = \frac{1}{2}H_0\sqrt{\Omega_{\rm r}}$, the scale factor is:

$$a(\eta) = \sqrt{\Omega_{\rm r}}H_0\eta + \frac{1}{4}\Omega_{\rm m}H_0^2\eta^2. \tag{7}$$

At matter-radiation equality ($y = 1$), Eq. (6) requires:

$$C^2\eta_{\rm eq}^2 + 2C\eta_{\rm eq} - 1 = 0 \implies C\eta_{\rm eq} = \sqrt{2} - 1. \tag{5.1.8}$$

We define the characteristic timescale η_3 as the epoch when $y = 3$ (where matter density is three times radiation density, $\bar{\rho}_{\rm m} = 3\bar{\rho}_{\rm r}$):

$$C^2\eta_3^2 + 2C\eta_3 - 3 = 0 \implies C\eta_3 = 1 \implies \eta_3 \equiv \frac{1}{C} = (\sqrt{2}+1)\eta_{\rm eq} = \frac{2\sqrt{\Omega_{\rm r}}}{H_0\Omega_{\rm m}}. \tag{5.1.9}$$

Introducing the dimensionless conformal time variable x:

$$x \equiv \frac{\eta}{\eta_3} = C\eta = \frac{\omega\eta}{2}, \qquad \omega \equiv \frac{H_0\Omega_{\rm m}}{\sqrt{\Omega_{\rm r}}} = \frac{2}{\eta_3} = 2C, \tag{5.1.10}$$

the expansion parameter y simplifies to the elegant algebraic relation:

$$\boxed{y(x) = 2x + x^2.} \tag{5.1.11}$$

The conformal Hubble expansion parameter $\mathcal{H}(\eta) \equiv a'/a = y'/y$ evaluates to:

$$\mathcal{H}(x) = \frac{1}{\eta_3}\frac{2+2x}{2x+x^2} = \frac{1}{\eta_3}\frac{1+x}{x\left(1+\frac{1}{2}x\right)} = \frac{\omega}{2}\frac{1+x}{x\left(1+\frac{1}{2}x\right)}. \tag{5.1.12}$$

In terms of y, the conformal Hubble parameter is:

$$\mathcal{H}(y) = \frac{2}{\eta_3}\frac{\sqrt{1+y}}{y} = \mathcal{H}_{\rm eq}\frac{\sqrt{1+y}}{\sqrt{2}\,y}, \qquad \mathcal{H}_{\rm eq} \equiv \mathcal{H}(\eta_{\rm eq}) = \sqrt{2}\,\omega = \frac{2\sqrt{2}}{\eta_3}. \tag{5.1.13}$$

5.2 Two-Fluid Cosmological Perturbations

In an inhomogeneous universe, each fluid component possesses its own fractional energy density contrast $\delta_i \equiv \delta\rho_i/\bar{\rho}_i$ and peculiar velocity potential v_i ($u_i^j = a^{-1}\partial^j v_i$).

5.2.1 Total and Relative Perturbation Variables

The total energy density perturbation δ and total momentum velocity v of the combined matter-radiation system are given by:

$$\delta = \frac{\bar{\rho}_{\rm r}\delta_{\rm r} + \bar{\rho}_{\rm m}\delta_{\rm m}}{\bar{\rho}} = \frac{1}{1+y}\delta_{\rm r} + \frac{y}{1+y}\delta_{\rm m}, \tag{5.2.1}$$

$$v = \frac{(\bar{\rho}_{\rm r} + \bar{P}_{\rm r})v_{\rm r} + (\bar{\rho}_{\rm m} + \bar{P}_{\rm m})v_{\rm m}}{\bar{\rho} + \bar{P}} = \frac{4}{4+3y}v_{\rm r} + \frac{3y}{4+3y}v_{\rm m}. \tag{5.2.2}$$

The gauge-invariant **relative entropy (isocurvature) perturbation** between matter and radiation is defined as:

$$\mathcal{S} \equiv \mathcal{S}_{\rm mr} = \delta_{\rm m} - \frac{3}{4}\delta_{\rm r}. \tag{5.2.3}$$

Under an arbitrary infinitesimal coordinate transformation $\eta \to \eta + T$, the component density contrasts transform as $\delta_i \to \delta_i - \frac{\bar{\rho}_i'}{\bar{\rho}_i}T$. Since $\bar{\rho}_{\rm m}'/\bar{\rho}_{\rm m} = -3\mathcal{H}$ and $\bar{\rho}_{\rm r}'/\bar{\rho}_{\rm r} = -4\mathcal{H}$, we have:

$$\delta_{\rm m} \to \delta_{\rm m} + 3\mathcal{H}T, \qquad \delta_{\rm r} \to \delta_{\rm r} + 4\mathcal{H}T \implies \mathcal{S} \to (\delta_{\rm m} + 3\mathcal{H}T) - \frac{3}{4}(\delta_{\rm r} + 4\mathcal{H}T) = \mathcal{S}. \tag{5.2.4}$$

Thus, $\mathcal{S}$ is strictly gauge invariant and takes identical values in comoving, Newtonian, and synchronous gauges:

$$\mathcal{S} = \delta_{\rm m}^{\rm C} - \frac{3}{4}\delta_{\rm r}^{\rm C} = \delta_{\rm m}^{\rm N} - \frac{3}{4}\delta_{\rm r}^{\rm N} = \delta_{\rm m}^{\rm syn} - \frac{3}{4}\delta_{\rm r}^{\rm syn}. \tag{5.2.5}$$

Inverting Eqs. (5.2.1), (5.2.2), and (5.2.3) expresses the individual component perturbations in terms of total and relative quantities:

$$\delta_{\rm m} = \frac{3+3y}{4+3y}\delta + \frac{4}{4+3y}\mathcal{S}, \qquad \delta_{\rm r} = \frac{4+4y}{4+3y}\delta - \frac{4y}{4+3y}\mathcal{S}, \tag{5.2.6}$$

$$v_{\rm m} = v + \frac{4}{4+3y}(v_{\rm m} - v_{\rm r}), \qquad v_{\rm r} = v - \frac{3y}{4+3y}(v_{\rm m} - v_{\rm r}). \tag{5.2.7}$$

5.2.2 The Coupled Bardeen and Kodama–Sasaki Equations

We formulate the perturbation evolution in the **comoving gauge** ($v - B = 0$), where the total comoving density contrast is $\delta^{\rm C}$. In Chapter 4, we established the exact relativistic Bardeen equation for the comoving density contrast $\delta^{\rm C}$ in the presence of non-adiabatic pressure perturbations $\delta P_{\rm nad}$:

$$\delta^{\rm C\prime\prime} + \mathcal{H}(1 - 6w + 3c_{\rm s}^2)\delta^{\rm C\prime} - \frac{3}{2}\mathcal{H}^2(1 + 8w - 6c_{\rm s}^2 - 3w^2)\delta^{\rm C} = -c_{\rm s}^2 k^2 \delta^{\rm C} - \frac{k^2}{\bar{\rho}}\delta P_{\rm nad}. \tag{5.2.8}$$

For a two-fluid mixture of matter and radiation, the non-adiabatic pressure perturbation is sourced entirely by the relative entropy perturbation $\mathcal{S}$:

$$\delta P_{\rm nad} = \delta P - c_{\rm s}^2 \delta\rho = \frac{1}{3}\bar{\rho}_{\rm r}\delta_{\rm r} - c_{\rm s}^2(\bar{\rho}_{\rm r}\delta_{\rm r} + \bar{\rho}_{\rm m}\delta_{\rm m}) = \bar{\rho}c_{\rm s}^2\frac{y}{4+3y}\mathcal{S} = \frac{4}{3}\frac{y}{(4+3y)^2}\bar{\rho}\,\mathcal{S}. \tag{5.2.9}$$

Dividing the entire equation by $\mathcal{H}^2$, the Bardeen equation takes the standard form:

$$\mathcal{H}^{-2}\delta^{\rm C\prime\prime} + (1 - 6w + 3c_{\rm s}^2)\mathcal{H}^{-1}\delta^{\rm C\prime} - \frac{3}{2}(1 + 8w - 6c_{\rm s}^2 - 3w^2)\delta^{\rm C} = -c_{\rm s}^2\left(\frac{k}{\mathcal{H}}\right)^2\left[\delta^{\rm C} + \frac{y}{1+y}\mathcal{S}\right]. \tag{5.2.10}$$

To obtain the dynamical equation for the entropy perturbation $\mathcal{S}$, we differentiate Eq. (5.2.3) with respect to conformal time. The individual continuity equations for non-interacting matter and radiation give:

$$\delta_{\rm m}' = -kv_{\rm m} + 3\psi', \qquad \delta_{\rm r}' = -\frac{4}{3}kv_{\rm r} + 4\psi' \implies \mathcal{S}' = -k(v_{\rm m} - v_{\rm r}). \tag{5.2.11}$$

Differentiating once more and substituting the Euler momentum equations for matter ($v_{\rm m}' + \mathcal{H}v_{\rm m} = k\Phi$) and radiation ($v_{\rm r}' = k\Phi + \frac{1}{4}k\delta_{\rm r}^{\rm N}$) yields the Kodama–Sasaki equation:

$$\mathcal{H}^{-2}\mathcal{S}'' + 3c_{\rm s}^2\mathcal{H}^{-1}\mathcal{S}' = \left(\frac{k}{\mathcal{H}}\right)^2\frac{1+y}{4+3y}\left[\delta^{\rm C} - \frac{y}{1+y}\mathcal{S}\right]. \tag{5.2.12}$$

Equations (5.2.10) and (5.2.12) constitute a closed linear system of two coupled second-order ordinary differential equations for $\delta^{\rm C}(\eta)$ and $\mathcal{S}(\eta)$.

5.2.3 Relation to Metric and Gauge-Invariant Variables

Once $\delta^{\rm C}$ and $\mathcal{S}$ are determined, all other cosmological perturbation fields are obtained algebraically:

$$\Phi = -\frac{3}{2}\left(\frac{\mathcal{H}}{k}\right)^2\delta^{\rm C}, \tag{5.2.13}$$

$$\Psi = \Phi \qquad \text{(for zero anisotropic stress, } \pi_{ij} = 0\text{)}, \tag{5.2.14}$$

$$v^{\mathrm{N}} = \frac{2}{3(1+w)}\frac{k}{\mathcal{H}}\left[\mathcal{H}^{-1}\Phi' + \Phi\right] = \frac{2(1+y)}{4+3y}\frac{k}{\mathcal{H}}\left[\mathcal{H}^{-1}\Phi' + \Phi\right], \tag{5.2.15}$$

$$\delta^{\mathrm{N}} = \delta^{\mathrm{C}} - 3(1+w)\frac{\mathcal{H}}{k}v^{\mathrm{N}} = \delta^{\mathrm{C}} - \frac{4+3y}{1+y}\frac{\mathcal{H}}{k}v^{\mathrm{N}}, \tag{5.2.16}$$

$$\mathcal{R} = -\Phi - \frac{2}{3(1+w)}\left[\mathcal{H}^{-1}\Phi' + \Phi\right] = -\Phi - \frac{\mathcal{H}}{k}v^{\mathrm{N}}. \tag{5.2.17}$$

5.3 Asymptotic Expansions in Conformal Time

Numerical codes like CAMB use conformal time η as the integration coordinate and begin deep in the radiation era at $x \equiv \eta/\eta_3 \ll 1$ on superhorizon scales $k\eta \ll 1$. To achieve high precision, initial values cannot be approximated as mere constants; they must be supplied as truncated power series in x.

5.3.1 Rescaling the System to the Dimensionless Variable x

Using the exact relations $y = 2x + x^2$, $\mathrm{d}/\mathrm{d}\eta = \eta_3^{-1}\mathrm{d}/\mathrm{d}x$, and $\mathcal{H}(x) = \eta_3^{-1}\frac{1+x}{x(1+\frac{1}{2}x)}$, the coupled system of Eqs. (5.2.10) and (5.2.12) is rewritten purely in terms of x:

$$\begin{aligned}
&\eta_3^2 x^2\delta^{\mathrm{C}\prime\prime}(x) + \frac{1+\frac{6}{5}x+\frac{3}{5}x^2}{1+\frac{5}{2}x+\frac{9}{4}x^2+\frac{3}{4}x^3}\frac{5}{2}x\,\eta_3\delta^{\mathrm{C}\prime}(x) \\
&\quad - \frac{1+\frac{19}{4}x+\frac{79}{8}x^2+12x^3+\frac{69}{8}x^4+\frac{27}{8}x^5+\frac{9}{16}x^6}{\left(1+\frac{3}{2}x+\frac{3}{2}x^2+\frac{9}{8}x^3+\frac{17}{16}x^4+\frac{69}{2}x^5+\frac{79}{16}x^6\right)}2\,\delta^{\mathrm{C}}(x) \\
&\quad = -\frac{1}{3}(k\eta_3)^2\frac{1+\frac{1}{2}x}{1+\frac{3}{2}x+\frac{3}{4}x^2}x^2\delta^{\mathrm{C}}(x) + \frac{2}{3}(k\eta_3)^2\frac{1+\frac{1}{2}x}{1+\frac{7}{2}x+\frac{19}{4}x^2+3x^3+\frac{3}{4}x^4}x^3\mathcal{S}(x),
\end{aligned} \tag{5.3.1}$$

$$\begin{aligned}
&\eta_3^2 x\,\mathcal{S}''(x) + \frac{1+x}{1+2x+\frac{3}{2}x^2+\frac{3}{8}x^3}\eta_3\mathcal{S}'(x) \\
&\quad = \frac{1}{4}(k\eta_3)^2\frac{1+2x+x^2}{1+\frac{3}{2}x+\frac{3}{4}x^2}x\,\delta^{\mathrm{C}}(x) - \frac{1}{2}(k\eta_3)^2\frac{1+\frac{1}{2}x}{1+\frac{3}{2}x+\frac{3}{4}x^2}x^2\mathcal{S}(x).
\end{aligned} \tag{5.3.2}$$

Expanding each rational background coefficient as a Taylor series in $x \ll 1$ up to fourth order produces the simplified differential system:

$$\begin{aligned}
&x^2\eta_3^2\delta^{\mathrm{C}\prime\prime} + \frac{5}{2}\left(1-\frac{13}{10}x+\frac{8}{5}x^2\right)x\eta_3\delta^{\mathrm{C}\prime} - 2\left(1+\frac{1}{4}x+\frac{1}{4}x^2+\frac{1}{8}x^3-\frac{37}{32}x^4\right)\delta^{\mathrm{C}} \\
&\quad \simeq -\frac{1}{3}(k\eta_3)^2\left(1-\frac{3}{2}x+\frac{3}{2}x^2\right)x^2\delta^{\mathrm{C}} + \frac{2}{3}(k\eta_3)^2(1-3x)x^3\mathcal{S},
\end{aligned} \tag{5.3.3}$$

$$\begin{aligned}
&x\eta_3^2\mathcal{S}'' + \left(1-x+\frac{1}{2}x^2+\frac{1}{8}x^3-\frac{5}{8}x^4\right)\eta_3\mathcal{S}' \\
&\quad \simeq \frac{1}{4}(k\eta_3)^2\left(1+\frac{1}{2}x-\frac{1}{2}x^2+\frac{3}{8}x^3\right)x\delta^{\mathrm{C}} - \frac{1}{2}(k\eta_3)^2\left(1-x+\frac{3}{4}x^2\right)x^2\mathcal{S}.
\end{aligned} \tag{5.3.4}$$

5.3.2 Power Series Recurrence Solution

We seek non-decaying regular solutions as power series in x:

$$\delta^{\mathrm{C}}(x) = A + Bx + Cx^2 + Dx^3 + Jx^4 + \mathcal{O}(x^5), \tag{5.3.5}$$

$$\mathcal{S}(x) = E + Fx + Gx^2 + Hx^3 + Kx^4 + Lx^5 + \mathcal{O}(x^6). \tag{5.3.6}$$

Negative powers of x are excluded because they correspond to decaying modes that have attenuated to zero long before the initial epoch.

Derivation: Order-by-Order Solution of Series Coefficients
Substituting the series expansions (5.3.5) and (5.3.6) into the differential equations (5.3.3) and (5.3.4) and collecting matching powers of x:

1. **Order x^0:**
   - In Eq. (5.3.3): $-2A = 0 \implies A = 0$.
   - In Eq. (5.3.4): $1 \cdot F = 0 \implies F = 0$.
2. **Order x^1:**
   - In Eq. (5.3.3): $\left(1 \cdot 0 + \frac{5}{2} \cdot 1 - 2\right) B - \frac{1}{2}A = \frac{1}{2}B = 0 \implies B = 0$.
   - In Eq. (5.3.4): $(2 \cdot 1 + 1 \cdot 2)G - F - \frac{1}{4}(k\eta_3)^2 A = 4G = 0 \implies G = 0$.
3. **Order x^2:**
   - In Eq. (5.3.3): $\left(2 \cdot 1 + \frac{5}{2} \cdot 2 - 2\right) C = 5C$, which leaves $C \equiv C_{\boldsymbol{k}}$ completely unconstrained as an arbitrary integration constant!
   - In Eq. (5.3.4): $(3 \cdot 2 + 1 \cdot 3)H - 2G + \frac{1}{2}F + \frac{1}{2}(k\eta_3)^2 E = 9H + \frac{1}{2}(k\eta_3)^2 E = 0 \implies H = -\frac{1}{18}(k\eta_3)^2 E$.
4. **Order x^3:**
   - In Eq. (5.3.3): $\left(3 \cdot 2 + \frac{5}{2} \cdot 3 - 2\right) D - \frac{13}{4}(2C) - \frac{1}{2}C = \frac{23}{2}D - 7C = \frac{2}{3}(k\eta_3)^2 E \implies D = -\frac{9}{8}C + \frac{1}{6}(k\eta_3)^2 E$.
   - In Eq. (5.3.4): $(4 \cdot 3 + 1 \cdot 4)K - 3H + G + \cdots = 16K - \frac{1}{4}(k\eta_3)^2 C - \frac{1}{3}(k\eta_3)^2 E = 0 \implies K = \frac{1}{64}(k\eta_3)^2 C + \frac{1}{48}(k\eta_3)^2 E$.
5. **Order x^4 and x^5:** Solving the next order algebraic relations yields:

$$J = \frac{119}{80}C - \frac{1}{30}(k\eta_3)^2 C - \frac{19}{60}(k\eta_3)^2 E, \tag{1}$$

$$L = -\frac{3}{800}(k\eta_3)^2 C - \frac{1}{120}(k\eta_3)^2 E + \frac{1}{600}(k\eta_3)^4 E. \tag{2}$$

Thus, the general regular solution is parameterized by exactly **two independent integration constants**, $C_{\boldsymbol{k}}$ and $E_{\boldsymbol{k}}$:

$$\begin{aligned}
\delta^{\mathrm{C}}_{\boldsymbol{k}}(x) &= C_{\boldsymbol{k}}\left[x^2 - \frac{9}{8}x^3 + \left(\frac{119}{80} - \frac{1}{30}(k\eta_3)^2\right)x^4\right] + E_{\boldsymbol{k}}\left[\frac{1}{6}(k\eta_3)^2 x^3 - \frac{19}{60}(k\eta_3)^2 x^4\right] + \mathcal{O}(x^5), \\
\mathcal{S}_{\boldsymbol{k}}(x) &= E_{\boldsymbol{k}}\left[1 - \frac{1}{18}(k\eta_3)^2 x^3 + \frac{1}{48}(k\eta_3)^2 x^4 + \left(-\frac{1}{120}(k\eta_3)^2 + \frac{1}{600}(k\eta_3)^4\right)x^5\right] \\
&\quad + C_{\boldsymbol{k}}\left[\frac{1}{64}(k\eta_3)^2 x^4 - \frac{3}{800}(k\eta_3)^2 x^5\right] + \mathcal{O}(x^6).
\end{aligned} \tag{5.3.7}$$

5.4 The Physical Initial Modes

The two integration constants in Eq. (5.3.7) define the two fundamental physical modes of primordial cosmological perturbations:

1. **The Growing Adiabatic Mode ($E_{\boldsymbol{k}} = 0, C_{\boldsymbol{k}} \neq 0$):** Characterized by $\delta^{\mathrm{C}} \sim \mathcal{O}(x^2)$ and an initially vanishing entropy perturbation $\mathcal{S} \sim \mathcal{O}(x^4) \to 0$ as $x \to 0$.
2. **The Growing Isocurvature Mode ($C_{\boldsymbol{k}} = 0, E_{\boldsymbol{k}} \neq 0$):** Characterized by a non-zero primordial entropy fluctuation $\mathcal{S} = E_{\boldsymbol{k}} = \text{const} \sim \mathcal{O}(1)$ and an initially vanishing comoving density perturbation $\delta^{\mathrm{C}} \sim \mathcal{O}(x^3) \to 0$ as $x \to 0$.

5.4.1 The Growing Adiabatic Mode

For the adiabatic mode, the primordial comoving curvature perturbation $\mathcal{R}_{\boldsymbol{k}}$ is related to the series amplitude $C_{\boldsymbol{k}}$ in the early radiation era ($w = 1/3$) by:

$$\begin{aligned}\mathcal{R}_{\boldsymbol{k}} = -\frac{3}{2}\Phi_{\boldsymbol{k}} = \frac{9}{4}\left(\frac{\mathcal{H}}{k}\right)^2 \delta^{\mathrm{C}}_{\boldsymbol{k}} = \frac{9}{4}\left(\frac{1}{k\eta}\right)^2 C_{\boldsymbol{k}} \left(\frac{\omega\eta}{2}\right)^2 = \frac{9}{16}\frac{\omega^2}{k^2} C_{\boldsymbol{k}} \\ \implies C_{\boldsymbol{k}} = \frac{16}{9}\frac{k^2}{\omega^2}\mathcal{R}_{\boldsymbol{k}} = \frac{4}{9}(k\eta_3)^2 \mathcal{R}_{\boldsymbol{k}}. \end{aligned} \tag{5.4.1}$$

Normalizing to unit primordial curvature ($\mathcal{R}_{\boldsymbol{k}} = 1$), the expansions in powers of conformal time η (with $\omega \equiv H_0\Omega_{\mathrm{m}}/\sqrt{\Omega_{\mathrm{r}}} = 2/\eta_3$) evaluate to:

$$\delta^{\mathrm{C}}(\eta) = \frac{4}{9}k^2\eta^2 - \frac{1}{4}\omega k^2\eta^3 + \left(\frac{119}{720}\omega^2 k^2 - \frac{2}{135}k^4\right)\eta^4 + \mathcal{O}(\eta^5), \tag{5.4.2}$$

$$\mathcal{S}(\eta) = \frac{1}{144}k^4\eta^4 - \frac{1}{1200}\omega k^4\eta^5 + \mathcal{O}(\eta^6), \tag{5.4.3}$$

$$\Phi(\eta) = -\frac{2}{3} + \frac{1}{24}\omega\eta + \left(-\frac{1}{80}\omega^2 + \frac{1}{45}k^2\right)\eta^2 + \mathcal{O}(\eta^3), \tag{5.4.4}$$

$$\Psi(\eta) = -\frac{2}{3} + \frac{1}{24}\omega\eta + \left(-\frac{1}{80}\omega^2 + \frac{1}{45}k^2\right)\eta^2 + \mathcal{O}(\eta^3), \tag{5.4.5}$$

$$v^{\mathrm{N}}(\eta) = -\frac{1}{3}k\eta + \frac{1}{24}\omega k\eta^2 + \left(-\frac{1}{160}\omega^2 k + \frac{1}{90}k^3\right)\eta^3 + \mathcal{O}(\eta^4), \tag{5.4.6}$$

$$\delta^{\mathrm{N}}_{\mathrm{m}}(\eta) = 1 + \frac{1}{8}\omega\eta + \left(-\frac{1}{40}\omega^2 - \frac{1}{18}k^2\right)\eta^2 + \mathcal{O}(\eta^3), \tag{5.4.7}$$

$$\delta^{\mathrm{N}}_{\mathrm{r}}(\eta) = \frac{4}{3} + \frac{1}{6}\omega\eta + \left(-\frac{1}{30}\omega^2 - \frac{2}{27}k^2\right)\eta^2 + \mathcal{O}(\eta^3). \tag{5.4.8}$$

Physical Features of the Adiabatic Mode

- **Primordial Amplitude Ratio:** In the $x \to 0$ limit, the component density contrasts satisfy $\delta^{\mathrm{N}}_{\mathrm{m}} = \frac{3}{4}\delta^{\mathrm{N}}_{\mathrm{r}}$, reflecting identical fractional spatial compressions across all particle species (uniform matter-to-radiation ratio, $\delta(\rho_{\mathrm{m}}/\rho_{\mathrm{r}}) = 0$).
- **Superhorizon Curvature Conservation:** The comoving curvature perturbation $\mathcal{R}_{\boldsymbol{k}}$ remains constant on superhorizon scales throughout both radiation and matter dominated eras:

$$\mathcal{R}_{\boldsymbol{k}}(\eta) = 1 - \frac{1}{144}k^2\eta^2 + \mathcal{O}(\eta^3) \approx \text{constant} \qquad (k\eta \ll 1). \tag{1}$$

- **Entropy Generation:** Although $\mathcal{S} = 0$ at the initial singularity, horizon-scale pressure gradients dynamically induce a minuscule secondary entropy perturbation

$\mathcal{S} \propto k^4\eta^4$.

5.4.2 The Growing Isocurvature Mode

For the isocurvature mode ($C_{\boldsymbol{k}} = 0$), the perturbation is initialized with unit primordial entropy amplitude $E_{\boldsymbol{k}} = \mathcal{S}_{\boldsymbol{k}}^{\rm prim} = 1$. The series expansions in powers of η are:

$$\mathcal{S}(\eta) = 1 - \frac{1}{36}\omega k^2\eta^3 + \frac{1}{192}\omega^2 k^2\eta^4 + \left(-\frac{1}{960}\omega^3 k^2 + \frac{1}{4800}\omega k^4\right)\eta^5 + \mathcal{O}(\eta^6), \tag{5.4.9}$$

$$\delta^{\rm C}(\eta) = \frac{1}{12}\omega k^2\eta^3 - \frac{19}{480}\omega^2 k^2\eta^4 + \mathcal{O}(\eta^5), \tag{5.4.10}$$

$$\Phi(\eta) = -\frac{1}{8}\omega\eta + \frac{1}{16}\omega^2\eta^2 - \frac{9}{320}\omega^3\eta^3 + \frac{1}{72}\omega k^2\eta^3 + \mathcal{O}(\eta^4), \tag{5.4.11}$$

$$\Psi(\eta) = -\frac{1}{8}\omega\eta + \frac{1}{16}\omega^2\eta^2 - \frac{9}{320}\omega^3\eta^3 + \frac{1}{72}\omega k^2\eta^3 + \mathcal{O}(\eta^4), \tag{5.4.12}$$

$$v^{\rm N}(\eta) = -\frac{1}{8}\omega k\eta^2 + \frac{3}{80}\omega^2 k\eta^3 + \mathcal{O}(\eta^4), \tag{5.4.13}$$

$$\delta_{\rm m}^{\rm C}(\eta) = 1 - \frac{3}{4}\omega\eta + \frac{3}{8}\omega^2\eta^2 - \frac{9}{64}\omega^3\eta^3 + \frac{5}{144}\omega k^2\eta^3 + \mathcal{O}(\eta^4), \tag{5.4.14}$$

$$\delta_{\rm r}^{\rm C}(\eta) = -\omega\eta + \frac{1}{2}\omega^2\eta^2 - \frac{3}{16}\omega^3\eta^3 + \frac{1}{12}\omega k^2\eta^3 + \mathcal{O}(\eta^4). \tag{5.4.15}$$

Physical Features and Superhorizon Evolution of the Isocurvature Mode

In the isocurvature mode, initial spatial curvature vanishes ($\Phi = \Psi = \mathcal{R} = 0$ as $\eta \to 0$) because the matter overdensity is compensated by a radiation underdensity:

$$\delta\rho = \bar{\rho}_{\rm m}\delta_{\rm m} + \bar{\rho}_{\rm r}\delta_{\rm r} = \bar{\rho}_{\rm r}\left(y\,\delta_{\rm m} + \delta_{\rm r}\right) = \bar{\rho}_{\rm r}\left(y(1) + (-\omega\eta)\right) = \bar{\rho}_{\rm r}(2x - 2x) = 0. \tag{1}$$

However, as the Universe expands and matter begins to dominate over radiation ($y > 0$), this initial entropy perturbation generates curvature. For superhorizon scales ($k \ll \mathcal{H}$), the comoving curvature perturbation evolves according to:

$$\mathcal{R}' = \mathcal{H}c_{\rm s}^2(1 - 3c_{\rm s}^2)\mathcal{S} = \mathcal{H}\frac{4y}{(4+3y)^2}\mathcal{S} \implies \frac{{\rm d}\mathcal{R}}{{\rm d}y} = \frac{4}{(4+3y)^2}\mathcal{S}. \tag{2}$$

Integrating from $y = 0$ where $\mathcal{R}(0) = 0$:

$$\mathcal{R}(y) = \frac{y}{4+3y}\mathcal{S}_{\boldsymbol{k}}^{\rm prim}. \tag{3}$$

In the matter-dominated epoch ($y \gg 1$), the generated curvature perturbation reaches the asymptotic plateau:

$$\mathcal{R}_{\rm matter} = \frac{1}{3}\mathcal{S}_{\boldsymbol{k}}^{\rm prim}. \tag{4}$$

5.5 Gauge Transformation to Synchronous Gauge

Boltzmann codes such as CAMB and COSMICS perform time integration in the **synchronous gauge**, where the perturbed FLRW metric is:

$${\rm d}s^2 = a^2(\eta)\left[-{\rm d}\eta^2 + (\delta_{ij} + h_{ij}(\boldsymbol{x},\eta)){\rm d}x^i{\rm d}x^j\right]. \tag{5.5.1}$$

In Fourier space, the spatial metric perturbation $h_{ij}(\boldsymbol{k},\eta)$ is decomposed into trace $h(\boldsymbol{k},\eta)$ and longitudinal trace-free $\eta_{\rm syn}(\boldsymbol{k},\eta)$ components:

$$h_{ij}(\boldsymbol{k},\eta) = \hat{k}_i\hat{k}_j h(\boldsymbol{k},\eta) + \left(\hat{k}_i\hat{k}_j - \frac{1}{3}\delta_{ij}\right) 6\eta_{\rm syn}(\boldsymbol{k},\eta). \tag{5.5.2}$$

5.5.1 Synchronous Gauge Fixing

The synchronous gauge condition $A = B = 0$ leaves a residual coordinate gauge freedom generated by scalar gauge parameters satisfying $T' + \mathcal{H}T = 0 \implies T(\boldsymbol{x}) = C_1(\boldsymbol{x})/a(\eta)$ and $L' = T$. To fix this residual freedom completely, standard practice sets the **cold dark matter velocity to zero identically**:

$$v_{\rm c}^{\rm syn}(\boldsymbol{k},\eta) \equiv 0. \tag{5.5.3}$$

Because cold dark matter is collisionless and pressureless, CDM particles follow geodesics of spacetime, ensuring that comoving CDM coordinates define an unperturbed synchronous time slicing.

5.5.2 Transformation from Comoving to Synchronous Gauge

Under a gauge transformation from synchronous gauge ($B^{\rm syn} = 0$) to comoving gauge, the density contrasts and velocities transform as:

$$\delta_i^{\rm C} = \delta_i^{\rm syn} + 3(1+w_i)\frac{\mathcal{H}}{k}v^{\rm syn}, \qquad v_i^{\rm C} = v_i^{\rm syn} - v^{\rm syn}, \tag{5.5.4}$$

where $v^{\rm syn}$ is the total fluid velocity in synchronous gauge:

$$v^{\rm syn} = \frac{4}{4+3y}v_{\rm r}^{\rm syn} + \frac{3y}{4+3y}v_{\rm m}^{\rm syn} = \frac{4}{4+3y}v_{\rm r}^{\rm syn}. \tag{5.5.5}$$

Since $v_{\rm m}^{\rm syn} \equiv 0$, the relative velocity is simply $v_{\rm r}^{\rm syn} = v_{\rm r}^{\rm syn} - v_{\rm m}^{\rm syn} = k^{-1}\mathcal{S}'$. Inverting Eq. (5.5.4) gives:

$$\delta_{\rm m}^{\rm syn} = \delta_{\rm m}^{\rm C} - 3\frac{\mathcal{H}}{k}v^{\rm syn}, \qquad \delta_{\rm r}^{\rm syn} = \delta_{\rm r}^{\rm C} - 4\frac{\mathcal{H}}{k}v^{\rm syn}. \tag{5.5.6}$$

5.5.3 Explicit Synchronous Initial Series

Substituting the series solutions for $\delta^{\rm C}$ and $\mathcal{S}$ yields the explicit initial expansions in synchronous gauge:

Adiabatic Mode ($\mathcal{R}_{\boldsymbol{k}} = 1$):

$$\delta_{\rm m}^{\rm syn}(\eta) = -\frac{1}{6}k^2\eta^2 + \frac{1}{48}\omega k^2\eta^3 + \left(-\frac{1}{240}\omega^2k^2 + \frac{1}{216}k^4\right)\eta^4 + \mathcal{O}(\eta^5), \tag{5.5.7}$$

$$\delta_{\rm r}^{\rm syn}(\eta) = -\frac{2}{9}k^2\eta^2 + \frac{1}{36}\omega k^2\eta^3 + \left(-\frac{1}{180}\omega^2k^2 + \frac{1}{162}k^4\right)\eta^4 + \mathcal{O}(\eta^5), \tag{5.5.8}$$

$$v_{\rm r}^{\rm syn}(\eta) = -\frac{1}{36}k^3\eta^3 + \frac{1}{240}\omega k^3\eta^4 + \mathcal{O}(\eta^5), \tag{5.5.9}$$

$$h(\eta) = \frac{1}{2}k^2\eta^2 - \frac{1}{16}\omega k^2\eta^3 + \mathcal{O}(\eta^4), \tag{5.5.10}$$

$$\eta_{\rm syn}(\eta) = 1 - \frac{1}{12}k^2\eta^2 + \frac{1}{96}\omega k^2\eta^3 + \mathcal{O}(\eta^4). \tag{5.5.11}$$

Isocurvature Mode ($\mathcal{S}_{\boldsymbol{k}}^{\mathrm{prim}} = 1$):

$$\delta_{\mathrm{m}}^{\mathrm{syn}}(\eta) = 1 - \frac{1}{2}\omega\eta + \frac{3}{16}\omega^2\eta^2 - \frac{5}{96}\omega^3\eta^3 + \frac{1}{72}\omega k^2\eta^3 + \mathcal{O}(\eta^4), \tag{5.5.12}$$

$$\delta_{\mathrm{r}}^{\mathrm{syn}}(\eta) = -\frac{2}{3}\omega\eta + \frac{1}{4}\omega^2\eta^2 - \frac{5}{72}\omega^3\eta^3 + \frac{1}{54}\omega k^2\eta^3 + \mathcal{O}(\eta^4), \tag{5.5.13}$$

$$v_{\mathrm{r}}^{\mathrm{syn}}(\eta) = -\frac{1}{12}\omega k\eta^2 + \frac{1}{48}\omega^2 k\eta^3 + \mathcal{O}(\eta^4). \tag{5.5.14}$$

These series expressions match the analytical formulations implemented in the CAMB codebase.

5.6 Multi-Fluid Primordial Dynamics in the Real Universe

In the realistic early Universe, the cosmic fluid cannot be approximated as a single radiation component and a single matter component. The radiation sector is partitioned into tightly coupled photons (γ) and free-streaming relativistic neutrinos (ν), while the matter sector comprises tightly coupled baryons (b) and collisionless cold dark matter (c):

$$\bar{\rho}_{\mathrm{r}}(\eta) = \bar{\rho}_\gamma(\eta) + \bar{\rho}_\nu(\eta), \qquad \bar{\rho}_{\mathrm{m}}(\eta) = \bar{\rho}_{\mathrm{b}}(\eta) + \bar{\rho}_{\mathrm{c}}(\eta). \tag{5.6.1}$$

We define the background radiation and matter fractional energy densities:

$$f_\nu \equiv \frac{\bar{\rho}_\nu}{\bar{\rho}_{\mathrm{r}}} = \frac{N_{\mathrm{eff}}\frac{7}{8}\left(\frac{4}{11}\right)^{4/3}}{1 + N_{\mathrm{eff}}\frac{7}{8}\left(\frac{4}{11}\right)^{4/3}} \approx 0.4052, \qquad f_\gamma \equiv \frac{\bar{\rho}_\gamma}{\bar{\rho}_{\mathrm{r}}} = 1 - f_\nu \approx 0.5948, \tag{5.6.2}$$

$$f_{\mathrm{b}} \equiv \frac{\bar{\rho}_{\mathrm{b}}}{\bar{\rho}_{\mathrm{m}}} = \frac{\Omega_{\mathrm{b}}}{\Omega_{\mathrm{m}}}, \qquad f_{\mathrm{c}} \equiv \frac{\bar{\rho}_{\mathrm{c}}}{\bar{\rho}_{\mathrm{m}}} = 1 - f_{\mathrm{b}} = \frac{\Omega_{\mathrm{c}}}{\Omega_{\mathrm{m}}}. \tag{5.6.3}$$

Deep in the radiation-dominated era before recombination, Thomson scattering between photons and free electrons couples baryons tightly to the photon plasma. The photon-baryon momentum ratio (drag parameter) is given by:

$$R \equiv \frac{4\bar{\rho}_{\mathrm{b}}}{3\bar{\rho}_\gamma} = \frac{4}{3}\frac{f_{\mathrm{b}}\bar{\rho}_{\mathrm{m}}}{f_\gamma\bar{\rho}_{\mathrm{r}}} = \frac{3}{4}\frac{f_{\mathrm{b}}}{f_\gamma}y = \frac{3}{4}\frac{f_{\mathrm{b}}}{1-f_\nu}(2x + x^2), \tag{5.6.4}$$

and enforces the tight-coupling velocity equality:

$$v_{\mathrm{b}}^{\mathrm{N}} = v_\gamma^{\mathrm{N}}. \tag{5.6.5}$$

5.6.1 Multi-Fluid Density and Velocity Perturbations

In terms of individual component density contrasts $\delta_i^{\mathrm{N}} \equiv \delta\rho_i/\bar{\rho}_i$ and peculiar velocity potentials v_i^{N}, the total energy density perturbation δ and total momentum velocity v of the combined four-fluid system are:

$$\delta = \frac{1}{1+y}(f_\gamma\delta_\gamma^{\mathrm{N}} + f_\nu\delta_\nu^{\mathrm{N}}) + \frac{y}{1+y}(f_{\mathrm{b}}\delta_{\mathrm{b}}^{\mathrm{N}} + f_{\mathrm{c}}\delta_{\mathrm{c}}^{\mathrm{N}}), \tag{5.6.6}$$

$$v = \frac{4}{4+3y}(f_\gamma v_\gamma^{\mathrm{N}} + f_\nu v_\nu^{\mathrm{N}}) + \frac{3y}{4+3y}(f_{\mathrm{b}}v_{\mathrm{b}}^{\mathrm{N}} + f_{\mathrm{c}}v_{\mathrm{c}}^{\mathrm{N}}). \tag{5.6.7}$$

The relative entropy perturbations among the four species are defined relative to the total radiation fluid or photons:

$$\mathcal{S}_{\mathrm{mr}} \equiv \delta_{\mathrm{m}} - \frac{3}{4}\delta_{\mathrm{r}} = f_{\mathrm{b}}\mathcal{S}_{\mathrm{br}} + f_{\mathrm{c}}\mathcal{S}_{\mathrm{cr}}, \qquad \mathcal{S}_{\nu\mathrm{r}} \equiv \frac{3}{4}(\delta_\nu - \delta_{\mathrm{r}}) = (1-f_\nu)\mathcal{S}_{\nu\gamma}, \tag{5.6.8}$$

where $\mathcal{S}_{\mathrm{br}} \equiv \delta_{\mathrm{b}} - \frac{3}{4}\delta_{\mathrm{r}}$ and $\mathcal{S}_{\mathrm{cr}} \equiv \delta_{\mathrm{c}} - \frac{3}{4}\delta_{\mathrm{r}}$.

5.6.2 The Neutrino Boltzmann Hierarchy and Anisotropic Stress

While photons remain isotropic in their rest frame due to rapid Thomson scattering ($\Pi_\gamma = 0$), collisionless neutrinos free-stream freely across the horizon, developing anisotropic shear stress Π_ν and higher multipole moments. The expansion of the neutrino phase-space perturbation into Legendre multipoles Θ_ℓ^ν obeys the standard Boltzmann hierarchy:

$$\Theta_\ell' = \frac{\ell}{2\ell+1} k\Theta_{\ell-1} - \frac{\ell+1}{2\ell+1} k\Theta_{\ell+1}, \tag{5.6.9}$$

with the lowest moments identified physically in Conformal Newtonian gauge as:

$$\Theta_0^\nu = \frac{1}{4}\delta_\nu^{\rm N}, \qquad \Theta_1^\nu = \frac{1}{3}v_\nu^{\rm N}, \qquad \Theta_2^\nu = \frac{1}{12}\Pi_\nu, \qquad \Theta_3^\nu = \Theta_3. \tag{5.6.10}$$

The dynamical equations governing the lowest four neutrino multipoles are:

$$\delta_\nu^{\rm N\prime} = -\frac{4}{3}kv_\nu^{\rm N} + 4\Psi', \tag{5.6.11}$$

$$v_\nu^{\rm N\prime} = \frac{1}{4}k\delta_\nu^{\rm N} - \frac{1}{6}k\Pi_\nu + k\Phi, \tag{5.6.12}$$

$$\Pi_\nu' = \frac{8}{5}kv_\nu^{\rm N} - \frac{36}{5}k\Theta_3, \tag{5.6.13}$$

$$\Theta_3' = \frac{1}{28}k\Pi_\nu - \frac{4}{7}k\Theta_4. \tag{5.6.14}$$

On superhorizon scales during the primordial era ($k\eta \ll 1$), higher multipoles $\ell \geq 3$ are strongly suppressed by successive powers of $(k\eta)$. Truncating the hierarchy at $\ell = 3$ with $\Theta_4 = \mathcal{O}(x^4)$ allows us to express the neutrino shear and octupole as asymptotic expansions in $x \equiv \eta/\eta_3$:

$$\Pi_\nu(x) = Qx^2 + Sx^3 + Tx^4, \qquad \Theta_3(x) = Ux^3 + \mathcal{O}(x^4). \tag{5.6.15}$$

Superhorizon Neutrino Hierarchy Recurrence Relations

Using the conformal time transformation $\mathrm{d}/\mathrm{d}\eta = \eta_3^{-1}\mathrm{d}/\mathrm{d}x$, the octupole evolution equation (5.6.14) becomes:

$$\eta_3^{-1}\frac{\mathrm{d}}{\mathrm{d}x}(Ux^3) = \frac{1}{28}k(Qx^2) \implies 3Ux^2 = \frac{1}{28}(k\eta_3)Qx^2 \implies \boxed{U = \frac{1}{84}(k\eta_3)Q.} \tag{1}$$

Similarly, substituting the velocity expansion $v_\nu^{\rm N} = L_\nu x + M_\nu x^2 + N_\nu x^3$ and the shear expansion into Eq. (5.6.13):

$$\eta_3^{-1}(2Qx + 3Sx^2 + 4Tx^3) = \frac{8}{5}k(L_\nu x + M_\nu x^2 + N_\nu x^3) - \frac{36}{5}k(Ux^3). \tag{2}$$

Matching powers of x order-by-order yields the recurrence relations:

$$\text{Order } x^1: \quad 2Q = \frac{8}{5}(k\eta_3)L_\nu \implies \boxed{Q = \frac{4}{5}(k\eta_3)L_\nu,} \tag{3}$$

$$\text{Order } x^2: \quad 3S = \frac{8}{5}(k\eta_3)M_\nu \implies \boxed{S = \frac{8}{15}(k\eta_3)M_\nu,} \tag{4}$$

$$\text{Order } x^3: \quad 4T = \frac{8}{5}(k\eta_3)N_\nu - \frac{36}{5}(k\eta_3)U \implies \boxed{T = \frac{2}{5}(k\eta_3)N_\nu - \frac{9}{5}(k\eta_3)U.} \tag{5}$$

5.7 Conformal Time Expansion for the Multi-Fluid Universe

Following Section 22.4 of Kurki-Suonio's formulation, we seek asymptotic solutions for all metric potentials, fluid density contrasts, and fluid velocity potentials as power series in the dimensionless conformal time variable $x \equiv \eta/\eta_3 = \frac{1}{2}\omega\eta$:

$$\Phi(x) = A + Bx + Cx^2, \tag{5.7.1}$$

$$\Psi(x) = D + Ex + Fx^2, \tag{5.7.2}$$

$$\delta_i^{\mathrm{N}}(x) = G_i + H_i x + J_i x^2 \qquad (i \in \{\mathrm{b}, \mathrm{c}, \gamma, \nu\}), \tag{5.7.3}$$

$$v_i^{\mathrm{N}}(x) = L_i x + M_i x^2 + N_i x^3 \qquad (i \in \{\mathrm{b}, \mathrm{c}, \gamma, \nu\}), \tag{5.7.4}$$

where tight coupling enforces $L_{\mathrm{b}} = L_\gamma$, $M_{\mathrm{b}} = M_\gamma$, and $N_{\mathrm{b}} = N_\gamma$.

5.7.1 Primordial Metric Amplitudes

In the presence of primordial curvature $\mathcal{R}_{\boldsymbol{k}}(\mathrm{rad})$ and neutrino density isocurvature $\mathcal{S}_{\nu\mathrm{r},\boldsymbol{k}}(\mathrm{rad})$, the constant leading-order gravitational potential $A \equiv \Phi(0)$ is fixed by the superhorizon Einstein constraint:

$$\boxed{A = -\frac{2}{3}\left(1 + \frac{4}{15}f_\nu\right)^{-1}\left[\mathcal{R}_{\boldsymbol{k}}(\mathrm{rad}) + \frac{4}{15}f_\nu \mathcal{S}_{\nu\mathrm{r},\boldsymbol{k}}(\mathrm{rad})\right].} \tag{5.7.5}$$

For adiabatic perturbations ($\mathcal{S}_{\nu\mathrm{r}} = 0$) with $\mathcal{R}_{\boldsymbol{k}} = 1$, this reproduces the renowned neutrino shear damping factor:

$$A = -\frac{2}{3}\frac{1}{1 + \frac{4}{15}f_\nu} = -\frac{10}{15 + 4f_\nu} \approx -0.6017. \tag{5.7.6}$$

5.7.2 The Complete Multi-Fluid Linear Matching System

To determine the unknown expansion coefficients $\{B, C, D, E, F, G_i, H_i, J_i, L_i, M_i, N_i\}$, we substitute the power series expansions into the exact relativistic fluid and Einstein equations:

1. **Fluid Continuity Equations to $\mathcal{O}(x)$:** From $\delta_i^{\mathrm{N}\prime} = 3(1 + w_i)\Psi' - (1 + w_i)kv_i^{\mathrm{N}}$, matching powers of x gives:

$$H_{\mathrm{b}} + 2J_{\mathrm{b}}x = -(k\eta_3)L_\gamma x + 3E + 6Fx, \tag{5.7.7}$$

$$H_{\mathrm{c}} + 2J_{\mathrm{c}}x = -(k\eta_3)L_{\mathrm{c}} x + 3E + 6Fx, \tag{5.7.8}$$

$$H_\gamma + 2J_\gamma x = -\frac{4}{3}(k\eta_3)L_\gamma x + 4E + 8Fx, \tag{5.7.9}$$

$$H_\nu + 2J_\nu x = -\frac{4}{3}(k\eta_3)L_\nu x + 4E + 8Fx. \tag{5.7.10}$$

2. **Fluid Momentum (Euler) Equations to $\mathcal{O}(x^2)$:**

   - **Cold Dark Matter (collisionless):**

$$L_{\mathrm{c}} + 2M_{\mathrm{c}}x + 3N_{\mathrm{c}}x^2 = -\mathcal{H}\eta_3 v_{\mathrm{c}}^{\mathrm{N}} + (k\eta_3)\Phi. \tag{5.7.11}$$

   - **Coupled Photon-Baryon Fluid (with drag R):**

$$(1 + R)(L_\gamma + 2M_\gamma x + 3N_\gamma x^2) = \frac{1}{4}(k\eta_3)\delta_\gamma^{\mathrm{N}} - R\,\mathcal{H}\eta_3 v_\gamma^{\mathrm{N}} + (1 + R)(k\eta_3)\Phi. \tag{5.7.12}$$

   - **Free-Streaming Neutrinos (with shear Π_ν):**

$$L_\nu + 2M_\nu x + 3N_\nu x^2 = \frac{1}{4}(k\eta_3)\delta_\nu^{\mathrm{N}} - \frac{1}{6}(k\eta_3)Qx^2 + (k\eta_3)\Phi. \tag{5.7.13}$$

3. **Einstein Constraint Equations to $\mathcal{O}(x^2)$:**

   - **Traceless Spatial Constraint (Anisotropic Stress):**

$$(k\eta_3)^2(\Phi-\Psi) = 3(\mathcal{H}\eta_3)^2 w(x) f_\nu \Pi_\nu(x). \tag{5.7.14}$$

   - **Energy Constraint (00 Einstein Equation):**

$$x^2(k\eta_3)^2\Psi = -x^2\frac{3}{2}(\mathcal{H}\eta_3)^2\left[\delta + 3(1+w(x))\frac{\mathcal{H}\eta_3}{k\eta_3}v\right]. \tag{5.7.15}$$

4. **Primordial Entropy Constraints:**

$$\mathcal{S}_{\mathrm{mr},\boldsymbol{k}}(\mathrm{rad}) = G_\mathrm{c} - \frac{3}{4}(f_\gamma G_\gamma + f_\nu G_\nu), \tag{5.7.16}$$

$$\mathcal{S}_{\nu\mathrm{r},\boldsymbol{k}}(\mathrm{rad}) = \frac{3}{4}(1-f_\nu)(G_\nu - G_\gamma), \tag{5.7.17}$$

$$G_\mathrm{b} = G_\mathrm{c}. \tag{5.7.18}$$

5.7.3 Transformation to Synchronous Gauge

In the synchronous gauge ($\mathrm{d}s^2 = a^2[-\mathrm{d}\eta^2 + (\delta_{ij}+h_{ij})\mathrm{d}x^i\mathrm{d}x^j]$), the residual gauge freedom is fixed unambiguously by setting the cold dark matter peculiar velocity to zero:

$$v_\mathrm{c}^\mathrm{Z} \equiv 0. \tag{5.7.19}$$

Under the gauge transformation generated by the CDM trajectory, the perturbation variables transform as:

$$\delta_i^\mathrm{Z}(x) = \delta_i^\mathrm{N}(x) + 3(1+w_i)\frac{\mathcal{H}}{k}(v^\mathrm{N}-v^\mathrm{Z}) = \delta_i^\mathrm{N}(x) + 3(1+w_i)\frac{\mathcal{H}\eta_3}{k\eta_3}v_\mathrm{c}^\mathrm{N}(x), \tag{5.7.20}$$

$$v_i^\mathrm{Z}(x) = v_i^\mathrm{N}(x) - v_\mathrm{c}^\mathrm{N}(x). \tag{5.7.21}$$

Expanding the synchronous variables as power series:

$$\delta_i^\mathrm{Z}(x) = G_i^\mathrm{Z} + H_i^\mathrm{Z}x + J_i^\mathrm{Z}x^2, \qquad v_i^\mathrm{Z}(x) = L_i^\mathrm{Z}x + M_i^\mathrm{Z}x^2 + N_i^\mathrm{Z}x^3, \tag{5.7.22}$$

the matching equations (5.7.20) and (5.7.21) fully determine all synchronous coefficients.

5.8 The Five Primordial Initial Condition Modes

A four-component cosmic medium $(\gamma, \nu, \mathrm{b}, \mathrm{c})$ supports exactly **five fundamental linear perturbation modes** (Bucher, Moodley & Turok, 2000):

1. **Adiabatic Growing Mode (ADI):** $\mathcal{R}_{\boldsymbol{k}}(\mathrm{rad}) = 1$, $\mathcal{S}_\mathrm{mr} = 0$, $\mathcal{S}_{\nu\mathrm{r}} = 0$.
2. **Cold Dark Matter Density Isocurvature Mode (CDI):** $\mathcal{R}_{\boldsymbol{k}}(\mathrm{rad}) = 0$, $\mathcal{S}_\mathrm{cr} = 1/f_\mathrm{c}$, $\mathcal{S}_\mathrm{br} = 0$, $\mathcal{S}_{\nu\mathrm{r}} = 0$.
3. **Baryon Density Isocurvature Mode (BDI):** $\mathcal{R}_{\boldsymbol{k}}(\mathrm{rad}) = 0$, $\mathcal{S}_\mathrm{br} = 1/f_\mathrm{b}$, $\mathcal{S}_\mathrm{cr} = 0$, $\mathcal{S}_{\nu\mathrm{r}} = 0$.
4. **Neutrino Density Isocurvature Mode (NDI):** $\mathcal{R}_{\boldsymbol{k}}(\mathrm{rad}) = 0$, $\mathcal{S}_\mathrm{mr} = 0$, $\mathcal{S}_{\nu\mathrm{r}} = 3/4$.
5. **Neutrino Velocity Isocurvature Mode (NVI):** Relative dipole velocity between free-streaming neutrinos and photons with $f_\nu v_\nu + f_\gamma v_\gamma = 0$.

5.8.1 The Multi-Fluid Adiabatic Mode (ADI)

Exact Multi-Fluid Expansions: Adiabatic Mode ($\mathcal{R}_{\boldsymbol{k}}(\mathbf{rad}) = 1$)

Solving the complete multi-fluid matching system with $\mathcal{R}_{\boldsymbol{k}}(\mathrm{rad}) = 1$ and all primordial entropy perturbations set to zero yields the explicit conformal time expansions (with $\omega \equiv H_0\Omega_{\mathrm{m}}/\sqrt{\Omega_{\mathrm{r}}}$):

Conformal Newtonian Gauge:

$$\Phi(\eta) = -\frac{2}{3}\frac{1}{1+\frac{4}{15}f_\nu} - \frac{25}{8}\frac{8f_\nu - 3}{(15+2f_\nu)(15+4f_\nu)}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.1}$$

$$\Psi(\eta) = -\frac{2}{3}\frac{1+\frac{2}{5}f_\nu}{1+\frac{4}{15}f_\nu} + \frac{5}{8}\frac{15+16f_\nu}{(15+2f_\nu)(15+4f_\nu)}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.2}$$

$$\delta_{\mathrm{b}}^{\mathrm{N}}(\eta) = \delta_{\mathrm{c}}^{\mathrm{N}}(\eta) = \frac{1}{1+\frac{4}{15}f_\nu} + \frac{1}{8}\frac{1+\frac{16}{15}f_\nu}{\left(1+\frac{2}{15}f_\nu\right)\left(1+\frac{4}{15}f_\nu\right)}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.3}$$

$$\delta_{\gamma}^{\mathrm{N}}(\eta) = \delta_{\nu}^{\mathrm{N}}(\eta) = \frac{4}{3}\frac{1}{1+\frac{4}{15}f_\nu} + \frac{1}{6}\frac{1+\frac{16}{15}f_\nu}{\left(1+\frac{2}{15}f_\nu\right)\left(1+\frac{4}{15}f_\nu\right)}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.4}$$

$$v_{\mathrm{c}}^{\mathrm{N}}(\eta) = -\frac{1}{3}\frac{1}{1+\frac{4}{15}f_\nu}k\eta + \frac{1}{8}\frac{5-4f_\nu}{(15+2f_\nu)\left(1+\frac{4}{15}f_\nu\right)}\omega k\eta^2 + \mathcal{O}(\eta^3), \tag{5.8.5}$$

$$v_{\mathrm{b}}^{\mathrm{N}}(\eta) = v_{\gamma}^{\mathrm{N}}(\eta) = v_{\mathrm{c}}^{\mathrm{N}}(\eta) + \mathcal{O}(\eta^3), \tag{5.8.6}$$

$$v_{\nu}^{\mathrm{N}}(\eta) = v_{\mathrm{c}}^{\mathrm{N}}(\eta) + \mathcal{O}(\eta^3), \tag{5.8.7}$$

$$\Pi_\nu(\eta) = -\frac{4}{15}\frac{1}{1+\frac{4}{15}f_\nu}(k\eta)^2 + \frac{1}{45}\frac{1-\frac{4}{5}f_\nu}{\left(1+\frac{4}{15}f_\nu\right)\left(1+\frac{2}{15}f_\nu\right)}\omega k^2\eta^3 + \mathcal{O}(\eta^4). \tag{5.8.8}$$

Synchronous Gauge ($v_{\mathrm{c}}^{\mathrm{Z}} \equiv 0$):

$$\delta_{\mathrm{b}}^{\mathrm{Z}}(\eta) = \delta_{\mathrm{c}}^{\mathrm{Z}}(\eta) = \frac{1}{4}(k\eta)^2 + \mathcal{O}(\eta^3), \tag{5.8.9}$$

$$\delta_{\gamma}^{\mathrm{Z}}(\eta) = \delta_{\nu}^{\mathrm{Z}}(\eta) = \frac{1}{3}(k\eta)^2 + \mathcal{O}(\eta^3), \tag{5.8.10}$$

$$v_{\mathrm{c}}^{\mathrm{Z}}(\eta) = 0, \tag{5.8.11}$$

$$v_{\mathrm{b}}^{\mathrm{Z}}(\eta) = v_{\gamma}^{\mathrm{Z}}(\eta) = \frac{1}{36}(k\eta)^3 + \mathcal{O}(\eta^4), \tag{5.8.12}$$

$$v_{\nu}^{\mathrm{Z}}(\eta) = \frac{1}{36}\frac{23+4f_\nu}{15+4f_\nu}(k\eta)^3 + \mathcal{O}(\eta^4). \tag{5.8.13}$$

5.8.2 Matter Density Isocurvature Modes (CDI and BDI)

The matter isocurvature modes are initialized by setting $\mathcal{R}_{\boldsymbol{k}}(\mathrm{rad}) = 0$, $\mathcal{S}_{\mathrm{mr},\boldsymbol{k}}(\mathrm{rad}) = 1$, and $\mathcal{S}_{\nu\mathrm{r},\boldsymbol{k}}(\mathrm{rad}) = 0$.

- For the **CDM density isocurvature (CDI)** mode: $S_{\mathrm{cr},\boldsymbol{k}} = 1/f_{\mathrm{c}}$ and $S_{\mathrm{br},\boldsymbol{k}} = 0$. All perturbation variables are multiplied by f_{c}, except the constant primordial term in δ_{c} which remains 1 (while the constant term in δ_{b} vanishes).
- For the **baryon density isocurvature (BDI)** mode: $S_{\mathrm{br},\boldsymbol{k}} = 1/f_{\mathrm{b}}$ and $S_{\mathrm{cr},\boldsymbol{k}} = 0$. All perturbation variables are multiplied by f_{b}, except the constant primordial term in δ_{b} which remains 1 (while the constant term in δ_{c} vanishes).

Exact Multi-Fluid Expansions: Matter Density Isocurvature Modes

Conformal Newtonian Gauge:

$$\Phi(\eta) = -\frac{1}{8}\frac{1-\frac{4}{15}f_\nu}{1+\frac{2}{15}f_\nu}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.14}$$

$$\Psi(\eta) = -\frac{1}{8}\frac{1+\frac{4}{15}f_\nu}{1+\frac{2}{15}f_\nu}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.15}$$

$$\delta_{\rm b}^{\rm N}(\eta) = \delta_{\rm c}^{\rm N}(\eta) = 1 - \frac{3}{8}\frac{1+\frac{4}{15}f_\nu}{1+\frac{2}{15}f_\nu}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.16}$$

$$\delta_\gamma^{\rm N}(\eta) = \delta_\nu^{\rm N}(\eta) = -\frac{1}{2}\frac{1+\frac{4}{15}f_\nu}{1+\frac{2}{15}f_\nu}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.17}$$

$$v_{\rm c}^{\rm N}(\eta) = -\frac{1}{24}\frac{1-\frac{4}{15}f_\nu}{1+\frac{2}{15}f_\nu}\omega k\eta^2 + \mathcal{O}(\eta^3), \tag{5.8.18}$$

$$v_\nu^{\rm N}(\eta) = -\frac{1}{8}\frac{1}{1+\frac{2}{15}f_\nu}\omega k\eta^2 + \mathcal{O}(\eta^3), \tag{5.8.19}$$

$$v_{\rm b}^{\rm N}(\eta) = v_\gamma^{\rm N}(\eta) = v_\nu^{\rm N}(\eta) + \mathcal{O}(\eta^3), \tag{5.8.20}$$

$$\Pi_\nu(\eta) = -\frac{1}{15}\frac{1}{1+\frac{2}{15}f_\nu}\omega k^2\eta^3 + \mathcal{O}(\eta^4). \tag{5.8.21}$$

Synchronous Gauge ($v_{\rm c}^{\rm Z} \equiv 0$):

$$\delta_{\rm b}^{\rm Z}(\eta) = \delta_{\rm c}^{\rm Z}(\eta) = 1 - \frac{1}{2}\omega\eta + \frac{3}{16}\omega^2\eta^2 + \mathcal{O}(\eta^3), \tag{5.8.22}$$

$$\delta_\gamma^{\rm Z}(\eta) = \delta_\nu^{\rm Z}(\eta) = -\frac{2}{3}\omega\eta + \frac{1}{4}\omega^2\eta^2 + \mathcal{O}(\eta^3), \tag{5.8.23}$$

$$v_{\rm c}^{\rm Z}(\eta) = 0, \tag{5.8.24}$$

$$v_{\rm b}^{\rm Z}(\eta) = v_\gamma^{\rm Z}(\eta) = -\frac{1}{12}\omega k\eta^2 + \frac{1}{48}\frac{1+3f_{\rm b}-f_\nu}{1-f_\nu}\omega^2 k\eta^3 + \mathcal{O}(\eta^4), \tag{5.8.25}$$

$$v_\nu^{\rm Z}(\eta) = -\frac{1}{12}\omega k\eta^2 + \frac{1}{48}\omega^2 k\eta^3 + \mathcal{O}(\eta^4). \tag{5.8.26}$$

5.8.3 Neutrino Density Isocurvature Mode (NDI)

The neutrino density isocurvature mode represents a primordial entropy perturbation between neutrinos and photons ($\mathcal{S}_{\nu\gamma} \equiv \delta_\nu - \delta_\gamma \neq 0$), normalized to $\mathcal{S}_{\nu{\rm r},\boldsymbol{k}}({\rm rad}) = 3/4$ with $\mathcal{R}_{\boldsymbol{k}}({\rm rad}) = 0$ and $\mathcal{S}_{{\rm mr},\boldsymbol{k}}({\rm rad}) = 0$.

Although $\mathcal{R}_{\boldsymbol{k}}({\rm rad}) = 0$, the metric potentials Φ and Ψ are **non-zero at** $\eta = 0$ because the relative density fluctuation between photons and neutrinos creates an imbalance in free-streaming anisotropic stress Π_ν. The traceless Einstein equation requires:

$$k^2(\Phi - \Psi) = 3\mathcal{H}^2 w f_\nu \Pi_\nu = \frac{1}{\eta^2} f_\nu \left[\frac{4}{15}(k\eta)^2 \mathcal{S}_{\nu{\rm r}}\right] = \frac{4}{15} f_\nu k^2 \mathcal{S}_{\nu{\rm r}} \implies \Phi - \Psi \neq 0. \tag{5.8.27}$$

Enforcing $\mathcal{R} = -\Psi - \frac{2}{3(1+w)}(\mathcal{H}^{-1}\Phi' + \Phi) = 0$ at $x = 0$ determines the initial amplitudes:

$$\Phi(0) = -\frac{2}{15}\frac{f_\nu}{1+\frac{4}{15}f_\nu}, \qquad \Psi(0) = \frac{1}{15}\frac{f_\nu}{1+\frac{4}{15}f_\nu}. \tag{5.8.28}$$

Exact Multi-Fluid Expansions: Neutrino Density Isocurvature Mode (NDI)

Conformal Newtonian Gauge:

$$\Phi(\eta) = -\frac{2}{15}\frac{f_\nu}{1+\frac{4}{15}f_\nu} - \frac{1}{12}\frac{\left(1-\frac{2}{75}f_\nu\right)f_\nu}{\left(1+\frac{2}{15}f_\nu\right)\left(1+\frac{4}{15}f_\nu\right)}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.29}$$

$$\Psi(\eta) = \frac{1}{15}\frac{f_\nu}{1+\frac{4}{15}f_\nu} - \frac{1}{60}\frac{\left(1-\frac{2}{15}f_\nu\right)f_\nu}{\left(1+\frac{2}{15}f_\nu\right)\left(1+\frac{4}{15}f_\nu\right)}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.30}$$

$$\delta_{\rm b}^{\rm N}(\eta) = \delta_{\rm c}^{\rm N}(\eta) = \frac{3}{15}\frac{f_\nu}{1+\frac{4}{15}f_\nu} - \frac{f_\nu}{20}\frac{1-\frac{2}{15}f_\nu}{\left(1+\frac{2}{15}f_\nu\right)\left(1+\frac{4}{15}f_\nu\right)}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.31}$$

$$\delta_\gamma^{\rm N}(\eta) = -\frac{11}{15}\frac{\left(1+\frac{8}{11}f_\nu\right)f_\nu}{\left(1+\frac{4}{15}f_\nu\right)(1-f_\nu)} - \frac{1}{15}\frac{\left(1-\frac{2}{15}f_\nu\right)f_\nu}{\left(1+\frac{2}{15}f_\nu\right)\left(1+\frac{4}{15}f_\nu\right)}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.32}$$

$$\delta_\nu^{\rm N}(\eta) = \frac{1+\frac{8}{15}f_\nu}{1+\frac{4}{15}f_\nu} - \frac{1}{15}\frac{\left(1-\frac{2}{15}f_\nu\right)f_\nu}{\left(1+\frac{2}{15}f_\nu\right)\left(1+\frac{4}{15}f_\nu\right)}\omega\eta + \mathcal{O}(\eta^2), \tag{5.8.33}$$

$$v_{\rm c}^{\rm N}(\eta) = -\frac{1}{15}\frac{f_\nu}{1+\frac{4}{15}f_\nu}k\eta + \frac{1}{30}\frac{f_\nu}{\left(1+\frac{2}{15}f_\nu\right)\left(1+\frac{4}{15}f_\nu\right)}\omega k\eta^2 + \mathcal{O}(\eta^3), \tag{5.8.34}$$

$$v_\nu^{\rm N}(\eta) = \frac{1}{4}\frac{1}{1+\frac{4}{15}f_\nu}k\eta + \frac{1}{30}\frac{f_\nu}{\left(1+\frac{2}{15}f_\nu\right)\left(1+\frac{4}{15}f_\nu\right)}\omega k\eta^2 + \mathcal{O}(\eta^3), \tag{5.8.35}$$

$$v_{\rm b}^{\rm N}(\eta) = v_\gamma^{\rm N}(\eta) = -\frac{19}{60}\frac{f_\nu}{(1-f_\nu)\left(1+\frac{4}{15}f_\nu\right)}k\eta + \mathcal{O}(\eta^2), \tag{5.8.36}$$

$$\Pi_\nu(\eta) = \frac{1}{5}\left(1+\frac{4}{15}f_\nu\right)(k\eta)^2 + \mathcal{O}(\eta^3). \tag{5.8.37}$$

Synchronous Gauge ($v_{\rm c}^{\rm Z} \equiv 0$):

$$\delta_{\rm b}^{\rm Z}(\eta) = \delta_{\rm c}^{\rm Z}(\eta) = \frac{f_\nu}{8(1-f_\nu)}(k\eta)^2 + \mathcal{O}(\eta^3), \tag{5.8.38}$$

$$\delta_\gamma^{\rm Z}(\eta) = -\frac{f_\nu}{1-f_\nu} + \frac{f_\nu}{6(1-f_\nu)}(k\eta)^2 + \mathcal{O}(\eta^3), \tag{5.8.39}$$

$$\delta_\nu^{\rm Z}(\eta) = 1 - \frac{1}{6}(k\eta)^2 + \mathcal{O}(\eta^3), \tag{5.8.40}$$

$$v_{\rm c}^{\rm Z}(\eta) = 0, \tag{5.8.41}$$

$$v_{\rm b}^{\rm Z}(\eta) = v_\gamma^{\rm Z}(\eta) = -\frac{f_\nu}{4(1-f_\nu)}k\eta + \frac{3}{16}\frac{f_{\rm b}f_\nu}{(1-f_\nu)^2}\omega k\eta^2 + \mathcal{O}(\eta^3), \tag{5.8.42}$$

$$v_\nu^{\rm Z}(\eta) = \frac{1}{4}k\eta - \frac{1}{40}\frac{1+\frac{4}{27}f_\nu}{1+\frac{4}{15}f_\nu}(k\eta)^3 + \mathcal{O}(\eta^4). \tag{5.8.43}$$

5.9 Summary and Comparison of Multi-Fluid Initial Conditions

Table 5.1 and Table 5.2 provide a complete reference of the leading-order asymptotic expansions across all physical primordial modes in Conformal Newtonian and Synchronous gauges.

Table 5.1: Asymptotic Conformal Newtonian gauge initial expansions for the multi-fluid Universe ($x \equiv \frac{1}{2}\omega\eta \ll 1, k\eta \ll 1$).

Quantity	**Adiabatic (ADI, $\mathcal{R}_{\boldsymbol{k}} = 1$)**	**Matter Isocurvature (CDI/BDI, $\mathcal{S}_{\rm mr} = 1$)**	**Neutrino Density Iso (NDI, $\mathcal{S}_{\nu r} = 3/4$)**
Φ	$-\frac{2}{3}\frac{1}{1+\frac{1}{15}f_\nu} - \frac{25}{8}\frac{8f_\nu-3}{(15+2f_\nu)(15+4f_\nu)}\omega\eta$	$-\frac{1}{8}\frac{1-\frac{4}{15}f_\nu}{1+\frac{2}{15}f_\nu}\omega\eta$	$-\frac{2}{15}\frac{f_\nu}{1+\frac{4}{15}f_\nu} - \frac{1}{12}\frac{(1-\frac{2}{15}f_\nu)f_\nu}{(1+\frac{2}{15}f_\nu)(1+\frac{4}{15}f_\nu)}\omega\eta$
Ψ	$-\frac{2}{3}\frac{1+\frac{2}{5}f_\nu}{1+\frac{4}{15}f_\nu} + \frac{5}{8}\frac{15+16f_\nu}{(15+2f_\nu)(15+4f_\nu)}\omega\eta$	$-\frac{1}{8}\frac{1+\frac{4}{15}f_\nu}{1+\frac{2}{15}f_\nu}\omega\eta$	$\frac{1}{15}\frac{f_\nu}{1+\frac{4}{15}f_\nu} - \frac{1}{60}\frac{(1-\frac{2}{15}f_\nu)f_\nu}{(1+\frac{2}{15}f_\nu)(1+\frac{4}{15}f_\nu)}\omega\eta$
$\delta_{\rm c}^{\rm N}$	$\frac{1}{1+\frac{4}{15}f_\nu} + \frac{1}{8}\frac{1+\frac{16}{15}f_\nu}{(1+\frac{2}{15}f_\nu)(1+\frac{4}{15}f_\nu)}\omega\eta$	$1 - \frac{3}{8}\frac{1+\frac{4}{15}f_\nu}{1+\frac{2}{15}f_\nu}\omega\eta$	$\frac{3}{15}\frac{f_\nu}{1+\frac{4}{15}f_\nu} - \frac{f_\nu}{20}\frac{1-\frac{2}{15}f_\nu}{(1+\frac{2}{15}f_\nu)(1+\frac{4}{15}f_\nu)}\omega\eta$
$\delta_\gamma^{\rm N}$	$\frac{4}{3}\frac{1}{1+\frac{4}{15}f_\nu} + \frac{1}{6}\frac{1+\frac{16}{15}f_\nu}{(1+\frac{2}{15}f_\nu)(1+\frac{4}{15}f_\nu)}\omega\eta$	$-\frac{1}{2}\frac{1+\frac{4}{15}f_\nu}{1+\frac{2}{15}f_\nu}\omega\eta$	$-\frac{11}{15}\frac{(1+\frac{8}{11}f_\nu)f_\nu}{(1+\frac{4}{15}f_\nu)(1-f_\nu)} - \frac{1}{15}\frac{(1-\frac{2}{15}f_\nu)f_\nu}{(1+\frac{2}{15}f_\nu)(1+\frac{4}{15}f_\nu)}\omega\eta$
$\delta_\nu^{\rm N}$	$\frac{4}{3}\frac{1}{1+\frac{4}{15}f_\nu} + \frac{1}{6}\frac{1+\frac{16}{15}f_\nu}{(1+\frac{2}{15}f_\nu)(1+\frac{4}{15}f_\nu)}\omega\eta$	$-\frac{1}{2}\frac{1+\frac{4}{15}f_\nu}{1+\frac{2}{15}f_\nu}\omega\eta$	$\frac{1+\frac{8}{15}f_\nu}{1+\frac{4}{15}f_\nu} - \frac{1}{15}\frac{(1-\frac{2}{15}f_\nu)f_\nu}{(1+\frac{2}{15}f_\nu)(1+\frac{4}{15}f_\nu)}\omega\eta$
$v_{\rm c}^{\rm N}$	$-\frac{1}{3}\frac{1}{1+\frac{4}{15}f_\nu}k\eta + \frac{1}{8}\frac{5-4f_\nu}{(15+2f_\nu)(1+\frac{4}{15}f_\nu)}\omega k\eta^2$	$-\frac{1}{24}\frac{1-\frac{4}{15}f_\nu}{1+\frac{2}{15}f_\nu}\omega k\eta^2$	$-\frac{1}{15}\frac{f_\nu}{1+\frac{4}{15}f_\nu}k\eta + \frac{1}{30}\frac{f_\nu}{(1+\frac{2}{15}f_\nu)(1+\frac{4}{15}f_\nu)}\omega k\eta^2$
$v_\gamma^{\rm N}$	$-\frac{1}{3}\frac{1}{1+\frac{4}{15}f_\nu}k\eta + \mathcal{O}(\omega k\eta^2)$	$-\frac{1}{8}\frac{1}{1+\frac{2}{15}f_\nu}\omega k\eta^2$	$-\frac{19}{60}\frac{f_\nu}{(1-f_\nu)(1+\frac{4}{15}f_\nu)}k\eta$
$v_\nu^{\rm N}$	$-\frac{1}{3}\frac{1}{1+\frac{4}{15}f_\nu}k\eta + \mathcal{O}(\omega k\eta^2)$	$-\frac{1}{8}\frac{1}{1+\frac{2}{15}f_\nu}\omega k\eta^2$	$\frac{1}{4}\frac{1}{1+\frac{4}{15}f_\nu}k\eta + \frac{1}{30}\frac{f_\nu}{(1+\frac{2}{15}f_\nu)(1+\frac{4}{15}f_\nu)}\omega k\eta^2$
Π_ν	$-\frac{4}{15}\frac{1}{1+\frac{4}{15}f_\nu}(k\eta)^2$	$-\frac{1}{15}\frac{1}{1+\frac{2}{15}f_\nu}\omega k^2\eta^3$	$\frac{1}{5}\left(1+\frac{4}{15}f_\nu\right)(k\eta)^2$

Table 5.2: Asymptotic Synchronous gauge initial expansions ($v_{\rm c}^{\rm Z} \equiv 0$).

Quantity	**Adiabatic (ADI, $\mathcal{R}_{\boldsymbol{k}} = 1$)**	**Matter Isocurvature (CDI/BDI, $\mathcal{S}_{\rm mr} = 1$)**	**Neutrino Density Iso (NDI, $\mathcal{S}_{\nu r} = 3/4$)**
$\delta_{\rm c}^{\rm Z}$	$\frac{1}{4}(k\eta)^2 + \mathcal{O}(\eta^3)$	$1 - \frac{1}{2}\omega\eta + \frac{3}{16}\omega^2\eta^2$	$\frac{f_\nu}{8(1-f_\nu)}(k\eta)^2$
$\delta_\gamma^{\rm Z}$	$\frac{1}{3}(k\eta)^2 + \mathcal{O}(\eta^3)$	$-\frac{2}{3}\omega\eta + \frac{1}{4}\omega^2\eta^2$	$-\frac{f_\nu}{1-f_\nu} + \frac{f_\nu}{6(1-f_\nu)}(k\eta)^2$
$\delta_\nu^{\rm Z}$	$\frac{1}{3}(k\eta)^2 + \mathcal{O}(\eta^3)$	$-\frac{2}{3}\omega\eta + \frac{1}{4}\omega^2\eta^2$	$1 - \frac{1}{6}(k\eta)^2$
$v_{\rm c}^{\rm Z}$	0	0	0
$v_\gamma^{\rm Z}$	$\frac{1}{36}(k\eta)^3$	$-\frac{1}{12}\omega k\eta^2 + \frac{1}{48}\frac{1+3f_{\rm b}-f_\nu}{1-f_\nu}\omega^2 k\eta^3$	$-\frac{f_\nu}{4(1-f_\nu)}k\eta + \frac{3}{16}\frac{f_{\rm b}f_\nu}{(1-f_\nu)^2}\omega k\eta^2$
$v_\nu^{\rm Z}$	$\frac{1}{36}\frac{23+4f_\nu}{15+4f_\nu}(k\eta)^3$	$-\frac{1}{12}\omega k\eta^2 + \frac{1}{48}\omega^2 k\eta^3$	$\frac{1}{4}k\eta - \frac{1}{40}\frac{1+\frac{4}{27}f_\nu}{1+\frac{4}{15}f_\nu}(k\eta)^3$
Π_ν	$-\frac{4}{15}\frac{1}{1+\frac{4}{15}f_\nu}(k\eta)^2$	$-\frac{1}{15}\frac{1}{1+\frac{2}{15}f_\nu}\omega k^2\eta^3$	$\frac{1}{5}\left(1+\frac{4}{15}f_\nu\right)(k\eta)^2$

Chapter 6

The cosmic microwave background

The primary objective of this chapter is to lay down the theoretical foundations governing the physics of the cosmic microwave background (CMB) and to elucidate the detailed derivations of the angular temperature anisotropies predicted across various cosmological frameworks.

We begin in Section 6.1 with an accessible, intuitive derivation of the angular power spectrum of temperature anisotropies, dissecting the distinct physical mechanisms that shape the observed multipoles. We establish the physical origin of CMB fluctuations and quantify how the temperature field mapped across the contemporary sky directly connects to primordial fluctuations imprinted on the last-scattering surface at decoupling. This foundational analysis equips us to analyze the qualitative and quantitative morphology of the angular power spectrum in Section 6.2.

In Section 6.3, we transition to the rigorous, modern kinetic description of the cosmic microwave background. Grounded in relativistic kinetic theory, this formulation necessitates solving the linearized relativistic Boltzmann equations for coupled photons, baryons, and collisionless components within an expanding spacetime. Finally, Section 6.5 investigates how cosmological parameters (including the respective matter energy fractions, the Hubble expansion rate, and baryonic content) alongside primordial initial conditions dictate the detailed structure and acoustic peaks of the CMB power spectrum. General foundational treatments and comprehensive reviews can be found in Refs. [1–4].

6.1 Origin of the cosmic microwave background anisotropies

Following the recombination and decoupling epoch, photons effectively cease to interact with matter and propagate freely along null geodesics to leading order. A transparent, physically intuitive picture of cosmic microwave background anisotropies can therefore be constructed by evaluating the total gravitational and kinematic redshift accumulated by a photon traversing a perturbed cosmological metric from the last-scattering surface to the present-day observer.

Within this streamlined framework, matter is treated in the hydrodynamic fluid approximation, and the last-scattering surface is idealized as an infinitely thin screen—corresponding to an instantaneous decoupling transition (as detailed in Section 3.6 of Chapter 3). This perspective clarifies the underlying gravitational and Doppler effects without sacrificing physical insight, while the complete kinetic treatment incorporating finite thickness and Thomson visibility will be fully developed in Section 6.3.

6.1.1 Sachs–Wolfe formula

The Sachs–Wolfe relation provides the foundational bridge connecting the observed energy of a CMB photon today to its local emission energy at the decoupling surface (Refs. [5, 6]).

Propagation of a photon in a perturbed space-time

Consider the propagation of a photon in a perturbed spacetime whose metric $g_{\mu\nu}$ takes the general form introduced in Eq. (4.2.1). The trajectory of the photon is defined by a null worldline $x^\mu(\lambda)$, parameterized by an affine parameter λ. The null tangent vector $k^\mu \equiv \mathrm{d}x^\mu/\mathrm{d}\lambda$ satisfies the fundamental null constraint and the geodesic equation (see Chapter 1):

$$k^\mu k_\mu = 0, \qquad k^\mu \nabla_\mu k^\nu = 0.$$

Recall that two conformally related spacetimes possess identical causal structures, meaning that their null paths coincide identically. Introducing the conformally rescaled metric $\hat{g}_{\mu\nu}$ defined by $g_{\mu\nu} = a^2 \hat{g}_{\mu\nu}$, if k^μ is the tangent vector to an affinely parameterized null geodesic of $g_{\mu\nu}$, then the vector $\hat{k}^\mu = a^2 k^\mu$ is the tangent vector to a null geodesic of $\hat{g}_{\mu\nu}$, satisfying

$$\hat{g}_{\mu\nu}\hat{k}^\mu \hat{k}^\nu = 0, \qquad \hat{k}^\mu \hat{\nabla}_\mu \hat{k}^\nu = 0,$$

where $\hat{\nabla}_\mu$ denotes the Levi-Civita connection compatible with $\hat{g}_{\mu\nu}$.

We decompose the conformal tangent four-vector $\hat{k}^\mu$ as

$$\hat{k}^\mu = E\left(1 + M,\, \hat{e}^i + \delta\hat{e}^i\right), \tag{6.1.1}$$

where E is an unperturbed constant,[1] while $(M, \delta\hat{e}^i)$ represent first-order perturbation fields. The null constraint $\hat{g}_{\mu\nu}\hat{k}^\mu\hat{k}^\nu = 0$ reduces the four perturbations to three independent degrees of freedom:

$$\gamma_{ij}\hat{e}^i\hat{e}^j = 1, \qquad \hat{e}_j\delta\hat{e}^j = A + M - B_i\hat{e}^i - \frac{1}{2}h_{ij}\hat{e}^i\hat{e}^j. \tag{6.1.2}$$

Starting from $\hat{k}^\mu\hat{k}_\mu = 0$, we find

$$\begin{aligned} 0 &= -(1+2A)[E(1+M)]^2 + 2B_iE^2(\hat{e}^i + \delta\hat{e}^i)(1+M) + (\gamma_{ij} + h_{ij})E^2(\hat{e}^i + \delta\hat{e}^i)(\hat{e}^j + \delta\hat{e}^j) \\ &= -1 - 2A - 2M - 4AM + 2B_i(\hat{e}^i + \delta\hat{e}^i + \hat{e}^i M + \delta\hat{e}^i M) \\ &\quad + \cancel{\gamma_{ij}\hat{e}^i\hat{e}^j} + 2\gamma_{ij}\hat{e}^i\delta\hat{e}^j + \gamma_{ij}\delta\hat{e}^i\delta\hat{e}^j + h_{ij}\hat{e}^i\hat{e}^j + 2h_{ij}\hat{e}^i\delta\hat{e}^j + h_{ij}\delta\hat{e}^i\delta\hat{e}^j \end{aligned} \tag{1}$$

At first order, we find

$$\hat{e}_j\delta\hat{e}^j = A + M - B_i\hat{e}^i - \frac{1}{2}h_{ij}\hat{e}^i\hat{e}^j \tag{2}$$

To first order in metric and kinematic fluctuations, the geodesic equation for $\hat{k}^\mu$ simplifies to

$$\frac{\mathrm{d}\delta\hat{k}^\nu}{\mathrm{d}s} \equiv \hat{k}^\mu\partial_\mu\delta\hat{k}^\nu = -\delta\hat{\Gamma}^\nu_{\mu\rho}\hat{k}^\mu\hat{k}^\rho. \tag{6.1.3}$$

Evaluating the timelike component ($\nu = 0$) using the decomposition (6.1.1) yields the evolution equation for the energy perturbation M:

$$\frac{\mathrm{d}M}{\mathrm{d}s} = -A' - 2\hat{e}^i\partial_i A - \frac{1}{2}h'_{ij}\hat{e}^i\hat{e}^j + D_iB_j\hat{e}^i\hat{e}^j, \tag{6.1.4}$$

where we have defined the total derivative along the background path as $\mathrm{d}M/\mathrm{d}s \equiv \hat{e}^i\partial_i M + \partial_0 M$.

[1] As an elementary consistency check, note that in an unperturbed Friedmann–Lemaître universe, the comoving observer four-velocity satisfies $u^0 \propto a^{-1}$, so that the physical photon energy scales as $k^\mu u_\mu = k^0 u_0 = (\hat{k}^0/a^2)a \propto 1/a$, reflecting the cosmological redshift since $E = \hat{k}^0$ remains constant.

Sachs–Wolfe formula

As established in Chapter 1 [see Eq. (1.4.10)], the frequency and energy of a photon registered by an observer with four-velocity u^μ is given by $E = -k^\mu u_\mu$. Hence, the ratio of the photon energy detected today at η_0 to that emitted at $(\boldsymbol{x}_\mathrm{E}, \eta_\mathrm{E})$ is

$$\frac{E_0(\boldsymbol{e})}{E_\mathrm{E}(\boldsymbol{x}_\mathrm{E}, \eta_\mathrm{E})} = \frac{(k^\mu u_\mu)_0}{(k^\mu u_\mu)_\mathrm{E}}, \tag{6.1.5}$$

where u^μ is the baryon fluid four-velocity. In the unperturbed background, the emission and reception coordinates are linked by the light-cone trajectory

$$\boldsymbol{x}_\mathrm{E} = \boldsymbol{x}_0 + \boldsymbol{e}(\eta_0 - \eta_\mathrm{E}), \tag{6.1.6}$$

with $\boldsymbol{e} \equiv -\hat{\boldsymbol{e}}$ designating the line-of-sight unit vector pointing from the observer back toward the source. Throughout this analysis, we adopt the Born approximation, evaluating all perturbative expressions along this unperturbed background geodesic.

From Eq. (6.1.5), we deduce that the radiation temperature observed along direction $\boldsymbol{e}$ is connected to the emission temperature by $f(\eta_0) = f(\eta_\mathrm{E})$

$$\frac{E_0}{T_0} = \frac{E_\mathrm{E}}{T_\mathrm{E}} \implies \frac{T_0(\boldsymbol{e})}{T_\mathrm{E}(\boldsymbol{x}_\mathrm{E}, \eta_\mathrm{E})} = \frac{a(\eta_\mathrm{E})}{a(\eta_0)} \left\{1 + \left[M + A + e^i(v_{\mathrm{b}i} + B_i)\right]_\mathrm{E}^0\right\}, \tag{6.1.7}$$

we used

$$(k^\mu)_0 = \frac{E}{a^2(\eta_0)}(1 + M, \hat{e}^i + \delta\hat{e}^i)_0 \tag{1}$$

$$(k^\mu)_\mathrm{E} = \frac{E}{a^2(\eta_\mathrm{E})}(1 + M, \hat{e}^i + \delta\hat{e}^i)_\mathrm{E} \tag{2}$$

Then,

$$\frac{E_0}{E_\mathrm{E}} = \frac{a(\eta_\mathrm{E})}{a(\eta_0)} \frac{(1 + M, -\hat{e}^i + \delta\hat{e}^i) \cdot (-1 - A, v_k + B_k)|_0}{(1 + M, -\hat{e}^i + \delta\hat{e}^i) \cdot (-1 - A, v_k + B_k)|_\mathrm{E}} \tag{3}$$

$$= \frac{a(\eta_\mathrm{E})}{a(\eta_0)} \frac{[-1 - M_0 - A_0 - \hat{e}^i(v_k + B_i)_0 + \mathcal{O}(\epsilon^2)]}{[-1 - M_\mathrm{E} - A_\mathrm{E} - \hat{e}^i(v_k + B_i)_\mathrm{E} + \mathcal{O}(\epsilon^2)]} \tag{4}$$

$$= \frac{a(\eta_\mathrm{E})}{a(\eta_0)} \left\{1 + \left[M + A + e^i(v_{\mathrm{b}i} + B_i)\right]_\mathrm{E}^0\right\} \tag{5}$$

where we substituted the perturbed four-velocity $u_\mu = a(-1 - A, v_k + B_k)$, with $v_{\mathrm{b}i}$ being the velocity perturbation of the baryon fluid (see Chapter 4).

Because the recombination epoch is dictated by the photon–electron interaction, governed by the interaction coefficient $\sigma_\mathrm{T} n_\mathrm{e}$ (see Chapter 3, Section 3.6), the last-scattering surface can be modeled as a hypersurface of constant electron number density, $n_\mathrm{e} = \text{const}$. Owing to local density fluctuations, recombination does not transpire synchronously across all points in the Universe (as illustrated in Fig. 6.1). The conformal time of emission is accordingly decomposed as

$$\eta_\mathrm{E} = \bar{\eta}_\mathrm{E} + \delta\eta_\mathrm{E}.$$

The emission and observed temperature fields are decomposed into spatial averages plus fractional contrasts:

$$T_\mathrm{E}(\boldsymbol{x}_\mathrm{E}, \eta_\mathrm{E}) = \bar{T}_\mathrm{E}(\eta_\mathrm{E})\left[1 + \Theta_\mathrm{E}(\boldsymbol{x}_\mathrm{E}, \eta_\mathrm{E})\right],$$

$$T_0(\boldsymbol{e}) = \bar{T}_0(\eta_0)\left[1 + \Theta_0(\boldsymbol{e})\right],$$

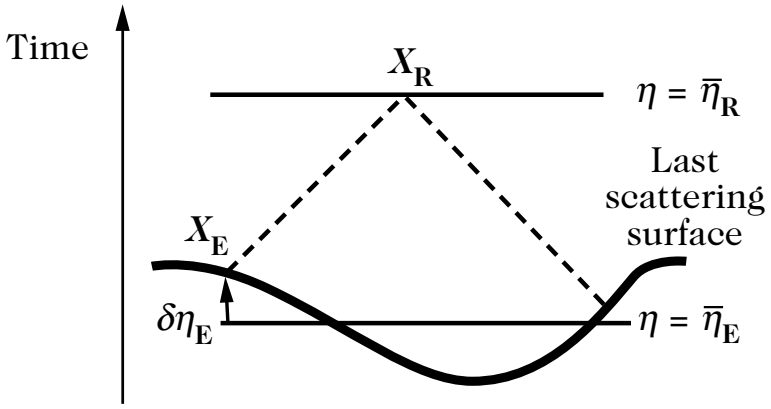


Figure 6.1: The last-scattering surface is a surface of constant electron number density, $n_e = \text{const}$. Thus, it is not a constant-time hypersurface. The time of emission is related to the mean time by $\eta_E = \bar{\eta}_E + \delta\eta_E$.

where $\bar{T}_E$ is the spatial average across the emission hypersurface at time η_E, and $\Theta_0(\boldsymbol{e})$ denotes the temperature contrast observed in direction $\boldsymbol{e}$ compared to the average sky temperature $\bar{T}_0(\eta_0)$.

Substituting these two expansions into Eq. (6.1.7), we find that at lowest order in perturbations

$$\frac{\text{present day CMB temp.}}{\text{CMB temp. during emission}} = \frac{\bar{T}_0}{\bar{T}_E} = \frac{a(\bar{\eta}_E)}{a(\eta_0)}, \tag{6.1.8}$$

which is simply the standard cosmological redshift scaling of blackbody radiation in a homogeneous and isotropic expanding spacetime (see Chapter 3).

Using the relation $\bar{T}_E(\eta_E)a(\eta_E) = \bar{T}_E(\bar{\eta}_E)[1 + (T'/T)\delta\eta_E]a(\bar{\eta}_E)[1 + (a'/a)\delta\eta_E] = \bar{T}_E(\bar{\eta}_E)a(\bar{\eta}_E)$, valid to first order in perturbations, we infer that the observed temperature contrast in the direction $\boldsymbol{e}$ is given by

$$\frac{T_0}{T_E} = \frac{1 + \Theta_0(\mathbf{e})}{1 + \Theta_E(\boldsymbol{x}_E, \bar{\eta}_E)} \frac{a(\bar{\eta}_E)}{a(\eta_0)} \implies \Theta_0(\boldsymbol{e}) = \Theta_E(\boldsymbol{x}_E, \bar{\eta}_E) + \Big[M + A + e^i(v_{bi} + B_i)\Big]_E^0. \tag{6.1.9}$$

Again, within the Born approximation, the bracketed first-order quantity may be evaluated either at $\bar{\eta}_E$ or $\bar{\eta}_R$.

To evaluate the intrinsic emission fluctuation Θ_E, we observe that the free electron density n_e is a function of the photon and baryon energy densities, ρ_γ and ρ_b. As evidenced by the Saha ionization equation (Ref. [6]), the effect of baryons can be neglected to a good approximation, so that the last-scattering surface coincides closely with the surface $\rho_\gamma = \text{const}$ (see Chapter 3). Applying the Stefan–Boltzmann law $\rho_\gamma \propto T^4$, we obtain

$$\frac{\delta\rho_r}{\rho_r} = 4\frac{\delta T}{T} \implies \Theta_E(\boldsymbol{x}_E, \bar{\eta}_E) = \frac{1}{4}\delta_\gamma(\boldsymbol{x}_E, \bar{\eta}_E). \tag{6.1.10}$$

To calculate the contribution $[M]_E^0$, we integrate the geodesic equation (6.1.4) along the unperturbed ray:

$$[M]_E^0 = -2[A]_E^0 + \int_E^0 \left(A' - \frac{1}{2}h'_{ij}e^ie^j + D_iB_je^ie^j\right) \mathrm{d}s\,, \tag{6.1.11}$$

where the integration is performed along the unperturbed null path.

From (6.1.4)

$$\frac{\mathrm{d}M}{\mathrm{d}s} = A' - 2\frac{\mathrm{d}A}{\mathrm{d}s} - \frac{1}{2}h'_{ij}\hat{e}^i\hat{e}^j + D_iB_j\hat{e}^i\hat{e}^j, \tag{1}$$

using (4.2.5)

$$\frac{1}{2}h'_{ij} = C'\gamma_{ij} + D_iD_jE' + \frac{1}{2}(D_i\bar{E}'_j + D_j\bar{E}'_i) + \bar{E}'_{ij} \qquad \text{with} \qquad D_i\bar{E}^{ij} = 0, \quad \bar{E}^i{}_i = 0. \tag{2}$$

we find:

$$M|_{\mathrm{E}}^0 = -2A|_{\mathrm{E}}^0 + \int_{\mathrm{E}}^0 \left[A' - \left(C'\gamma_{ij} + D_iD_jE' + D_i\bar{E}'_j + \bar{E}'_{ij} + D_iB_j\right)\hat{e}^i\hat{e}^j\right] ds, \tag{3}$$

Combining Eqs. (6.1.9), (6.1.10), and (6.1.11), we deduce:

$$\begin{aligned}\Theta_0(\boldsymbol{e}) = &\left[\frac{1}{4}\delta_\gamma + A - e^i(v_{\mathrm{b}i} + B_i)\right](\boldsymbol{x}_{\mathrm{E}}, \bar{\eta}_{\mathrm{E}}) \\ &+ \int_{\mathrm{E}}^0 \left[A' - C' - e^ie^j\left(D_iD_jE' + D_i\overline{E}'_j - D_iB_j + \overline{E}'_{ij}\right)\right]\mathrm{d}s + f(0),\end{aligned} \tag{6.1.12}$$

where $f(0)$ gathers terms evaluated today at the observer position. Rewriting this in terms of gauge-invariant cosmological perturbations, and utilizing the identity $e^jD_j(\overline{E}'_i - \overline{B}_i) = e^jD_j\overline{\Phi}_i = -\overline{\Phi}'_i + \mathrm{d}\overline{\Phi}_i/\,\mathrm{d}s$, this expression takes the form

$$\boxed{\begin{aligned}\Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e}) = &\left[\frac{1}{4}\delta^{\mathrm{N}}_\gamma + \Phi - e^i\left(D_iV_{\mathrm{b}} + \overline{V}_{\mathrm{b}i} + \overline{\Phi}_i\right)\right](\boldsymbol{x}_{\mathrm{E}}, \bar{\eta}_{\mathrm{E}}) \\ &+ \int_{\mathrm{E}}^0 (\Phi' + \Psi')[\boldsymbol{x}(\eta), \eta]\mathrm{d}\eta - \int_{\mathrm{E}}^0 e^i\overline{\Phi}'_i[\boldsymbol{x}(\eta), \eta]\mathrm{d}\eta \\ &- \int_{\mathrm{E}}^0 e^ie^j\overline{E}'_{ij}[\boldsymbol{x}(\eta), \eta]\mathrm{d}\eta,\end{aligned}} \tag{6.1.13}$$

where we integrate along the unperturbed geodesic parameterized in conformal time as

$$\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{e}(\eta_0 - \eta), \tag{6.1.14}$$

so that the integral runs directly over η. We have intentionally dropped the local monopole shift $f(0)$, which depends solely on the observer location and cannot be measured observationally. Equation (6.1.13) is celebrated as the Sachs–Wolfe equation.

Interpretation

The general Sachs–Wolfe formula (6.1.13) decomposes naturally into the sum of three distinct physical contributions:

$$\Theta_{\mathrm{SW}} = \left(\frac{1}{4}\delta^{\mathrm{N}}_\gamma + \Phi\right)(\boldsymbol{x}_{\mathrm{E}}, \bar{\eta}_{\mathrm{E}}) = \left(\frac{1}{4}\delta^{\mathrm{F}}_\gamma + \Phi + \Psi\right)(\boldsymbol{x}_{\mathrm{E}}, \bar{\eta}_{\mathrm{E}}), \tag{6.1.15}$$

$$\Theta_{\mathrm{dop}} = -e^i\left(V_{\mathrm{b}i} + \overline{\Phi}_i\right)(\boldsymbol{x}_{\mathrm{E}}, \bar{\eta}_{\mathrm{E}}), \tag{6.1.16}$$

$$\Theta_{\mathrm{ISW}} = \int_{\mathrm{E}}^0 \left[(\Phi' + \Psi') - e^i\overline{\Phi}'_i - e^ie^j\overline{E}'_{ij}\right]\mathrm{d}\eta. \tag{6.1.17}$$

The first term receives exclusively scalar contributions and is termed the ordinary Sachs–Wolfe effect; the second contains scalar and vector contributions representing the Doppler term; while

the third term integrates across all three perturbation species along the photon path and is known as the integrated Sachs–Wolfe (ISW) term.

The ordinary Sachs–Wolfe term $\Theta_{\rm SW}$ is evaluated locally at decoupling $(\boldsymbol{x}_{\rm E}, \bar{\eta}_{\rm E})$ and reflects two competing physical effects:

- the intrinsic density contrast of the radiation fluid, reflecting that overdense regions are hotter by virtue of the Stefan–Boltzmann law;
- the local gravitational potential, indicating that a photon emitted inside a gravitational potential well experiences an additional gravitational redshift (the Einstein effect).

The Doppler term $\Theta_{\rm dop}$ arises purely from the relative kinematic velocity between the emitting plasma and the observing reference frame.

The integral term $\Theta_{\rm ISW}$ depends upon the photon's propagation history between emission and reception, containing the conformal time derivatives of the gravitational potentials in the SVT decomposition. These terms receive significant contributions from:

- cosmic structures in the process of gravitational formation traversed by the photon: as a structure collapses, its gravitational potential deepens, breaking time symmetry as the photon enters and exits the well, imprinting a net frequency shift;
- late-time cosmic expansion in a universe with non-vanishing dark energy or spatial curvature, wherein the gravitational potentials decay in the recent cosmological epoch.

6.1.2 Angular power spectrum

The Sachs–Wolfe equation relates the observed directional temperature fluctuations to primordial cosmological perturbations. Thus, $\Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e})$ constitutes a stochastic field on the celestial sphere, characterized statistically by its two-point angular correlation function:

$$C(\vartheta) = \langle \Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e}_1)\Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e}_2)\rangle. \tag{6.1.18}$$

Statistical isotropy, guaranteed by the cosmological principle, dictates that this correlation function depends only on the separation angle ϑ between the observation unit vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, with $\cos\vartheta = \boldsymbol{e}_1 \cdot \boldsymbol{e}_2$. It is accordingly convenient to expand $C(\vartheta)$ in the basis of Legendre polynomials:

$$C(\vartheta) = \sum_\ell \frac{2\ell+1}{4\pi} C_\ell P_\ell(\boldsymbol{e}_1 \cdot \boldsymbol{e}_2), \tag{6.1.19}$$

which defines the angular power spectrum C_ℓ. A multipole moment ℓ corresponds approximately to an angular aperture π/ℓ on the sky (see Appendix B), making C_ℓ a direct probe of the temperature fluctuation variance at that angular scale. For Gaussian temperature fluctuations, this spectrum completely specifies the statistical distribution.

Expanding $\Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e})$ in spherical harmonics,

$$\Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e}) = \sum_{\ell\geq 1} \sum_{m=-\ell}^{\ell} a_{\ell m}(\boldsymbol{x}_0, \eta_0) Y_{\ell m}(\boldsymbol{e}), \tag{6.1.20}$$

the expansion multipole coefficients are given by

$$a_{\ell m}(\boldsymbol{x}_0, \eta_0) = \int \mathrm{d}^2\boldsymbol{e}\, \Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e}) Y^*_{\ell m}(\boldsymbol{e}). \tag{6.1.21}$$

Using the addition theorem [Eq. (B.17)] to represent the Legendre polynomial in Eq. (6.1.19) and inverting using the orthogonality of spherical harmonics [Eq. (B.15)], we find

$$(2\ell+1)C_\ell = \sum_m \langle a_{\ell m}(\boldsymbol{x}_0, \eta_0) a^*_{\ell m}(\boldsymbol{x}_0, \eta_0)\rangle. \tag{6.1.22}$$

Since $a_{00} = 0$, and the $\ell = 1$ multipole is dominated by the Doppler effect arising from the motion of the Solar System with respect to the surface of last scattering, which cannot be cleanly separated from the observed signal, we restrict our analysis to $\ell \geq 2$.

We now proceed to compute this angular power spectrum as a function of the cosmological perturbations within the SVT decomposition. For each perturbation mode, we Fourier expand $\Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e})$ as

$$\Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e}) = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^{3/2}} \hat{\Theta}(\eta_0, \boldsymbol{k}, \boldsymbol{e}), \tag{6.1.23}$$

where the phase $\exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}_0)$ has been absorbed into the definition of $\hat{\Theta}$. It follows that

$$a_{\ell m}(\boldsymbol{x}_0, \eta_0) = \int \mathrm{d}^2\boldsymbol{e} \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^{3/2}} \hat{\Theta}(\eta_0, \boldsymbol{k}, \boldsymbol{e}) Y^*_{\ell m}(\boldsymbol{e}). \tag{6.1.24}$$

Henceforth, the explicit dependence on $(\boldsymbol{x}_0, \eta_0)$ in $a_{\ell m}$ is omitted.

Scalar modes

After substituting $\boldsymbol{x}_\mathrm{E}$ via Eq. (6.1.6), the scalar sector of Eq. (6.1.13) indicates that

$$\hat{\Theta}(\eta_0, \boldsymbol{k}, \boldsymbol{e}) = \left[\hat{\Theta}_\mathrm{SW}(\boldsymbol{k}) - \mathrm{i}k\mu\hat{V}_\mathrm{b}(\boldsymbol{k})\right] \mathrm{e}^{\mathrm{i}k\mu\Delta\eta_\mathrm{E}} + \int (\hat{\Phi}' + \hat{\Psi}')\mathrm{e}^{\mathrm{i}k\mu\Delta\eta}\mathrm{d}\eta, \tag{6.1.25}$$

with $\Delta\eta \equiv \eta_0 - \eta$, $\Delta\eta_\mathrm{E} \equiv \eta_0 - \bar{\eta}_\mathrm{E}$, and $\mu \equiv (\boldsymbol{k}\cdot\boldsymbol{e})/k$. The exponential factor $\exp(\mathrm{i}k\mu\Delta\eta)$ appears under the integral because the spatial Fourier transform contains $(\hat{\Phi}' + \hat{\Psi}')\exp[\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}(\eta)]$, while Eq. (6.1.14) implies $\exp[\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}(\eta)] = \exp(\mathrm{i}k\mu\Delta\eta)\exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}_0)$, with the constant phase absorbed into the Fourier amplitude [see Eq. (6.1.23)]. An identical mechanism accounts for the factor $\exp(\mathrm{i}k\mu\Delta\eta_\mathrm{E})$ in the first two terms.

Each term in Eq. (6.1.25) is a stochastic variable factorizing as $X(\boldsymbol{k}, \eta) = X(k, \eta)a(\boldsymbol{k})$, where $a(\boldsymbol{k})$ satisfies the random Gaussian property

$$\langle a(\boldsymbol{k})a^*(\boldsymbol{k}')\rangle = \delta^{(3)}(\boldsymbol{k} - \boldsymbol{k}'), \tag{6.1.26}$$

such that $\hat{\Theta}(\eta_0, \boldsymbol{k}, \boldsymbol{e}) = \hat{\Theta}(\eta_0, k, \mu)a(\boldsymbol{k})$ with

$$\hat{\Theta}(\eta_0, k, \mu) = \left[\hat{\Theta}_\mathrm{SW}(k) - \hat{V}_\mathrm{b}(k)\partial_{\Delta\eta_\mathrm{E}}\right] \mathrm{e}^{\mathrm{i}k\mu\Delta\eta_\mathrm{E}} + \int (\hat{\Phi}' + \hat{\Psi}')\mathrm{e}^{\mathrm{i}k\mu\Delta\eta}\mathrm{d}\eta. \tag{6.1.27}$$

Using the spherical harmonic plane-wave expansion [Eq. (B.21)] for the exponential factor, we obtain

$$\hat{\Theta}(\eta_0, \boldsymbol{k}, \mu) = 4\pi \sum_{\ell m} \hat{\Theta}^{(\mathrm{S})}_\ell(k) Y^*_{\ell m}(\hat{\boldsymbol{k}}) Y_{\ell m}(\boldsymbol{e}), \tag{6.1.28}$$

with the scalar multipole transfer function

$$\boxed{\hat{\Theta}^{(\mathrm{S})}_\ell(k) = \hat{\Theta}_\mathrm{SW}(k) j_\ell(k\Delta\eta_\mathrm{E}) - \frac{\hat{V}_\mathrm{b}(k)}{k} j'_\ell(k\Delta\eta_\mathrm{E}) + \int (\hat{\Phi}' + \hat{\Psi}') j_\ell(k\Delta\eta)\mathrm{d}\eta.} \tag{6.1.29}$$

Inserting these expressions into Eq. (6.1.22) and applying Eq. (6.1.26), we arrive at

$$\begin{aligned}(2\ell+1)C^\mathrm{S}_\ell &= \sum_m \int \mathrm{d}^2\boldsymbol{e}_1 \mathrm{d}^2\boldsymbol{e}_2 \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} \Theta(\eta_0, \boldsymbol{k}, \mu_1)\Theta^*(\eta_0, \boldsymbol{k}, \mu_2) Y_{\ell m}(\boldsymbol{e}_1) Y^*_{\ell m}(\boldsymbol{e}_2) \\ &= \frac{2}{\pi} \int k^2\,\mathrm{d}k \sum_{\ell_1 m_1 \ell_2 m_2 m} \hat{\Theta}^{(\mathrm{S})}_{\ell_1}(k) \hat{\Theta}^{(\mathrm{S})*}_{\ell_2}(k) \int \mathrm{d}^2\hat{\boldsymbol{k}} \mathrm{d}^2\boldsymbol{e}_1 \mathrm{d}^2\boldsymbol{e}_2\end{aligned}$$

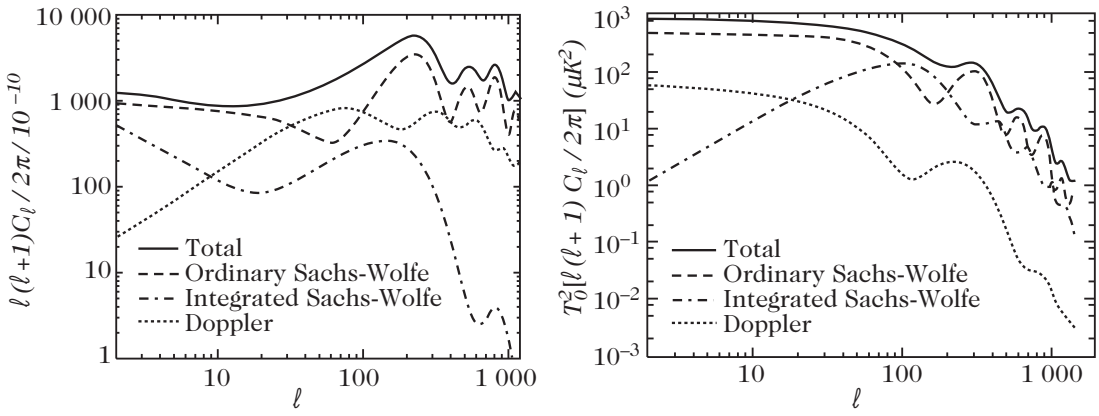


Figure 6.2: The angular power spectrum and its three contributions, ordinary Sachs–Wolfe, Doppler and integrated Sachs–Wolfe for adiabatic (left) and isocurvature (right) initial conditions and a scale invariant primordial spectrum ($n_{\rm S} = 1$), as predicted by inflation (see Chapter 9).

$$\times\, Y_{\ell_1 m_1}(\hat{\boldsymbol{k}}) Y^*_{\ell_2 m_2}(\hat{\boldsymbol{k}}) Y_{\ell m}(\boldsymbol{e}_1) Y^*_{\ell_1 m_1}(\boldsymbol{e}_1) Y_{\ell_2 m_2}(\boldsymbol{e}_2) Y^*_{\ell m}(\boldsymbol{e}_2).$$

Integration over the unit wavevector $\hat{\boldsymbol{k}} = \boldsymbol{k}/k$ yields the factor $\delta_{\ell_1\ell_2}\delta_{m_1 m_2}$, while integrating over $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ produces $\delta_{\ell_1\ell}\delta_{m_1 m}$ and $\delta_{\ell\ell_2}\delta_{m m_2}$, respectively, yielding the boxed scalar angular power spectrum

$$\boxed{C_\ell^{\rm S} = \frac{2}{\pi}\int \left|\hat{\Theta}_\ell^{(\rm S)}(k)\right|^2 k^2\,{\rm d}k\,.} \tag{6.1.30}$$

Figure 6.2 decomposes the scalar angular power spectrum into its ordinary Sachs–Wolfe, Doppler, and integrated Sachs–Wolfe contributions.

Vector modes

We now perform the corresponding computation for the vector contributions. The Fourier expansion (6.1.23) yields (from (6.1.13))

$$\hat{\Theta}(\eta_0, \boldsymbol{k}, \boldsymbol{e}) = -e^i\left[\overline{V}_{{\rm b}i}(\boldsymbol{k}) + \overline{\Phi}_i(\boldsymbol{k})\right]{\rm e}^{{\rm i}k\mu\Delta\eta_{\rm E}} - \int e^i\overline{\Phi}'_i{\rm e}^{{\rm i}k\mu\Delta\eta}{\rm d}\eta. \tag{6.1.31}$$

Because vector perturbations are transverse, their Fourier coefficients possess no components longitudinal to the wavevector $\boldsymbol{k}$. They can thus be expanded in an orthonormal basis $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ perpendicular to $\boldsymbol{k}$ as

$$\overline{X}_i(\boldsymbol{k}, \eta) = \sum_{\lambda=\pm}\overline{X}_\lambda(k,\eta)e_i^\lambda(\boldsymbol{k})a_\lambda(\boldsymbol{k}), \tag{6.1.32}$$

with the complex circular polarization vectors $e^i_\pm \equiv e^i_1 \pm {\rm i}e^i_2$, satisfying $e^i_\pm e_i = \sin\theta\exp(\pm{\rm i}\varphi)$. The two helicity states are statistically independent, with the random amplitudes satisfying

$$\langle a_\lambda(\boldsymbol{k})a^*_{\lambda'}(\boldsymbol{k}')\rangle = \delta^{(3)}(\boldsymbol{k}-\boldsymbol{k}')\delta_{\lambda\lambda'}. \tag{6.1.33}$$

Using Eq. (B.21) to expand the plane-wave exponential in spherical harmonics, we obtain

$$\hat{\Theta}(\eta_0, \boldsymbol{k}, \boldsymbol{e}) = 4\pi\sum_{\lambda=\pm}\sum_{\ell m}\left(\sin\vartheta{\rm e}^{{\rm i}\lambda\varphi}\right)\widetilde{\Theta}^{(\rm V)}_{\lambda,\ell}(k)Y_{\ell m}(\boldsymbol{k})Y^*_{\ell m}(\boldsymbol{e})a_\lambda(\boldsymbol{k}), \tag{6.1.34}$$

with

$$\widetilde{\Theta}^{(\rm V)}_{\lambda,\ell}(k) = -\left[\overline{V}_\lambda(k) + \overline{\Phi}_\lambda(k)\right]j_\ell(k\Delta\eta_{\rm E}) - \int\overline{\Phi}'_\lambda(k)j_\ell(k\Delta\eta){\rm d}\eta. \tag{6.1.35}$$

Inserting these expressions into Eq. (6.1.22), using the statistical orthogonality relation (6.1.33), and performing the integration over $\mathrm{d}^2\hat{\boldsymbol{k}}$, we obtain

$$(2\ell+1)C_\ell^{\mathrm{V}} = \frac{2}{\pi}\int k^2\,\mathrm{d}k \sum_{\ell_1,\lambda}\left|\widetilde{\Theta}^{(\mathrm{V})}_{\lambda,\ell_1}(k)\right|^2 \sum_{mm_1}\left|\int \mathrm{d}^2\boldsymbol{e}\,\sin\vartheta \mathrm{e}^{\mathrm{i}\lambda\varphi}Y_{\ell m}(\boldsymbol{e})Y^*_{\ell_1 m_1}(\boldsymbol{e})\right|^2.$$

To calculate the integral over $\boldsymbol{e}$, we first note that $\sin\vartheta\mathrm{e}^{\mathrm{i}\lambda\varphi} = -\lambda 2\sqrt{2\pi/3}Y_{1\lambda}(\boldsymbol{e})$ and evaluate the triple spherical harmonic integral using Eq. (B.23). The summation over m and m_1 is then performed via the orthogonality relations of the Wigner $3j$-symbols [Eq. (B.25)]. This sum restricts the non-vanishing contributions strictly to $\ell_1 = \ell \pm 1$, causing the spherical Bessel functions j_{ℓ_1} appearing in $\widetilde{\Theta}^{(\mathrm{V})}_{\lambda,\ell_1}$ to recombine into the composite radial function [see Eq. (B.46)]:

$$j_\ell^{(11)}(x) = \sqrt{\frac{\ell(\ell+1)}{2}}\frac{j_\ell(x)}{x}. \tag{6.1.36}$$

We finally obtain the boxed vector angular power spectrum and transfer function:

$$\boxed{\begin{aligned} C_\ell^{\mathrm{V}} &= \frac{2}{\pi}\int \sum_\lambda \left|\hat{\Theta}^{(\mathrm{V})}_{\lambda,\ell}(k)\right|^2 k^2\,\mathrm{d}k\,, && (6.1.37)\\ \hat{\Theta}^{(\mathrm{V})}_{\lambda,\ell}(k) &= [\overline{V}_\lambda(k)+\overline{\Phi}_\lambda(k)]j_\ell^{(11)}(k\Delta\eta_{\mathrm{E}}) + \int \overline{\Phi}'_\lambda(k)j_\ell^{(11)}(k\Delta\eta)\mathrm{d}\eta. && (6.1.38)\end{aligned}}$$

Tensor modes

The computation for tensor modes proceeds along the same mathematical lines as that for vector modes. The Fourier expansion (6.1.23) yields

$$\hat{\Theta}(\eta_0,\boldsymbol{k},\boldsymbol{e}) = -\int e^j e^i \overline{E}'_{ij}\mathrm{e}^{\mathrm{i}k\mu\Delta\eta}\mathrm{d}\eta. \tag{6.1.39}$$

The transverse-traceless metric tensor $\overline{E}_{ij}$ is expanded as

$$\overline{E}_{ij}(\boldsymbol{k},\eta) = \sum_\lambda \overline{E}_\lambda(k,\eta)\varepsilon^\lambda_{ij}a_\lambda(\boldsymbol{k}), \tag{6.1.40}$$

where the random amplitude $a_\lambda(\boldsymbol{k})$ satisfies Eq. (6.1.33) and the polarization tensors are defined by

$$\varepsilon^\lambda_{ij} = e^+_i e^+_j \delta^\lambda_+ + e^-_i e^-_j \delta^\lambda_\times. \tag{6.1.41}$$

In particular, this implies that $\varepsilon^\lambda_{ij}e^i e^j = \sin^2\theta\exp(2\mathrm{i}\lambda\varphi)$.

Applying Eq. (B.21) to expand the plane-wave exponential factor, we find

$$\hat{\Theta}(\eta_0,\boldsymbol{k},\boldsymbol{e}) = 4\pi\sum_{\lambda=\pm}\sum_{\ell m}\left(\sin^2\vartheta\mathrm{e}^{2\mathrm{i}\lambda\varphi}\right)\widetilde{\Theta}^{(\mathrm{T})}_{\lambda,\ell}(k)Y_{\ell m}(\boldsymbol{k})Y^*_{\ell m}(\boldsymbol{e})a_\lambda(\boldsymbol{k}), \tag{6.1.42}$$

with

$$\widetilde{\Theta}^{(\mathrm{T})}_{\lambda,\ell}(k) = -\int \overline{E}'_\lambda(k)j_\ell(k\Delta\eta)\mathrm{d}\eta. \tag{6.1.43}$$

Inserting these expressions into Eq. (6.1.22), using Eq. (6.1.33), and performing the integration over $\mathrm{d}^2\hat{\boldsymbol{k}}$, we find that

$$(2\ell+1)C_\ell^{\mathrm{T}} = \frac{2}{\pi}\int k^2\,\mathrm{d}k \sum_{\ell_1,\lambda}\left|\widetilde{\Theta}^{(\mathrm{T})}_{\lambda,\ell_1}(k)\right|^2 \sum_{mm_1}\left|\int \mathrm{d}^2\boldsymbol{e}\,\sin^2\vartheta \mathrm{e}^{2\mathrm{i}\lambda\varphi}Y_{\ell m}(\boldsymbol{e})Y^*_{\ell_1 m_1}(\boldsymbol{e})\right|^2.$$

To evaluate the angular integral over $\boldsymbol{e}$, the procedure mirrors the vector derivation, using the identity $\sin^2\vartheta\mathrm{e}^{\pm 2\mathrm{i}\varphi} = 4\sqrt{2\pi/15}Y_{2\pm2}(\boldsymbol{e})$. Summation over the Wigner $3j$-symbols restricts

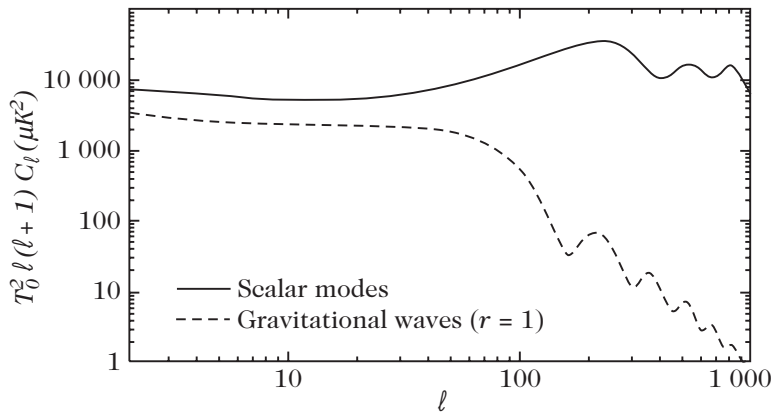


Figure 6.3: Angular power spectrum for the tensor modes, compared to that of scalars for a ratio $r = 1$ between the primordial spectra of the two modes.

surviving terms to $\ell_1 = \ell \pm 2$ and $\ell_1 = \ell$. Using the Bessel relation (B.47), the j_{ℓ_1} factors combine into the tensor radial function [see Eq. (B.46)]:

$$\tilde{j}_\ell^{(22)}(x) = \sqrt{\frac{(\ell+2)!}{(\ell-2)!}}\frac{j_\ell(x)}{x^2}. \tag{6.1.44}$$

We finally obtain the boxed tensor angular power spectrum and transfer function:

$$\boxed{C_\ell^{\mathrm{T}} = \frac{1}{\pi}\int \sum_\lambda \left|\hat{\Theta}_{\lambda,\ell}^{(\mathrm{T})}(k)\right|^2 k^2\,\mathrm{d}k\,,} \tag{6.1.45}$$

$$\boxed{\hat{\Theta}_{\lambda,\ell}^{(\mathrm{T})}(k) = -\int \overline{E}'_\lambda(k)\tilde{j}_\ell^{(22)}(k\Delta\eta)\mathrm{d}\eta.} \tag{6.1.46}$$

6.2 Properties of the angular power spectrum

The previous description allows us to understand the general shape of the power spectra of the scalar and tensor modes. The vector modes will be neglected in the rest of this chapter. Complementary references on the study of the properties of the temperature power spectrum can be found in Refs. [2, 7].

6.2.1 Large angular scales

On sufficiently large angular scales, corresponding to perturbation wavelengths that were super-Hubble at the recombination epoch, the gravitational potential remains effectively constant throughout the matter-dominated era (see Chapter 4). For such super-horizon modes, one can analytically evaluate the low-ℓ angular power spectrum directly in terms of the primordial power spectrum of metric fluctuations. This low-multipole regime of the angular spectrum constitutes the celebrated Sachs–Wolfe plateau.

Adiabatic perturbations

The primordial adiabatic initial conditions [Eq. (4.4.19)] dictate that initially $\delta_\gamma^{\mathrm{N}}(0) = -2\Psi(0)$ in the Newtonian gauge, meaning that radiation overdensities coincide precisely with potential wells, whilst the initial radiation velocity field vanishes, $V_\gamma(0) = 0$. On scales extending far beyond the Hubble radius at decoupling, causal causal transport prevents V_γ from undergoing

appreciable dynamical growth. Equation (5.255) then demonstrates that the linear combination $\delta_\gamma^{\rm N} - 4\Psi$ is conserved. During the subsequent matter-dominated era, the fluid relations $\delta_\gamma^{\rm N}/4 = \delta_{\rm m}^{\rm N}/3 = -2\Phi/3$ [cf. Eqs. (4.4.16) and (4.4.23)] yield the net ordinary Sachs–Wolfe fluctuation:

$$\Theta_{\rm SW} = \frac{\delta_{\rm r}^{\rm N}}{4} + \Phi = \left(-\frac{2}{3} + 1\right)\Phi = \frac{1}{3}\Phi, \tag{6.2.1}$$

which represents the dominant contribution to large-scale temperature anisotropies, in full accordance with the numerical decomposition presented in Fig. 6.2.

Upon substituing (6.1.40) in (6.1.20) and using equation (6.1.30) subsequently implies that the scalar angular power spectrum is given by the convolution of the primordial potential spectrum with spherical Bessel functions:

$$C_\ell^{\rm S} = \frac{2}{\pi}\int j_\ell^2[k(\eta_0 - \eta_{\rm LSS})]\frac{1}{9}\mathcal{P}_\Phi(k)\frac{k^3\,{\rm d}k}{k}, \tag{6.2.2}$$

inasmuch as only the first term in Eq. (6.1.29) contributes significantly on super-horizon scales, with $\Delta\eta_{\rm E} = \eta_0 - \eta_{\rm LSS}$, where $\eta_{\rm LSS}$ denotes the conformal time of last scattering. Deep in the matter era, the gravitational potential is static and connects to the gauge-invariant comoving curvature perturbation through $\Phi = \frac{3}{5}\zeta$ for super-Hubble modes [Eq. (4.3.19)]. Consequently, $\mathcal{P}_\Phi = \frac{9}{25}\mathcal{P}_\zeta$. Defining the dimensionless power spectrum $\mathcal{P}_X \equiv k^3 P_X/2\pi^2$, we infer

$$C_\ell^{\rm S} = \pi\int A_{\rm S}^2(k) j_\ell^2[k(\eta_0 - \eta_{\rm LSS})]\frac{{\rm d}k}{k}, \tag{6.2.3}$$

for a primordial spectrum of the form predicted by inflation [Eq. (9.4.120)], with the normalization $A_{\rm S}^2 \equiv \frac{4}{25}\mathcal{P}_\zeta$ [Eq. (9.4.126)].

The integral in Eq. (6.2.3) is heavily weighted around modes satisfying $k\eta_0 \sim \ell$, so that one naturally expects $C_\ell \propto A_{\rm S}^2(k \sim \ell/\eta_0)$. More rigorously, for a power-law primordial spectrum (see Appendix B):

$$C_\ell^{\rm S} \simeq \pi A_{\rm S}^2(k_0)\left[\frac{\pi}{2}\frac{\Gamma(3-n_{\rm S})}{2^{3-n_{\rm S}}}\frac{\Gamma\left(\frac{\ell+n_{\rm S}-1}{2}\right)}{\Gamma^2\left(\frac{2-n_{\rm S}}{2}\right)\Gamma\left(\frac{\ell+5-n_{\rm S}}{2}\right)}\right], \tag{6.2.4}$$

Derived using (6.1.40) with $k = 2 - n_s$ in

$$\int_0^\infty j_\ell(t) j_\ell(t) t^{-k} dt = \frac{\pi}{2}\frac{\Gamma(k+1)\Gamma(\frac{2\ell+1-k}{2})}{2^{k+1}\Gamma^2\left(\frac{k+2}{2}\right)\Gamma\left(\frac{2\ell+k+3}{2}\right)} \tag{1}$$

adopting the pivot scale $k_* = k_0 = 1/\eta_0$ and invoking the approximation $\eta_{\rm LSS} \ll \eta_0$. For an exact scale-invariant Harrison–Zel'dovich spectrum ($n_{\rm S} = 1$), this reduces to

$$\ell(\ell+1)C_\ell^{\rm S} = \frac{\pi}{2}A_{\rm S}^2(k_0), \tag{6.2.5}$$

which defines the horizontal plateau of the angular power spectrum at low multipoles (Fig. 6.2). In the asymptotic limit of large ℓ, Stirling's formula $\Gamma(\ell + a) \propto \ell^a$ establishes that

$$\ell(\ell+1)C_\ell^{\rm S} \propto \ell^{n_{\rm S}-1}. \tag{6.2.6}$$

The primordial spectral index $n_{\rm S}$ therefore directly governs the tilt of the Sachs–Wolfe plateau. Furthermore, Eq. (6.2.1) shows that $\Theta_{\rm SW} = -\delta_\gamma^{\rm N}/6$, revealing that cold spots on large angular scales correspond physically to primordial overdense regions at the last-scattering surface due to the dominant gravitational redshift.

Isocurvature perturbations

For primordial isocurvature (entropy) perturbations, the initial conditions in the radiation era satisfy $\delta_\gamma^{\rm N}(0) = \Phi(0) = 0$ and $V_\gamma(0) = 0$ (Chapter 4). If $S(0)$ denotes the conserved primordial entropy fluctuation, then during the matter-dominated epoch the gravitational potential settles to $\Phi = -S/5$. On super-Hubble scales, the quantity $\delta_\gamma^{\rm N} - 4\Psi$ remains conserved, which implies $\delta_\gamma^{\rm N}/4 + \Phi = \Phi + \Psi \simeq 2\Phi$. The leading-order Sachs–Wolfe temperature fluctuation in the matter era is therefore

$$\Theta_{\rm SW} = 2\Phi = -\frac{2}{5}S; \tag{6.2.7}$$

see Eq. (5.229). Paralleling the derivation for adiabatic perturbations, we find

$$C_\ell^{\rm S} \simeq 36\pi A_{\rm S}^2(k_0)\left[\frac{\pi}{2}\frac{\Gamma(3-n_{\rm S})}{2^{3-n_{\rm S}}}\frac{\Gamma\left(\frac{\ell+n_{\rm S}-1}{2}\right)}{\Gamma^2\left(\frac{2-n_{\rm S}}{2}\right)\Gamma\left(\frac{\ell+5-n_{\rm S}}{2}\right)}\right]. \tag{6.2.8}$$

For a scale-invariant isocurvature spectrum ($n_{\rm S} = 1$), this produces an equivalent low-ℓ plateau:

$$\ell(\ell+1)C_\ell^{\rm S} = 36\pi A_{\rm S}^2(k_0). \tag{6.2.9}$$

6.2.2 Intermediate scales

The physical origin, morphological shape, and acoustic peak structure of the angular power spectrum at intermediate multipoles can be understood from first principles by analyzing the coupled photon–baryon plasma prior to recombination.

The linear evolution of photons tightly coupled to baryons is governed by the perturbation equations derived in Chapter 4 [Eqs. (4.4.49)–(4.4.52)]:

$$\delta_\gamma^{\rm N\prime} = \frac{4}{3}k^2 V_\gamma + 4\Psi', \qquad V_\gamma' = -\frac{1}{4}\delta_\gamma^{\rm N} - \Phi + \frac{1}{6}k^2\pi_\gamma + \tau'(V_{\rm b} - V_\gamma), \tag{6.2.10}$$

$$\delta_{\rm b}^{\rm N\prime} = k^2 V_{\rm b} + 3\Psi', \qquad V_{\rm b}' = -\mathcal{H}V_{\rm b} - \Phi + \frac{\tau'}{R}(V_\gamma - V_{\rm b}), \tag{6.2.11}$$

where $\tau' \equiv a n_{\rm e}\sigma_{\rm T}$ is the differential optical depth per unit conformal time. The residual radiation quadrupole π_γ is directly proportional to the photon velocity divergence:

$$\frac{k^2}{12}\pi_\gamma = -\frac{8}{45}\frac{k^2}{\tau'}V_\gamma. \tag{6.2.12}$$

This relation expresses the fact that an anisotropic stress quadrupole is locally generated by the shear gradient of the velocity field (look up (6.3.73)). As established in Eq. (6.3.153) below, incorporating the linear polarization of Thomson scattering fixes the numerical prefactor to 8/45; if polarization is omitted, this coefficient simplifies to 2/15.

Dynamics of the photon–baryon fluid

Consider perturbation modes whose spatial wavelengths are substantially larger than the photon mean free path, $k/\tau' \ll 1$. Expanding the photon and baryon Euler equations to leading order in k/τ' under negligible anisotropic stress yields the tight coupling conditions

$$V_{\rm b} = V_\gamma, \qquad \delta_\gamma' - \frac{4}{3}\delta_{\rm b}' = 0.$$

Even though photon-baryon anisotropic stress tensor can be ignored during radiation era. The anisotropic stress tensor from neutrino stays relevant fir precision cosmology. This is happening because the neutrinos are not undergoing collision and hence their velocity is no longer being isotropized.

The second relation guarantees that the entropy perturbation remains constant in time. In this tight-coupling regime, combining the two momentum equations eliminates the Thomson drag term:

$$[(1+R)V_\gamma]' = -\frac{1}{4}\delta_\gamma^{\rm N} - (1+R)\Phi. \tag{6.2.13}$$

Differentiating the photon continuity equation and substituting Eq. (6.2.13), we obtain the fundamental governing equation for the photon density contrast:

$$\delta_\gamma^{\rm N\prime\prime} + \frac{R'}{1+R}\delta_\gamma^{\rm N\prime} + k^2 c_{\rm s}^2 \delta_\gamma^{\rm N} = 4\mathcal{F}(\Phi,\Psi) = 4\left[\Psi'' + \frac{R'}{1+R}\Psi' - \frac{1}{3}k^2\Phi\right], \tag{6.2.14}$$

where the baryon-to-photon momentum density ratio R and the adiabatic sound speed $c_{\rm s}$ are defined by

$$R \equiv \frac{3}{4}\frac{\rho_{\rm b}}{\rho_\gamma}, \qquad c_{\rm s}^2 \equiv \frac{1}{3}\frac{1}{1+R}. \tag{6.2.15}$$

Equation (6.2.14) describes a forced, damped harmonic oscillator. The driving force $\mathcal{F}$ is governed by the gravitational potentials, which become dominated by cold dark matter as R increases, shifting the zero-point of the acoustic oscillations. Equation (6.2.14) can be recast in the self-adjoint form:

$$\left[(1+R)\delta_\gamma^{\rm N\prime}\right]' + \frac{1}{3}k^2\delta_\gamma^{\rm N} = 4\left\{\left[(1+R)\Psi'\right]' - \frac{1}{3}k^2(1+R)\Phi\right\}. \tag{6.2.16}$$

Equation (6.2.16) lies at the heart of acoustic oscillations in the primordial plasma: radiation pressure supplies the restoring force against gravitational compression, driving sound waves at frequency $\omega_{\rm s} = kc_{\rm s}$. In the WKB limit, where the oscillation period is much shorter than the background evolution timescale ($\delta_\gamma^{\rm N\prime} \ll \omega_{\rm s}\delta_\gamma^{\rm N}$, while accounting for $\omega_{\rm s}'/\omega_{\rm s} = -R'/[2(1+R)]$), the homogeneous equation admits the two independent basis solutions

$$\delta_a = (1+R)^{-1/4}\cos(kr_{\rm s}), \qquad \delta_b = (1+R)^{-1/4}\sin(kr_{\rm s}), \tag{6.2.17}$$

where the sound horizon at conformal time η is given by

$$r_{\rm s}(\eta) \equiv \int_0^\eta c_{\rm s}{\rm d}\eta' = \int_0^\eta \frac{{\rm d}\eta'}{\sqrt{3(1+R)}}. \tag{6.2.18}$$

The Green's function associated with the driven oscillator is

$$\mathcal{G}(\eta,\eta') \equiv \frac{\delta_a(\eta')\delta_b(\eta) - \delta_a(\eta)\delta_b(\eta')}{\delta_a(\eta')\delta_b'(\eta') - \delta_a'(\eta')\delta_b(\eta')} = \frac{\sqrt{3}}{k}\frac{[1+R(\eta')]^{3/4}}{[1+R(\eta)]^{1/4}}\sin\left[kr_{\rm s}(\eta) - kr_{\rm s}(\eta')\right], \tag{6.2.19}$$

enabling the full inhomogeneous solution to be written as

$$\begin{aligned}[1+R(\eta)]^{1/4}\delta_\gamma^{\rm N} = \delta_\gamma(0)\cos[kr_{\rm s}(\eta)] + \frac{\sqrt{3}}{k}\left[\delta_\gamma'(0) + \frac{1}{4}R'(0)\delta_\gamma(0)\right]\sin[kr_{\rm s}(\eta)] \\ + 4\frac{\sqrt{3}}{k}\int_0^\eta [1+R(\eta')]^{3/4}\sin\left[kr_{\rm s}(\eta) - kr_{\rm s}(\eta')\right]\mathcal{F}(\eta'){\rm d}\eta'.\end{aligned} \tag{6.2.20}$$

Sachs–Wolfe term

Assuming to leading order that the Bardeen potentials are quasi-static (as in the matter-dominated epoch) and neglecting the slow time evolution of R, the solution (6.2.20) simplifies to

$$\Theta_{\rm SW} = [\Theta_{\rm SW}(0) + R\Phi]\cos[kr_{\rm s}(\eta)] + \frac{\Theta'_{\rm SW}(0)}{kc_{\rm s}}\sin[kr_{\rm s}(\eta)] - R\Phi, \tag{6.2.21}$$

with $r_{\rm s} = c_{\rm s}\eta$. The Sachs–Wolfe term exhibits acoustic oscillations in conformal time representing sound waves propagating through the plasma. Crucially, the equilibrium center of oscillation is displaced from zero to $-R\Phi$ by the gravitational drag of the baryons. Over many periods, R and Φ evolve dynamically, causing the oscillation envelope of $\Theta_{\rm SW}$ to attenuate as $(1+R)^{-1/4}$. Figure 6.4 displays the temporal evolution of $\Theta_{\rm SW}$ for different modes prior to decoupling.

Doppler term

Neglecting the time derivatives of the metric potentials and R, the photon continuity equation sets $k^2 V_\gamma/3 = \Theta'_{\rm SW} + (\Psi - \Phi)'$, leading in this approximation to

$$\frac{kV_\gamma}{3} = -c_{\rm s}[\Theta_{\rm SW}(0) + R\Phi]\sin[kr_{\rm s}(\eta)] + \frac{\Theta'_{\rm SW}(0)}{k}\cos[kr_{\rm s}(\eta)]. \tag{6.2.22}$$

Under tight coupling, $V_\gamma \simeq V_{\rm b}$, generating the kinematic Doppler anisotropy $\Theta_{\rm dop} \sim kV_\gamma/\sqrt{3}$. This term oscillates symmetrically about zero in temporal quadrature ($\pi/2$ out of phase) with $\Theta_{\rm SW}$, with an amplitude reduced by $\sqrt{3}c_{\rm s} = (1+R)^{-1/2}$. Consequently, the Doppler peaks naturally fill in the troughs of the Sachs–Wolfe power spectrum. In the pure radiation limit ($R \to 0$), the Doppler and Sachs–Wolfe amplitudes are comparable, whereas for realistic baryon fractions ($R > 0$), the Doppler contribution is suppressed relative to $\Theta_{\rm SW}$, preserving pronounced acoustic peaks (Fig. 6.2).

Positions of the acoustic peaks

The multipole locations of the acoustic peaks depend decisively on the primordial initial conditions. Inside the sound horizon, the Bardeen potentials are heavily suppressed. Accounting for photon diffusion damping via the factor $D(k,\eta) \equiv \exp(-k^2/k_{\rm D}^2)$ [Eq. (6.2.34)], the acoustic solution takes the form

$$\Theta_{\rm SW} = C_{\rm A}\frac{D(k,\eta)}{(1+R)^{1/4}}\cos(kr_{\rm s}) + C_{\rm I}\frac{D(k,\eta)}{(1+R)^{1/4}}\sin(kr_{\rm s}), \tag{6.2.23}$$

where $C_{\rm A}$ and $C_{\rm I}$ are determined by the initial conditions:

- **Adiabatic perturbations**: Here $\delta^{\rm N}_\gamma(0) = -2\Phi(0)$. Neglecting early velocity perturbations, $\delta^{\rm N\prime}_\gamma = 4\Phi'$, so that $\delta^{\rm N}_\gamma(\eta) = -2\Phi(0) + 4\Phi(\eta) - 4\Phi(0)$. When the perturbation crosses the Hubble radius, the potential decays, leaving $\delta^{\rm N}_\gamma(\eta) \simeq -6\Phi(0)$. At sound horizon entry, $\delta^{\rm N}_\gamma/4 = -3\Phi(0)/2$ and $\delta^{\rm N\prime}_\gamma = 0$, selecting $C_{\rm A} = -\frac{3}{2}\Phi(0)$ and $C_{\rm I} = 0$:

$$\Theta_{\rm SW} = -\frac{3}{2}\Phi(0)\frac{D(k,\eta)}{(1+R)^{1/4}}\cos(kr_{\rm s}). \tag{6.2.24}$$

  The constant gravitational drive excites exclusively the cosine mode.

- **Isocurvature perturbations**: Here $\delta^{\rm N}_\gamma(0) = \Phi(0) = 0$. As shown in Eq. (5.220), $\delta^{\rm N}_\gamma \propto -aS$ while $S = S(0)$ remains constant. The potential then grows as $\Phi \propto -aS$ [Eq. (5.228)]. Since $a \propto \eta$ in the radiation era, $\mathcal{F} \propto \eta$ in Eq. (6.2.14), which excites the sine mode as $kr_{\rm s} \to 0$. Solving the dynamical equations yields

$$C_{\rm A} = 0, \qquad C_{\rm I} = \frac{\sqrt{6}}{4}\frac{k_{\rm eq}}{k}S(0). \tag{6.2.25}$$

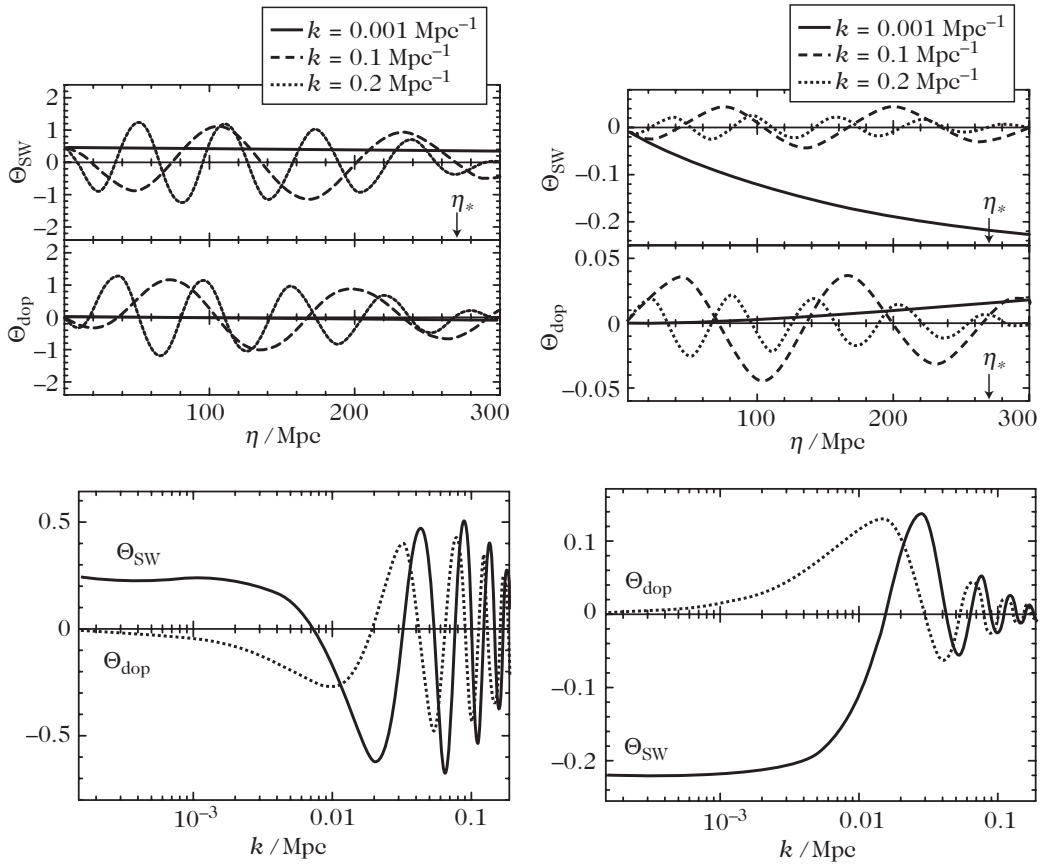


Figure 6.4: Evolution of $\Theta_{\rm SW} = \delta_\gamma^{\rm N}/4 + \Phi$ and of $\Theta_{\rm dop} \sim kV_\gamma/\sqrt{3}$ as a function of time for three modes with wavenumber $k = 10^{-3}$ (solid line), 10^{-1} (dashed) and $2 \times 10^{-1}\,{\rm Mpc}^{-1}$ (dotted). The contribution at the time of decoupling as a function of k is presented at the bottom. We compare the adiabatic (left) to the isocurvature (right) initial conditions.

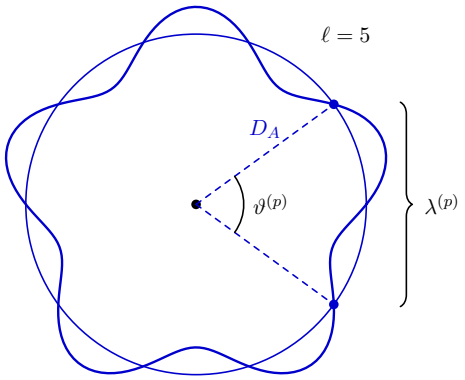
The factor $k_{\rm eq}/k$ reflects the growth of $\delta_\gamma^{\rm N} \propto -aS$ until the mode enters the Jeans scale at $a \sim k_{\rm eq}/k$.

At the decoupling epoch ($\eta = \eta_{\rm LSS}$), the temperature contrast $\Theta_{\rm SW}(k, \eta_{\rm LSS})$ achieves local extrema at wavenumbers satisfying

$$k^{(p)} r_{\rm s}(\eta_{\rm LSS}) = \begin{cases} p\pi, & \text{(adiabatic)} \\ \left(p - \frac{1}{2}\right)\pi, & \text{(isocurvature)} \end{cases} \tag{6.2.26}$$

for integers $p = 1, 2, \ldots$. The physical wavelength $\lambda^{(p)} = a(\eta_{\rm LSS})\pi/k^{(p)}$ subtends an observed angular scale $\vartheta^{(p)} \sim \lambda^{(p)}/D_{\rm A}(\eta_{\rm LSS})$, where $D_{\rm A}$ is the comoving angular diameter distance (given by (2.4.14) and (2.4.16)):

$$\vartheta^{(p)} = \frac{\lambda^{(p)}}{D_A} = \frac{a_0}{1+z}\frac{\pi}{k^{(p)}}\frac{1}{D_A} = \frac{\pi}{k^{(p)} f_K(z_{\rm LSS})}. \tag{6.2.27}$$



Relating the angular separation ϑ to multipole order $\ell \sim \pi/\vartheta$, the acoustic peaks are situated at

$$\ell^{(p)} = \frac{\pi}{\vartheta^{(p)}} = \frac{f_K(z_{\rm LSS})}{r_s(\eta_{\rm LSS})} \times k^{(p)} r_s(\eta_{\rm LSS}) = \frac{\pi f_K(z_{\rm LSS})}{r_{\rm s}(z_{\rm LSS})} \times \begin{cases} p, & \text{(adiabatic)} \\ p - \frac{1}{2}, & \text{(isocurvature).} \end{cases} \tag{6.2.28}$$

The peak positions are therefore dictated primarily by the background cosmological geometry through $z_{\rm LSS}$ and $r_{\rm s}(z_{\rm LSS})$. Adiabatic models predict a harmonic series of peak locations in ratios $1:2:3:\ldots$, whereas isocurvature models produce odd-harmonic ratios $1:3:5:\ldots$. In isocurvature models, the $p=1$ mode is heavily suppressed, placing the first observable peak at

$$\ell^{(1)} \sim \begin{cases} 220, & \text{(adiabatic)} \\ 330, & \text{(isocurvature)} \end{cases} \tag{6.2.29}$$

in a spatially flat universe with $\Omega_\Lambda = 0$. For an adiabatic model with arbitrary curvature, $\ell^{(1)} \sim 220/\sqrt{\Omega_0}$, providing an exquisitely sensitive geometric probe of spatial curvature. These acoustic signatures were first predicted by Sakharov in 1965 (Ref. [8]) and are widely known as Sakharov oscillations.

Question. A guitar string produces a discrete set of harmonics because it is pinned down at both ends, restricting it to a discrete spectrum of allowed frequencies. There is no analogous boundary condition on the perturbations of the early Universe, so why do we observe CMB anisotropies concentrated at certain angular scales, the acoustic peaks, rather than a continuum?

Answer. The acoustic peaks should not be read as a discrete set of allowed modes in the sense of a string's harmonics. Every wavenumber k is present and evolving in the perturbation field; the discreteness visible in C_ℓ comes from two facts that have nothing to do with k itself being restricted.

The primordial spectrum is genuinely continuous in k. There is no boundary condition analogous to $y(0,t) = y(L,t) = 0$ that restricts the comoving curvature perturbation $\zeta(\boldsymbol{k})$ to a discrete set of wavenumbers: ζ is Fourier-decomposed over a continuum of k, with a nearly scale-invariant dimensionless spectrum $A_{\rm S}^2(k) \propto k^{n_{\rm S}-1}$ [Eq. (6.2.3)] built from $\mathcal{P}_\zeta(k)$ [Eq. (4.3.19) ff.]. Every k therefore carries primordial power; the existence of power at a given k does not by itself dictate how strongly that mode shows up in the observed temperature field.

Every k-mode oscillates, and different k simply reach different phases by decoupling. Equation (6.2.14) shows that each Fourier mode of the photon–baryon fluid is a driven, damped harmonic oscillator with its own frequency $\omega_k = kc_{\rm s}$, with homogeneous solutions $\propto \cos(kr_{\rm s})$, $\sin(kr_{\rm s})$ [Eq. (6.2.17)]. Modes are not created or destroyed according to the value of k: they all evolve continuously, but by decoupling they have accumulated

different phases $kr_\mathrm{s}(\eta_\mathrm{LSS})$. From Eq. (6.2.24), $\Theta_\mathrm{SW}(k,\eta_\mathrm{LSS}) \propto \cos[kr_\mathrm{s}(\eta_\mathrm{LSS})]$: modes at $kr_\mathrm{s} = p\pi$ [Eq. (6.2.26)] sit at extrema of the density oscillation and dominate the observed temperature contrast, while modes at $kr_\mathrm{s} = (p-\frac{1}{2})\pi$ are near zero density contrast at that instant. This is exactly the statement that a harmonic oscillator $x(t) = A\cos(\omega t)$, at any instant where $x = 0$, has maximal velocity $\dot{x} = -A\omega\sin(\omega t)$ rather than having ceased to exist: the corresponding photon–baryon modes have a suppressed density contrast but an enhanced velocity perturbation V_γ instead, which is precisely why the Doppler term fills in the troughs left by Θ_SW (Eq. (6.2.22) and the surrounding discussion). Every k between the acoustic extrema remains present in the perturbation field; it is simply not at the phase that maximizes the temperature contrast at η_LSS.

The discreteness of ℓ, separately, is a property of the spherical harmonic basis, not of k. Expanding the observed sky on the celestial two-sphere necessarily uses the discrete basis $Y_{\ell m}$ [Eq. (6.1.20)], since only integer ℓ index normalizable functions on S^2; this has nothing to do with whether k is continuous. Each multipole is an integral over all continuous k,

$$C_\ell^\mathrm{S} = \pi \int A_\mathrm{S}^2(k)\, j_\ell^2[k(\eta_0 - \eta_\mathrm{LSS})]\, \frac{\mathrm{d}k}{k}$$

[Eq. (6.2.3)], and the spherical Bessel function j_ℓ peaks broadly around $k(\eta_0 - \eta_\mathrm{LSS}) \sim \ell$, so a whole range of k contributes to any single ℓ: there is no one-to-one correspondence between a discrete k and a discrete ℓ.

Putting the two facts together: the acoustic extrema at $k^{(p)}r_\mathrm{s} = p\pi$ project, via $\ell \sim kD_\mathrm{A}$ [Eq. (6.2.28)], onto the multipoles $\ell^{(p)}$ where the temperature power is locally *enhanced*. The peaks in C_ℓ are therefore preferred scales in a continuously populated power spectrum, produced by every k-mode reaching a temperature-favorable phase of its acoustic oscillation at the same cosmic time η_LSS, not a discrete set of allowed frequencies selected by a boundary condition, as for the guitar string.

6.2.3 Small scales

On small spatial scales, k/τ' is no longer negligible, necessitating corrections beyond the tight-coupling approximation as the photon mean free path becomes comparable to the perturbation wavelength.

Silk damping

Consider perturbation modes that cross inside the Hubble radius during the radiation-dominated era. Because radiation dominates the cosmic energy density, the gravitational potentials decay rapidly during acoustic oscillations, allowing metric perturbations and gauge ambiguities to be neglected. Furthermore, because the oscillation period is much shorter than the expansion timescale, the expansion of the universe can be frozen ($R \approx$ const).

The baryon and radiation Euler equations [Eqs. (4.4.60) and (4.3.61)] reduce to

$$V_\mathrm{b}' = \frac{\tau'}{R}(V_\gamma - V_\mathrm{b}), \qquad V_\gamma' = -\frac{1}{4}\delta_\gamma + \frac{1}{6}k^2\pi_\gamma - \tau'(V_\gamma - V_\mathrm{b}).$$

Because $\pi_\gamma \sim \mathcal{O}(1/\tau')$, expanding the velocity difference to order $(\tau')^0$ yields

$$\frac{R+1}{R}\tau'(V_\gamma - V_\mathrm{b}) = -\frac{1}{4}\delta_\gamma. \tag{6.2.30}$$

Combining both momentum equations into $(R+1)V_\gamma' + R(V_\mathrm{b}' - V_\gamma')$ and employing Eq. (6.2.30), we find

$$(R+1)V_\gamma' = -\frac{1}{4}\delta_\gamma + \frac{1}{6}k^2\pi_\gamma - \frac{1}{4}\frac{R^2}{1+R}\frac{\delta_\gamma'}{\tau'}. \tag{6.2.31}$$

Expressing π_γ via Eq. (6.2.12) and combining with the radiation continuity equation produces the damped wave equation:

$$\delta_\gamma'' + \frac{k^2 c_\mathrm{s}^2}{\tau'}\left[\frac{16}{15} + \frac{R^2}{1+R}\right]\delta_\gamma' + k^2 c_\mathrm{s}^2 \delta_\gamma = 0. \tag{6.2.32}$$

(If polarization is ignored, the factor 16/15 reverts to 4/5). The WKB solution to this equation exhibits exponential dissipation:

$$\delta_\gamma \propto \exp\left(-\frac{k^2}{k_\mathrm{D}^2}\right)\exp(\pm \mathrm{i} k r_\mathrm{s}), \tag{6.2.33}$$

where the Silk diffusion scale k_D is given by

$$k_\mathrm{D}^{-2} = \frac{1}{6}\int_0^\eta \frac{1}{1+R}\left[\frac{16}{15} + \frac{R^2}{1+R}\right]\frac{\mathrm{d}\eta'}{\tau'}. \tag{6.2.34}$$

The bracketed factor varies only between 16/15 (for $R = 0$) and 1 (as $R \to \infty$).

The same result can be reached directly from the kinetic (Boltzmann) hierarchy, by solving for the dispersion relation $\omega(k)$ of the coupled photon–baryon system, rather than by passing through the second-order equation (6.2.32). As above, we freeze the expansion of the universe and drop the gravitational potentials, and for simplicity we first omit the polarization correction to the quadrupole; its effect is restored at the end. The starting point is the monopole and dipole hierarchy equations (6.3.69)–(6.3.70), the quadrupole relation (6.2.12) together with the dictionary (6.3.73), and the baryon Euler equation (6.2.11) (with its gravitational term dropped):

$$\Theta_0' = -\frac{k}{3}\Theta_1, \tag{1}$$

$$\Theta_1' = k\left(\Theta_0 - \frac{2}{5}\Theta_2\right) - \tau'(kV_\mathrm{b} + \Theta_1), \tag{2}$$

$$V_\mathrm{b}' = \frac{\tau'}{R}(V_\gamma - V_\mathrm{b}), \qquad V_\gamma = -\frac{\Theta_1}{k}. \tag{3}$$

The quadrupole itself is eliminated using Eq. (6.2.12) together with the dictionary (6.3.73) ($\Theta_2 = \frac{5}{12}k^2\pi_\gamma$, $V_\gamma = -\Theta_1/k$), in the unpolarized limit ($8/45 \to 2/15$):

$$\frac{k^2}{12}\pi_\gamma = -\frac{2}{15}\frac{k^2}{\tau'}V_\gamma = \frac{2k}{15\tau'}\Theta_1 \quad\Longrightarrow\quad \Theta_2 = \frac{5}{12}k^2\pi_\gamma \cdot \frac{12}{5k^2}\cdot\frac{2k}{15\tau'}\cdot\frac{5}{2} = \frac{2k}{3\tau'}\Theta_1. \tag{4}$$

Because $k\eta \gg 1$, many collisions occur per oscillation period, and we model the time dependence harmonically, $\Theta_0, \Theta_1, V_\mathrm{b} \propto \mathrm{e}^{-\mathrm{i}\omega\eta}$, so that $X' \to -\mathrm{i}\omega X$ for each variable; we write $x \equiv \omega/\tau' \ll 1$ for the small expansion parameter. Equation (1) becomes $-\mathrm{i}\omega\Theta_0 = -\frac{k}{3}\Theta_1$, i.e.

$$\Theta_1 = \frac{3\mathrm{i}\omega}{k}\Theta_0. \tag{5}$$

Substituting Eq. (4) into (2) and applying the harmonic ansatz gives

$$-\mathrm{i}\omega\Theta_1 = k\Theta_0 - \frac{2k}{5}\cdot\frac{2k}{3\tau'}\Theta_1 - \tau' k V_\mathrm{b} - \tau'\Theta_1 \quad\Longrightarrow\quad \tau' k V_\mathrm{b} = k\Theta_0 + \Theta_1\left(\mathrm{i}\omega - \tau' - \frac{4k^2}{15\tau'}\right). \tag{6}$$

Separately, the harmonic baryon equation (3) reads $-\mathrm{i}\omega V_\mathrm{b} = \frac{\tau'}{R}\left(-\frac{\Theta_1}{k} - V_\mathrm{b}\right)$, i.e. $V_\mathrm{b}\left(\frac{\tau'}{R} - \mathrm{i}\omega\right) = -\frac{\tau'\Theta_1}{kR}$, so that

$$\tau' k V_\mathrm{b} = \frac{-\tau'\Theta_1}{1 - \mathrm{i}xR}. \tag{7}$$

Equating (6) and (7) gives

$$k\Theta_0 + \Theta_1\left(\mathrm{i}\omega - \tau' - \frac{4k^2}{15\tau'}\right) = \frac{-\tau'\Theta_1}{1 - \mathrm{i}xR},$$

and multiplying through by $(1 - \mathrm{i}xR)$ collects every term on one side:

$$k\Theta_0(1 - \mathrm{i}xR) + \Theta_1(1 - \mathrm{i}xR)\left(\mathrm{i}\omega - \tau' - \frac{4k^2}{15\tau'}\right) + \tau'\Theta_1 = 0.$$

Substituting Θ_1 from Eq. (5), dividing by Θ_0, and multiplying by k turns this into a relation for ω alone. Expanding the product, the two terms proportional to $\mathrm{i}\omega\tau'$ (one from the substituted Θ_1 inside the bracket, one from the lone $\tau'\Theta_1$) cancel exactly, and the single resulting term of order x^2 (from $\mathrm{i}xR$ multiplying the already order-x quadrupole piece) is dropped as subleading. What remains is

$$3\omega^2\left[(1+R) - \mathrm{i}xR\right] + \frac{4\mathrm{i}k^2}{5\tau'}\,\omega - k^2\,(1 - \mathrm{i}xR) = 0. \tag{8}$$

Dropping the quadrupole term ($k^2/\tau'^2 \to 0$) reduces this to the clean form $\omega^2 = \frac{k^2}{3}(1 - \mathrm{i}xR)/[(1+R) - \mathrm{i}xR]$, whose $x \to 0$ limit is exactly the adiabatic sound speed $\omega_0 \equiv kc_\mathrm{s} = k/\sqrt{3(1+R)}$ of Eq. (6.2.15); keeping the quadrupole does not change this leading-order limit, since its contribution to Eq. (8) is already first order in x.

To find the correction, write $\omega = \omega_0 + \delta\omega$ with $\delta\omega = O(x\omega_0)$, and expand Eq. (8) to first order in x. Writing $x_0 \equiv \omega_0/\tau'$, every term already carrying an explicit factor of x or $1/\tau'$ may be evaluated at $\omega = \omega_0$, since a further $\delta\omega$ correction inside it would be second order:

$$3\left[(1+R) - \mathrm{i}x_0 R\right](\omega_0 + \delta\omega)^2 + \frac{4\mathrm{i}k^2}{5\tau'}\omega_0 - k^2\,(1 - \mathrm{i}x_0 R) = 0.$$

Expanding $(\omega_0 + \delta\omega)^2 \approx \omega_0^2 + 2\omega_0\delta\omega$ and dropping the second-order term $x_0\delta\omega$,

$$\underbrace{3(1+R)\omega_0^2 - k^2}_{=0} \; - 3\mathrm{i}x_0 R\,\omega_0^2 \; + \; 6(1+R)\omega_0\,\delta\omega \; + \; \frac{4\mathrm{i}k^2}{5\tau'}\omega_0 \; + \; \mathrm{i}k^2 x_0 R = 0,$$

where the underbraced term vanishes by the definition of ω_0. The two terms linear in R combine using $k^2 - 3\omega_0^2 = 3R\omega_0^2$ (from $k^2 = 3(1+R)\omega_0^2$):

$$-3\mathrm{i}x_0 R\,\omega_0^2 + \mathrm{i}k^2 x_0 R = \mathrm{i}x_0 R\left(k^2 - 3\omega_0^2\right) = 3\mathrm{i}x_0 R^2\omega_0^2 = \frac{3\mathrm{i}\omega_0^3 R^2}{\tau'},$$

using $x_0 = \omega_0/\tau'$ in the last step. Substituting back leaves

$$6(1+R)\omega_0\,\delta\omega = -\frac{3\mathrm{i}\omega_0^3 R^2}{\tau'} - \frac{4\mathrm{i}k^2\omega_0}{5\tau'},$$

and dividing by $6(1+R)\omega_0$, then substituting $\omega_0^2 = k^2/[3(1+R)]$ in the first term, gives

$$\delta\omega = -\frac{\mathrm{i}k^2}{6\tau'(1+R)^2}\left[R^2 + \frac{4}{5}(1+R)\right]. \tag{9}$$

Hence

$$\omega(k) = k\left[\frac{1}{\sqrt{3(R+1)}} - \mathrm{i}\,\frac{k}{6\tau'}\,\frac{R^2+\frac{4}{5}(R+1)}{(R+1)^2}\right]. \tag{10}$$

The real part reproduces the sound speed already found from Eq. (6.2.15). For the imaginary part, $|\mathrm{Im}(\omega)|/k^2$ is exactly $\frac{1}{6}\frac{1}{1+R}\left[\frac{4}{5}+\frac{R^2}{1+R}\right]\tau'^{-1}$, i.e. k^2 times the unpolarized (4/5) integrand of k_D^{-2} in Eq. (6.2.34); integrating $|\mathrm{Im}(\omega)|$ over conformal time therefore reproduces k^2/k_D^2, and hence the same WKB envelope $\exp(-k^2/k_\mathrm{D}^2)$ of Eq. (6.2.33). At fixed τ' and R, over the time interval

$$t_\mathrm{d} = \frac{1}{|\mathrm{Im}(\omega)|} = \frac{6\tau'}{k^2}\,\frac{(R+1)^2}{R^2+\frac{4}{5}(R+1)}, \tag{11}$$

the oscillation amplitude falls by one e-fold. This is Silk damping: it originates in two distinct effects entering at the same order in k/τ', the imperfect coupling between the baryon and photon velocities (the R^2 term, from the finite baryon-drag time in Eq. (3)) and the residual photon quadrupole sourced by the velocity gradient (the $\frac{4}{5}(1+R)$ term, from Eq. (6.2.12)); it vanishes as $k/\tau' \to 0$.

Restoring the polarization correction changes only the quadrupole elimination step: using the polarized coefficient 8/45 in place of the unpolarized 2/15 in Eq. (4) rescales Θ_2 by $\frac{8/45}{2/15} = \frac{4}{3}$, so that $\Theta_2 = \frac{8}{9}\frac{k}{\tau'}\Theta_1$ in place of $\frac{2}{3}\frac{k}{\tau'}\Theta_1$. This is the only place Θ_2 enters the derivation, so the same factor of $\frac{4}{3}$ carries through every later step untouched, and every subsequent $\frac{4}{5}$ is simply replaced by $\frac{4}{5}\times\frac{4}{3} = \frac{16}{15}$:

$$\omega(k) = k\left[\frac{1}{\sqrt{3(R+1)}} - \mathrm{i}\,\frac{k}{6\tau'}\,\frac{R^2+\frac{16}{15}(R+1)}{(R+1)^2}\right], \qquad t_\mathrm{d} = \frac{6\tau'}{k^2}\,\frac{(R+1)^2}{R^2+\frac{16}{15}(R+1)}, \tag{12}$$

which is exactly the polarized result of Eq. (6.2.34). The 16/15 versus 4/5 is therefore not a discrepancy between the two derivations: it is the same enhancement of the photon quadrupole by Thomson polarization feedback that already distinguishes 8/45 from 2/15 in Eq. (4), propagated through to the damping rate.

Physically, photons execute a random walk through Thomson scattering with mean free path $\lambda_\mathrm{C} \sim 1/\tau'$. In conformal time η, the accumulated diffusion distance is $\lambda_\mathrm{D} \sim \sqrt{N}\lambda_\mathrm{C}$ with $N \sim \eta/\lambda_\mathrm{C}$ steps, yielding $\lambda_\mathrm{D} \sim \sqrt{\eta/\tau'} \sim k_\mathrm{D}^{-1}$. In a flat universe without dark energy, assuming $z_\mathrm{eq} \sim z_\mathrm{LSS}$, the diffusion wavenumber satisfies

$$k_\mathrm{D}\eta_\mathrm{LSS} \sim (1+z_\mathrm{LSS})\sqrt{6\frac{\sigma_\mathrm{T}}{m_\mathrm{p}}\sqrt{\frac{3H_0}{8\pi G_\mathrm{N}}}\Omega_{\mathrm{b}0}h}. \tag{6.2.35}$$

This photon diffusion effect (Silk damping, Ref. [9]) induces an exponential suppression of the angular power spectrum starting around $\ell \sim 1400\sqrt{\Omega_{\mathrm{b}0}h^2/0.019}$, well before the first acoustic peak, almost completely extinguishing power beyond $\ell \sim 2000$ (as demonstrated in Fig. 6.5).

Effect of the thickness of the last-scattering surface

Because recombination is not instantaneous, the last-scattering surface has a finite physical thickness. The visibility function $g(\eta)$ defines the probability distribution that a photon underwent its final Thomson scattering at conformal time η. It is related to the optical depth $\tau(\eta)$ by

$$\tau(\eta) = \int_\eta^{\eta_0} \tau'(u)\,\mathrm{d}u\,, \tag{6.2.36}$$

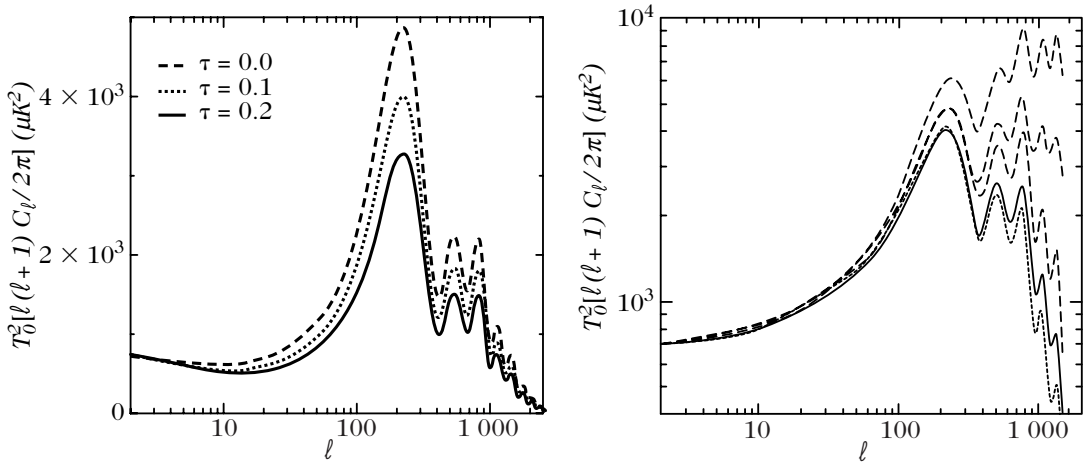


Figure 6.5: (left): Effect of the reionization on the temperature angular power spectrum. (right): The dashes represent the anisotropies of a standard cold dark matter model without taking into account the Silk damping and the thickness of the surface of last-scattering surface (top) or the thickness of the surface of last scattering alone (middle) or the Silk effect alone (bottom). The solid line represents the exact computation and the dotted line corresponds to a computation including the effect of the thickness of the last scattering surface and the Silk effect by an analytical correction deduced from our estimates.

Consider a monochromatic electromagnetic wave propagating through a dilute medium of scatterers. The electric field may be written as

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 e^{i(\tilde{k}z-\omega t)}, \tag{1}$$

where the wave number is, in general, complex. Writing

$$\tilde{k} = k + i\kappa, \tag{2}$$

the electric field becomes

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 e^{ikz} e^{-\kappa z} e^{-i\omega t}. \tag{3}$$

Since the intensity of an electromagnetic wave is proportional to the squared magnitude of the electric-field amplitude,

$$I(z) \propto |\mathbf{E}(z)|^2, \tag{4}$$

we obtain

$$I(z) = I_0 e^{-2\kappa z}. \tag{5}$$

For a dilute medium with number density n, the complex refractive index is related to the forward scattering amplitude $f(0)$ by

$$\tilde{n} - 1 = \frac{2\pi n}{k^2} f(0), \tag{6}$$

where $k = \omega/c$. Writing

$$\tilde{n} = n_r + i n_i, \tag{7}$$

and using

$$\tilde{k} = k\tilde{n}, \tag{8}$$

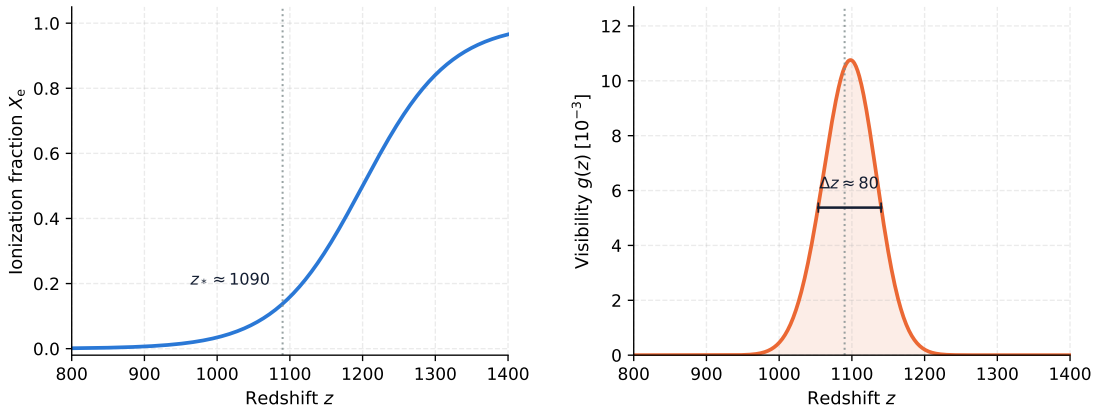


Figure 6.6: Recombination history and the visibility function. *Left:* Ionization fraction $X_e(z)$ dropping precipitously from fully ionized plasma ($X_e \to 1$) at $z > 1400$ to freeze-out residual levels at $z < 800$. *Right:* Visibility function $g(z) = -\dot{\tau}e^{-\tau}$, representing the probability distribution of photon decoupling. The peak defines the last-scattering redshift $z_* \approx 1090$, while its full width at half maximum $\Delta z \approx 80$ defines the finite physical thickness of the last-scattering surface.

we find

$$\kappa = kn_i = \frac{2\pi n}{k} \operatorname{Im} f(0). \tag{9}$$

The optical theorem relates the imaginary part of the forward scattering amplitude to the total scattering cross section:

$$\sigma_{\text{tot}} = \frac{4\pi}{k} \operatorname{Im} f(0). \tag{10}$$

Consequently,

$$\kappa = \frac{n\sigma_{\text{tot}}}{2}, \tag{11}$$

Substituting this relation into Eq. (5) gives

$$I(z) = I_0 e^{-n\sigma_{\text{tot}} z}. \tag{12}$$

For a spatially varying medium, this generalizes to

$$I = I_0 \exp\left(-\int n\sigma_{\text{tot}} dl\right). \tag{13}$$

such that $\mathrm{d}\tau/\mathrm{d}\eta = -\tau'$. The visibility function is given by

$$g(\eta) \equiv \tau' \exp(-\tau), \tag{6.2.37}$$

normalized to $\int g(\eta)\mathrm{d}\eta = 1$. The recombination history and the peaked visibility window defining the finite thickness of the last-scattering surface are displayed in Fig. 6.6.

The observed temperature field along direction $\boldsymbol{e}$ is therefore a visibility-weighted line-of-sight average:

$$\Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e}) = \int g(\eta) \left[\Theta_{\text{SW}} + \Theta_{\text{dop}}\right](\eta, [\eta_0 - \eta]\boldsymbol{e})\mathrm{d}\eta. \tag{6.2.38}$$

Moreover, gravitational potential variations only affect photons that have already decoupled. The decoupled fraction is given by

$$g_\gamma(\eta) = \int_{\eta_{\text{in}}}^{\eta} g(u)\,\mathrm{d}u\,, \tag{6.2.39}$$

which increases monotonically from 0 at early times to 1 today. The integrated Sachs–Wolfe contribution is accordingly modified to

$$\Theta_{\rm ISW} = \int g_\gamma(\eta)[\Phi' + \Psi'](\eta, [\eta_0 - \eta]\boldsymbol{e})\mathrm{d}\eta. \tag{6.2.40}$$

In Fourier space, approximating the visibility function by a Gaussian centered on $\eta_{\rm LSS}$ with dispersion $\sigma_{\rm LSS}$:

$$g(\eta) \sim \frac{1}{\sqrt{2\pi\sigma^2_{\rm LSS}}} \exp\left[-\frac{1}{2}\frac{(\eta - \eta_{\rm LSS})^2}{\sigma^2_{\rm LSS}}\right],$$

its Fourier transform suppresses small-scale anisotropies according to

$$\hat{g}(k) \propto \exp\left(-\frac{1}{2}k^2\sigma^2_{\rm LSS}\right). \tag{6.2.41}$$

The finite thickness of the last-scattering surface thus washes out temperature fluctuations on spatial scales smaller than $\sigma_{\rm LSS}$, operating with comparable magnitude to Silk damping.

6.2.4 Other effects

Following decoupling, photons travel essentially unhindered, so that only gravitational field variations and late-time reionization can alter their temperature distribution.

Integrated Sachs–Wolfe

The temporal variation of the metric potentials and primordial gravitational waves contributes to the temperature contrast via $\Theta_{\rm ISW}$ [Eq. (6.1.17)]. This term encompasses two distinct physical regimes:

- **Early Sachs–Wolfe effect**: The residual evolution of the gravitational potentials during the transition from radiation to matter domination leaves an imprint on modes entering the sound horizon between equality and decoupling. This effect enhances the power spectrum in the vicinity of the first acoustic peak. Because $\Theta_{\rm SW} \sim -\Phi/2$ carries the same sign as $\Phi' < 0$ during this transition, $\Theta_{\rm ISW}$ adds constructively to $\Theta_{\rm SW}$, boosting the height of the first peak (Fig. 6.2). For isocurvature modes, potential growth on scales larger than the sound horizon provides a substantial contribution below the first peak.
- **Late Sachs–Wolfe effect**: At low redshifts ($z \lesssim 1$), cosmic acceleration driven by dark energy (Λ) or non-zero spatial curvature reactivates the time evolution of the gravitational potentials. This effect operates primarily on very large angular scales, modifying the low-ℓ Sachs–Wolfe plateau.

Because Φ decays in a Universe dominated by dark energy or spatial curvature at low redshifts (Chapter 4), an additional net blueshift or redshift is imparted to photons traversing time-varying potential wells. This late integrated Sachs–Wolfe effect provides one of the few direct cosmological probes of dark energy through CMB anisotropies, detectable primarily via its cross-correlation with tracers of the large-scale gravitational potential in the local Universe (such as galaxy catalogs).

Furthermore, gravitational lensing by intervening large-scale structures deflects CMB photons along the line of sight. While the distortion induced in the temperature power spectrum is modest (smoothing the acoustic peaks and generating a small non-Gaussian signal at very high ℓ), weak gravitational lensing plays a central role in generating B-mode polarization from primordial E-modes (Chapter 8).

Reionization

Once the earliest generations of stars and quasars ignite, ultraviolet radiation reionizes the neutral intergalactic medium. Free electrons liberated during this epoch can Thomson-scatter CMB photons long after recombination. This rescattering tends to suppress primordial temperature anisotropies by averaging over multiple intersecting lines of sight arriving at each scattering center, while simultaneously generating secondary anisotropies through the motion of the scattering electrons.

Under the approximation of homogeneous reionization, the visibility function can be idealized as the sum of two Dirac delta distributions centered at decoupling ($\eta_{\rm LSS}$) and reionization ($\eta_{\rm re}$). Neglecting the integrated Sachs–Wolfe contribution, the observed temperature fluctuation along direction $\boldsymbol{e}$ is approximated by

$$\Theta(\boldsymbol{x}_0,\eta_0,\boldsymbol{e}) \simeq (1-\mathrm{e}^{-\tau_{\rm re}})(\Theta+\boldsymbol{e}\cdot\boldsymbol{V}_{\rm b})[\boldsymbol{e}(\eta_0-\eta_{\rm re}),\eta_{\rm re}]+\mathrm{e}^{-\tau_{\rm re}}(\Theta+\boldsymbol{e}\cdot\boldsymbol{V}_{\rm b})[\boldsymbol{e}(\eta_0-\eta_{\rm LSS}),\eta_{\rm LSS}], \quad (6.2.42)$$

where $\tau_{\rm re}$ denotes the optical depth from the reionization era to the present. The first term represents the line-of-sight averaging and the generation of secondary Doppler anisotropies due to the motion of scattering electrons at $\eta_{\rm re}$. The second term corresponds to the unscattered fraction ($\mathrm{e}^{-\tau_{\rm re}}$) of the primordial temperature anisotropy that originated at the primary last-scattering surface.

Dropping the ISW term, the general line-of-sight formula (6.2.38) reads, with $\Theta_{\rm SW}+\Theta_{\rm dop}\to\Theta+\boldsymbol{e}\cdot\boldsymbol{V}_{\rm b}$,

$$\Theta(\boldsymbol{x}_0,\eta_0,\boldsymbol{e}) = \int_0^{\eta_0} g(\eta)\,(\Theta+\boldsymbol{e}\cdot\boldsymbol{V}_{\rm b})(\eta,[\eta_0-\eta]\boldsymbol{e})\,\mathrm{d}\eta. \quad (1)$$

Because $g=\tau'\mathrm{e}^{-\tau}=\mathrm{d}(\mathrm{e}^{-\tau})/\mathrm{d}\eta$ [Eq. (6.2.37), using $\mathrm{d}\tau/\mathrm{d}\eta=-\tau'$], integrating g over any interval $[\eta_1,\eta_2]$ gives the exact identity

$$\int_{\eta_1}^{\eta_2} g(\eta)\,\mathrm{d}\eta = \mathrm{e}^{-\tau(\eta_2)}-\mathrm{e}^{-\tau(\eta_1)}, \quad (2)$$

with $\tau(\eta)=\int_\eta^{\eta_0}\tau'(u)\mathrm{d}u$ as in Eq. (6.2.36). Under homogeneous reionization the free-electron fraction, and hence τ', is negligible throughout the dark ages between decoupling and reionization, so $\tau(\eta)$ stays essentially constant there, equal to $\tau_{\rm re}\equiv\tau(\eta_{\rm re})=\int_{\eta_{\rm re}}^{\eta_0}\tau'\,\mathrm{d}\eta'$, for all $\eta_{\rm LSS}<\eta<\eta_{\rm re}$. Idealizing both transitions as instantaneous collapses this into the step profile

$$\tau(\eta)\simeq\begin{cases}\tau_{\rm LSS}\gg 1, & \eta<\eta_{\rm LSS},\\ \tau_{\rm re}, & \eta_{\rm LSS}<\eta<\eta_{\rm re},\\ 0, & \eta>\eta_{\rm re},\end{cases} \quad (3)$$

so that $\mathrm{e}^{-\tau(\eta)}$ rises from 0 to $\mathrm{e}^{-\tau_{\rm re}}$ to 1 across these two jumps. Writing $H(x)$ for the Heaviside step function ($H=0$ for $x<0$ and $H=1$ for $x>0$), Eq. (3) is equivalent to the closed-form expression

$$\mathrm{e}^{-\tau(\eta)} \simeq \mathrm{e}^{-\tau_{\rm re}}\,H(\eta-\eta_{\rm LSS})+(1-\mathrm{e}^{-\tau_{\rm re}})\,H(\eta-\eta_{\rm re}), \quad (4)$$

$$g(\eta)=\frac{\mathrm{d}}{\mathrm{d}\eta}\mathrm{e}^{-\tau(\eta)} \simeq \mathrm{e}^{-\tau_{\rm re}}\,\delta_{\rm D}(\eta-\eta_{\rm LSS})+(1-\mathrm{e}^{-\tau_{\rm re}})\,\delta_{\rm D}(\eta-\eta_{\rm re}), \quad (5)$$

where the second line differentiates the first using $\mathrm{d}H(\eta-a)/\mathrm{d}\eta=\delta_{\rm D}(\eta-a)$ (equivalently, this is exactly the identity (2) applied across each of the two infinitesimal jumps). The two weights sum to 1, consistent with the normalization $\int g\,\mathrm{d}\eta=1$ of Eq. (6.2.37). Substituting Eq. (5) into Eq. (1) and evaluating the two delta functions reproduces Eq. (6.2.42) term by term: the $\eta_{\rm LSS}$ spike gives the $\mathrm{e}^{-\tau_{\rm re}}$-weighted primordial term, and the $\eta_{\rm re}$ spike gives the $(1-\mathrm{e}^{-\tau_{\rm re}})$-weighted rescattering term.

The quantity $\Theta[\boldsymbol{e}(\eta_0 - \eta_{\rm re}), \eta_{\rm re}]$ represents the directional average of $\Theta - \boldsymbol{e}' \cdot \boldsymbol{V}_{\rm b}$ across the electron scattering surface at reionization. For long-wavelength modes satisfying $k(\eta_{\rm re} - \eta_0) \ll 1$, this average reproduces $\Theta[\boldsymbol{e}(\eta_0 - \eta_{\rm LSS}), \eta_{\rm LSS}]$, while on sub-horizon scales it averages to zero. Consequently, for super-Hubble modes at the reionization epoch, the observed temperature perturbation transforms as

$$\Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e}) \to \Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e}) + (1 - \mathrm{e}^{-\tau_{\rm re}})\boldsymbol{e} \cdot \Delta\boldsymbol{V}_{\rm b},$$

where $\Delta\boldsymbol{V}_{\rm b}$ is the velocity differential of the electron fluid between reionization and the earlier scattering event. Because the Doppler contribution is negligible on super-Hubble scales, large-angle temperature anisotropies are left virtually unchanged. Conversely, on scales smaller than the horizon at reionization,

$$\Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e}) \to \mathrm{e}^{-\tau_{\rm re}}\Theta(\boldsymbol{x}_0, \eta_0, \boldsymbol{e}) + (1 - \mathrm{e}^{-\tau_{\rm re}})\boldsymbol{e} \cdot \boldsymbol{V}_{\rm b}(\eta_{\rm re}).$$

The net consequence is an exponential damping by $\mathrm{e}^{-\tau_{\rm re}}$ of the primordial acoustic peak amplitudes (as illustrated in Fig. 6.5). Cross-correlation measurements of CMB temperature and polarization by the WMAP mission constrain the optical depth to $\tau_{\rm re} \sim 0.17$, indicating a reionization redshift of $z_{\rm re} \sim 15$.

Sunyaev–Zel'dovich effect

Interactions between CMB photons and high-temperature astrophysical plasmas alter the photon energy distribution. In particular, the intracluster gas within massive galaxy clusters reaches temperatures of order $T_{\rm e} \sim 10^7 - 10^8$ K, emitting X-rays via thermal bremsstrahlung. Inverse Compton scattering between the free thermal electrons and the passing CMB photons shifts low-energy photons toward higher frequencies. This spectral distortion is known as the thermal Sunyaev–Zel'dovich (tSZ) effect and constitutes the dominant secondary anisotropy on arcminute angular scales.

To order of magnitude, the resulting temperature modification is given by $\Delta T_{\rm SZ} \sim L\overline{n_{\rm e}T_{\rm e}}$, where the overline denotes an average along the line of sight through the cluster, L is the physical path length, and $T_{\rm e}$, $n_{\rm e}$ are the electron temperature and number density. By comparison, the thermal bremsstrahlung X-ray surface brightness of the cluster scales as

$$S_X \sim \frac{V}{4\pi D_{\rm L}^2}\overline{\left(n_{\rm e} n_{\rm p} T_{\rm e}^{1/2}\right)},$$

where $V = D_{\rm A}^2\theta^2 L$ is the cluster volume subtended by an angular diameter θ, and $D_{\rm L}$, $D_{\rm A}$ are the luminosity and angular diameter distances, respectively.

The $n_{\rm e}n_{\rm p}T_{\rm e}^{1/2}$ scaling of thermal bremsstrahlung follows from a single Coulomb encounter. An electron of charge $-e$ and mass $m_{\rm e}$ passing an ion of charge Ze at impact parameter b and speed v suffers an impulsive acceleration $a \sim Ze^2/(m_{\rm e}b^2)$ over a deflection time $\Delta t \sim b/v$. The Larmor formula, $P = 2e^2a^2/3c^3$, then gives the energy radiated in the encounter,

$$\Delta E(b, v) \sim P\,\Delta t \sim \frac{2Z^2e^6}{3m_{\rm e}^2c^3\,b^3v}. \tag{1}$$

Because the pulse has duration Δt, its power is spread roughly evenly over frequencies up to $\omega_{\rm max} \sim v/b$, so

$$\frac{{\rm d}E}{{\rm d}\omega}(b, v) \sim \frac{\Delta E}{\omega_{\rm max}} \sim \frac{2Z^2e^6}{3m_{\rm e}^2c^3\,b^2v^2}. \tag{2}$$

Summing Eq. (2) over all impact parameters, weighted by the flux $n_{\rm e}v\,2\pi b\,{\rm d}b$ of electrons

past each of the n_i ions per unit volume, gives the volume emissivity at fixed electron speed v:

$$\epsilon_\omega(v) \sim n_\mathrm{e} n_\mathrm{i} v \int_{b_\mathrm{min}}^{b_\mathrm{max}} \frac{\mathrm{d}E}{\mathrm{d}\omega}(b,v)\, 2\pi b\, \mathrm{d}b \sim n_\mathrm{e} n_\mathrm{i} \frac{Z^2 e^6}{m_\mathrm{e}^2 c^3 v} \ln\Lambda, \tag{3}$$

where $\ln\Lambda \equiv \ln(b_\mathrm{max}/b_\mathrm{min})$ is the (slowly varying) Coulomb logarithm. For a thermal plasma at temperature T_e, the characteristic electron speed is $v \sim \sqrt{k_\mathrm{B} T_\mathrm{e}/m_\mathrm{e}}$, so Eq. (3) scales as $\epsilon_\omega \propto n_\mathrm{e} n_\mathrm{i} Z^2 T_\mathrm{e}^{-1/2}$ per unit frequency. Integrating over frequency up to the thermal cutoff $\hbar\omega \sim k_\mathrm{B} T_\mathrm{e}$ supplies one power of T_e, so the bolometric (frequency-integrated) volume emissivity is

$$\epsilon^{ff} \equiv \int_0^\infty \epsilon_\omega\, \mathrm{d}\omega \;\propto\; n_\mathrm{e} n_\mathrm{i} Z^2 T_\mathrm{e}^{-1/2} \times T_\mathrm{e} \;=\; n_\mathrm{e} n_\mathrm{i} Z^2 T_\mathrm{e}^{1/2}, \tag{4}$$

which for a hydrogen plasma ($Z = 1$, $n_\mathrm{i} = n_\mathrm{p}$) is exactly the scaling quoted above. The cluster's X-ray luminosity is the volume integral of Eq. (4), $L_X = \int_V \epsilon^{ff}\, \mathrm{d}V \sim V\, \overline{\epsilon^{ff}}$, and the flux reaching the observer follows the usual cosmological inverse-square law, $F_X = L_X/(4\pi D_\mathrm{L}^2)$. Identifying the surface brightness with this flux reproduces

$$S_X \sim \frac{L_X}{4\pi D_\mathrm{L}^2} \sim \frac{V}{4\pi D_\mathrm{L}^2} \overline{\left(n_\mathrm{e} n_\mathrm{p} T_\mathrm{e}^{1/2}\right)}. \tag{5}$$

As a concrete example, consider an isothermal cluster described by a standard β-model electron density profile:

$$n_\mathrm{e}(r) = n_0 \left[1 + \left(\frac{r}{r_\mathrm{c}}\right)^2\right]^{-3\beta/2}, \tag{6.2.43}$$

for $0 < r < R_\mathrm{cluster}$ (and zero otherwise), where r_c is the cluster core radius. In the non-relativistic limit, the temperature decrement in the Rayleigh–Jeans regime is given by

$$\Delta T_\mathrm{SZ} = -2 \frac{k_\mathrm{B} T_0}{m_\mathrm{e} c^2} \sigma_\mathrm{T} \int_{-L_\mathrm{max}}^{L_\mathrm{max}} n_\mathrm{e}\, \mathrm{d}\ell\,, \tag{6.2.44}$$

assuming that the electron gas is isothermal throughout the cluster. In the limit where the cluster integration boundary $R_\mathrm{cluster} \to \infty$, performing the line-of-sight integral yields

$$\Delta T_\mathrm{SZ} = -2 \frac{k_\mathrm{B} T_0}{m_\mathrm{e} c^2} \sigma_\mathrm{T} n_0 r_\mathrm{c} B\left(\frac{3\beta - 1}{2}, \frac{1}{2}\right) \left[1 + \left(\frac{\theta}{\theta_\mathrm{c}}\right)^2\right]^{(1-3\beta)/2}, \tag{6.2.45}$$

where $B(x, y)$ is the Euler beta function and $\theta_\mathrm{c} = r_\mathrm{c}/D_\mathrm{A}$ is the angular core radius.

Unlike Thomson scattering—which preserves the blackbody spectral shape—the thermal Sunyaev–Zel'dovich effect generates a unique spectral signature characterized by a temperature decrease at long wavelengths (Rayleigh–Jeans region) and an increment at short wavelengths ($\nu > 217$ GHz). Because the surface brightness decrement is redshift-independent, the tSZ effect provides a powerful observational tool for detecting distant clusters and probing cosmic structure evolution (see Ref. [10] for an extensive review). Furthermore, if the cluster possesses a peculiar velocity $\boldsymbol{v}$ relative to the CMB rest frame, the Doppler effect introduces a secondary modification known as the kinetic Sunyaev–Zel'dovich (kSZ) effect, which is proportional to the line-of-sight radial component of the cluster velocity.

Gravitational waves

Deriving a tensor-mode counterpart to the scalar plateau [Eq. (6.2.4)] is more delicate because the tensor Sachs–Wolfe integral [Eq. (6.1.45)] depends explicitly on the time derivative of

the gravitational wave amplitude E'_λ. During the matter-dominated era, the propagation of primordial gravitational waves is governed by Eq. (4.3.5), whose mode solution yields

$$E'_\lambda(k,\eta) = -3kE_\lambda(k,\eta_\mathrm{i})\frac{j_2(k\eta)}{k\eta}. \tag{6.2.46}$$

Substituting this result into Eq. (6.1.45) for an initial tensor power spectrum parametrized as $\mathcal{P}_\mathrm{T}(k) \propto k^{n_\mathrm{T}}$ (as predicted by standard slow-roll inflation, Eq. (8.234)) leads to the tensor angular power spectrum:

$$C_\ell^\mathrm{T} \propto \frac{(\ell+2)!}{(\ell-2)!}\int \frac{\mathrm{d}y_0}{y_0^{1-n_\mathrm{T}}}\left[\int_{y_0\eta_\mathrm{LSS}/\eta_0}^{y_0} \frac{j_\ell(y_0-y)}{(y_0-y)^2}\frac{j_2(y)}{y}\,\mathrm{d}y\right]^2, \tag{6.2.47}$$

where we have set the dimensionless integration variables $y = k\eta$ and $y_0 = k\eta_0$ with pivot scale $k_0 = 1/\eta_0$. The integrand is appreciable only when the arguments of the spherical Bessel functions simultaneously satisfy $y \sim 2$ and $y_0 - y \sim \ell$, as the Bessel functions rapidly decay away from these peaks. Given the domain of integration $y_0 \leq y\eta_0/\eta_\mathrm{LSS}$, both conditions can be met only for multipoles fulfilling

$$\ell \lesssim 2\frac{\eta_0}{\eta_\mathrm{LSS}} \sim 2\sqrt{z_\mathrm{LSS}+1} \sim 60. \tag{6.2.48}$$

This analytical estimate matches the numerical behavior depicted in Fig. 6.3, where C_ℓ^T drops off precipitously beyond $\ell \sim 100$. Consequently, the tensor spectrum C_ℓ^T is primarily sensitive to large angular scales and depends only weakly on background cosmological parameters.

To evaluate the integral in Eq. (6.2.47) analytically, we approximate the sharply peaked factor $j_2(y)/y$ by an effective delta function $j_2(y)/y \sim 2^{-1/2}\delta(y-2)$. Integrating over y under the assumption that $y_0 \gg 2$, the spectrum simplifies to

$$C_\ell^\mathrm{T} \propto \frac{(\ell+2)!}{(\ell-2)!}\int \frac{\mathrm{d}y_0}{y_0^{1-n_\mathrm{T}}}\left[\frac{j_\ell(y_0)}{y_0^2}\right]^2. \tag{6.2.49}$$

Evaluating the remaining integral over y_0 leads to the closed-form expression

$$C_\ell^\mathrm{T} \propto \frac{(\ell+2)!}{(\ell-2)!}\frac{\Gamma\left(\frac{\ell-2+n_\mathrm{T}}{2}\right)}{\Gamma\left(\frac{\ell+4+n_\mathrm{T}}{2}\right)}. \tag{6.2.50}$$

Applying Stirling's asymptotic expansion in an analogous manner to Eq. (6.2.4), we obtain the low-multipole scaling

$$\ell(\ell+1)C_\ell^\mathrm{T} \propto \ell^{n_\mathrm{T}} \tag{6.2.51}$$

for small multipoles ($\ell \gtrsim 2$). Thus, a scale-invariant primordial tensor background ($n_\mathrm{T} \simeq 0$) produces a flat quadrupole and low-ℓ plateau in $\ell(\ell+1)C_\ell^\mathrm{T}$.

6.3 Kinetic description

To achieve a rigorous and quantitatively accurate description of photon propagation, Thomson scattering, and the decoupling process from the baryonic plasma, one must move beyond the ideal fluid approximation and develop a relativistic kinetic framework. A kinetic approach is essential because photons possess a long mean free path in the transition regime and undergo directional and spectral evolution. Detailed presentations of relativistic kinetic theory can be found in Refs. [11–13], with foundational properties of distribution functions addressed in Ref. [14]. The application of the relativistic Boltzmann equation to cosmological perturbations is treated in Refs. [15–19].

6.3.1 Perturbed Boltzmann equation

Distribution function

As in Chapter 4, the state of the photon gas is described by a single-particle phase-space distribution function $f(x^\mu, p^\mu)$, which depends on the space-time position coordinates and 4-momentum components treated as independent variables. For particles of mass m, the distribution function is restricted to the mass shell

$$\mathcal{P}_m = \left\{ (x^\mu, p^\mu) \in TM \;\middle|\; p_\mu p^\mu = -m^2 \right\}, \tag{6.3.1}$$

defined on the tangent bundle $TM = \{(x^\mu, p^\mu) \mid x^\mu \in \mathcal{M}, p^\mu \in T_x\mathcal{M}\}$, where $\mathcal{M}$ is the four-dimensional space-time manifold and $T_x\mathcal{M}$ is the tangent space at point x.

To decompose the phase space, it is advantageous to introduce the unit timelike vector field normal to the spatial hypersurfaces of constant conformal time $\{\eta = \text{const.}\}$. For the linearly perturbed metric of Eq. (5.52), this normal vector and its dual are given by

$$N_\mu = -a(1 + A, \mathbf{0}), \qquad N^\mu = \frac{1}{a}(1 - A, -B^i), \tag{6.3.2}$$

to first order in cosmological perturbations. The photon 4-momentum k^μ (reserving p^μ for general massive particles and k^μ strictly for null geodesics) can then be parametrized as

$$k^\mu = \frac{E}{a}\left(1 - A,\; n^i - \left[B_j n^j + \frac{1}{2} h_{jk} n^j n^k\right] n^i\right), \tag{6.3.3}$$

where the unit spatial direction vector satisfies $\gamma_{ij} n^i n^j = 1$. This parameterization identically satisfies the null condition $k_\mu k^\mu = 0$ and identifies E as the physical photon energy measured by a comoving observer whose four-velocity is N^μ, since $N_\mu k^\mu = -E$.

In perturbation theory, the full distribution function is split into an unperturbed, isotropic and homogeneous background $\bar{f}(\eta, E)$ and a first-order perturbation $\delta f(\eta, x^i, E, n^j)$ depending on space-time position and direction:

$$\boxed{f(x^\mu, p_\mu) = f(\eta, x^i, E, n^j) = \bar{f}(\eta, E) + \delta f(\eta, x^i, E, n^j).} \tag{6.3.4}$$

The dynamical evolution of f is governed by the relativistic Boltzmann equation

$$\mathcal{L}[f] = C[f], \tag{6.3.5}$$

where $\mathcal{L} \equiv \mathrm{d}/\mathrm{d}\eta$ denotes the Liouville operator along the phase-space trajectory and $C[f]$ represents the collision term, implying that phase-space density is conserved in the absence of scattering interactions.

Collisionless Boltzmann equation

We first analyze the collisionless Liouville operator:

$$\mathcal{L}[f] = \frac{\mathrm{d}f}{\mathrm{d}\eta} = \left[k^\mu \frac{\partial f}{\partial x^\mu} + \frac{\mathrm{d}k^\mu}{\mathrm{d}\lambda}\frac{\partial f}{\partial k^\mu}\right]\frac{\mathrm{d}\lambda}{\mathrm{d}\eta}, \qquad k^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda} \tag{6.3.6}$$

where λ is an affine parameter along the geodesic with tangent vector k^μ, so that $\mathrm{d}\eta/\mathrm{d}\lambda = k^0$. Utilizing the geodesic equation [Eq. (1.37)],

$$\frac{\mathrm{d}k^\mu}{\mathrm{d}\lambda} = -\Gamma^\mu_{\alpha\beta} k^\alpha k^\beta, \tag{6.3.7}$$

the Liouville operator takes the general form

$$\mathcal{L}[f] = \left[k^\mu \frac{\partial f}{\partial x^\mu} - \Gamma^\mu_{\alpha\beta} k^\alpha k^\beta \frac{\partial f}{\partial k^\mu}\right] \frac{1}{k^0}. \tag{6.3.8}$$

In terms of the coordinates (η, x^i, E, n^i) introduced in Eq. (6.3.4), the total derivative expands as

$$\mathcal{L}[f] = \frac{\partial f}{\partial \eta} + \frac{\mathrm{d}x^i}{\mathrm{d}\eta}\frac{\partial f}{\partial x^i} + \frac{\mathrm{d}E}{\mathrm{d}\eta}\frac{\partial f}{\partial E} + \frac{\mathrm{d}n^i}{\mathrm{d}\eta}\frac{\partial f}{\partial n^i}. \tag{6.3.9}$$

We evaluate each term in Eq. (6.3.9) systematically to linear order:

1. Because the directional derivative $\partial f/\partial n^i$ vanishes at zeroth order for an isotropic background, the rate $\mathrm{d}n^i/\mathrm{d}\eta$ is only required at the unperturbed level. To zeroth order, the geodesic equation gives (from (2.2.7))

$$E' = -\mathcal{H}E, \qquad n^{i\prime} = -{}^{(3)}\Gamma^i_{jk} n^j n^k, \tag{6.3.10}$$

where ${}^{(3)}\Gamma^i_{jk}$ are the Christoffel symbols of the spatial metric γ_{ij} (Appendix A). Thus, the directional advection term evaluates to

$$\frac{\mathrm{d}n^i}{\mathrm{d}\eta}\frac{\partial f}{\partial n^i} = -{}^{(3)}\Gamma^i_{jk} n^j n^k \frac{\partial f}{\partial n^i}.$$

2. For the spatial gradient term, $\mathrm{d}x^i/\mathrm{d}\eta = k^i/k^0$. Since $\partial_i f$ is itself first order in perturbations, we can evaluate $\mathrm{d}x^i/\mathrm{d}\eta$ at background order, yielding $k^i/k^0 = \mathrm{d}x^i/\mathrm{d}\eta = n^i$, so that

$$\frac{\mathrm{d}x^i}{\mathrm{d}\eta}\frac{\partial f}{\partial x^i} = n^i \partial_i f.$$

3. To evaluate the energy evolution rate $\mathrm{d}E/\mathrm{d}\eta$, we examine the temporal component of the geodesic equation [Eq. (6.3.7)]:

$$\frac{\mathrm{d}k^0}{\mathrm{d}\eta} = -\Gamma^0_{\alpha\beta}\frac{k^\alpha k^\beta}{k^0} \implies (1-A)\frac{\mathrm{d}}{\mathrm{d}\eta}\left(\frac{E}{a}\right) - \frac{E}{a}(A' + n^i\partial_i A) = -\Gamma^0_{\alpha\beta}\frac{k^\alpha k^\beta}{E/a(1-A)}.$$

Recalling that $k^0 = \frac{E}{a}(1-A)$, its total derivative gives

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\left(\frac{E}{a}\right) = \frac{E}{a}\left(A' + n^i\partial_i A\right) - \Gamma^0_{\alpha\beta}\frac{k^\alpha k^\beta}{E/a}(1+2A) + \mathcal{O}\left(A^2\right).$$

Substituting the perturbed Christoffel symbols, we find

$$\frac{\mathrm{d}E}{\mathrm{d}\eta} = -E\left[\mathcal{H} + n^i\partial_i A + \left(\frac{1}{2}h'_{ij} - D_{(i}B_{j)}\right)n^i n^j\right].$$

Combining all four contributions yields the complete Liouville operator to first order in perturbations:

$$\boxed{\mathcal{L}[f] = f' + n^k\partial_k f - \left[\mathcal{H} + n^k\partial_k A + \left(\frac{1}{2}h'_{ij} - D_{(i}B_{j)}\right)n^i n^j\right]E\frac{\partial f}{\partial E} - {}^{(3)}\Gamma^k_{ij} n^i n^j \frac{\partial f}{\partial n^k},} \tag{6.3.11}$$

which can equivalently be written in compact form as

$$\boxed{\mathcal{L}[f] = f' + n^k\partial_k f - \left[\mathcal{H} + n^k\partial_k A + \left(\frac{1}{2}h'_{ij} - D_{(i}B_{j)}\right)n^i n^j\right]E\frac{\partial f}{\partial E} - {}^{(3)}\Gamma^k_{ij} n^i n^j \frac{\partial f}{\partial n^k}.} \tag{6.3.12}$$

Decomposing the distribution function according to Eq. (6.3.4) into background and perturbation components, we obtain the corresponding evolution equations:

$$\mathcal{L}[\bar{f}] = \bar{f}' - \mathcal{H}E\frac{\partial \bar{f}}{\partial E}, \tag{6.3.13}$$

$$\mathcal{L}[\delta f] = \delta f' + n^k\partial_k\delta f - \mathcal{H}E\frac{\partial \delta f}{\partial E} - \left[n^k\partial_k A + \left(\frac{1}{2}h'_{ij} - D_{(i}B_{j)}\right)n^i n^j\right]E\frac{\partial \bar{f}}{\partial E} - {}^{(3)}\Gamma^k_{ij}n^i n^j\frac{\partial \delta f}{\partial n^k}. \tag{6.3.14}$$

At the background level, the unperturbed photon distribution is a blackbody Bose–Einstein distribution (Chapter 4):

$$\bar{f}(E,\eta) = \left[\exp\left(\frac{E}{T(\eta)}\right) - 1\right]^{-1} = \bar{f}\left(\frac{E}{T}\right). \tag{6.3.15}$$

The background Liouville equation $\mathcal{L}[\bar{f}] = 0$ then guarantees that the background temperature redshifts as $T(\eta) \propto 1/a$, in complete harmony with the geodesic energy redshifting of individual photons.

At linear order, we consider perturbed distribution functions that maintain a local Planckian spectral shape with a direction-dependent temperature $T(\eta, x^i, n^j) = \bar{T}(\eta)[1 + \theta(\eta, x^i, n^j)]$, namely $f = \bar{f}(E/T)$. Taylor-expanding to first order in the fractional temperature perturbation θ, we obtain

$$f(\eta, x^i, E, n^j) = \bar{f}(\eta, E) - E\frac{\partial \bar{f}}{\partial E}\theta(\eta, x^i, n^j). \tag{6.3.16}$$

The temperature perturbation $\theta(\eta, x^i, n^j)$ is both inhomogeneous (dependent on position x^i) and anisotropic (dependent on direction n^j). This ansatz implicitly assumes that θ is energy-independent—a hypothesis justified by the energy independence of Thomson scattering in the Thomson regime, which introduces no spectral distortions.

Macroscopic quantities

Macroscopic fluid quantities are obtained as velocity moments of the phase-space distribution function. To carry out momentum integrations covariantly, we construct an orthonormal tetrad (vierbein) $(e^a_\mu)_{a=0\ldots3}$ representing the local Minkowski rest frame:

$$e^0_\mu = N_\mu, \qquad e^i_\mu e^j_\nu g^{\mu\nu} = \delta^{ij}, \qquad e^i_\mu N^\mu = 0.$$

Any null 4-momentum can then be projected onto the tetrad basis as $k^a = k^\mu e^a_\mu$. The energy-momentum tensor is defined by integrating over the mass shell in momentum space:

$$T^{\mu\nu}(x) = \int_{\mathcal{P}_m(x)} k^\mu k^\nu f(x^\alpha, k^\beta)\frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3 E_k}, \tag{6.3.17}$$

where $k^2 = \delta_{ab}k^a k^b$ and $E_k = \sqrt{k^2 + m^2} = k$ for massless photons.

Consider a collection of non-interacting point particles of mass m, with worldlines $z^\mu_i(\tau_i)$ parameterized by proper time τ_i. The microscopic stress-energy tensor is

$$T^{\mu\nu}(x) = \sum_i \int \frac{k^\mu_i k^\nu_i}{m}\frac{\delta^{(4)}(x - z_i(\tau_i))}{\sqrt{-g(x)}}\,\mathrm{d}\tau_i, \tag{1}$$

where

$$k_i^\mu = m\frac{\mathrm{d}z_i^\mu}{\mathrm{d}\tau_i} \tag{2}$$

is the four-momentum of the ith particle. Using

$$\mathrm{d}\tau_i = \frac{m}{k_i^0}\,\mathrm{d}t \tag{3}$$

and integrating over the temporal part of the four-dimensional delta function gives

$$T^{\mu\nu}(x) = \frac{1}{\sqrt{-g}}\sum_i \frac{k_i^\mu k_i^\nu}{k_i^0}\delta^{(3)}(\boldsymbol{x} - \boldsymbol{z}_i(t))\,. \tag{4}$$

In kinetic theory, the discrete distribution of particles is replaced by a smooth phase-space distribution function $f(x^\alpha, k^\beta)$

$$\sum_i \delta^{(3)}(\mathbf{x} - \mathbf{x}_i(t))\,A(\mathbf{p}_i) \;\longrightarrow\; g_* \int \frac{d^3p}{(2\pi)^3}\,f(t,\mathbf{x},\mathbf{p})\,A(\mathbf{p}), \tag{5}$$

where g_* is the degrees of freedom. The four-momentum is restricted to the mass shell,

$$g_{\alpha\beta}k^\alpha k^\beta + m^2 = 0. \tag{6}$$

It is convenient to impose this constraint using a delta function. The invariant momentum-space measure is

$$\mathrm{d}\Pi = \frac{\sqrt{-g}}{(2\pi)^3}\theta(k^0)\delta\Big(g_{\alpha\beta}k^\alpha k^\beta + m^2\Big)\,\mathrm{d}^4k, \tag{7}$$

where $\theta(k^0)$ selects the positive-energy branch.
For the scalar-perturbed FLRW metric in Newtonian gauge,

$$\mathrm{d}s^2 = a^2\left[-(1+2\Psi)\mathrm{d}\eta^2 + (1+2\Phi)\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j\right], \tag{8}$$

the determinant is, to first order in the perturbations,

$$\sqrt{-g} = a^4(1+\Psi+3\Phi). \tag{9}$$

Let p denote the physical momentum magnitude measured by a local comoving observer, and let n^i be its direction, satisfying

$$\delta_{ij}n^i n^j = 1. \tag{10}$$

Writing

$$k^i = Cn^i, \tag{11}$$

the definition of the physical momentum gives

$$p^2 = g_{ij}k^i k^j = a^2(1+2\Phi)C^2. \tag{12}$$

Hence, to first order,

$$k^i = \frac{p}{a}(1-\Phi)n^i. \tag{13}$$

Defining

$$E_k = \sqrt{p^2 + m^2}, \tag{14}$$

the mass-shell condition becomes

$$-|g_{00}|(k^0)^2 + E_k^2 = 0. \tag{15}$$

Thus, on the positive-energy branch,

$$k^0 = \frac{E_k}{\sqrt{|g_{00}|}}, \tag{16}$$

and

$$\delta\Big(-|g_{00}|(k^0)^2 + E_k^2\Big)\,\theta(k^0) = \frac{\delta\left(k^0 - \dfrac{E_k}{\sqrt{|g_{00}|}}\right)}{2E_k\sqrt{|g_{00}|}}. \tag{17}$$

Substituting this into (7) and performing the k^0 integration gives

$$\mathrm{d}\Pi = \frac{\sqrt{-g}}{(2\pi)^3}\frac{\mathrm{d}^3\boldsymbol{k}}{2E_k\sqrt{|g_{00}|}}. \tag{18}$$

It remains to express the coordinate momentum measure in terms of the physical momentum. From (13),

$$\mathrm{d}^3\boldsymbol{k} = \frac{(1-\Phi)^3}{a^3}\mathrm{d}^3\boldsymbol{p} = \frac{1-3\Phi}{a^3}\mathrm{d}^3\boldsymbol{p}. \tag{19}$$

Furthermore,

$$|g_{00}| = a^2(1+2\Psi), \tag{20}$$

so that

$$\sqrt{|g_{00}|} = a(1+\Psi) \tag{21}$$

to first order. Therefore,

$$\begin{aligned}\mathrm{d}\Pi &= \frac{1}{(2\pi)^3}\frac{a^4(1+\Psi+3\Phi)}{2E_k a(1+\Psi)}\frac{1-3\Phi}{a^3}\mathrm{d}^3\boldsymbol{p}\\ &= \frac{\mathrm{d}^3\boldsymbol{p}}{(2\pi)^3 2E_k} + \mathcal{O}(\Phi^2,\Psi^2,\Phi\Psi).\end{aligned} \tag{22}$$

Thus, to first order in the metric perturbations, the invariant measure takes the familiar flat-space form when expressed in terms of physical momentum,

$$\mathrm{d}\Pi = \frac{\mathrm{d}^3\boldsymbol{p}}{(2\pi)^3 2E_k}. \tag{23}$$

The energy-momentum tensor is obtained by integrating the momentum dyad over the mass shell:

$$T^{\mu\nu}(x) = 2\int_{\mathcal{P}_m(x)} \mathrm{d}\Pi\, k^\mu k^\nu f(x^\alpha, k^\beta). \tag{24}$$

Using the invariant measure after performing the k^0 integration, this becomes

$$\boxed{T^{\mu\nu}(x) = \int_{\mathcal{P}_m(x)} k^\mu k^\nu f(x^\alpha, k^\beta)\frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3 E_k}}. \tag{25}$$

Here $\mathcal{P}_m(x)$ denotes the positive-energy mass shell at the spacetime point x, and

$$E_k = \sqrt{p^2+m^2}, \qquad k^0 = \frac{E_k}{\sqrt{|g_{00}|}}, \qquad k^i = \frac{p}{a}(1-\Phi)n^i. \tag{26}$$

Absorbing the conventional factor $(2\pi)^3$ into the definition of the distribution function, the radiation energy-momentum tensor becomes

$$T^{\mu\nu}(x) = \int_{\mathcal{P}_m(x)} k^\mu k^\nu f(x^\alpha, k^\beta) E \, \mathrm{d}E \, \mathrm{d}^2\boldsymbol{n}. \tag{6.3.18}$$

Comparing this expression with the fluid decomposition of Eq. (4.2.18), we extract the radiation energy density, pressure, and anisotropic stress tensor:

$$\rho = \int (k^\mu u_\mu)^2 f E \, \mathrm{d}E \, \mathrm{d}^2\boldsymbol{n}, \tag{6.3.19}$$

$$P = \frac{1}{3} \int (k^\mu k^\nu \perp_{\mu\nu}) f E \, \mathrm{d}E \, \mathrm{d}^2\boldsymbol{n}, \tag{6.3.20}$$

$$\Pi^{\mu\nu} = \int k^\alpha k^\beta \left[\perp^\mu_\alpha \perp^\nu_\beta - \frac{1}{3} \perp_{\alpha\beta} \perp^{\mu\nu} \right] f E \, \mathrm{d}E \, \mathrm{d}^2\boldsymbol{n}, \tag{6.3.21}$$

where $\perp_{\mu\nu} = g_{\mu\nu} + u_\mu u_\nu$ is the spatial projector orthogonal to the fluid four-velocity u^μ. The four-velocity is defined by Eq. (5.70) as $u_\mu = a(-1 - A, v_k + B_k)$. Because photons are null ($k_\mu k^\mu = 0$), contracting Eq. (6.3.18) yields $T^\mu_{\ \mu} = 0$, immediately recovering the radiation equation of state $P = \rho/3$.

Evaluating the contraction $k^\mu u_\mu$ to first order:

$$k^\mu u_\mu = -E \left[1 - n^i (v_i + B_i) \right].$$

Since the angular integral of the background distribution $\bar{f} n^i$ vanishes by isotropy, we obtain the perturbed macroscopic fluid quantities:

$$\delta\rho = \int \delta f E^3 \, \mathrm{d}E \, \mathrm{d}^2\boldsymbol{n}, \tag{6.3.22}$$

$$(\rho + P)(v^i + B^i) = \int n^i \delta f E^3 \, \mathrm{d}E \, \mathrm{d}^2\boldsymbol{n}, \tag{6.3.23}$$

$$\Pi^{ij} = P\pi^{ij} = \int n^{ij} \delta f E^3 \, \mathrm{d}E \, \mathrm{d}^2\boldsymbol{n}, \tag{6.3.24}$$

where the trace-free symmetric tensor product of the direction vector is defined by

$$n^{ij} \equiv n^i n^j - \frac{1}{3}\gamma^{ij}. \tag{6.3.25}$$

Brightness and temperature

The radiation brightness (or specific intensity integrated over frequency) I is obtained by integrating the distribution function weighted by E^3 over energy:

$$I(\eta, x^i, n^j) \equiv 4\pi \int E^3 f(\eta, x^i, E, n^j) \, \mathrm{d}E \, . \tag{6.3.26}$$

The brightness represents the directional energy density per unit solid angle propagating along direction n^i at space-time position (η, x^j). Splitting into background and perturbation, $I = \bar{I} + \delta I$:

$$\bar{I}(\eta) \equiv 4\pi \int E^3 \bar{f}(\eta, E) \, \mathrm{d}E \, , \qquad \delta I(\eta, x^i, n^j) \equiv 4\pi \int E^3 \delta f(\eta, x^i, E, n^j) \, \mathrm{d}E \, . \tag{6.3.27}$$

From Eq. (6.3.19), the total photon energy density ρ_γ is precisely the angular monopole of the brightness:

$$\rho_\gamma(\eta, x^j) = \int \frac{\mathrm{d}^2\boldsymbol{n}}{4\pi} I(\eta, x^i, n^j). \tag{6.3.28}$$

The background equation (6.3.13) immediately implies

$$\bar{I}' + 4\mathcal{H}\bar{I} = 0, \qquad \bar{\rho}'_\gamma + 4\mathcal{H}\bar{\rho}_\gamma = 0, \tag{6.3.29}$$

with $\bar{I} = \bar{\rho}_\gamma$, in agreement with background spatial isotropy.

Using boltzmann equation (6.3.13)

$$\begin{aligned}\bar{I}' &= 4\pi \int E^3 \bar{f}'(\eta, E) dE \\ &= 4\pi \int \mathcal{H} E^4 \frac{\partial \bar{f}(\eta, E)}{\partial E} dE \qquad \text{(using } f' = \mathcal{H} E \tfrac{\partial \bar{f}(\eta,E)}{\partial E}\text{)} \\ &= -12\pi \mathcal{H} \int E^3 \bar{f}(\eta, E) dE \qquad (1)\end{aligned}$$

Similarly, we define the energy-integrated collision operator acting on brightness perturbations:

$$C[\delta I] = 4\pi \int C[\delta f] E^3 \, \mathrm{d}E \,. \tag{6.3.30}$$

Integrating the first-order Liouville equation (6.3.14) over energy weighted by $4\pi E^3 \, \mathrm{d}E$ yields the brightness Boltzmann equation:

$$\delta I' + n^k \partial_k \delta I + 4\mathcal{H}\delta I + 4\left[n^k \partial_k A + \left(\frac{1}{2} h'_{ij} - D_{(i} B_{j)}\right) n^i n^j\right] \bar{I} - {}^{(3)}\Gamma^k_{ij} n^i n^j \frac{\partial \delta I}{\partial n^k} = C[\delta I]. \tag{6.3.31}$$

The fractional temperature perturbation is defined from the brightness fluctuation via

$$\theta(\eta, x^i, n^i) = \frac{1}{4} \frac{\delta I}{\bar{I}}, \tag{6.3.32}$$

in direct correspondence with the ansatz of Eq. (6.3.16). The angular monopole of θ is directly proportional to the photon density contrast δ_γ:

$$\theta_0(\eta, x^i) = \int \theta(\eta, x^i, n^i) \frac{\mathrm{d}^2 \boldsymbol{n}}{4\pi} = \frac{1}{4} \delta_\gamma. \tag{6.3.33}$$

Using the background scaling $\bar{I}' = -4\mathcal{H}\bar{I}$ [Eq. (6.3.29)], the evolution equation for θ follows from Eq. (6.3.31):

$$\boxed{\theta' + n^k \partial_k \theta - {}^{(3)}\Gamma^k_{ij} n^i n^j \frac{\partial \theta}{\partial n^k} + \left[n^k \partial_k A + \left(\frac{1}{2} h'_{ij} - D_{(i} B_{j)}\right) n^i n^j\right] = C[\theta],} \tag{6.3.34}$$

where the normalized collision term is defined by $C[\theta] = C[\delta I]/(4\bar{I})$. This equation can alternatively be derived by inserting the temperature ansatz (6.3.16) directly into Eq. (6.3.14). Because the Thomson collision operator is energy-independent, the two formulations are identical.

Multipole expansion

In analogy with Eq. (6.3.28), the macroscopic fluid perturbations emerge as the low-order angular moments of the brightness fluctuation from (6.3.23) and (6.3.24) using (6.3.26):

$$(\rho + P)(v^i + B^i) = \int \delta I n^i \frac{\mathrm{d}^2 \boldsymbol{n}}{4\pi}, \qquad \Pi^{ij} = \int n^{ij} \delta I \frac{\mathrm{d}^2 \boldsymbol{n}}{4\pi}. \tag{6.3.35}$$

To convert the directional partial differential equation (6.3.31) into a hierarchy of spatial tensor equations, we expand the brightness perturbation in symmetric trace-free tensor moments:

$$\delta I(\eta, x^i, n^j) = \sum_{\ell=0}^{\infty} \left(n^{i_1} \dots n^{i_\ell}\right) \mathcal{I}_{i_1 \dots i_\ell}(\eta, x^i). \tag{6.3.36}$$

Gauge transformations in linear perturbation theory affect only the scalar and vector metric perturbations; hence, only the monopole $\mathcal{I}_0$ and dipole $\mathcal{I}_{i_1}$ transform under a change of coordinates, while all higher multipoles ($\ell \geq 2$) are automatically gauge invariant. To make the

decomposition unique, the moment tensors $\mathcal{I}_{i_1\ldots i_\ell}$ are chosen to be completely symmetric and trace-free (STF). Under these conditions, an STF tensor of rank ℓ in three spatial dimensions possesses exactly $2\ell+1$ independent degrees of freedom:

$$\frac{(\ell+1)(\ell+2)}{2} - \frac{\ell(\ell-1)}{2} = 2\ell+1.$$

Only the first three multipoles ($\ell = 0, 1, 2$) represent fluid quantities that couple directly into Einstein's field equations. A completely parallel STF multipole expansion can be applied to the temperature perturbation field $\theta(\eta, x^i, n^j)$, defining the moment tensors $\theta_{i_1\ldots i_\ell}(\eta, x^i)$.

The expansion (6.3.36) substantially streamlines the Boltzmann equation: in the advection term $n^k\partial_k\delta I$, the spatial derivative acts exclusively on the tensor coefficients $\mathcal{I}_{i_1\ldots i_\ell}$, enabling it to be combined with the spatial connection coefficients ${}^{(3)}\Gamma^k_{ij}$ into a spatial covariant derivative D_k. Eq. (6.3.34) then adopts the concise form:

$$\theta' + \sum_{\ell=0}^{\infty} n^k \left(n^{i_1}\ldots n^{i_\ell}\right) D_k\theta_{i_1\ldots i_\ell} + \left[n^k\partial_k A + \left(\frac{1}{2}h'_{ij} - D_{(i}B_{j)}\right) n^i n^j\right] = C[\theta]. \tag{6.3.37}$$

Because the metric perturbation terms depend on direction only up to quadratic order in n^i, they enter exclusively into the monopole and dipole evolution equations, matching the fluid equations of motion.

The scalar-vector-tensor (SVT) decomposition established in Chapter 4 extends naturally to arbitrary rank-ℓ tensors. Recall that any symmetric, trace-free rank-2 tensor (such as the anisotropic stress π_{ij}) can be decomposed into scalar, vector, and tensor components as

$$\pi_{ij} = \Delta_{ij}\pi + D_{(i}\pi_{j)} + \bar{\pi}_{ij},$$

with the individual projections defined by

$$\pi^{(\mathrm{S})}_{ij} = [D_{ij}\pi]^{\mathrm{STF}}, \qquad \pi^{(\mathrm{V})}_{ij} = [D_i\pi_j]^{\mathrm{STF}}, \qquad \pi^{(\mathrm{T})}_{ij} = [\bar{\pi}_{ij}]^{\mathrm{STF}},$$

where STF denotes the extraction of the symmetric trace-free part and where $\pi^{(\mathrm{V})}_{ij}$ and $\bar{\pi}_{ij}$ are constrained to be transverse.

For a general tensor moment of rank ℓ, the component of order m decomposes as

$$\theta^{(m)}_{i_1\ldots i_\ell} = \left[D_{i_1\ldots i_{\ell-m}}\theta^{(m)}_{\ell,i_{\ell-m+1}\ldots i_\ell}\right]^{\mathrm{STF}}. \tag{6.3.38}$$

The indices $|m| = 0, 1, 2$ correspond to scalar (S), vector (V), and tensor (T) perturbations, respectively. Decomposing into Fourier modes $\hat{\theta}(\boldsymbol{k})$ with orientation $\boldsymbol{e}_3 = \hat{\boldsymbol{k}} = \boldsymbol{k}/k$, the scalar sector of $\hat{\theta}_{i_1\ldots i_\ell}$ is proportional to $k_{i_1}\ldots k_{i_\ell}\hat{\theta}^{(0)}_\ell$. Contracting this tensor with the STF product of direction vectors $n^{i_1}\ldots n^{i_\ell}$ yields the Legendre polynomial $P_\ell(\hat{\boldsymbol{k}}\cdot\boldsymbol{n})$, so that

$$n^{i_1}\ldots n^{i_\ell}\hat{\theta}^{(0)}_{i_1\ldots i_\ell} \propto Y_{\ell 0}\hat{\theta}^{(0)}_\ell.$$

Vector modes possess two transverse polarization degrees of freedom orthogonal to $\boldsymbol{k}$ [cf. Eq. (6.1.32)]. Choosing the helicity basis vectors $\boldsymbol{e}^\pm = \boldsymbol{e}_1 \pm \boldsymbol{e}_2$, the vector portion of $\hat{\theta}_{i_1\ldots i_\ell}$ is proportional to the STF part of $k_{i_1}\ldots k_{i_{\ell-1}}e^\pm_{i_\ell}\hat{\theta}^{(\pm1)}_\ell$. Contraction with $n^{i_1}\ldots n^{i_\ell}$ then leads to spherical harmonics of order $m = \pm1$:

$$n^{i_1}\ldots n^{i_\ell}\hat{\theta}^{(\pm1)}_{i_1\ldots i_\ell} \propto Y_{\ell,\pm1}\hat{\theta}^{(\pm1)}_\ell.$$

Generalizing this construction to all azimuthal indices m, the Fourier moment $\hat{\theta}_{i_1\ldots i_\ell}$ expands as

$$\hat{\theta}_{i_1\ldots i_\ell} = \sum_{m=-\ell}^{\ell} \hat{\theta}^{(m)}_{i_1\ldots i_\ell}, \tag{6.3.39}$$

where

$$\hat{\theta}^{(m)}_{i_1\dots i_\ell} \propto \left[k_{i_1}\dots k_{i_{\ell-m}} e^{\pm}_{i_{\ell-m+1}} \otimes \cdots \otimes e^{\pm}_{i_\ell}\right]^{\mathrm{STF}} \hat{\theta}^{(m)}_\ell, \tag{6.3.40}$$

taking the + sign for $m \geq 0$ and the − sign for $m < 0$. Projecting onto the directional line of sight yields (see Chapter 5 of Ref. [3] and Ref. [20] for explicit tensor properties):

$$n^{i_1}\dots n^{i_\ell}\hat{\theta}^{(m)}_{i_1\dots i_\ell} \propto Y_{\ell m}\hat{\theta}^{(m)}_\ell. \tag{6.3.41}$$

Each perturbation type carries 2 degrees of freedom (except for the scalar mode $m = 0$, which possesses only 1). For each multipole ℓ, there are thus $2\ell + 1$ independent degrees of freedom, precisely spanning the spherical harmonics $Y_{\ell m}$ on the celestial sphere.

This construction demonstrates that the radiation temperature perturbation $\theta(\eta, \boldsymbol{x}, \boldsymbol{n})$ at spatial position $\boldsymbol{x}$ propagating along direction $\boldsymbol{n}$ can be expanded in the complete basis functions $G_{\ell m}(\boldsymbol{x}, \boldsymbol{n})$ defined by

$$G_{\ell m}(\boldsymbol{x}, \boldsymbol{n}) = (-\mathrm{i})^\ell \sqrt{\frac{4\pi}{2\ell+1}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} Y_{\ell m}(\boldsymbol{n}), \tag{6.3.42}$$

yielding the modal expansion

$$\theta(\eta, \boldsymbol{x}, \boldsymbol{n}) = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^{3/2}} \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \hat{\theta}^{(m)}_\ell(\eta, \boldsymbol{k}) G_{\ell m}(\boldsymbol{x}, \boldsymbol{n}). \tag{6.3.43}$$

At linear order, cosmological perturbations of the matter and space-time geometry only excite the low-order azimuthal modes $m = 0$ (scalar), $m = \pm 1$ (vector), and $m = \pm 2$ (tensor).

Relation with the SVT decomposition

Expanding scalar plane-wave modes according to

$$Q^{(0)} = \exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}), \qquad Q^{(0)}_i = \partial_i Q^{(0)}, \qquad Q^{(0)}_{ij} = -\left[\partial_i\partial_j - \frac{1}{3}\delta_{ij}\Delta\right] Q^{(0)},$$

the corresponding angular projections evaluate to

$$Q^{(0)} = G_{00}, \qquad n^i Q^{(0)}_i = -kG_{10}, \qquad n^i n^j Q^{(0)}_{ij} \propto k^2 G_{20}. \tag{6.3.44}$$

Similarly, the vector modes are expanded as

$$Q^{(\pm1)}_i = -\frac{\mathrm{i}}{\sqrt{2}} e^{\pm}_i Q^{(0)}, \qquad Q^{(\pm1)}_{ij} = \partial_{(i} Q^{(\pm1)}_{j)},$$

so that contracting with the line-of-sight direction yields

$$n^i Q^{(\pm1)}_i = G_{1,\pm1}, \qquad n^i n^j Q^{(\pm1)}_{ij} \propto -kG_{2,\pm1}. \tag{6.3.45}$$

This decomposition directly parallels Eq. (6.1.32) introduced earlier. Finally, the transverse-traceless tensor modes are expanded as

$$Q^{(\pm2)}_{ij} = -\sqrt{\frac{3}{8}} e^{\pm}_i \otimes e^{\pm}_j Q^{(0)},$$

leading to the angular projection

$$n^i n^j Q^{(\pm2)}_{ij} = G_{2,\pm2}. \tag{6.3.46}$$

Note that this harmonic convention differs from the basis defined in Eq. (6.1.40) by a normalising prefactor of $\sqrt{3/8}$.

Collision term

We now incorporate the Thomson scattering of photons off free electrons into the relativistic Boltzmann equation. The collision term describes the net rate of change of the distribution function along the geodesic:

$$C[f] = \left(\frac{\mathrm{d}f}{\mathrm{d}\eta}\right)_{\rm coll} = \frac{\mathrm{d}f^+}{\mathrm{d}\eta} - \frac{\mathrm{d}f^-}{\mathrm{d}\eta} = \tau' \left[\int \frac{\mathrm{d}\Omega'}{4\pi} f(\boldsymbol{n}')\omega(\boldsymbol{n}', \boldsymbol{n}) - f(\boldsymbol{n})\right], \tag{6.3.47}$$

where $\mathrm{d}f^+/\mathrm{d}\eta$ denotes scattering into direction $\boldsymbol{n}$ (gain term) and $\mathrm{d}f^-/\mathrm{d}\eta$ represents scattering out of direction $\boldsymbol{n}$ (loss term). In the local rest frame of the baryon-electron plasma, these two rates are given by

$$\frac{\mathrm{d}f^+}{\mathrm{d}\eta} = \tau' \int f(\boldsymbol{n}')\omega(\boldsymbol{n}, \boldsymbol{n}')\frac{\mathrm{d}^2\boldsymbol{n}'}{4\pi}, \qquad \frac{\mathrm{d}f^-}{\mathrm{d}\eta} = \tau' f(\boldsymbol{n}), \qquad \mathrm{d}\boldsymbol{n}' = \sin\theta'\,\mathrm{d}\theta'\,\mathrm{d}\phi', \tag{6.3.48}$$

where the differential optical depth is defined as

$$\tau' \equiv a\sigma_{\rm T} n_{\rm e}, \tag{6.3.49}$$

with $\sigma_{\rm T} = 8\pi\alpha^2/(3m_{\rm e}^2)$ the Thomson cross section and $n_{\rm e}$ the number density of free electrons. The unpolarized Thomson scattering phase function $\omega(\boldsymbol{n}, \boldsymbol{n}')$ encodes the angular probability distribution:

$$\omega(\boldsymbol{n}, \boldsymbol{n}') = \frac{3}{4}\left[1 + (\boldsymbol{n}\cdot\boldsymbol{n}')^2\right] = 1 + \frac{3}{4} n_{ij} n'^{ij} = 1 + \frac{1}{2} P_2(\boldsymbol{n}\cdot\boldsymbol{n}'), \tag{6.3.50}$$

where $P_2(\mu) = \frac{1}{2}(3\mu^2 - 1)$ is the second Legendre polynomial. It is normalized such hat

$$\int \frac{\mathrm{d}\boldsymbol{n}'}{4\pi}\omega(\boldsymbol{n}', \boldsymbol{n}) = \frac{3}{4}\int \frac{\mathrm{d}\boldsymbol{n}'}{4\pi}(1 + (\boldsymbol{n}\cdot\boldsymbol{n}')) = 1. \tag{6.3.51}$$

One natural choice for $\boldsymbol{n}$ and $\boldsymbol{n}'$ is

$$\boldsymbol{n} = (0, 0, 1) \tag{6.3.52}$$

$$\boldsymbol{n}' = (\sin\theta'\cos\phi', \sin\theta'\sin\phi', \cos\theta'). \tag{6.3.53}$$

A crucial property of non-relativistic Thomson scattering is that $\omega(\boldsymbol{n}, \boldsymbol{n}')$ and the resulting collision operator are strictly energy-independent. Thomson scattering therefore induces no spectral distortion, justifying our ansatz that the fractional brightness fluctuation $\delta I/(4\bar{I})$ corresponds directly to a temperature perturbation θ [Eqs. (6.3.16), (6.3.32)].

At the background level, the angular symmetry of $\omega(\boldsymbol{n}, \boldsymbol{n}')$ ensures that

$$C[\bar{f}] = 0, \tag{6.3.54}$$

as required by the spatial isotropy of the unperturbed Friedmann–Lemaître background, where all fluid velocities vanish in comoving coordinates ($u^\mu \propto \delta_0^\mu$).

To first order in cosmological perturbations, transforming from the baryon rest frame to the general coordinate system yields

$$C[\delta f] = \tau' \left[\int \delta f(\eta, x^i, E, \boldsymbol{n}')\frac{\mathrm{d}^2\boldsymbol{n}'}{4\pi} - \delta f(\eta, x^i, E, \boldsymbol{n}) + \frac{3}{4} n_{ij}\int \delta f(\eta, x^i, E, \boldsymbol{n}') n'^{ij}\frac{\mathrm{d}^2\boldsymbol{n}'}{4\pi}\right], \tag{6.3.55}$$

where the photon energy E in the electron frame is related to the coordinate energy $\bar{E}$ by $E = -k^\mu u_\mu^{\rm b} = \bar{E}[1 - (B_i + v_i^{\rm b})n^i]$, giving phase-space measure transformation

$$E^3\,\mathrm{d}E = [1 + (B^i + v_b^i)n_i]^4 \bar{E}^3\,\mathrm{d}\bar{E} \approx \left[1 + 4(B_i + v_i^{\rm b})n^i\right]\bar{E}^3\mathrm{d}\bar{E}.$$

Integrating Eq. (6.3.55) over energy weighted by $4\pi E^3\,\mathrm{d}E$ yields the collision operator for brightness perturbations:

$$C[\delta I] = \tau' \left[\int \delta I \frac{\mathrm{d}^2\boldsymbol{n}'}{4\pi} - \delta I + 4\bar{I}(B_i + v_{\rm b}^i)n^i + \frac{3}{4} n_{ij} \int \delta I(\eta, x^i, \boldsymbol{n}') n'^{ij} \frac{\mathrm{d}^2\boldsymbol{n}'}{4\pi} \right]. \tag{6.3.56}$$

The first term simplifies to the brightness monopole $\bar{I}\delta_\gamma$, and the quadrupole term evaluates to $3n^{ij}\Pi_{ij} = (4\bar{I})n^i n^j \pi_{ij}/16$. Defining the normalized temperature collision term $C[\theta] = C[\delta I]/(4\bar{I})$, we obtain:

$$\boxed{C[\theta] = \tau' \left[\theta_0 - \theta + (B^i + v_{\rm b}^i)n_i + \frac{1}{16} n^i n^j \pi_{ij} \right],} \tag{6.3.57}$$

where $\theta_0 = \frac{1}{4}\delta_\gamma$ is the temperature monopole [Eq. (6.3.33)]. Since $C[\theta]$ vanishes identically in the unperturbed universe, the Stewart–Walker lemma (Chapter 4) guarantees that $C[\theta]$ is gauge invariant at linear order. Moreover, in the multipole decomposition, the Thomson collision operator couples to the fluid velocity and metric only up to $\ell = 2$; for all higher multipoles ($\ell > 2$), the collision term reduces simply to

$$C[\theta_\ell] = -\tau'\theta_\ell.$$

We derive the Thomson-scattering collision term keeping the photon momentum measured in the cosmological frame distinct from the physical energy. We will work in Newtonian Gauge and define

$$f = f(x^\mu, k^\mu) \tag{1}$$

to denote the photon distribution function. The photon energy associated with the cosmological coordinate frame is denoted by $\bar{p}$, so that

$$k^\mu = (\bar{p}, \bar{p}\, n^i) \tag{2}$$

to the required order.
The microscopic collision term is written using physical momenta:

$$C_\gamma[f(k)] = \frac{1}{2\bar{p}} \int d\Pi_q\, d\Pi_{k'}\, d\Pi_{q'} (2\pi)^4 \delta^{(4)}(k + q - k' - q') |\mathcal{M}|^2 \left\{ f' f'_e (1 + f) - f f_e (1 + f') \right\}, \tag{3}$$

where

$$d\Pi_k = \frac{d^3k}{(2\pi)^3\, 2E_k}. \tag{4}$$

For non-relativistic electrons,

$$E_q \simeq m_e + \frac{q^2}{2m_e}, \qquad d\Pi_q \simeq \frac{d^3q}{(2\pi)^3\, 2m_e}. \tag{5}$$

In the Thomson limit,

$$|\mathcal{M}|^2 = 24\pi m_e^2 \sigma_T \left[1 + (\mathbf{n} \cdot \mathbf{n}')^2 \right], \qquad \sigma_T = \frac{8\pi\alpha^2}{3m_e^2}. \tag{6}$$

Writing $\mathbf{k} = \bar{p}\mathbf{n}$ and $\mathbf{k}' = \bar{p}'\mathbf{n}'$, spatial momentum conservation gives $\mathbf{q}' = \mathbf{q} + \mathbf{k} - \mathbf{k}'$. Neglecting quadratic recoil, the energy-conserving delta function expands as

$$\delta(\bar{p} + E_q - \bar{p}' - E_{q'}) \simeq \delta \left[\bar{p} - \bar{p}' - \frac{\mathbf{q} \cdot (\mathbf{k} - \mathbf{k}')}{m_e} \right]$$

$$= \delta(\bar{p}-\bar{p}') + \frac{\mathbf{q}\cdot(\mathbf{k}-\mathbf{k}')}{m_e}\frac{\partial}{\partial\bar{p}'}\delta(\bar{p}-\bar{p}'). \tag{7}$$

The electron distribution expansion $f(\boldsymbol{q}') \approx f_e(\boldsymbol{q}) + \frac{\partial f_e}{\partial \boldsymbol{q}}\cdot(\boldsymbol{k}-\boldsymbol{k}')$ yields

$$f'f'_e(1+f) - ff_e(1+f') = f_e(\mathbf{q})(f(\boldsymbol{k}')-f(\boldsymbol{k})) + \frac{\partial f_e}{\partial \mathbf{q}}\cdot(\boldsymbol{k}-\boldsymbol{k}')f(\boldsymbol{k}')(1+f(\boldsymbol{k})). \tag{8}$$

Since the second term corresponds to a boundary term, integrating over electron momentum $\mathbf{q}$, the electron density n_e and bulk baryon velocity v_b^i are defined by[a]

$$n_e = 2\int\frac{d^3q}{(2\pi)^3}f_e(\mathbf{q}), \qquad v_b^i = \frac{2}{n_e}\int\frac{d^3q}{(2\pi)^3}\frac{q^i}{m_e}f_e(\mathbf{q}). \tag{10}$$

Thus, the derivative of the delta function in Eq. (7) carries the first moment of the electron distribution $\mathbf{v}_b$:

$$C_\gamma[f] = \frac{3\sigma_T n_e}{16\pi\bar{p}}\int\frac{d^3k'}{\bar{p}'}\left[1+(\mathbf{n}\cdot\mathbf{n}')^2\right][f'-f]\left[\delta(\bar{p}-\bar{p}') + \mathbf{v}_b\cdot(\mathbf{k}-\mathbf{k}')\frac{\partial}{\partial\bar{p}'}\delta(\bar{p}-\bar{p}')\right]. \tag{11}$$

Performing the radial $\bar{p}'$ integration via integration by parts on the second term gives

$$\int\bar{p}'d\bar{p}'[f(\bar{p}',\mathbf{n}')-f(\bar{p},\mathbf{n})]\mathbf{v}_b\cdot(\bar{p}\mathbf{n}-\bar{p}'\mathbf{n}')\frac{\partial}{\partial\bar{p}'}\delta(\bar{p}-\bar{p}') = -\bar{p}^2 f^{(0)\prime}(\bar{p})\mathbf{v}_b\cdot(\mathbf{n}-\mathbf{n}'). \tag{12}$$

Therefore, Eq. (11) evaluates to

$$C_\gamma[f] = \frac{3\sigma_T n_e}{16\pi}\int d\Omega'\left[1+(\mathbf{n}\cdot\mathbf{n}')^2\right]\left[\delta f(\bar{p},\mathbf{n}') - \delta f(\bar{p},\mathbf{n}) - \bar{p}f^{(0)\prime}(\bar{p})\,\mathbf{v}_b\cdot(\mathbf{n}-\mathbf{n}')\right]. \tag{13}$$

Expanding $\delta f(\bar{p},\mathbf{n}) = -\bar{p}f^{(0)\prime}(\bar{p})\Theta(\mathbf{n})$ and integrating the incoming dipole term $\mathbf{v}_b\cdot\mathbf{n}'$ to zero by parity, we find

$$\boxed{C_\gamma[f] = -\frac{3\sigma_T n_e}{16\pi}\bar{p}f^{(0)\prime}(\bar{p})\int d\Omega'\left[1+(\mathbf{n}\cdot\mathbf{n}')^2\right]\left[\Theta(\mathbf{n}')-\Theta(\mathbf{n})+v_i^b n^i\right].} \tag{14}$$

Using the identity $1+(\mathbf{n}\cdot\mathbf{n}')^2 = \frac{4}{3}+n_{ij}n'^{ij}$ with $n^{ij} = n^in^j - \frac{1}{3}\delta^{ij}$, the collision integral becomes

$$C_\gamma[\delta f] = \sigma_T n_e\left[\int\frac{d\Omega'}{4\pi}\delta f(\mathbf{n}') - \delta f(\mathbf{n}) - \bar{p}f^{(0)\prime}(\bar{p})v_i^b n^i\right] + \frac{3}{4}\sigma_T n_e\, n_{ij}\int\frac{d\Omega'}{4\pi}\delta f(\mathbf{n}')n'^{ij}. \tag{15}$$

Transforming to conformal time via $C_\eta[\delta f] = aC_t[\delta f]$ and defining optical depth $\tau' = a\sigma_T n_e$, the collision term for the temperature perturbation $\Theta(\mathbf{n})$ becomes

$$C[\Theta] = \tau'\left[\int\frac{d\Omega'}{4\pi}\Theta(\mathbf{n}') - \Theta(\mathbf{n}) + v_i^b n^i + \frac{3}{4}n_{ij}\int\frac{d\Omega'}{4\pi}\Theta(\mathbf{n}')n'^{ij}\right]. \tag{16}$$

[a] In general form where $u_\mu^b = a(-1-A, v_i^b + B_i)$ the second equation becomes

$$v_i^b + B_i = \frac{2}{n_e}\int\frac{d^3q}{(2\pi)^3}\frac{q_i}{m_e}f_e(\mathbf{q}). \tag{9}$$

6.3.2 Gauge invariant expressions

A comprehensive treatment of the gauge transformation properties of relativistic distribution functions is given in Refs. [19, 21].

Gauge invariant distribution function

The split $f = \bar{f} + \delta f$ depends on the choice of background spacetime mapping. In linear perturbation theory, defining δf requires choosing an isomorphism ψ between the physical and background mass shells:

$$\psi : \mathcal{P}_m \to \mathcal{P}_m : (x^\alpha, p^\beta) \mapsto (\bar{x}^\alpha, \bar{p}^\beta),$$

so that

$$f = \bar{f} \circ \psi + \delta f. \tag{6.3.58}$$

Under an infinitesimal coordinate transformation $x^\alpha \to y^\alpha = x^\alpha - \xi^\alpha$, the 4-momentum transforms according to

$$p^\mu \to p^\mu - p^\nu \partial_\nu \xi^\mu. \tag{6.3.59}$$

This phase-space transformation is generated by the tangent vector field

$$T_\xi = (\xi^\mu,\ p^\nu \partial_\nu \xi^\mu) \tag{6.3.60}$$

in the natural coordinate basis $(\partial_\mu, \partial/\partial p^\mu)$. To first order, the distribution function transforms via the Lie derivative along T_ξ:

$$f \to f + \pounds_{T_\xi} f. \tag{6.3.61}$$

Because the background $\bar{f}$ is defined strictly on the mass shell $\mathcal{P}_m$, and the generator T_ξ is not necessarily tangent to $\mathcal{P}_m$, the gauge variation of δf is properly defined by projecting T_ξ along the mass shell ($T_\xi^{\parallel}$), as shown in Ref. [21]. Under a general gauge transformation, the perturbation δf transforms as

$$\boxed{\delta f \to \delta f + (\mathcal{H}T + n^i \partial_i T) E \frac{\partial \bar{f}}{\partial E},} \tag{6.3.62}$$

where $T \equiv \xi^0$ is the temporal gauge shift parameter. This result can also be understood physically: the photon energy $E = -N_\mu k^\mu = a(1+A)k^0$ shifts by $E \to E - E n^i \partial_i T$, which combined with the background Liouville relation $\bar{f}' = \mathcal{H} E \partial \bar{f}/\partial E$ [Eq. (6.3.13)] reproduces Eq. (6.3.62).

We can therefore construct a gauge invariant distribution function $\mathcal{F}$ by combining δf with scalar metric perturbations:

$$\boxed{\mathcal{F} \equiv \delta f - \Big[C + n^i \partial_i (E' - B)\Big] E \frac{\partial \bar{f}}{\partial E},} \tag{6.3.63}$$

where E in $(E' - B)$ represents the scalar metric shear perturbation, while outside the brackets E denotes photon energy. Integrating $\mathcal{F}$ over momentum space gives

$$\int \mathcal{F} E^3 \, \mathrm{d}E \, \mathrm{d}^2\boldsymbol{n} = \delta\rho_\gamma + 4C\bar{\rho}_\gamma = \bar{\rho}_\gamma \delta_\gamma^{\mathcal{F}},$$

revealing that $\mathcal{F}$ represents the distribution perturbation in the flat-slicing gauge ($C = 0$).

Alternatively, one can define the gauge invariant distribution function in the conformal Newtonian (longitudinal) gauge:

$$\bar{\mathcal{F}} \equiv \mathcal{F} + \Psi E \frac{\partial \bar{f}}{\partial E} = \mathcal{F} + \bar{f}' \frac{\Psi}{\mathcal{H}}. \tag{6.3.64}$$

Integrating over energy, the corresponding gauge invariant temperature perturbations are denoted by $\mathcal{M}$ (flat-slicing gauge) and Θ (Newtonian gauge). They are related by

$$\mathcal{M} = \Theta - \Psi,$$

and satisfy the respective gauge invariant Boltzmann equations:

$$\mathcal{M}' + n^k\partial_k\mathcal{M} - {}^{(3)}\Gamma^k_{ij}n^in^j\frac{\partial\mathcal{M}}{\partial n^k} + \left[n^k\partial_k(\Phi+\Psi) + \left(\bar{E}'_{ij} + D_{(i}\bar{\Phi}_{j)}\right)n^in^j\right] = C[\mathcal{M}], \tag{6.3.65}$$

$$\Theta' + n^k\partial_k(\Theta+\Phi) - {}^{(3)}\Gamma^k_{ij}n^in^j\frac{\partial\Theta}{\partial n^k} + \left[-\Psi' + \left(\bar{E}'_{ij} + D_{(i}\bar{\Phi}_{j)}\right)n^in^j\right] = C[\Theta]. \tag{6.3.66}$$

Because the metric perturbation source terms depend on direction only up to quadratic order in n^i, these two equations differ only at the level of the monopole and dipole.

Boltzmann equation for the temperature

Specializing to a spatially flat FLRW background ($K = 0$) and restricting attention to scalar perturbations, the Newtonian-gauge Boltzmann equation (6.3.66) in Fourier space becomes

$$\Theta' + \mathrm{i}k\mu(\Theta+\Phi) = \Psi' + \overbrace{\tau'\left[\Theta_0 - \Theta + \mathrm{i}k\mu V_\mathrm{b} + \frac{1}{16}n_in_j\Pi^{ij}\right]}^{\text{from eq. (6.3.57)}}, \tag{6.3.67}$$

where $\mu \equiv \hat{\boldsymbol{k}}\cdot\boldsymbol{n}$ is the cosine of the angle between the wavevector and photon direction. We expand the directional dependence of $\Theta(\boldsymbol{k},\eta,\mu)$ in Legendre polynomials:

$$\Theta(\boldsymbol{k},\eta,\mu) = \sum_{\ell=0}^{\infty}(-\mathrm{i})^\ell\Theta_\ell(k,\eta)P_\ell(\mu)\mathrm{e}^{+\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}. \tag{6.3.68}$$

> Some people also use the definition
>
> $$\Theta(\boldsymbol{k},\eta,\mu) = \sum_{\ell=0}^{\infty}(-\mathrm{i})^\ell\Theta_\ell(k,\eta)(2\ell+1)P_\ell(\mu)\mathrm{e}^{+\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}. \tag{1}$$

Using the recurrence relation for Legendre polynomials,

$$(2\ell+1)\mu P_\ell(\mu) = (\ell+1)P_{\ell+1}(\mu) + \ell P_{\ell-1}(\mu),$$

the streaming term $\mathrm{i}k\mu\Theta$ is transformed into

$$k\sum_{\ell=0}^{\infty}\left[\frac{\ell+1}{2\ell+3}\Theta_{\ell+1} - \frac{\ell}{2\ell-1}\Theta_{\ell-1}\right](-\mathrm{i})^\ell P_\ell(\mu).$$

> Starting from
>
> $$\mu\Theta(k,\eta,\mu) = \sum_{\ell=0}^{\infty}(-i)^\ell\Theta_\ell(k,\eta)\left[\frac{\ell+1}{2\ell+1}P_{\ell+1}(\mu) + \frac{\ell}{2\ell+1}P_{\ell-1}(\mu)\right]. \tag{1}$$

we can simplify

$$\sum_{\ell=0}^{\infty}(-i)^{\ell}\Theta_{\ell}\frac{\ell+1}{2\ell+1}P_{\ell+1}(\mu) = i\sum_{\ell=1}^{\infty}(-i)^{\ell}\Theta_{\ell-1}\frac{\ell}{2\ell-1}P_{\ell}(\mu) \tag{2}$$

$$\sum_{\ell=0}^{\infty}(-i)^{\ell}\Theta_{\ell}\frac{\ell}{2\ell+1}P_{\ell-1}(\mu) = -i\sum_{\ell=0}^{\infty}(-i)^{\ell}\Theta_{\ell+1}\frac{\ell+1}{2\ell+3}P_{\ell}(\mu) \tag{3}$$

Since P_{-1}, Θ_{-1} doesn't exist, we can start each sum from $\ell = 0$ itself by setting them to zero.

Projecting onto individual Legendre multipoles using $\int_{-1}^{1} P_\ell(\mu)P_{\ell'}(\mu)\mathrm{d}\mu = \frac{2}{2\ell+1}\delta_{\ell\ell'}$, and noting that $\Pi_{ij}n^i n^j/16 = (-\mathrm{i})^2 P_2(\mu)\Theta_2/10$, the infinite hierarchy of coupled Boltzmann moment equations emerges after comparing the coefficient of P_ℓ for $\ell = 0, 1\ldots$:

$$\Theta_0' = -\frac{k}{3}\Theta_1 + \Psi', \tag{6.3.69}$$

$$\Theta_1' = k\left(\Theta_0 - \frac{2}{5}\Theta_2 + \Phi\right) - \tau'(kV_\mathrm{b} + \Theta_1), \tag{6.3.70}$$

$$\Theta_2' = k\left(\frac{2}{3}\Theta_1 - \frac{3}{7}\Theta_3\right) - \frac{9}{10}\tau'\Theta_2, \tag{6.3.71}$$

$$\Theta_\ell' = k\left(\frac{\ell}{2\ell-1}\Theta_{\ell-1} - \frac{\ell+1}{2\ell+3}\Theta_{\ell+1}\right) - \tau'\Theta_\ell \qquad (\ell \geq 3). \tag{6.3.72}$$

As anticipated, metric perturbations and baryonic velocity appear exclusively in the monopole and dipole equations ($\ell = 0, 1$), while the anisotropic shear quadrupole damping enters at $\ell = 2$ with coefficient $\frac{9}{10}\tau'$. For all multipoles $\ell \geq 3$, the collision operator reduces strictly to $-\tau'\Theta_\ell$.

The lowest-order multipoles connect directly to macroscopic fluid variables:

$$4\Theta_0 = \delta_\gamma^{\mathrm{N}}, \qquad \Theta_1 = -kV_\gamma, \qquad \Theta_2 = \frac{5}{12}k^2\pi_\gamma. \tag{6.3.73}$$

The factor $-k$ between Θ_1 and V_γ reflects the harmonic normalization of Eq. (6.3.44). Thus, the first two equations of the hierarchy (6.3.69)–(6.3.70) are precisely the continuity and Euler equations of the radiation fluid.

We can derive the relations in (6.3.73) directly from the definitions of the perturbed photon stress tensor. For massless particles, the perturbation of the distribution function is written as

$$f = \bar{f} - \frac{\mathrm{d}\bar{f}}{\mathrm{d}\ln\epsilon}\Theta = \bar{f} - \epsilon\frac{\mathrm{d}\bar{f}}{\mathrm{d}\epsilon}\Theta, \tag{1}$$

where $\epsilon = aE$ is the comoving energy. We use

$$P^\mu = \frac{E}{a}\left[1 - \Psi, (1+\Phi)\hat{\mathbf{p}}\right],$$
$$P_\mu = aE\left[1 + \Psi, -(1-\Phi)\hat{\mathbf{p}}\right].$$

- For the time-time component of the stress tensor,

$$\begin{aligned} T^0{}_0 &= \int \frac{\mathrm{d}^3\mathbf{p}}{E} f\, P^0 P_0 = \int \mathrm{d}^3\mathbf{p}\, f E \\ &= \frac{1}{a^4}\int \mathrm{d}\hat{\mathbf{p}} \int \mathrm{d}\epsilon\, f\, \epsilon^3 \\ &= \frac{4\pi}{a^4}\int \mathrm{d}\epsilon\, \bar{f}\, \epsilon^3 - \frac{1}{a^4}\int \mathrm{d}\epsilon\, \epsilon^4 \frac{\mathrm{d}\bar{f}}{\mathrm{d}\epsilon}\int \mathrm{d}\hat{\mathbf{p}}\, \Theta. \end{aligned} \tag{2}$$

  Integrating the second term by parts gives

$$-\int_0^\infty \epsilon^4 \frac{\mathrm{d}\bar{f}}{\mathrm{d}\epsilon}\,\mathrm{d}\epsilon = 4\int_0^\infty \bar{f}\,\epsilon^3\,\mathrm{d}\epsilon, \tag{3}$$

  where the boundary term vanishes. Hence

$$T^0{}_0 = \bar{\rho}_\gamma \left[1 + 4\int \frac{\mathrm{d}\hat{\mathbf{p}}}{4\pi}\Theta\right]. \tag{4}$$

  Using the multipole expansion

$$\Theta(\eta, \mathbf{k}, \hat{\mathbf{p}}) = \sum_{\ell=0}^{\infty} (-i)^\ell \Theta_\ell(\eta, \mathbf{k}) P_\ell(\hat{\mathbf{k}}\cdot\hat{\mathbf{p}}), \tag{5}$$

  the angular average selects the monopole,

$$\int \frac{\mathrm{d}\hat{\mathbf{p}}}{4\pi}\Theta = \Theta_0. \tag{6}$$

  Therefore

$$\boxed{\delta_\gamma^{\mathrm{N}} = 4\Theta_0}. \tag{7}$$

- Next, consider the momentum density. Following the definition of the book,

$$(\bar{\rho}_\gamma + \bar{P}_\gamma)(v_\gamma^i + B^i) = \int n^i\, \delta f\, E^3\, \mathrm{d}E\, \mathrm{d}^2 n, \tag{8}$$

  where $n^i = \hat{p}^i$. Since

$$\bar{P}_\gamma = \frac{1}{3}\bar{\rho}_\gamma,$$

  the energy integral gives

$$\begin{aligned} (\bar{\rho}_\gamma + \bar{P}_\gamma)(v_\gamma^i + B^i) &= -\int n^i \frac{\mathrm{d}\bar{f}}{\mathrm{d}\ln E}\Theta\, E^3\, \mathrm{d}E\, \mathrm{d}^2 n \\ &= 4\bar{\rho}_\gamma \int \frac{\mathrm{d}\hat{\mathbf{p}}}{4\pi}\Theta\, \hat{p}^i. \end{aligned} \tag{9}$$

  Thus

$$v_\gamma^i + B^i = 3\int \frac{\mathrm{d}\hat{\mathbf{p}}}{4\pi}\Theta\, \hat{p}^i. \tag{10}$$

  Substituting the multipole expansion,

$$v_\gamma^i + B^i = 3\sum_\ell (-i)^\ell \Theta_\ell \int \frac{\mathrm{d}\hat{\mathbf{p}}}{4\pi} P_\ell(\hat{\mathbf{k}}\cdot\hat{\mathbf{p}})\hat{p}^i. \tag{11}$$

By rotational symmetry,

$$\int \frac{\mathrm{d}\hat{\mathbf{p}}}{4\pi} P_\ell(\hat{\mathbf{k}}\cdot\hat{\mathbf{p}})\hat{p}^i = A_\ell \hat{k}^i, \tag{12}$$

with

$$\begin{aligned} A_\ell &= \hat{k}_i \int \frac{\mathrm{d}\hat{\mathbf{p}}}{4\pi} P_\ell(\hat{\mathbf{k}}\cdot\hat{\mathbf{p}})\hat{p}^i \\ &= \frac{1}{2}\int_{-1}^{1} \mathrm{d}\mu\, \mu P_\ell(\mu) = \frac{1}{3}\delta_{\ell 1}. \end{aligned} \tag{13}$$

Therefore

$$v_\gamma^i + B^i = -i\Theta_1 \hat{k}^i. \tag{14}$$

We define the scalar velocity potential V_γ through

$$v_\gamma^i + B^i = ik^i V_\gamma = ikV_\gamma \hat{k}^i. \tag{15}$$

Comparison with the previous expression gives

$$\boxed{\Theta_1 = -kV_\gamma}. \tag{16}$$

- Finally, consider the anisotropic stress. Taking the symmetric trace-free part of the spatial stress tensor gives

$$\Pi_\gamma^{ij} = \int \hat{p}^{\langle i}\hat{p}^{j\rangle}\delta f\, E^3\, \mathrm{d}E\, \mathrm{d}^2 n. \tag{17}$$

Using

$$\delta f = -E\frac{\mathrm{d}\bar{f}}{\mathrm{d}E}\Theta, \tag{18}$$

we obtain

$$\begin{aligned} \Pi_\gamma^{ij} &= -\int \hat{p}^{\langle i}\hat{p}^{j\rangle} E^4 \frac{\mathrm{d}\bar{f}}{\mathrm{d}E}\Theta\, \mathrm{d}E\, \mathrm{d}^2 n \\ &= 4\bar{\rho}_\gamma \int \frac{\mathrm{d}\hat{\mathbf{p}}}{4\pi}\Theta\, \hat{p}^{\langle i}\hat{p}^{j\rangle}. \end{aligned} \tag{19}$$

Substituting the multipole expansion gives

$$\Pi_\gamma^{ij} = 4\bar{\rho}_\gamma \sum_\ell (-i)^\ell \Theta_\ell \int \frac{\mathrm{d}\hat{\mathbf{p}}}{4\pi} P_\ell(\hat{\mathbf{k}}\cdot\hat{\mathbf{p}})\hat{p}^{\langle i}\hat{p}^{j\rangle}. \tag{20}$$

By rotational symmetry, the angular integral must have the form

$$\int \frac{\mathrm{d}\hat{\mathbf{p}}}{4\pi} P_\ell(\hat{\mathbf{k}}\cdot\hat{\mathbf{p}})\hat{p}^{\langle i}\hat{p}^{j\rangle} = B_\ell \hat{k}^{\langle i}\hat{k}^{j\rangle}. \tag{21}$$

Contracting both sides with $\hat{k}_i\hat{k}_j$ gives

$$\frac{2}{3}B_\ell = \frac{1}{2}\int_{-1}^{1} \mathrm{d}\mu\, P_\ell(\mu)\left(\mu^2 - \frac{1}{3}\right). \tag{22}$$

Since

$$\mu^2 - \frac{1}{3} = \frac{2}{3}P_2(\mu), \tag{23}$$

orthogonality of the Legendre polynomials yields

$$\frac{2}{3}B_\ell = \frac{1}{3}\int_{-1}^{1} \mathrm{d}\mu\, P_\ell(\mu)P_2(\mu) = \frac{2}{15}\delta_{\ell 2}, \tag{24}$$

and hence

$$B_\ell = \frac{1}{5}\delta_{\ell 2}. \tag{25}$$

Only the quadrupole therefore contributes:

$$\begin{aligned}\Pi_\gamma^{ij} &= 4\bar{\rho}_\gamma(-i)^2\Theta_2\frac{1}{5}\hat{k}^{\langle i}\hat{k}^{j\rangle} \\ &= -\frac{4}{5}\bar{\rho}_\gamma\Theta_2\hat{k}^{\langle i}\hat{k}^{j\rangle}.\end{aligned} \tag{26}$$

We now introduce the scalar anisotropic-stress potential π_γ by

$$\pi_\gamma^{ij} = -k^2\pi_\gamma\hat{k}^{\langle i}\hat{k}^{j\rangle}. \tag{27}$$

Since

$$\Pi_\gamma^{ij} = \bar{P}_\gamma\pi_\gamma^{ij}, \qquad \bar{P}_\gamma = \frac{1}{3}\bar{\rho}_\gamma, \tag{28}$$

we have

$$\Pi_\gamma^{ij} = -\frac{1}{3}\bar{\rho}_\gamma k^2\pi_\gamma\hat{k}^{\langle i}\hat{k}^{j\rangle}. \tag{29}$$

Comparing this with the result from the momentum integral gives

$$-\frac{4}{5}\bar{\rho}_\gamma\Theta_2 = -\frac{1}{3}\bar{\rho}_\gamma k^2\pi_\gamma, \tag{30}$$

and therefore

$$\boxed{\Theta_2 = \frac{5}{12}k^2\pi_\gamma}. \tag{31}$$

Consequently, the first three photon temperature multipoles are related to the fluid variables by

$$\boxed{4\Theta_0 = \delta_\gamma^{\rm N}, \qquad \Theta_1 = -kV_\gamma, \qquad \Theta_2 = \frac{5}{12}k^2\pi_\gamma}. \tag{32}$$

For primordial tensor perturbations, the Boltzmann equation simplifies to

$$\Theta' + \mathrm{i}k\mu\Theta = -\bar{E}'_{ij}n^in^j - \tau'\left(\Theta - \frac{1}{16}n_in_j\Pi^{ij}\right). \tag{6.3.74}$$

Decomposing $\bar{E}_{ij} = H^{(\pm2)}Q_{ij}^{(\pm2)}$ [cf. Eq. (6.3.44)], the tensor hierarchy begins at the quadrupole $\ell = 2$:

$$\boxed{\Theta'_2 = -k\frac{\sqrt{5}}{7}\Theta_2 - \frac{9}{10}\tau'\Theta_2 - \dot{H},} \tag{6.3.75}$$

$$\boxed{\Theta'_\ell = k\left(\frac{\sqrt{\ell^2-4}}{2\ell-1}\Theta_{\ell-1} - \frac{\sqrt{(\ell+1)^2-4}}{2\ell+3}\Theta_{\ell+1}\right) - \tau'\Theta_\ell \qquad (\ell \geq 3).} \tag{6.3.76}$$

6.3.3 Polarization

Relativistic kinetic theory and polarization formalisms for the CMB are reviewed in Refs. [22–26], with detailed multipole approaches presented in Refs. [3, 27].

Stokes parameters

A monochromatic electromagnetic wave propagating along unit direction $\boldsymbol{e}_3$ possesses an electric field vector lying in the orthogonal transverse plane $(\boldsymbol{e}_1, \boldsymbol{e}_2)$:

$$\boldsymbol{E}(t, \boldsymbol{x}) = \mathrm{Re} \begin{pmatrix} a_1 \exp[\mathrm{i}\theta_1] \\ a_2 \exp[\mathrm{i}\theta_2] \\ 0 \end{pmatrix}. \tag{6.3.77}$$

It is essential to clarify that Eq. (6.3.77) is expressed not in a global Cartesian basis $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$, but rather in the local orthonormal spherical basis $(\boldsymbol{e}_\theta, \boldsymbol{e}_\phi, \hat{\boldsymbol{n}})$ adapted to the photon line of sight $\hat{\boldsymbol{n}} = \boldsymbol{e}_3$. Because the transverse plane $(\boldsymbol{e}_1, \boldsymbol{e}_2) \equiv (\boldsymbol{e}_\theta, \boldsymbol{e}_\phi)$ varies across the celestial sphere, the components a_1 and a_2 represent the physical field amplitudes along the local meridian $(\boldsymbol{e}_\theta)$ and parallel $(\boldsymbol{e}_\phi)$.

When $\theta_1 = \theta_2$, the electric field oscillates in a constant plane at angle $\tan\alpha = a_2/a_1$, describing linear polarization. When the phases differ, the wave is elliptically polarized. A statistical ensemble of radiation is completely described by the four Stokes parameters:

$$I \equiv \langle a_1^2 \rangle + \langle a_2^2 \rangle = I_1 + I_2, \tag{6.3.78}$$

$$Q \equiv \langle a_1^2 \rangle - \langle a_2^2 \rangle = I_1 - I_2, \tag{6.3.79}$$

$$U \equiv 2\langle a_1 a_2 \cos(\theta_1 - \theta_2) \rangle, \tag{6.3.80}$$

$$V \equiv 2\langle a_1 a_2 \sin(\theta_1 - \theta_2) \rangle, \tag{6.3.81}$$

where angle brackets denote time averages over durations much longer than the optical period $2\pi/\omega$. For fully polarized monochromatic radiation, these parameters obey

$$I^2 = Q^2 + U^2 + V^2. \tag{6.3.82}$$

I represents the total intensity; Q quantifies the excess of linear polarization along $\boldsymbol{e}_1$ relative to $\boldsymbol{e}_2$; U measures linear polarization along the diagonal axes at 45°; and V measures the degree of circular polarization (the asymmetry between right- and left-handed helicity states, $V = \langle a_+^2 \rangle - \langle a_-^2 \rangle$). Thomson scattering generates no circular polarization from an initially unpolarized background, so $V = 0$ throughout CMB cosmology.

Stokes parameters in quantum mechanics The Stokes parameters can also be formulated quantum mechanically. Consider photons propagating along the line-of-sight unit vector $\hat{\boldsymbol{n}}$, with two transverse polarization modes along $\boldsymbol{e}_\theta$ and $\boldsymbol{e}_\phi$, described by annihilation operators a_θ and a_ϕ satisfying the canonical commutation relations $[a_i, a_j^\dagger] = \delta_{ij}$ (where $i, j \in \{\theta, \phi\}$). The Stokes operators are defined by

$$\hat{I} = a_\theta^\dagger a_\theta + a_\phi^\dagger a_\phi, \tag{6.3.83}$$

$$\hat{Q} = a_\theta^\dagger a_\theta - a_\phi^\dagger a_\phi, \tag{6.3.84}$$

$$\hat{U} = a_\theta^\dagger a_\phi + a_\phi^\dagger a_\theta, \tag{6.3.85}$$

$$\hat{V} = \mathrm{i}(a_\theta^\dagger a_\phi - a_\phi^\dagger a_\theta). \tag{6.3.86}$$

The observed Stokes parameters correspond to the quantum expectation values of these operators: $I = \langle \hat{I} \rangle$, $Q = \langle \hat{Q} \rangle$, $U = \langle \hat{U} \rangle$, and $V = \langle \hat{V} \rangle$. Here $\hat{I}$ measures the total photon number operator, $\hat{Q}$ measures the difference between θ- and ϕ-polarized populations, $\hat{U}$ isolates linear polarization along axes rotated by 45°, and $\hat{V}$ measures circular polarization.

To express these observables in the circular helicity basis, define the right- and left-circularly polarized modes

$$a_R = \frac{a_\theta + \mathrm{i}a_\phi}{\sqrt{2}}, \qquad a_L = \frac{a_\theta - \mathrm{i}a_\phi}{\sqrt{2}}. \tag{6.3.87}$$

In terms of these helicity operators, the Stokes operators take the form

$$\hat{I} = a_R^\dagger a_R + a_L^\dagger a_L, \tag{6.3.88}$$

$$\hat{Q} = a_R^\dagger a_L + a_L^\dagger a_R, \tag{6.3.89}$$

$$\hat{U} = \mathrm{i}(a_R^\dagger a_L - a_L^\dagger a_R), \tag{6.3.90}$$

$$\hat{V} = a_R^\dagger a_R - a_L^\dagger a_L. \tag{6.3.91}$$

Thus, circular polarization measures the population imbalance between the two helicities ($V = \langle a_R^\dagger a_R\rangle - \langle a_L^\dagger a_L\rangle$), whereas linear polarization is governed entirely by the off-diagonal cross-correlations $\langle a_R^\dagger a_L\rangle$ and $\langle a_L^\dagger a_R\rangle$ between the right- and left-helicity states. Importantly, having equal populations of the two helicities ($\langle a_R^\dagger a_R\rangle = \langle a_L^\dagger a_L\rangle$) only guarantees that circular polarization vanishes ($V = 0$); it does *not* imply that the radiation is unpolarized. If a definite phase relation or coherence exists between the two opposite-helicity modes, the cross-correlations remain nonzero, giving rise to net linear polarization ($Q \neq 0$ or $U \neq 0$). For example, a pure state of 100% linear polarization along $\boldsymbol{e}_\theta$ corresponds to an equal, in-phase quantum superposition of right- and left-circular modes ($a_\theta = (a_R + a_L)/\sqrt{2}$), possessing equal helicity populations yet maximum linear polarization.

Truly unpolarized radiation requires the complete absence of net linear and circular polarization while maintaining nonzero total intensity:

$$\boxed{Q = U = V = 0, \qquad I \neq 0.} \tag{6.3.92}$$

In the helicity basis, this demands not only equal mean occupations across both circular helicities, but also the vanishing of all coherent phase cross-correlations between them:

$$\langle a_R^\dagger a_R\rangle = \langle a_L^\dagger a_L\rangle \neq 0, \qquad \langle a_R^\dagger a_L\rangle = \langle a_L^\dagger a_R\rangle = 0. \tag{6.3.93}$$

Only under both conditions simultaneously do the expectation values of all three polarization operators vanish identically:

$$\langle\hat{Q}\rangle = \langle\hat{U}\rangle = \langle\hat{V}\rangle = 0, \qquad \langle\hat{I}\rangle > 0. \tag{6.3.94}$$

In the early Universe, Thomson scattering of anisotropic CMB radiation generates linear polarization from the quadrupole anisotropy of the incident radiation. However, the standard Thomson-scattering collision term does not generate circular polarization from initially unpolarized radiation.

Under a right-handed rotation of the polarization basis $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ by angle ψ around the propagation axis $\boldsymbol{e}_3$:

$$\boldsymbol{e}_1' = \cos\psi\boldsymbol{e}_1 + \sin\psi\boldsymbol{e}_2, \qquad \boldsymbol{e}_2' = -\sin\psi\boldsymbol{e}_1 + \cos\psi\boldsymbol{e}_2,$$

In the matrix form:

$$\begin{pmatrix}\boldsymbol{e}_1'\\ \boldsymbol{e}_2'\end{pmatrix} = \begin{pmatrix}\cos\psi & \sin\psi\\ -\sin\psi & \cos\psi\end{pmatrix}\begin{pmatrix}\boldsymbol{e}_1\\ \boldsymbol{e}_2\end{pmatrix}; \qquad \begin{pmatrix}a_1'\\ a_2'\end{pmatrix} = \begin{pmatrix}\cos\psi & \sin\psi\\ -\sin\psi & \cos\psi\end{pmatrix}\begin{pmatrix}a_1\\ a_2\end{pmatrix}. \tag{6.3.95}$$

Notice that this is not a rotation on the sphere: it does not move points across the celestial sphere, nor does it rotate the line-of-sight vector $\boldsymbol{n}$. Rather, it is strictly a rotation in the two-dimensional tangent plane $T_{\boldsymbol{n}}(S^2)$, whose axis of rotation is precisely the normal line-of-sight propagation vector $\boldsymbol{n} = \boldsymbol{e}_3$.

Derivation of basis rotation under parallel transport

The origin of the frame rotation angle ψ can be understood geometrically by considering parallel transport of the orthonormal polarization basis on the celestial 2-sphere S^2 (with metric $\mathrm{d}s^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$). The non-zero Christoffel symbols on the sphere are

$$\Gamma^{\theta}_{\phi\phi} = -\sin\theta\cos\theta, \qquad \Gamma^{\phi}_{\theta\phi} = \Gamma^{\phi}_{\phi\theta} = \cot\theta. \tag{1}$$

For the standard orthonormal coordinate basis $(\boldsymbol{e}_1, \boldsymbol{e}_2) \equiv (\partial_\theta, \frac{1}{\sin\theta}\partial_\phi)$, the covariant derivatives $\nabla_a\partial_b = \Gamma^c_{ab}\partial_c$ evaluate to:

$$\nabla_\theta \boldsymbol{e}_1 = \nabla_\theta\partial_\theta = 0, \tag{2}$$

$$\nabla_\phi \boldsymbol{e}_1 = \nabla_\phi\partial_\theta = \Gamma^{\phi}_{\phi\theta}\partial_\phi = \cot\theta\,\partial_\phi = \cos\theta\,\boldsymbol{e}_2, \tag{3}$$

$$\nabla_\theta \boldsymbol{e}_2 = \partial_\theta\left(\frac{1}{\sin\theta}\right)\partial_\phi + \frac{1}{\sin\theta}\nabla_\theta\partial_\phi = -\frac{\cos\theta}{\sin^2\theta}\partial_\phi + \frac{\cot\theta}{\sin\theta}\partial_\phi = 0, \tag{4}$$

$$\nabla_\phi \boldsymbol{e}_2 = \frac{1}{\sin\theta}\nabla_\phi\partial_\phi = \frac{1}{\sin\theta}\Gamma^{\theta}_{\phi\phi}\partial_\theta = -\cos\theta\,\boldsymbol{e}_1. \tag{5}$$

The covariant differential of the basis vectors is therefore

$$\nabla \boldsymbol{e}_1 = \cos\theta\,\mathrm{d}\phi\,\boldsymbol{e}_2, \qquad \nabla \boldsymbol{e}_2 = -\cos\theta\,\mathrm{d}\phi\,\boldsymbol{e}_1. \tag{6}$$

For any vector $\boldsymbol{V} = V^1\boldsymbol{e}_1 + V^2\boldsymbol{e}_2$, the parallel transport condition $\nabla\boldsymbol{V} = 0$ along a path on the sphere implies

$$\begin{aligned}\nabla\boldsymbol{V} &= (\mathrm{d}V^1)\boldsymbol{e}_1 + (\mathrm{d}V^2)\boldsymbol{e}_2 + V^1\nabla\boldsymbol{e}_1 + V^2\nabla\boldsymbol{e}_2 \\ &= (\mathrm{d}V^1 - V^2\cos\theta\,\mathrm{d}\phi)\boldsymbol{e}_1 + (\mathrm{d}V^2 + V^1\cos\theta\,\mathrm{d}\phi)\boldsymbol{e}_2 = 0,\end{aligned} \tag{7}$$

which yields the coupled system for the vector components:

$$\mathrm{d}V^1 = (\cos\theta\,\mathrm{d}\phi)V^2, \qquad \mathrm{d}V^2 = -(\cos\theta\,\mathrm{d}\phi)V^1. \tag{8}$$

Defining the complex linear combination $Z \equiv V^1 + \mathrm{i}V^2$:

$$\mathrm{d}Z = -\mathrm{i}(\cos\theta\,\mathrm{d}\phi)Z \implies \frac{\mathrm{d}Z}{Z} = -\mathrm{i}\cos\theta\,\mathrm{d}\phi. \tag{9}$$

Integrating from an initial state $Z_0 = V^1_0 + \mathrm{i}V^2_0$ along a trajectory with rotation angle $\delta\psi \equiv -\cos\theta\,\mathrm{d}\phi$:

$$\int_{Z_0}^{Z}\frac{\mathrm{d}Z'}{Z'} = \mathrm{i}\int\mathrm{d}\psi \implies Z = Z_0\,\mathrm{e}^{\mathrm{i}\delta\psi}. \tag{10}$$

Expanding via Euler's identity gives

$$V^1 + \mathrm{i}V^2 = (V^1_0 + \mathrm{i}V^2_0)(\cos\delta\psi + \mathrm{i}\sin\delta\psi), \tag{11}$$

so the vector components in the fixed basis rotate as:

$$\begin{pmatrix} V^1 \\ V^2 \end{pmatrix} = \begin{pmatrix} \cos\delta\psi & -\sin\delta\psi \\ \sin\delta\psi & \cos\delta\psi \end{pmatrix}\begin{pmatrix} V^1_0 \\ V^2_0 \end{pmatrix}. \tag{12}$$

Because $\boldsymbol{V} = V^1_0\tilde{\boldsymbol{e}}_1 + V^2_0\tilde{\boldsymbol{e}}_2 = V^1\boldsymbol{e}_1 + V^2\boldsymbol{e}_2$, the parallel-transported basis vectors $\tilde{\boldsymbol{e}}_i$ transform reciprocally:

$$\begin{pmatrix} \tilde{\boldsymbol{e}}_1 \\ \tilde{\boldsymbol{e}}_2 \end{pmatrix} = \begin{pmatrix} \cos\delta\psi & \sin\delta\psi \\ -\sin\delta\psi & \cos\delta\psi \end{pmatrix}\begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{pmatrix}, \tag{13}$$

matching precisely the rotation matrix (6.3.95) with $\psi = \delta\psi$.

Under such a rotation of the polarization basis, the intensity I and circular polarization V remain invariant, whereas the linear polarization parameters (Q, U) rotate by 2ψ:

$$\begin{pmatrix} Q' \\ U' \end{pmatrix} = \begin{pmatrix} \cos 2\psi & \sin 2\psi \\ -\sin 2\psi & \cos 2\psi \end{pmatrix} \begin{pmatrix} Q \\ U \end{pmatrix}. \tag{6.3.96}$$

One can define (6.3.79) and (6.3.80) in the matrix notation as follows:

$$\begin{pmatrix} Q \\ U \end{pmatrix} = \begin{pmatrix} a_1 & -a_2 \\ a_2 & a_1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} a_1^2 - a_2^2 \\ 2a_1 a_2 \end{pmatrix}. \tag{1}$$

Using (6.3.95) and the fact that $\begin{pmatrix} a_1 & -a_2 \\ a_2 & a_1 \end{pmatrix}$ and $R(\theta)$ commutes, we can express.

$$\begin{pmatrix} Q' \\ U' \end{pmatrix} = \begin{pmatrix} a_1' & -a_2' \\ a_2' & a_1' \end{pmatrix} \begin{pmatrix} a_1' \\ a_2' \end{pmatrix} \tag{2}$$

$$= \begin{pmatrix} \cos\psi\, a_1 + \sin\psi\, a_2 & \sin\psi\, a_1 - \cos\psi\, a_2 \\ -\sin\psi\, a_1 + \cos\psi\, a_2 & \cos\psi\, a_1 + \sin\psi\, a_2 \end{pmatrix} \begin{pmatrix} \cos\psi\, a_1 + \sin\psi\, a_2 \\ -\sin\psi\, a_1 + \cos\psi\, a_2 \end{pmatrix} \tag{3}$$

$$= \begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix} \begin{pmatrix} a_1 & -a_2 \\ a_2 & a_1 \end{pmatrix} \begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \tag{4}$$

$$= \begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix}^2 \begin{pmatrix} Q \\ U \end{pmatrix} \tag{5}$$

$$= \begin{pmatrix} \cos 2\psi & \sin 2\psi \\ -\sin 2\psi & \cos 2\psi \end{pmatrix} \begin{pmatrix} Q \\ U \end{pmatrix}. \tag{6}$$

Equivalently, the transformation of the vector $(I_1, I_2, U, V)^{\mathrm{T}}$ takes the matrix form:

$$\begin{pmatrix} I_1' \\ I_2' \\ U' \\ V' \end{pmatrix} = \begin{pmatrix} \cos^2\psi & \sin^2\psi & \frac{1}{2}\sin 2\psi & 0 \\ \sin^2\psi & \cos^2\psi & -\frac{1}{2}\sin 2\psi & 0 \\ -\sin 2\psi & \sin 2\psi & \cos 2\psi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \\ U \\ V \end{pmatrix}. \tag{6.3.97}$$

From Eq. (6.3.96), the complex linear combinations $(Q \pm \mathrm{i}U)$ transform under basis rotation as spin-± 2 quantities:

$$(Q \pm \mathrm{i}U)'(\boldsymbol{e}_3) = \mathrm{e}^{\mp 2\mathrm{i}\psi}(Q \pm \mathrm{i}U)(\boldsymbol{e}_3). \tag{6.3.98}$$

Because the unperturbed universe is strictly unpolarized, Q and U are first-order gauge invariant perturbations that propagate without linear coupling to gravitational metric perturbations.

Polarization tensor

Let us define the complex spin combinations $P_\pm \equiv Q \pm \mathrm{i}U$. Under a passive rotation of the orthonormal polarization basis $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ by an angle ψ, the Stokes parameters transform according to Eq. (6.3.96), which gives:

$$P_+ \to \mathrm{e}^{-2\mathrm{i}\psi} P_+, \qquad P_- \to \mathrm{e}^{2\mathrm{i}\psi} P_-. \tag{1}$$

Comparing with the standard spin-weight transformation convention ${}_sf' = \mathrm{e}^{-\mathrm{i}s\psi}\,{}_sf$ [Eq. (6.3.103)], this identifies their respective spin weights as:

$$s(P_+) = +2, \qquad s(P_-) = -2. \tag{2}$$

We now introduce the orthonormal basis $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ on the 2-dimensional tangent screen, with contravariant components $(e_1)^a, (e_2)^a$ and covariant dual 1-forms $(e_1)_a = g_{ab}(e_1)^b, (e_2)_a = g_{ab}(e_2)^b$, satisfying:

$$g_{ab}(e_i)^a(e_j)^b = \delta_{ij}, \qquad g^{ab}(e_i)_a(e_j)_b = \delta_{ij} \qquad (i,j \in \{1,2\}). \tag{3}$$

The complex null vector basis on the screen is defined by:

$$m^a = \frac{1}{\sqrt{2}}\left[(e_1)^a + \mathrm{i}(e_2)^a\right], \qquad \bar{m}^a = \frac{1}{\sqrt{2}}\left[(e_1)^a - \mathrm{i}(e_2)^a\right], \tag{4}$$

with dual null 1-forms $m_a = g_{ab}m^b$ and $\bar{m}_a = g_{ab}\bar{m}^b$, satisfying the null and normalization conditions:

$$m^a m_a = 0, \qquad \bar{m}^a \bar{m}_a = 0, \qquad m^a \bar{m}_a = 1. \tag{5}$$

Under a passive basis rotation by angle ψ, the basis vectors rotate as:

$$\begin{pmatrix} (e_1)'^a \\ (e_2)'^a \end{pmatrix} = \begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix} \begin{pmatrix} (e_1)^a \\ (e_2)^a \end{pmatrix}, \tag{6}$$

which induces the transformations:

$$m^a \to \mathrm{e}^{-\mathrm{i}\psi} m^a, \qquad \bar{m}^a \to \mathrm{e}^{\mathrm{i}\psi}\bar{m}^a, \qquad m_a \to \mathrm{e}^{-\mathrm{i}\psi} m_a, \qquad \bar{m}_a \to \mathrm{e}^{\mathrm{i}\psi}\bar{m}_a. \tag{7}$$

We now construct the physical rank-2 linear polarization tensor field P_{ab} associated with $P_\pm$, demanding manifest invariance under local basis rotations:

$$P'_{ab} = P_{ab}. \tag{8}$$

To cancel the transformation phases, P_+ must pair with $\bar{m}_a\bar{m}_b$, while P_- pairs with $m_a m_b$:

$$P_{ab} \equiv \frac{1}{2}\left(P_+ \bar{m}_a\bar{m}_b + P_- m_a m_b\right). \tag{9}$$

Under a rotation by ψ:

$$P'_{ab} = \frac{1}{2}\left[(\mathrm{e}^{-2\mathrm{i}\psi}P_+)(\mathrm{e}^{2\mathrm{i}\psi}\bar{m}_a\bar{m}_b) + (\mathrm{e}^{2\mathrm{i}\psi}P_-)(\mathrm{e}^{-2\mathrm{i}\psi}m_a m_b)\right] = P_{ab}, \tag{10}$$

ensuring frame invariance. Expanding the null tensor products in the real orthonormal basis:

$$m_a m_b = \frac{1}{2}\left\{\left[(e_1)_a(e_1)_b - (e_2)_a(e_2)_b\right] + \mathrm{i}\left[(e_1)_a(e_2)_b + (e_2)_a(e_1)_b\right]\right\}, \tag{11}$$

$$\bar{m}_a\bar{m}_b = \frac{1}{2}\left\{\left[(e_1)_a(e_1)_b - (e_2)_a(e_2)_b\right] - \mathrm{i}\left[(e_1)_a(e_2)_b + (e_2)_a(e_1)_b\right]\right\}. \tag{12}$$

Substituting $P_+ = Q + \mathrm{i}U$ and $P_- = Q - \mathrm{i}U$:

$$\begin{aligned} P_{ab} &= \frac{1}{2}(Q+\mathrm{i}U)\bar{m}_a\bar{m}_b + \frac{1}{2}(Q-\mathrm{i}U)m_a m_b \\ &= \frac{1}{2}Q(\bar{m}_a\bar{m}_b + m_a m_b) + \frac{\mathrm{i}}{2}U(\bar{m}_a\bar{m}_b - m_a m_b) \\ &= \frac{1}{2}Q\left[(e_1)_a(e_1)_b - (e_2)_a(e_2)_b\right] + \frac{1}{2}U\left[(e_1)_a(e_2)_b + (e_2)_a(e_1)_b\right]. \end{aligned} \tag{13}$$

Projecting onto the tetrad basis $(e_i)^a(e_j)^b$, the matrix components $P_{ij} \equiv P_{ab}(e_i)^a(e_j)^b$ are:

$$P_{ij} = \frac{1}{2}\begin{pmatrix} Q & U \\ U & -Q \end{pmatrix}_{ij}. \tag{14}$$

The trace of P_{ab} vanishes identically:

$$g^{ab}P_{ab} = \frac{1}{2}\left(P_+ \bar{m}^a \bar{m}_a + P_- m^a m_a\right) = 0, \tag{15}$$

and $P_{ab} = P_{ba}$ is symmetric. Thus, P_{ab} is a genuine rank-2 symmetric trace-free tensor on the celestial 2-sphere, whose components in the local $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ polarization frame are precisely the Stokes parameters Q and U.

Their kinetic Boltzmann equation takes the form:

$$\mathcal{L}\begin{pmatrix} Q \\ U \end{pmatrix} = \begin{pmatrix} Q \\ U \end{pmatrix}' + n^i \partial_i \begin{pmatrix} Q \\ U \end{pmatrix} - \mathcal{H}E\frac{\partial}{\partial E}\begin{pmatrix} Q \\ U \end{pmatrix} = C\begin{pmatrix} Q \\ U \end{pmatrix}. \tag{6.3.99}$$

Covariant derivation of the polarization Boltzmann equation (6.3.99) from $P_{\mu\nu}$

Rather than postulating the Boltzmann equation (6.3.99) for the Stokes parameters (Q, U) heuristically, it can be derived from first principles using the projected symmetric trace-free (PSTF) transverse polarization tensor $P_{\mu\nu}$ on the celestial screen (Challinor 2000 [Ref. 27]). Writing $P_{\mu\nu}$ explicitly in terms of the orthonormal polarization basis $\{(\boldsymbol{e}_1)^\mu, (\boldsymbol{e}_2)^\mu\}$:

$$P_{\mu\nu} = P_{ij}(e_i)_\mu(e_j)_\nu, \qquad P_{ij} = \frac{1}{2}\begin{pmatrix} Q & U \\ U & -Q \end{pmatrix}_{ij} \qquad (i, j \in \{1, 2\}). \tag{1}$$

Applying the spacetime covariant derivative ∇_σ to this expansion, the Leibniz product rule yields

$$\nabla_\sigma P_{\mu\nu} = (\nabla_\sigma P_{ij})(e_i)_\mu(e_j)_\nu + P_{ij}\left[\nabla_\sigma(e_i)_\mu\right](e_j)_\nu + P_{ij}(e_i)_\mu\left[\nabla_\sigma(e_j)_\nu\right]. \tag{2}$$

Crucially, because the components P_{ij} are scalar fields under spacetime coordinate transformations (carrying internal tetrad labels i, j rather than spacetime indices), the covariant derivative acts on them purely as the partial derivative:

$$\nabla_\sigma P_{ij} = \partial_\sigma P_{ij}. \tag{3}$$

Meanwhile, the covariant derivative on each basis 1-form contains the Christoffel connection, $\nabla_\sigma(e_i)_\mu = \partial_\sigma(e_i)_\mu - \Gamma^\alpha_{\sigma\mu}(e_i)_\alpha$. Substituting this back into the expansion gives

$$\begin{aligned}
\nabla_\sigma P_{\mu\nu} &= (\partial_\sigma P_{ij})(e_i)_\mu(e_j)_\nu + P_{ij}\left[\partial_\sigma(e_i)_\mu - \Gamma^\alpha_{\sigma\mu}(e_i)_\alpha\right](e_j)_\nu \\
&\quad + P_{ij}(e_i)_\mu\left[\partial_\sigma(e_j)_\nu - \Gamma^\alpha_{\sigma\nu}(e_j)_\alpha\right] \\
&= \partial_\sigma\left[P_{ij}(e_i)_\mu(e_j)_\nu\right] - \Gamma^\alpha_{\sigma\mu}\left[P_{ij}(e_i)_\alpha(e_j)_\nu\right] - \Gamma^\alpha_{\sigma\nu}\left[P_{ij}(e_i)_\mu(e_j)_\alpha\right] \\
&= \partial_\sigma P_{\mu\nu} - \Gamma^\alpha_{\sigma\mu}P_{\alpha\nu} - \Gamma^\alpha_{\sigma\nu}P_{\mu\alpha},
\end{aligned} \tag{4}$$

confirming that the spacetime connection terms on $P_{\mu\nu}$ originate entirely from the basis vectors.

Now consider the directional derivative along a photon null geodesic with tangent vector $k^\sigma = \mathrm{d}x^\sigma/\mathrm{d}\lambda$. For any scalar function on phase space such as $P_{ij}(x^\alpha, k^\beta)$, the convective

derivative along the geodesic trajectory is

$$\frac{\mathrm{d}P_{ij}}{\mathrm{d}\lambda} = k^\sigma \partial_\sigma P_{ij} + \frac{\mathrm{d}k^\sigma}{\mathrm{d}\lambda}\frac{\partial P_{ij}}{\partial k^\sigma} = k^\sigma \partial_\sigma P_{ij} - \Gamma^\sigma_{\alpha\beta} k^\alpha k^\beta \frac{\partial P_{ij}}{\partial k^\sigma}. \tag{5}$$

The directional covariant derivative of the tensor $P_{\mu\nu}$ along the ray therefore reads

$$\begin{aligned}\frac{\mathrm{D}P_{\mu\nu}}{\mathrm{d}\lambda} \equiv k^\sigma \nabla_\sigma P_{\mu\nu} = \left(\frac{\mathrm{d}P_{ij}}{\mathrm{d}\lambda}\right)(e_i)_\mu (e_j)_\nu + P_{ij}\left(\frac{\mathrm{D}(e_i)_\mu}{\mathrm{d}\lambda}\right)(e_j)_\nu \\ + P_{ij}(e_i)_\mu \left(\frac{\mathrm{D}(e_j)_\nu}{\mathrm{d}\lambda}\right),\end{aligned} \tag{6}$$

where $\mathrm{D}(e_i)_\mu/\mathrm{d}\lambda \equiv k^\sigma \nabla_\sigma (e_i)_\mu$. Projecting this onto the polarization frame by contracting with $(e_k)^\mu (e_l)^\nu$ and dividing by $k^0 = \mathrm{d}\eta/\mathrm{d}\lambda$, we obtain

$$\frac{1}{k^0}(e_k)^\mu (e_l)^\nu \frac{\mathrm{D}P_{\mu\nu}}{\mathrm{d}\lambda} = \frac{\mathrm{d}P_{kl}}{\mathrm{d}\eta} + P_{il}\omega_{ki} + P_{kj}\omega_{lj}, \tag{7}$$

where $\omega_{ki} \equiv (e_k)^\mu \frac{1}{k^0}\frac{\mathrm{D}(e_i)_\mu}{\mathrm{d}\lambda}$ is the screen spin connection per unit conformal time. Because the polarization basis remains orthonormal along the ray ($g^{\mu\nu}(e_i)_\mu (e_k)_\nu = \delta_{ik}$), metric compatibility implies

$$k^\sigma \nabla_\sigma \left[g^{\mu\nu}(e_i)_\mu (e_k)_\nu\right] = 0 \implies \omega_{ki} = -\omega_{ik}. \tag{8}$$

In two dimensions, the antisymmetric matrix ω has $\omega_{11} = \omega_{22} = 0$, with a single independent rotation rate $\omega_{12} = -\omega_{21} \equiv \psi'$. Evaluating the projection for $(k,l) = (1,1)$ with $P_{11} = Q/2$:

$$P_{i1}\omega_{1i} + P_{1j}\omega_{1j} = 2P_{12}\omega_{12} = 2\left(\frac{1}{2}U\right)\psi' = U\psi', \tag{9}$$

which gives

$$\frac{\mathrm{d}Q}{\mathrm{d}\eta} = 2(e_1)^\mu (e_1)^\nu \mathcal{L}[P_{\mu\nu}] - 2\psi' U. \tag{10}$$

Similarly, for $(k,l) = (1,2)$ with $P_{12} = U/2$:

$$P_{i2}\omega_{1i} + P_{1j}\omega_{2j} = P_{22}\omega_{12} + P_{11}\omega_{21} = \left(-\frac{1}{2}Q\right)\psi' + \left(\frac{1}{2}Q\right)(-\psi') = -Q\psi', \tag{11}$$

which yields

$$\frac{\mathrm{d}U}{\mathrm{d}\eta} = 2(e_1)^\mu (e_2)^\nu \mathcal{L}[P_{\mu\nu}] + 2\psi' Q, \tag{12}$$

where $\mathcal{L}[P_{\mu\nu}] \equiv \frac{1}{k^0}\frac{\mathrm{D}P_{\mu\nu}}{\mathrm{d}\lambda}$. In terms of $P_\pm = Q \pm \mathrm{i}U$, this takes the compact form $\frac{\mathrm{d}P_\pm}{\mathrm{d}\eta} \mp 2\mathrm{i}\psi' P_\pm = 2(e_1 \pm \mathrm{i}e_2)^\mu (e_1)^\nu \mathcal{L}[P_{\mu\nu}]$, where $\mp 2\mathrm{i}\psi'$ is precisely the spin connection acting on fields of spin weight ± 2.

Under Sachs parallel transport of the polarization basis along the null geodesic ($H_{\mu\nu}k^\sigma \nabla_\sigma (e_i)^\nu = 0$), the screen projection of the basis derivative vanishes, so $\omega_{ki} = 0$ and $\psi' = 0$. The polarization frame does not rotate along the ray, decoupling the evolution:

$$\frac{\mathrm{d}Q}{\mathrm{d}\eta} = 2(e_1)^\mu (e_1)^\nu \mathcal{L}[P_{\mu\nu}], \qquad \frac{\mathrm{d}U}{\mathrm{d}\eta} = 2(e_1)^\mu (e_2)^\nu \mathcal{L}[P_{\mu\nu}]. \tag{13}$$

Equating $\mathcal{L}[P_{\mu\nu}]$ to the Thomson scattering tensor $K_{\mu\nu}$, and evaluating the scalar phase-space convective derivative $\mathrm{d}/\mathrm{d}\eta = \partial_\eta + n^k \partial_k - \mathcal{H}E\partial/\partial E$, directly yields the uncoupled Boltzmann equation (6.3.99):

$$\mathcal{L}\begin{pmatrix} Q \\ U \end{pmatrix} = \begin{pmatrix} Q \\ U \end{pmatrix}' + n^k \partial_k \begin{pmatrix} Q \\ U \end{pmatrix} - \mathcal{H}E\frac{\partial}{\partial E}\begin{pmatrix} Q \\ U \end{pmatrix} = C\begin{pmatrix} Q \\ U \end{pmatrix}. \tag{14}$$

If one instead adopts a fixed coordinate basis on the celestial sphere $(\boldsymbol{e}_\theta, \boldsymbol{e}_\phi)$, the non-vanishing spin connection $\psi' = -\cot\theta n^\phi$ combines with the directional gradient to form the spin-weighted covariant derivatives $\eth, \bar{\eth}$ and ladder operators $\partial^\pm$ of Subsection 6.3.3.

Normalizing the Stokes parameters by the background brightness $\bar{I}$ [Eq. (6.3.29)],

$$\begin{pmatrix} Q \\ U \end{pmatrix} \to \frac{1}{4\bar{I}} \int 4\pi E^3 \begin{pmatrix} Q \\ U \end{pmatrix} \mathrm{d}E\,, \tag{6.3.100}$$

we obtain the simplified Liouville equation for linear polarization:

$$\mathcal{L} \begin{pmatrix} Q \\ U \end{pmatrix} = \begin{pmatrix} Q \\ U \end{pmatrix}' + n^i \partial_i \begin{pmatrix} Q \\ U \end{pmatrix} = C \begin{pmatrix} Q \\ U \end{pmatrix}, \tag{6.3.101}$$

with normalized collision integral

$$C \begin{pmatrix} Q \\ U \end{pmatrix} \to \frac{1}{4\bar{I}} \int 4\pi E^3 C \begin{pmatrix} Q \\ U \end{pmatrix} \mathrm{d}E\,. \tag{6.3.102}$$

Spherical harmonics of spin s

For any celestial direction $\boldsymbol{e}_3$ specified by spherical coordinates (θ, ϕ), the tangent space is spanned by the orthonormal basis $(\boldsymbol{e}_\theta, \boldsymbol{e}_\phi)$, determined up to an internal rotation. A function ${}_sf(\theta, \phi)$ on the sphere is said to possess spin weight s if it transforms as ${}_sf'(\theta, \phi) = \mathrm{e}^{-\mathrm{i}s\psi}{}_sf(\theta, \phi)$ under a local basis rotation of angle ψ.

Angular Momentum Perspective: Local Triad Fields and Spin Weight

In standard quantum mechanics, a scalar field has no internal orientation; its angular behavior on the sphere is fully described by global rotations $\mathbf{J} = (J_x, J_y, J_z)$, yielding multipole states $|\ell, m\rangle$ with $J_z\,|\ell, m\rangle = m\,|\ell, m\rangle$.
For polarization, however, the field at direction $\boldsymbol{n}(\theta, \phi)$ lives in the 2D tangent plane (the Sachs screen basis $(\boldsymbol{e}_1, \boldsymbol{e}_2)$). If we were to project $\mathbf{J}$ along a fixed laboratory axis $\boldsymbol{n}_0$, the resulting operator $\boldsymbol{n}_0 \cdot \mathbf{J}$ would not commute with J_z: $[J_z, \boldsymbol{n}_0 \cdot \mathbf{J}] = \mathrm{i}(\boldsymbol{e}_z \times \boldsymbol{n}_0) \cdot \mathbf{J} \neq 0$.
Instead, we attach an orthonormal triad $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3 = \boldsymbol{n})$ to each point on the sphere, where the axes transform covariantly under spatial rotations: $[J_a, e_i^b] = \mathrm{i}\epsilon_{abc}e_i^c$. Projecting $\mathbf{J}$ onto this local triad defines the frame operators:

$$K_i \equiv \boldsymbol{e}_i \cdot \mathbf{J} = \sum_a e_i^a J_a \quad (i = 1, 2, 3). \tag{1}$$

Because both $\boldsymbol{e}_i$ and $\mathbf{J}$ rotate together under global rotations, their scalar product is a rotational invariant. The commutator vanishes identically:

$$[J_a, K_j] = [J_a, e_j^b]J_b + e_j^b[J_a, J_b] = 0 \quad \Longrightarrow \quad [J_z, K_3] = 0. \tag{2}$$

The normal projection $K_3 = \boldsymbol{n} \cdot \mathbf{J}$ does not move points across the sphere; rather, it generates internal rotations of the polarization axes $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ by the angle ψ in place about the local normal $\boldsymbol{n}$ at every point. States of definite spin weight s are eigenstates:

$$K_3\,|\ell, m, s\rangle = -s\,|\ell, m, s\rangle\,. \tag{3}$$

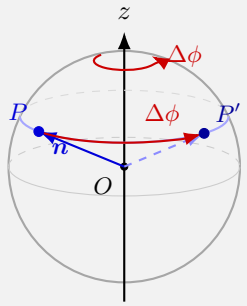
The frame projections satisfy the SU(2) algebra with an anomalous minus sign, $[K_i, K_j] = -\mathrm{i}\epsilon_{ijk}K_k$, and share the Casimir invariant $K^2 \equiv \sum_i K_i^2 = J^2 = \ell(\ell + 1)$.
Thus, (J^2, J_z, K_3) forms a complete set of commuting observables with joint eigenstates $|\ell, m, s\rangle$:

- m specifies the global celestial multipole orientation along J_z (transporting points $\phi \to \phi + \Delta\phi$ across the sphere),

- s specifies the local spin weight (helicity) along $K_3 = \boldsymbol{n} \cdot \mathbf{J}$ (rotating the local frame by ψ in place without moving the point).

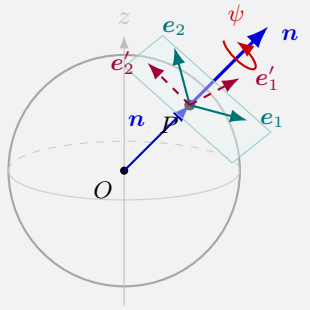
The coordinate-space overlap $\langle \theta, \phi; s | \ell, m, s \rangle = {}_sY_{\ell m}(\theta, \phi)$ is precisely the superposition amplitude—proportional to the Wigner rotation matrix element $D^{\ell *}_{-s,m}(\phi, \theta, 0)$—that resolves the global state into the local line-of-sight frame.

(a) Global rotation by J_z



Points move across sphere:
$(\theta, \phi) \to (\theta, \phi + \Delta\phi)$

(b) Local frame rotation by $K_3 = \boldsymbol{n} \cdot \mathbf{J}$



Point P is fixed; basis rotates:
$(\boldsymbol{e}_1, \boldsymbol{e}_2) \to (\boldsymbol{e}'_1, \boldsymbol{e}'_2)$ by angle ψ

Any spin-s field can be expanded in spin-weighted spherical harmonics ${}_sY_{\ell m}(\theta, \phi)$ [Ref. [29]], which satisfy the orthogonality and completeness relations:

$$\int_0^{2\pi} \mathrm{d}\phi \int_{-1}^{1} \mathrm{d}(\cos\theta)\; {}_sY^*_{\ell' m'}(\theta, \phi)\, {}_sY_{\ell m}(\theta, \phi) = \delta_{\ell\ell'}\delta_{mm'}, \tag{6.3.103}$$

$$\sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} {}_sY^*_{\ell m}(\theta, \phi)\, {}_sY_{\ell m}(\theta', \phi') = \delta(\phi - \phi')\delta(\cos\theta - \cos\theta'). \tag{6.3.104}$$

Direction Bras and Completeness of Spin-Weighted Harmonics

The connection between abstract state kets and spatial harmonics is made by defining localized direction bras $\langle \theta, \phi; s| \equiv \sqrt{2\pi}\, \langle \alpha = \phi, \beta = \theta, \gamma = 0 |\, P_s$ on the sphere at reference orientation $\gamma = 0$:

$$ {}_sY_{\ell m}(\theta, \phi) \equiv \langle \theta, \phi; s | \ell, m, s \rangle\,. \tag{1}$$

Within the subspace of spin weight s (projector P_s), the direction bras provide a resolution of the identity:

$$P_s = \int \mathrm{d}\Omega\, |\theta, \phi; s\rangle\, \langle \theta, \phi; s|\,, \qquad \mathrm{d}\Omega = \sin\theta\, \mathrm{d}\theta\, \mathrm{d}\phi. \tag{2}$$

Inserting this resolution of the identity between $\langle \ell, m, s|$ and $|\ell', m', s\rangle$ immediately yields orthogonality (6.3.103):

$$\delta_{\ell\ell'}\delta_{mm'} = \langle \ell, m, s | P_s | \ell', m', s \rangle = \int \mathrm{d}\Omega\, {}_sY^*_{\ell m}(\theta, \phi)\, {}_sY_{\ell' m'}(\theta, \phi). \tag{3}$$

Similarly, inserting the discrete completeness of states $\sum_{\ell \geq |s|, m} |\ell, m, s\rangle\, \langle \ell, m, s| = P_s$ between $\langle \theta, \phi; s|$ and $|\theta', \phi'; s\rangle$ yields the completeness relation (6.3.104):

$$\sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} {}_sY_{\ell m}(\theta, \phi)\, {}_sY^*_{\ell m}(\theta', \phi') = \langle \theta, \phi; s | \theta', \phi'; s \rangle = \delta(\phi - \phi')\delta(\cos\theta - \cos\theta'), \tag{4}$$

where terms with $\ell < |s|$ vanish identically.

The spin weight can be adjusted using the spin-raising (∂^+) and spin-lowering (∂^-) differential

operators:

$$\partial^{\pm}{}_s f(\theta,\phi) = -\sin^{\pm s}\theta\left(\partial_\theta \pm \frac{\mathrm{i}}{\sin\theta}\partial_\phi\right)\sin^{\mp s}\theta\,{}_s f(\theta,\phi). \tag{6.3.105}$$

For spin-2 functions with azimuthal dependence $\partial_\phi[{}_{\pm2}f(\mu,\phi)] = \mathrm{i}m\,{}_{\pm2}f(\mu,\phi)$ where $\mu=\cos\theta$, the second-order operators reduce to

$$(\partial^{\mp})^2\,{}_{\pm2}f(\mu,\phi) = \left(-\partial_\mu \pm \frac{m}{1-\mu^2}\right)^2\left[(1-\mu^2)\,{}_{\pm2}f(\mu,\phi)\right]. \tag{6.3.106}$$

Derivation of Differential Operators from Euler Angle Kinematics

The orientation of the local triad $(\boldsymbol{e}_1,\boldsymbol{e}_2,\boldsymbol{e}_3=\boldsymbol{n})$ relative to the fixed laboratory axes is specified by the state $|R\rangle = |\alpha,\beta,\gamma\rangle$, where $R(\alpha,\beta,\gamma) = R_z(\alpha)R_y(\beta)R_z(\gamma)$ is the standard z–y–z Euler decomposition. The global angular momentum $\mathbf{J}$ and the local frame operators $K_i = \boldsymbol{e}_i\cdot\mathbf{J}$ generate the left- and right-multiplications on SO(3):

$$J_i:\ R\longrightarrow gR \quad\text{(laboratory-frame rotations)},\qquad K_i:\ R\longrightarrow Rh \quad\text{(local frame rotations)}. \tag{1}$$

Under infinitesimal rotations $g=\mathrm{e}^{-\mathrm{i}\boldsymbol{\varepsilon}\cdot\mathbf{J}}$ and $h=\mathrm{e}^{-\mathrm{i}\boldsymbol{\varepsilon}\cdot\mathbf{K}}$, their unitary representations act on the orientation bras via the inverse group elements:

$$\langle R|\,U(g) = \left\langle g^{-1}R\right|,\qquad \langle R|\,V(h) = \left\langle Rh^{-1}\right|. \tag{2}$$

For an infinitesimal rotation about the local normal $\boldsymbol{e}_3$ by angle ε, $h=R_z(\varepsilon)$. Acting on the right,

$$Rh^{-1} = R_z(\alpha)R_y(\beta)R_z(\gamma)R_z(-\varepsilon) = R_z(\alpha)R_y(\beta)R_z(\gamma-\varepsilon), \tag{3}$$

which shifts the Euler angles by $(\alpha,\beta,\gamma)\to(\alpha,\beta,\gamma-\varepsilon)$. Expanding the bra to first order:

$$\langle R|\,V(h) = \langle\alpha,\beta,\gamma-\varepsilon| = \langle\alpha,\beta,\gamma| - \varepsilon\,\partial_\gamma\,\langle\alpha,\beta,\gamma| + \mathcal{O}(\varepsilon^2). \tag{4}$$

Comparing directly with the generator expansion $\langle R|\,V(h) = \langle R|\,(\mathbb{I}-\mathrm{i}\varepsilon K_3) = \langle R| - \mathrm{i}\varepsilon\,\langle R|\,K_3$ yields

$$K_3 = -\mathrm{i}\partial_\gamma. \tag{5}$$

Similarly, an infinitesimal space rotation on the left $g=R_z(\varepsilon)$ shifts $\alpha\to\alpha-\varepsilon$, giving $J_3=-\mathrm{i}\partial_\alpha$. The relationship between the triad angular velocity $(\omega_1,\omega_2,\omega_3)$ and the Euler rates is obtained algebraically from the pullback of the rotation matrix $R(t)=R_z(\alpha)R_y(\beta)R_z(\gamma)$ to the Lie algebra $\mathfrak{so}(3)$, $R^T\dot{R}=\sum_{a=1}^3\omega_a\mathcal{J}_a$:

$$R^T\dot{R} = \dot\alpha\,R_z(-\gamma)R_y(-\beta)\mathcal{J}_zR_y(\beta)R_z(\gamma) + \dot\beta\,R_z(-\gamma)\mathcal{J}_yR_z(\gamma) + \dot\gamma\,\mathcal{J}_z, \tag{6}$$

where $\mathcal{J}_a$ are the standard $\mathfrak{so}(3)$ generator matrices. Evaluating the similarity transformations under the adjoint action:

$$R_z(-\gamma)\mathcal{J}_yR_z(\gamma) = \sin\gamma\,\mathcal{J}_x + \cos\gamma\,\mathcal{J}_y, \tag{7}$$

$$\begin{aligned} R_z(-\gamma)R_y(-\beta)\mathcal{J}_zR_y(\beta)R_z(\gamma) &= R_z(-\gamma)\,(-\sin\beta\,\mathcal{J}_x + \cos\beta\,\mathcal{J}_z)\,R_z(\gamma)\\ &= -\sin\beta\cos\gamma\,\mathcal{J}_x + \sin\beta\sin\gamma\,\mathcal{J}_y + \cos\beta\,\mathcal{J}_z. \end{aligned} \tag{8}$$

Equating coefficients along $(\mathcal{J}_x,\mathcal{J}_y,\mathcal{J}_z)$ gives the kinematic relations:

$$\begin{pmatrix}\omega_1\\ \omega_2\end{pmatrix} = \begin{pmatrix}-\sin\beta\cos\gamma & \sin\gamma\\ \sin\beta\sin\gamma & \cos\gamma\end{pmatrix}\begin{pmatrix}\dot\alpha\\ \dot\beta\end{pmatrix},\qquad \omega_3 = \dot\alpha\cos\beta + \dot\gamma. \tag{9}$$

Inverting the 2×2 matrix gives the Euler-angle shifts:

$$\dot\alpha = -\frac{\cos\gamma}{\sin\beta}\,\omega_1 + \frac{\sin\gamma}{\sin\beta}\,\omega_2, \tag{10}$$

$$\dot\beta = \sin\gamma\,\omega_1 + \cos\gamma\,\omega_2, \tag{11}$$

$$\dot\gamma = \omega_3 - \dot\alpha\cos\beta = \cot\beta\cos\gamma\,\omega_1 - \cot\beta\sin\gamma\,\omega_2 + \omega_3. \tag{12}$$

Specializing to unit rotations about the frame 1-axis ($\omega_1 = 1, \omega_2 = 0, \omega_3 = 0$) and 2-axis ($\omega_1 = 0, \omega_2 = 1, \omega_3 = 0$), the generators $K_i = -\mathrm{i}(\dot{\alpha}\partial_\alpha + \dot{\beta}\partial_\beta + \dot{\gamma}\partial_\gamma)$ become:

$$K_1 = -\mathrm{i}\left(-\frac{\cos\gamma}{\sin\beta}\partial_\alpha + \sin\gamma\,\partial_\beta + \cot\beta\cos\gamma\,\partial_\gamma\right), \tag{13}$$

$$K_2 = -\mathrm{i}\left(\frac{\sin\gamma}{\sin\beta}\partial_\alpha + \cos\gamma\,\partial_\beta - \cot\beta\sin\gamma\,\partial_\gamma\right). \tag{14}$$

The ladder combinations $K_\pm \equiv K_1 \pm \mathrm{i}K_2$ and laboratory generators $J_\pm \equiv J_1 \pm \mathrm{i}J_2$ combine to:

$$K_\pm = \mathrm{e}^{\mp\mathrm{i}\gamma}\left(\pm\partial_\beta + \frac{\mathrm{i}}{\sin\beta}\partial_\alpha - \mathrm{i}\cot\beta\,\partial_\gamma\right), \tag{15}$$

$$J_\pm = \mathrm{e}^{\pm\mathrm{i}\alpha}\left(\pm\partial_\beta + \mathrm{i}\cot\beta\,\partial_\alpha - \frac{\mathrm{i}}{\sin\beta}\partial_\gamma\right). \tag{16}$$

Within the spin-s subspace (projector P_s), internal rotations about $\boldsymbol{e}_3$ shift the phase according to $K_3 P_s = -sP_s$:

$$\langle\alpha,\beta,\gamma|\,P_s = \langle\alpha,\beta,0|\,\mathrm{e}^{\mathrm{i}\gamma K_3}P_s = \mathrm{e}^{-\mathrm{i}s\gamma}\,\langle\alpha,\beta,0|\,P_s. \tag{17}$$

Differentiating with respect to γ yields $\partial_\gamma\langle\alpha,\beta,\gamma|\,P_s = -\mathrm{i}s\,\langle\alpha,\beta,\gamma|\,P_s$. Consequently, the connection term $-\mathrm{i}\cot\beta\,\partial_\gamma$ evaluates to $-s\cot\theta$ at the reference orientation $\gamma = 0$ with $(\beta,\alpha) = (\theta,\phi)$. Because $\eth \equiv -K_+$ raises the spin weight ($P_{s+1}\eth = \eth P_s$), evaluating its action on the direction bra $\langle\theta,\phi;s+1| \equiv \sqrt{2\pi}\,\langle\alpha=\phi,\beta=\theta,\gamma=0|\,P_{s+1}$ gives:

$$\begin{aligned}
\langle\theta,\phi;s+1|\,\eth &= -\sqrt{2\pi}\left[\mathrm{e}^{-\mathrm{i}\gamma}\left(\partial_\beta + \frac{\mathrm{i}}{\sin\beta}\partial_\alpha - \mathrm{i}\cot\beta\,\partial_\gamma\right)\langle\alpha,\beta,\gamma|\,P_s\right]_{\gamma=0} \\
&= -\left(\partial_\theta + \frac{\mathrm{i}}{\sin\theta}\partial_\phi - s\cot\theta\right)\langle\theta,\phi;s| \\
&= -(\sin\theta)^s\left(\partial_\theta + \frac{\mathrm{i}}{\sin\theta}\partial_\phi\right)(\sin\theta)^{-s}\,\langle\theta,\phi;s| \equiv \partial^+\,\langle\theta,\phi;s|\,.
\end{aligned} \tag{18}$$

Likewise, since $\bar{\eth} \equiv K_-$ lowers the spin weight ($P_{s-1}\bar{\eth} = \bar{\eth}P_s$), acting on $\langle\theta,\phi;s-1|$ gives:

$$\begin{aligned}
\langle\theta,\phi;s-1|\,\bar{\eth} &= \sqrt{2\pi}\left[\mathrm{e}^{\mathrm{i}\gamma}\left(-\partial_\beta + \frac{\mathrm{i}}{\sin\beta}\partial_\alpha - \mathrm{i}\cot\beta\,\partial_\gamma\right)\langle\alpha,\beta,\gamma|\,P_s\right]_{\gamma=0} \\
&= -\left(\partial_\theta - \frac{\mathrm{i}}{\sin\theta}\partial_\phi + s\cot\theta\right)\langle\theta,\phi;s| \\
&= -(\sin\theta)^{-s}\left(\partial_\theta - \frac{\mathrm{i}}{\sin\theta}\partial_\phi\right)(\sin\theta)^s\,\langle\theta,\phi;s| \equiv \partial^-\,\langle\theta,\phi;s|\,.
\end{aligned} \tag{19}$$

Projecting against $|\ell,m,s\rangle$ directly yields the differential ladder actions:

$$\partial^+\,{}_sY_{\ell m}(\theta,\phi) = \sqrt{(\ell-s)(\ell+s+1)}\,{}_{s+1}Y_{\ell m}(\theta,\phi), \tag{20}$$

$$\partial^-\,{}_sY_{\ell m}(\theta,\phi) = -\sqrt{(\ell+s)(\ell-s+1)}\,{}_{s-1}Y_{\ell m}(\theta,\phi). \tag{21}$$

Restricted to $\gamma = 0$, the laboratory generators act on the spin-weighted angular dependence as:

$$J_\pm = \mathrm{e}^{\pm\mathrm{i}\phi}\left(\pm\partial_\theta + \mathrm{i}\cot\theta\,\partial_\phi - \frac{s}{\sin\theta}\right), \tag{22}$$

where the geometric term $-s/\sin\theta$ represents the intrinsic spin contribution to the total angular momentum.

Acting with these ladder operators on ordinary spherical harmonics ($s = 0$) yields the

spin-weighted spherical harmonics:

$$ {}_sY_{\ell m} \equiv \sqrt{\frac{(\ell-s)!}{(\ell+s)!}}(\partial^+)^s Y_{\ell m} \quad (0 \le s \le \ell), \qquad {}_sY_{\ell m} \equiv \sqrt{\frac{(\ell+s)!}{(\ell-s)!}}(-1)^s(\partial^-)^{-s} Y_{\ell m} \quad (-\ell \le s \le 0). \tag{6.3.107} $$

Derivation of Ladder Operators and Normalization in Bra–Ket Form

The differential operators $\partial^\pm$ and the factorial prefactor in Eq. (6.3.107) originate from the ladder algebra of the frame angular momentum. Defining $K_\pm \equiv K_1 \pm \mathrm{i}K_2$, the commutator $[K_3, K_\pm] = \mp K_\pm$ yields:

$$ K_3\big(K_\pm |\ell,m,s\rangle\big) = (K_\pm K_3 \mp K_\pm)|\ell,m,s\rangle = (-s \mp 1)\big(K_\pm |\ell,m,s\rangle\big). \tag{1} $$

Because the eigenvalue of K_3 is $-s$, lowering the eigenvalue of K_3 by 1 *increases* the spin weight s by 1. Defining the Newman–Penrose operators $\eth \equiv -K_+$ and $\bar\eth \equiv K_-$ (with $\eth^\dagger = -\bar\eth$):

$$ \bar\eth\eth\,|\ell,m,s\rangle = -K_-K_+\,|\ell,m,s\rangle = -[K^2 - K_3^2 + K_3]\,|\ell,m,s\rangle = -(\ell-s)(\ell+s+1)\,|\ell,m,s\rangle\,, \tag{2} $$

$$ \eth\bar\eth\,|\ell,m,s\rangle = -K_+K_-\,|\ell,m,s\rangle = -[K^2 - K_3^2 - K_3]\,|\ell,m,s\rangle = -(\ell+s)(\ell-s+1)\,|\ell,m,s\rangle\,. \tag{3} $$

Taking the positive square root in the Condon–Shortley convention gives the exact ladder step:

$$ \eth\,|\ell,m,s\rangle = \sqrt{(\ell-s)(\ell+s+1)}\;|\ell,m,s+1\rangle\,. \tag{4} $$

Applying $\eth$ repeatedly s times from the scalar state $|\ell,m,0\rangle$, the product telescopes:

$$ \eth^s\,|\ell,m,0\rangle = \prod_{k=0}^{s-1}\sqrt{(\ell-k)(\ell+k+1)}\;|\ell,m,s\rangle = \sqrt{\frac{(\ell+s)!}{(\ell-s)!}}\;|\ell,m,s\rangle \quad (s \ge 0). \tag{5} $$

Inverting this relation gives $|\ell,m,s\rangle = \sqrt{\frac{(\ell-s)!}{(\ell+s)!}}\,\eth^s\,|\ell,m,0\rangle$. Projecting onto the direction bra $\langle\theta,\phi;s|$ where $\eth$ evaluates to the differential operator ∂^+, this derives Eq. (6.3.107) purely algebraically:

$$ {}_sY_{\ell m}(\theta,\phi) = \langle\theta,\phi;s|\ell,m,s\rangle = \sqrt{\frac{(\ell-s)!}{(\ell+s)!}}\,(\partial^+)^s Y_{\ell m}(\theta,\phi). \tag{6} $$

These harmonics satisfy the conjugation and differential relations:

$$ {}_sY^*_{\ell m} = (-1)^s\,{}_{-s}Y_{\ell,-m}, \tag{6.3.108} $$

$$ \partial^\pm\,{}_sY_{\ell m} = \pm\sqrt{(\ell\mp s)(\ell+1\pm s)}\,{}_{s\pm1}Y_{\ell m}, \tag{6.3.109} $$

$$ \partial^-\partial^+\,{}_sY_{\ell m} = -(\ell-s)(\ell+1+s)\,{}_sY_{\ell m}. \tag{6.3.110} $$

Algebraic Origin of Differential Eigenvalues, Conjugation, and Wigner $3j$ Symbols

The differential eigenvalue relation (6.3.110) is the coordinate projection of the operator product $\bar\eth\eth$:

$$ \partial^-\partial^+\,{}_sY_{\ell m}(\theta,\phi) = \langle\theta,\phi;s|\bar\eth\eth|\ell,m,s\rangle = -(\ell-s)(\ell+1+s)\,{}_sY_{\ell m}(\theta,\phi). \tag{1} $$

The commutator of the ladder operators is purely proportional to the spin weight:

$$ [\partial^-,\partial^+] = [\bar\eth,\eth] = -[K_-,K_+] = -2K_3 = 2s. \tag{2} $$

Under antiunitary time-reversal conjugation Θ and parity Π, both angular momenta invert ($\Theta\mathbf{J}\Theta^{-1} = -\mathbf{J}$, $\Theta\mathbf{K}\Theta^{-1} = -\mathbf{K}$), which maps $\Theta\,|\ell,m,s\rangle = (-1)^{m+s}\,|\ell,-m,-s\rangle$. Projection onto $\langle\theta,\phi;s|$ directly yields the conjugation relation (6.3.108):

$$ {}_sY^*_{\ell m}(\theta,\phi) = (-1)^{m+s}\,{}_{-s}Y_{\ell,-m}(\theta,\phi) = (-1)^s\,{}_{-s}Y_{\ell,-m}(\theta,\phi), \tag{3} $$

while parity inversion yields ${}_sY_{\ell m}(-\boldsymbol{n}) = (-1)^\ell {}_{-s}Y_{\ell m}(\boldsymbol{n})$, showing that parity inverts the spin weight.

When coupling states $|\ell_1, m_1, s_1\rangle$ and $|\ell_2, m_2, s_2\rangle$ or reducing products of spin-weighted harmonics, angular momentum addition produces Clebsch–Gordan coefficients $\langle \ell_1 m_1 \ell_2 m_2 | LM\rangle$:

$$ {}_{s_1}Y_{\ell_1 m_1}\, {}_{s_2}Y_{\ell_2 m_2} = \sum_L \sqrt{\frac{(2\ell_1+1)(2\ell_2+1)}{4\pi(2L+1)}}\, \langle \ell_1 m_1 \ell_2 m_2 | LM\rangle\, \langle \ell_1, -s_1, \ell_2, -s_2 | L, -S\rangle\ {}_SY_{LM}, \quad (4) $$

with $M = m_1 + m_2$ and $S = s_1 + s_2$. However, the Clebsch–Gordan coefficient carries an asymmetric square-root prefactor $\sqrt{2L+1}$ in the denominator and phase factors that break the manifest permutation symmetry between the participating angular momenta. It is therefore far more natural to define a new, fully symmetric notation—the Wigner $3j$ symbol:

$$ \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \equiv \frac{(-1)^{\ell_1 - \ell_2 - m_3}}{\sqrt{2\ell_3+1}}\, \langle \ell_1 m_1 \ell_2 m_2 | \ell_3, -m_3\rangle\,. \quad (5) $$

This definition absorbs the dimensional prefactors $\sqrt{2\ell_3+1}$ and phase factors directly into the coupling. Consequently, the integral of three spin-weighted harmonics takes the completely symmetric form:

$$ \int \mathrm{d}\Omega\, {}_{s_1}Y_{\ell_1 m_1}\, {}_{s_2}Y_{\ell_2 m_2}\, {}_{s_3}Y_{\ell_3 m_3} = \sqrt{\frac{(2\ell_1+1)(2\ell_2+1)(2\ell_3+1)}{4\pi}} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ -s_1 & -s_2 & -s_3 \end{pmatrix}, \quad (6) $$

where conservation of J_3 enforces $m_1 + m_2 + m_3 = 0$, while conservation of the frame angular momentum K_3 enforces the spin-weight sum rule $s_1 + s_2 + s_3 = 0$.

E and B modes

In complete analogy with the scalar temperature field expansion,

$$ \Theta(\boldsymbol{n}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\boldsymbol{n}), \quad (6.3.111) $$

the spin-±2 linear polarization fields expand in spin-weighted harmonics as

$$ (Q \pm \mathrm{i}U)(\boldsymbol{n}) = \sum_{\ell=2}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m}^{(\pm2)}\, {}_{\pm2}Y_{\ell m}(\boldsymbol{n}), \quad (6.3.112) $$

with $a_{\ell m}^{(-2)*} = a_{\ell,-m}^{(+2)}$. Applying the differential ladder relations yields

$$ (\partial^{\pm})^2 (Q \pm \mathrm{i}U)(\boldsymbol{n}) = \sqrt{\frac{(\ell+2)!}{(\ell-2)!}} \sum_{\ell=2}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m}^{(\pm2)} Y_{\ell m}(\boldsymbol{n}). \quad (6.3.113) $$

The expansion multipole moments are obtained by projection:

$$ a_{\ell m} = \int \mathrm{d}^2\boldsymbol{n}\, Y_{\ell m}^*(\boldsymbol{n}) \Theta(\boldsymbol{n}), \quad (6.3.114) $$

and

$$ a_{\ell m}^{(\pm2)} = \int \mathrm{d}^2\boldsymbol{n}\, {}_{\pm2}Y_{\ell m}^*(\boldsymbol{n}) (Q \pm \mathrm{i}U)(\boldsymbol{n}) = \sqrt{\frac{(\ell-2)!}{(\ell+2)!}} \int \mathrm{d}^2\boldsymbol{n}\, Y_{\ell m}^*(\boldsymbol{n}) (\partial^{\mp})^2 (Q \pm \mathrm{i}U)(\boldsymbol{n}). \quad (6.3.115) $$

Because the Stokes parameters Q and U depend on the coordinate orientation of the polarization basis, it is advantageous to decompose the spin-2 field into coordinate-independent scalar E

(electric) and pseudo-scalar B (magnetic) modes:

$$(Q \pm \mathrm{i}U)(\eta, \boldsymbol{x}, \boldsymbol{n}) = -\sum_{\ell=2}^{\infty} \sum_{m=-\ell}^{\ell} (E_{\ell m} \pm \mathrm{i}B_{\ell m})_{\pm 2}Y_{\ell m}(\boldsymbol{n}), \tag{6.3.116}$$

where the multipole coefficients are defined by

$$E_{\ell m} \equiv -\frac{1}{2}\left(a_{\ell m}^{(+2)} + a_{\ell m}^{(-2)}\right), \qquad B_{\ell m} \equiv -\frac{1}{2\mathrm{i}}\left(a_{\ell m}^{(+2)} - a_{\ell m}^{(-2)}\right). \tag{6.3.117}$$

The designations E and B reflect their definite parity behavior under spatial reflections: under a reflection across the plane containing the propagation vector $\boldsymbol{k}$, $Q \to Q$ and $U \to -U$.

Parity transformation of the polarization basis

Under spatial parity $\boldsymbol{x} \to -\boldsymbol{x}$, the Cartesian coordinates transform as

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{1}$$

The orthonormal spherical basis $(\boldsymbol{e}_\theta, \boldsymbol{e}_\phi, \hat{\boldsymbol{n}})$ relates to the Cartesian basis via

$$\begin{pmatrix} \boldsymbol{e}_\theta \\ \boldsymbol{e}_\phi \\ \hat{\boldsymbol{n}} \end{pmatrix} = R(\theta, \phi) \begin{pmatrix} \hat{\boldsymbol{x}} \\ \hat{\boldsymbol{y}} \\ \hat{\boldsymbol{z}} \end{pmatrix}, \qquad R(\theta, \phi) = \begin{pmatrix} \cos\theta\cos\phi & \cos\theta\sin\phi & -\sin\theta \\ -\sin\phi & \cos\phi & 0 \\ \sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta \end{pmatrix}. \tag{2}$$

Since $\hat{\boldsymbol{n}} \to -\hat{\boldsymbol{n}}$ corresponds to $(\theta, \phi) \to (\pi - \theta, \phi + \pi)$, evaluating the rotation matrix at the reflected angles gives

$$\begin{aligned} R(\pi - \theta, \phi + \pi) &= \begin{pmatrix} (-\cos\theta)(-\cos\phi) & (-\cos\theta)(-\sin\phi) & -\sin\theta \\ -(-\sin\phi) & -\cos\phi & 0 \\ (\sin\theta)(-\cos\phi) & (\sin\theta)(-\sin\phi) & -\cos\theta \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} R(\theta, \phi). \end{aligned} \tag{3}$$

The transformed spherical basis is therefore

$$\begin{pmatrix} \boldsymbol{e}'_\theta \\ \boldsymbol{e}'_\phi \\ \hat{\boldsymbol{n}}' \end{pmatrix} = R(\pi - \theta, \phi + \pi) R^{\mathrm{T}}(\theta, \phi) \begin{pmatrix} \boldsymbol{e}_\theta \\ \boldsymbol{e}_\phi \\ \hat{\boldsymbol{n}} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} \boldsymbol{e}_\theta \\ \boldsymbol{e}_\phi \\ \hat{\boldsymbol{n}} \end{pmatrix}. \tag{4}$$

Restricting to the transverse polarization basis $(\boldsymbol{e}_1, \boldsymbol{e}_2) \equiv (\boldsymbol{e}_\theta, \boldsymbol{e}_\phi)$:

$$\begin{pmatrix} \boldsymbol{e}'_1 \\ \boldsymbol{e}'_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{pmatrix} \implies \begin{pmatrix} a'_1 \\ a'_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}. \tag{5}$$

Therefore $a'_1 = a_1$ and $a'_2 = -a_2$. Substituting into the Stokes parameter definitions [Eqs. (6.3.79)–(6.3.80)]:

$$Q' = \langle a_1^2 \rangle - \langle (-a_2)^2 \rangle = \langle a_1^2 \rangle - \langle a_2^2 \rangle = Q, \tag{6}$$

$$U' = 2\langle a_1(-a_2)\cos(\theta_1 - \theta_2) \rangle = -2\langle a_1 a_2 \cos(\theta_1 - \theta_2) \rangle = -U. \tag{7}$$

The spin-weighted spherical harmonics transform as ${}_sY_{\ell m} \to (-1)^\ell {}_{-s}Y_{\ell m}$, implying that

$$E_{\ell m} \to (-1)^\ell E_{\ell m}, \qquad B_{\ell m} \to (-1)^{\ell+1} B_{\ell m}.$$

Thus, the E-mode behaves as a pure scalar (even parity under spatial inversion), while the B-mode behaves as a pseudo-scalar (odd parity). Linear scalar perturbations possess reflection symmetry and consequently generate pure E-modes with $B_{\ell m} \equiv 0$; primordial tensor gravitational waves break reflection symmetry and generate both E-modes and B-modes.

Relation with the Stokes parameters

The geometric meaning of E and B modes becomes particularly transparent in the flat-sky approximation. Over small angular patches, spherical coordinates (ϑ, ϕ) map onto Cartesian angular coordinates $\boldsymbol{\theta} = (\theta_x, \theta_y)$, with conjugate Fourier wavevector $\boldsymbol{\ell} = (\ell_x, \ell_y) = \ell(\cos\phi_\ell, \sin\phi_\ell)$. Here ϑ and ϕ parametrize the line-of-sight direction $\boldsymbol{n}$ itself: ϑ is the polar angle measured from the line of sight, and ϕ is the azimuthal angle measured around it, on the transverse circle (Fig. 6.7).

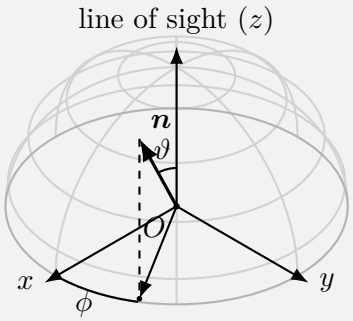


Figure 6.7: The polar angle ϑ is measured from the line of sight z, and the azimuthal angle ϕ is measured on the transverse circle from the x-axis. Dropping $\boldsymbol{n}$ onto the transverse plane gives local polar coordinates (ϑ, ϕ) on the sky patch.

Consider a point on the unit sphere close to the line-of-sight direction $\boldsymbol{n}_0 = (0, 0, 1)$, parametrized by polar and azimuthal angles (ϑ, ϕ) as

$$\begin{aligned} \boldsymbol{n} &= (\sin\vartheta\cos\phi,\ \sin\vartheta\sin\phi,\ \cos\vartheta) \\ &\simeq \left(\vartheta\cos\phi,\ \vartheta\sin\phi,\ 1 - \tfrac{1}{2}\vartheta^2\right), \end{aligned} \tag{1}$$

for $\vartheta \ll 1$. The tangent plane at $\boldsymbol{n}_0$ is the xy-plane, so the flat-sky coordinates on this local patch are read off from the transverse components:

$$\boldsymbol{\theta} = (\theta_x, \theta_y) = \vartheta(\cos\phi, \sin\phi), \quad |\boldsymbol{\theta}| = \vartheta. \tag{2}$$

It is convenient to also introduce the chord distance $r \equiv |\boldsymbol{n} - \boldsymbol{n}_0|$ between the two points, measured through the sphere rather than along it. Since $\boldsymbol{n}^2 = \boldsymbol{n}_0^2 = 1$ and $\boldsymbol{n} \cdot \boldsymbol{n}_0 = \cos\vartheta$,

$$r^2 = (\boldsymbol{n} - \boldsymbol{n}_0)^2 = 2 - 2\cos\vartheta = 4\sin^2\frac{\vartheta}{2},$$

so that exactly

$$\boxed{r = 2\sin\frac{\vartheta}{2}}, \qquad \text{and for } \vartheta \ll 1, \quad r = \vartheta + O(\vartheta^3).$$

At the order we work throughout – leading order in ϑ, with corrections controlled by the flat-sky expansion parameter $1/\ell$ introduced below – the three lengths r, ϑ, and $|\boldsymbol{\theta}|$ all coincide, and we use r for whichever of them is most convenient without further comment. In particular the tangent-plane vector is

$$\boldsymbol{\theta} \simeq r(\cos\phi, \sin\phi).$$

The two-dimensional wavevector conjugate to $\boldsymbol{\theta}$, already introduced above, is $\boldsymbol{\ell} = \ell(\cos\phi_\ell, \sin\phi_\ell)$, with magnitude ℓ playing the role of a multipole. Its scalar product

with the position vector is

$$\boldsymbol{\ell}\cdot\boldsymbol{\theta} = \ell r\big(\cos\phi_\ell\cos\phi + \sin\phi_\ell\sin\phi\big) = \ell r\cos(\phi-\phi_\ell),$$

so a flat-sky Fourier mode takes the form

$$e^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{\theta}} = e^{\mathrm{i}\ell r\cos(\phi-\phi_\ell)}.$$

This function of the single angle $\phi-\phi_\ell$ has the Fourier expansion

$$e^{\mathrm{i}z\cos\alpha} = \sum_{m=-\infty}^{\infty} \mathrm{i}^m J_m(z) e^{\mathrm{i}m\alpha},$$

which is one of the standard integral definitions of the Bessel functions J_m: it is precisely the statement that $\mathrm{i}^m J_m(z)$ is the m-th Fourier coefficient of $e^{\mathrm{i}z\cos\alpha}$ in α, and we take it as known rather than re-derive it here (the Jacobi–Anger identity). Setting $z=\ell r$, $\alpha=\phi-\phi_\ell$,

$$e^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{\theta}} = \sum_{m=-\infty}^{\infty} \mathrm{i}^m e^{-\mathrm{i}m\phi_\ell} J_m(\ell r) e^{\mathrm{i}m\phi}. \tag{3}$$

This is an exact rewriting of the plane wave, valid for any ℓ and r; no small-angle approximation has entered yet beyond the substitution $\boldsymbol{\ell}\cdot\boldsymbol{\theta}=\ell r\cos(\phi-\phi_\ell)$ already used to define the flat patch.

The objective of the flat-sky correspondence is to relate this Fourier plane-wave expansion to the spherical harmonic expansion on the full sky, both for scalar fields ($s=0$) and for spin-weighted fields ${}_sY_{\ell m}(\vartheta,\phi)$ such as linear polarization ($s=\pm2$).

For ordinary scalar spherical harmonics ($s=0$), write

$$Y_{\ell m}(\vartheta,\phi) = N_{\ell m}\, P_\ell^m(\cos\vartheta)\, e^{\mathrm{i}m\phi}, \qquad N_{\ell m} = \sqrt{\frac{2\ell+1}{4\pi}\,\frac{(\ell-m)!}{(\ell+m)!}},$$

with P_ℓ^m the associated Legendre function. In the joint high-multipole, small-angle limit

$$\ell\to\infty, \qquad \vartheta\to 0, \qquad x\equiv\ell\vartheta\simeq\ell r \text{ fixed},$$

with m held fixed, the associated Legendre function reduces to a cylindrical Bessel function via the Mehler–Heine theorem. Combined with $N_{\ell m}\to\sqrt{\ell/2\pi}$ at leading order in $1/\ell$, this yields the classical asymptotic relation

$$Y_{\ell m}(\vartheta,\phi) \;\simeq\; (-1)^m\sqrt{\frac{\ell}{2\pi}}\, J_m(\ell r)\, e^{\mathrm{i}m\phi}. \tag{4}$$

To extend this correspondence to spin-weighted spherical harmonics ${}_sY_{\ell m}$, recall that for $s>0$ they are generated from ordinary harmonics by the spin-raising operator $\eth$ [Eq. (6.3.107)]:

$${}_sY_{\ell m} = \sqrt{\frac{(\ell-s)!}{(\ell+s)!}}\,\eth^s Y_{\ell m}, \tag{5}$$

where in standard notation $\eth$ acting on a field η of spin weight s is

$$\begin{aligned}\eth\eta &= -(\sin\vartheta)^s\left(\partial_\vartheta + \frac{\mathrm{i}}{\sin\vartheta}\partial_\phi\right)\left[(\sin\vartheta)^{-s}\eta\right]\\ &= -\left(\partial_\vartheta + \frac{\mathrm{i}}{\sin\vartheta}\partial_\phi - s\cot\vartheta\right)\eta.\end{aligned} \tag{6}$$

In the small-angle limit $\vartheta \ll 1$, $\sin\vartheta \simeq \vartheta$ and $\cot\vartheta \simeq 1/\vartheta$. Starting at $s = 0$, the first action on $Y_{\ell m}$ gives

$$\begin{aligned} \eth Y_{\ell m} &= -\left(\partial_\vartheta + \frac{\mathrm{i}}{\sin\vartheta}\partial_\phi\right) Y_{\ell m} \\ &\simeq -(-1)^m \sqrt{\frac{\ell}{2\pi}} \left[\ell J'_m(\ell r) - \frac{m}{r} J_m(\ell r)\right] e^{\mathrm{i}m\phi}. \end{aligned} \tag{7}$$

Using the standard Bessel recurrence relation $J'_m(x) - \frac{m}{x}J_m(x) = -J_{m+1}(x)$, the bracket simplifies to $-\ell J_{m+1}(\ell r)$, exactly cancelling the leading minus sign:

$$\eth Y_{\ell m} \simeq (-1)^m \sqrt{\frac{\ell}{2\pi}}\, \ell J_{m+1}(\ell r)\, e^{\mathrm{i}m\phi}. \tag{8}$$

Because ${}_1Y_{\ell m} = [\ell(\ell+1)]^{-1/2}\eth Y_{\ell m} \simeq \ell^{-1}\eth Y_{\ell m}$ at leading order in $1/\ell$, the explicit factor of ℓ cancels, yielding

$${}_1Y_{\ell m}(\vartheta,\phi) \simeq (-1)^m \sqrt{\frac{\ell}{2\pi}}\, J_{m+1}(\ell r)\, e^{\mathrm{i}m\phi}. \tag{9}$$

Applying $\eth$ repeatedly s times shifts the Bessel index by $+s$ at each step, while the prefactor $\sqrt{(\ell-s)!/(\ell+s)!} \simeq \ell^{-s}$ cancels the factor ℓ^s generated by the radial derivatives:

$$\boxed{{}_sY_{\ell m}(\vartheta,\phi) \simeq (-1)^m \sqrt{\frac{\ell}{2\pi}}\, J_{m+s}(\ell r)\, e^{\mathrm{i}m\phi}} \qquad (s \geq 0). \tag{10}$$

For negative spin weight, the spin-lowering operator $\bar{\eth} \simeq -(\partial_\vartheta - \frac{\mathrm{i}}{\vartheta}\partial_\phi + \frac{s}{\vartheta})$ and the companion Bessel recurrence $J'_m(x) + \frac{m}{x}J_m(x) = J_{m-1}(x)$ introduce an additional phase factor $(-1)^s$:

$${}_sY_{\ell m}(\vartheta,\phi) \simeq (-1)^{m+s} \sqrt{\frac{\ell}{2\pi}}\, J_{m+s}(\ell r)\, e^{\mathrm{i}m\phi} \qquad (s < 0). \tag{11}$$

Crucially, for linear polarization ($s = \pm 2$), $(-1)^{\pm 2} = 1$, so both helicity states share the unified form:

$${}_{\pm 2}Y_{\ell m}(\vartheta,\phi) \simeq (-1)^m \sqrt{\frac{\ell}{2\pi}}\, J_{m\pm 2}(\ell r)\, e^{\mathrm{i}m\phi}. \tag{12}$$

Inverting this asymptotic relation provides the harmonic representation of the azimuthal Bessel modes:

$$J_{m+s}(\ell r)\, e^{\mathrm{i}m\phi} \simeq (-1)^m \sqrt{\frac{2\pi}{\ell}}\, {}_sY_{\ell m}(\vartheta,\phi) \qquad (s \geq 0 \text{ or } s = \pm 2). \tag{13}$$

Now substitute this into the exact Jacobi–Anger expansion of the plane wave $e^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{\theta}}$. Re-indexing the sum by setting $k = m + s$ (so $m = k - s$),

$$\begin{aligned} e^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{\theta}} &= \sum_{k=-\infty}^{\infty} \mathrm{i}^k e^{-\mathrm{i}k\phi_\ell} J_k(\ell r) e^{\mathrm{i}k\phi} \\ &= \sum_{m=-\infty}^{\infty} \mathrm{i}^{m+s} e^{-\mathrm{i}(m+s)\phi_\ell} J_{m+s}(\ell r) e^{\mathrm{i}(m+s)\phi} \\ &= e^{\mathrm{i}s(\phi-\phi_\ell)} \sum_{m=-\infty}^{\infty} \mathrm{i}^{m+s} e^{-\mathrm{i}m\phi_\ell} \left[J_{m+s}(\ell r) e^{\mathrm{i}m\phi}\right]. \end{aligned} \tag{14}$$

Inserting $J_{m+s}(\ell r)e^{\mathrm{i}m\phi} \simeq (-1)^m\sqrt{2\pi/\ell}\,{}_sY_{\ell m}$ and using $(-1)^m\mathrm{i}^{m+s} = (-\mathrm{i})^m\mathrm{i}^s$:

$$e^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{\theta}} \simeq \sqrt{\frac{2\pi}{\ell}}\,\mathrm{i}^s e^{\mathrm{i}s(\phi-\phi_\ell)} \sum_{m=-\ell}^{\ell} (-\mathrm{i})^m e^{-\mathrm{i}m\phi_\ell}\,{}_sY_{\ell m}(\vartheta,\phi). \tag{15}$$

For scalar perturbations ($s = 0$), this reduces directly to the standard flat-sky superposition:

$$e^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{\theta}} \simeq \sqrt{\frac{2\pi}{\ell}} \sum_{m=-\ell}^{\ell} (-\mathrm{i})^m e^{-\mathrm{i}m\phi_\ell} Y_{\ell m}(\vartheta,\phi). \tag{16}$$

For polarization ($s = \pm 2$), $\mathrm{i}^{\pm 2} = -1$, giving

$$e^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{\theta}} \simeq -\sqrt{\frac{2\pi}{\ell}}\,e^{\pm 2\mathrm{i}(\phi-\phi_\ell)} \sum_{m=-\ell}^{\ell} (-\mathrm{i})^m e^{-\mathrm{i}m\phi_\ell}\,{}_{\pm2}Y_{\ell m}(\vartheta,\phi). \tag{17}$$

Multiplying by $e^{\mp 2\mathrm{i}\phi}$ to transform from the celestial polarization basis $(\boldsymbol{e}_\theta, \boldsymbol{e}_\phi)$ to the spatially constant Cartesian frame $(\hat{\boldsymbol{e}}_x, \hat{\boldsymbol{e}}_y)$ across the tangent patch, the position-dependent angle ϕ drops out completely, leaving:

$${}_{\pm2}Y_{\ell m}(\boldsymbol{n}) \longleftrightarrow \mathrm{e}^{\pm 2\mathrm{i}\phi_\ell}\,\mathrm{e}^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n}}. \tag{18}$$

This index shift $m \to m \pm 2$ in the Bessel expansion is the precise mathematical origin of the $\mathrm{e}^{\pm 2\mathrm{i}\phi_\ell}$ phase factors that couple the Cartesian Stokes parameters $(Q \pm \mathrm{i}U)(\boldsymbol{\ell})$ to the rotationally invariant E and B mode Fourier amplitudes.

In this small-scale limit ($r = 2\sin(\vartheta/2) \approx \vartheta \ll 1$), $Y_{\ell m}(\boldsymbol{n}) \to \exp(\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n})$ and ${}_{\pm2}Y_{\ell m}(\boldsymbol{n}) \to \exp(\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n} \pm 2\mathrm{i}\phi_\ell)$, so that the fields expand in two-dimensional Fourier modes:

$$\Theta(\boldsymbol{n}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}}{2\pi}\Theta(\boldsymbol{\ell})\mathrm{e}^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n}}, \tag{6.3.118}$$

$$(Q \pm \mathrm{i}U)(\boldsymbol{n}) = -\int \frac{\mathrm{d}^2\boldsymbol{\ell}}{2\pi}\,[E(\boldsymbol{\ell}) \pm \mathrm{i}B(\boldsymbol{\ell})]\,\frac{1}{\ell^2}(\partial^\pm)^2\mathrm{e}^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n}}. \tag{6.3.119}$$

Evaluating the action of the spin-differential operators in the flat limit gives

$$(\partial^\pm)^2\mathrm{e}^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n}} = -\frac{(\ell_x \pm \mathrm{i}\ell_y)^2}{\ell^2}\mathrm{e}^{\mp 2\mathrm{i}\phi}\mathrm{e}^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n}} = -\ell^2\mathrm{e}^{\pm 2\mathrm{i}(\phi_\ell-\phi)}\mathrm{e}^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n}}.$$

While the natural spherical basis $(\boldsymbol{e}_r, \boldsymbol{e}_\theta, \boldsymbol{e}_\phi)$ varies across the celestial sphere, within a small patch one can fix a Cartesian coordinate grid (x, y), defining a constant polarization reference frame across the field of view (Fig. 6.8).

Referred to the local vertical axis z, the Stokes parameters in the rotated Cartesian coordinates are related to the celestial basis quantities by $(Q \pm \mathrm{i}U)_{\mathrm{Cart}} = \mathrm{e}^{\mp 2\mathrm{i}\phi}(Q \pm \mathrm{i}U)_{\mathrm{cel}}$. In this Cartesian frame, the Stokes fields take the explicit Fourier representations:

$$Q_{\mathrm{Cart}}(\boldsymbol{n}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}}{2\pi}\,[E(\boldsymbol{\ell})\cos 2\phi_\ell - B(\boldsymbol{\ell})\sin 2\phi_\ell]\,\mathrm{e}^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n}}, \tag{6.3.120}$$

$$U_{\mathrm{Cart}}(\boldsymbol{n}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}}{2\pi}\,[E(\boldsymbol{\ell})\sin 2\phi_\ell + B(\boldsymbol{\ell})\cos 2\phi_\ell]\,\mathrm{e}^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n}}. \tag{6.3.121}$$

(Where unambiguous, we suppress the subscript "Cart" and write simply Q and U throughout the flat-sky treatment.)

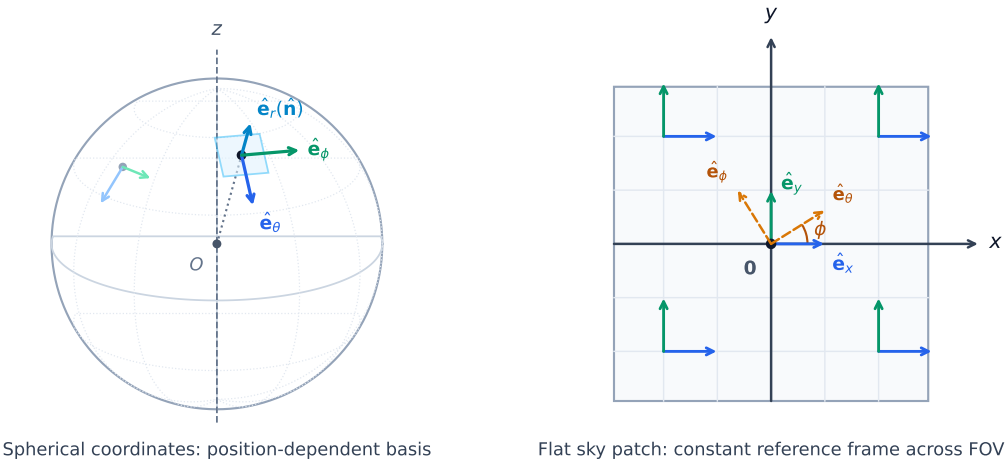


Figure 6.8: Comparison between celestial coordinates and the flat-sky approximation. Left: on the full celestial sphere, the orthonormal spherical basis $(\boldsymbol{e}_r, \boldsymbol{e}_\theta, \boldsymbol{e}_\phi)$ varies across lines of sight $\boldsymbol{n}$. Right: within a small angular patch around a chosen line of sight, the sphere is approximated by a flat tangent plane endowed with Cartesian coordinates (x, y), establishing a spatially uniform polarization reference frame $(\hat{\boldsymbol{e}}_x, \hat{\boldsymbol{e}}_y)$ across the entire field of view.

Polarization Reference Frames on the Flat Sky.
The Stokes parameters Q and U appearing in Eqs. (6.3.120)–(6.3.121) are defined with respect to the spatially uniform Cartesian frame $(\hat{\boldsymbol{e}}_x, \hat{\boldsymbol{e}}_y)$ on the tangent patch, denoted $Q_{\rm Cart}$ and $U_{\rm Cart}$, rather than the position-dependent celestial basis $(\boldsymbol{e}_\theta, \boldsymbol{e}_\phi)$. The transformation proceeds through three frames:

1. *Celestial spherical basis* $(\boldsymbol{e}_\theta, \boldsymbol{e}_\phi)$: On the curved sky, polarization is measured with respect to local meridians and parallels. In Eq. (6.3.119), the combination $(Q \pm \mathrm{i}U)_{\rm cel}(\boldsymbol{n})$ is defined in this basis. The spin-2 differential operator on the flat-sky patch generates:

$$(Q \pm \mathrm{i}U)_{\rm cel}(\boldsymbol{n}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}}{2\pi} \left[E(\boldsymbol{\ell}) \pm \mathrm{i}B(\boldsymbol{\ell})\right] \mathrm{e}^{\pm 2\mathrm{i}\phi_\ell}\, \mathrm{e}^{\pm 2\mathrm{i}\phi}\, \mathrm{e}^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n}}. \tag{1}$$

   Notice that the celestial Stokes parameters depend on *both* the wavevector orientation angle ϕ_ℓ and the spatial position coordinate angle ϕ. Because the celestial frame $(\boldsymbol{e}_\theta, \boldsymbol{e}_\phi)$ rotates with position angle ϕ, its components do not possess a simple Fourier plane-wave expansion.

2. *Fixed Cartesian frame* $(\hat{\boldsymbol{e}}_x, \hat{\boldsymbol{e}}_y)$ *on the sky patch*: To perform standard 2D Fourier analysis and compute derivatives like $\Delta E = (\partial_x^2 - \partial_y^2)Q + 2\partial_x\partial_y U$, one must work in a single, spatially constant reference frame. Rotating the polarization reference axes from $(\boldsymbol{e}_\theta, \boldsymbol{e}_\phi)$ to $(\hat{\boldsymbol{e}}_x, \hat{\boldsymbol{e}}_y)$ by the position angle ϕ gives

$$(Q \pm \mathrm{i}U)_{\rm Cart}(\boldsymbol{n}) = \mathrm{e}^{\mp 2\mathrm{i}\phi}(Q \pm \mathrm{i}U)_{\rm cel}(\boldsymbol{n}). \tag{2}$$

   The position phase $\mathrm{e}^{\mp 2\mathrm{i}\phi}$ cancels the spatial factor $\mathrm{e}^{\pm 2\mathrm{i}\phi}$ identically, leaving

$$(Q \pm \mathrm{i}U)_{\rm Cart}(\boldsymbol{n}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}}{2\pi} \left[E(\boldsymbol{\ell}) \pm \mathrm{i}B(\boldsymbol{\ell})\right] \mathrm{e}^{\pm 2\mathrm{i}\phi_\ell} \mathrm{e}^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n}}. \tag{3}$$

   Now the integrand has no position-dependent phase other than the plane-wave spatial dependence $\mathrm{e}^{\mathrm{i}\boldsymbol{\ell}\cdot\boldsymbol{n}}$.

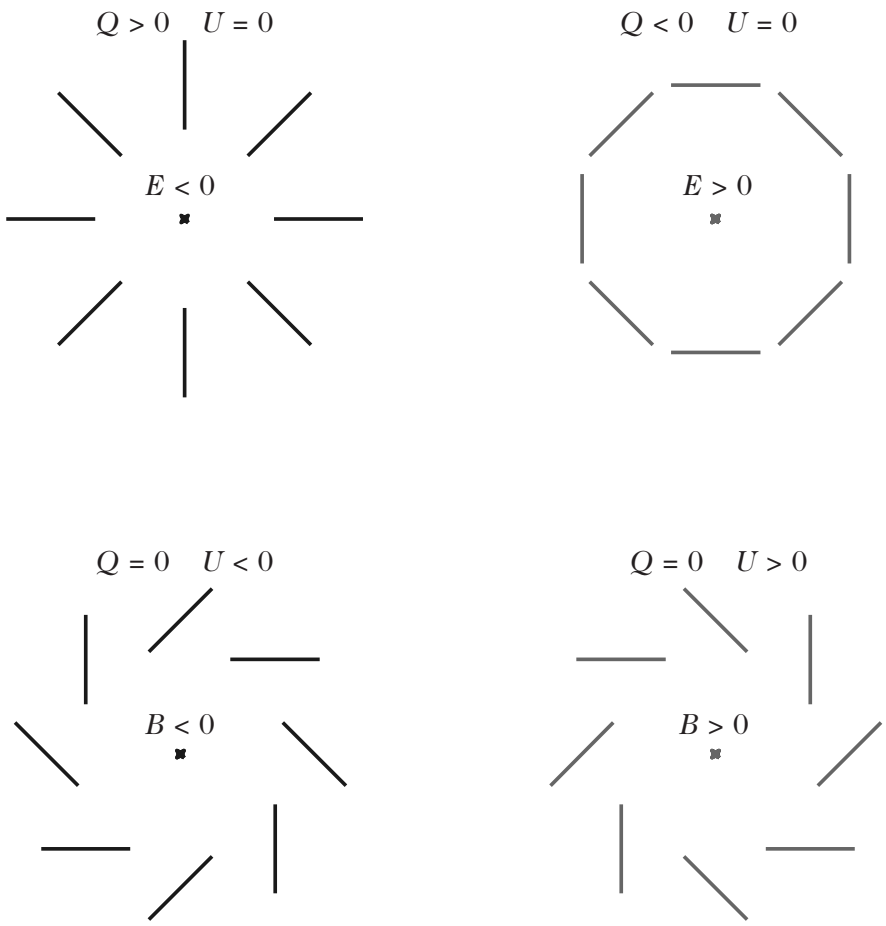


Figure 6.9: Polarization field configurations representing pure E-mode and pure B-mode patterns in the vicinity of a given central point. The line segments indicate the local orientation and amplitude of the linear polarization, directly depicting the Stokes parameters Q and U. These patterns repeat periodically across the plane. Because the scalar E and pseudo-scalar B fields are related to Q and U non-locally, determining their amplitudes at a specific point requires sampling the surrounding polarization texture. Under spatial reflection across a coordinate axis, the two upper configurations (E-modes) are invariant, whereas the two lower configurations (B-modes) are interchanged under reflection, illustrating their opposite parity properties.

3. *Fourier wavevector frame* $(\hat{\boldsymbol{\ell}}_\parallel, \hat{\boldsymbol{\ell}}_\perp)$: The scalar E and pseudoscalar B modes are physical geometric modes defined intrinsically with respect to each Fourier wavevector $\boldsymbol{\ell}$. The polarization basis aligned with $\boldsymbol{\ell}$ is rotated by angle ϕ_ℓ relative to the Cartesian x-axis $(\hat{\boldsymbol{e}}_x)$. Hence, the Cartesian Fourier modes and (E, B) modes are related by a pure rotation by $2\phi_\ell$:

$$(Q \pm \mathrm{i}U)_{\mathrm{Cart}}(\boldsymbol{\ell}) = \mathrm{e}^{\pm 2\mathrm{i}\phi_\ell} \left[E(\boldsymbol{\ell}) \pm \mathrm{i}B(\boldsymbol{\ell})\right]. \tag{4}$$

Separating into real and imaginary parts produces Eqs. (6.3.120)–(6.3.121).

Using the trigonometric identities connecting the wavevector orientation to its Cartesian components $\boldsymbol{\ell} = (\ell_x, \ell_y)$,

$$\cos 2\phi_\ell = \frac{\ell_x^2 - \ell_y^2}{\ell^2}, \qquad \sin 2\phi_\ell = \frac{2\ell_x \ell_y}{\ell^2}, \tag{6.3.122}$$

one can invert these expressions. In coordinate space, where $\ell_i \to -\mathrm{i}\partial_i$, this yields the spatial relations

$$\boxed{\Delta E = (\partial_x^2 - \partial_y^2)Q + 2\partial_x\partial_y U \qquad \text{and} \qquad \Delta B = (\partial_x^2 - \partial_y^2)U - 2\partial_x\partial_y Q.} \tag{6.3.123}$$

Derivation of the Real-Space Polarization Laplacians. Inverting the orthogonal rotation in Eqs. (6.3.120)–(6.3.121) on the Fourier plane gives

$$\begin{pmatrix} E(\boldsymbol{\ell}) \\ B(\boldsymbol{\ell}) \end{pmatrix} = \begin{pmatrix} \cos 2\phi_\ell & \sin 2\phi_\ell \\ -\sin 2\phi_\ell & \cos 2\phi_\ell \end{pmatrix} \begin{pmatrix} Q(\boldsymbol{\ell}) \\ U(\boldsymbol{\ell}) \end{pmatrix}, \tag{1}$$

which evaluates in components to

$$E(\boldsymbol{\ell}) = \cos(2\phi_\ell)Q(\boldsymbol{\ell}) + \sin(2\phi_\ell)U(\boldsymbol{\ell}), \tag{2}$$
$$B(\boldsymbol{\ell}) = -\sin(2\phi_\ell)Q(\boldsymbol{\ell}) + \cos(2\phi_\ell)U(\boldsymbol{\ell}). \tag{3}$$

Substituting the wavevector direction components from Eq. (6.3.122), $\cos 2\phi_\ell = (\ell_x^2 - \ell_y^2)/\ell^2$ and $\sin 2\phi_\ell = 2\ell_x\ell_y/\ell^2$:

$$E(\boldsymbol{\ell}) = \frac{\ell_x^2 - \ell_y^2}{\ell^2}Q(\boldsymbol{\ell}) + \frac{2\ell_x\ell_y}{\ell^2}U(\boldsymbol{\ell}), \tag{4}$$
$$B(\boldsymbol{\ell}) = -\frac{2\ell_x\ell_y}{\ell^2}Q(\boldsymbol{\ell}) + \frac{\ell_x^2 - \ell_y^2}{\ell^2}U(\boldsymbol{\ell}). \tag{5}$$

Multiplying both sides by $-\ell^2$:

$$-\ell^2 E(\boldsymbol{\ell}) = -(\ell_x^2 - \ell_y^2)Q(\boldsymbol{\ell}) - 2\ell_x\ell_y U(\boldsymbol{\ell}), \tag{6}$$
$$-\ell^2 B(\boldsymbol{\ell}) = 2\ell_x\ell_y Q(\boldsymbol{\ell}) - (\ell_x^2 - \ell_y^2)U(\boldsymbol{\ell}). \tag{7}$$

Under the 2D inverse Fourier transform, spatial derivatives correspond to $\partial_j \leftrightarrow \mathrm{i}\ell_j$, establishing the differential operator replacements:

$$-\ell^2 = -(\ell_x^2 + \ell_y^2) \longleftrightarrow \partial_x^2 + \partial_y^2 \equiv \Delta, \tag{8}$$

alongside the anisotropic quadrupole differential operators:

$$-(\ell_x^2 - \ell_y^2) \longleftrightarrow \partial_x^2 - \partial_y^2, \qquad -2\ell_x\ell_y \longleftrightarrow 2\partial_x\partial_y. \tag{9}$$

Transforming the Fourier-space system directly into real coordinate space yields

$$\Delta E(\boldsymbol{n}) = (\partial_x^2 - \partial_y^2)Q(\boldsymbol{n}) + 2\partial_x\partial_y U(\boldsymbol{n}), \tag{10}$$
$$\Delta B(\boldsymbol{n}) = (\partial_x^2 - \partial_y^2)U(\boldsymbol{n}) - 2\partial_x\partial_y Q(\boldsymbol{n}), \tag{11}$$

which completes the derivation of Eq. (6.3.123). Inverting the Laplacian introduces the inverse Green's operator Δ^{-1}:

$$E(\boldsymbol{n}) = \Delta^{-1}\left[(\partial_x^2 - \partial_y^2)Q + 2\partial_x\partial_y U\right], \qquad B(\boldsymbol{n}) = \Delta^{-1}\left[(\partial_x^2 - \partial_y^2)U - 2\partial_x\partial_y Q\right], \tag{12}$$

proving that the scalar E and pseudoscalar B fields are fundamentally non-local spatial integrals over the surrounding Stokes parameter textures.

These relations demonstrate that E and B are fundamentally non-local functions of the observable Stokes parameters Q and U. This non-locality is directly inherited from Eq. (6.3.113),

reflecting the differential connection between spin-0 and spin-2 spherical harmonics.

Temperature and polarization expansion

The preceding analysis indicates that the radiation temperature perturbation $\Theta(\eta, \boldsymbol{x}, \boldsymbol{n})$ at position $\boldsymbol{x}$ in direction $\boldsymbol{n}$ can be expanded in terms of the scalar basis functions $G_{\ell m}(\boldsymbol{x}, \boldsymbol{n})$ defined by

$$G_{\ell m}(\boldsymbol{x}, \boldsymbol{n}) = (-\mathrm{i})^\ell \sqrt{\frac{4\pi}{2\ell+1}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} Y_{\ell m}(\boldsymbol{n}), \tag{6.3.124}$$

while the spin-± 2 polarization field is expanded in terms of the spin-weighted basis functions

$$_{\pm 2}G_{\ell m}(\boldsymbol{x}, \boldsymbol{n}) = (-\mathrm{i})^\ell \sqrt{\frac{4\pi}{2\ell+1}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} {}_{\pm 2}Y_{\ell m}(\boldsymbol{n}). \tag{6.3.125}$$

Consequently, we decompose the spatial dependence of Θ and $(Q \pm \mathrm{i}U)$ in Fourier plane waves and their directional dependence in spin-weighted spherical harmonics:

$$\Theta(\eta, \boldsymbol{x}, \boldsymbol{n}) = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^{3/2}} \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \Theta_\ell^{(m)}(\eta, \boldsymbol{k}) G_{\ell m}(\boldsymbol{x}, \boldsymbol{n}), \tag{6.3.126}$$

$$(Q \pm \mathrm{i}U)(\eta, \boldsymbol{x}, \boldsymbol{n}) = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^{3/2}} \sum_{\ell=2}^{\infty} \sum_{m=-\ell}^{\ell} \left[E_\ell^{(m)} \pm \mathrm{i}B_\ell^{(m)}\right](\eta, \boldsymbol{k}) {}_{\pm 2}G_{\ell m}(\boldsymbol{x}, \boldsymbol{n}). \tag{6.3.127}$$

By virtue of the Stewart–Walker lemma, because the background FLRW spacetime is unpolarized, the first-order polarization perturbation variables $E_\ell^{(m)}$ and $B_\ell^{(m)}$ are automatically gauge-invariant.

Scattering cross-section

The differential Thomson scattering cross section for an incident electromagnetic wave with linear polarization vector $\boldsymbol{E}_{\mathrm{in}}$ scattered into an outgoing state with polarization $\boldsymbol{E}_{\mathrm{out}}$ is given by (see Ref. [30])

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{3}{8\pi}\sigma_{\mathrm{T}} |\boldsymbol{E}_{\mathrm{in}}^* \cdot \boldsymbol{E}_{\mathrm{out}}|^2, \tag{6.3.128}$$

where σ_{T} is the Thomson cross section. Upon averaging over all polarization

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{\alpha^2}{2m_e}(1 + \cos^2\theta) \tag{6.3.129}$$

It is convenient to introduce the orthogonal partial intensities

$$I_1 \equiv a_1^2 = \frac{I+Q}{2}, \qquad I_2 \equiv a_2^2 = \frac{I-Q}{2}. \tag{6.3.130}$$

For an incoming photon scattered through an angle θ in the plane defined by $(\boldsymbol{n}, \boldsymbol{E}_2)$ (see Fig. 6.10), the outgoing partial intensities are

$$I_1^{\mathrm{out}} = \frac{3\sigma_{\mathrm{T}}}{8\pi} I_1^{\mathrm{in}}, \qquad I_2^{\mathrm{out}} = \frac{3\sigma_{\mathrm{T}}}{8\pi} \cos^2\theta\, I_2^{\mathrm{in}}, \qquad U^{\mathrm{out}} = \frac{3\sigma_T}{8\pi} \cos\theta\, U^{\mathrm{in}}. \tag{6.3.131}$$

In the matrix form:

$$\begin{pmatrix} I_1^{\mathrm{out}} \\ I_2^{\mathrm{out}} \\ U^{\mathrm{out}} \end{pmatrix} = \frac{3\sigma_T}{8\pi} \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & (\hat{\boldsymbol{n}}' \cdot \hat{\boldsymbol{n}})^2 & 0 \\ 0 & 0 & \hat{\boldsymbol{n}}' \cdot \hat{\boldsymbol{n}} \end{pmatrix}}_{M_{\mathrm{I_1\,I_2\,U}}} \begin{pmatrix} I_1^{\mathrm{in}} \\ I_2^{\mathrm{in}} \\ U^{\mathrm{in}} \end{pmatrix} \tag{6.3.132}$$

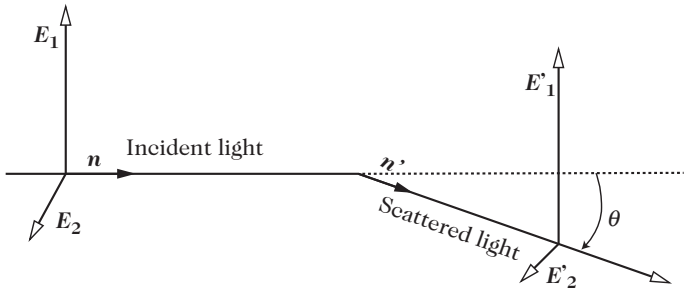


Figure 6.10: Geometry of Thomson scattering in the rest frame of the scattering electron. An incident photon propagating along $\boldsymbol{n}$ is scattered through an angle θ into direction $\boldsymbol{n}'$.

Expressed in terms of the Stokes parameters (I, Q), this scattering relation takes the matrix form

$$\begin{pmatrix} I_{\rm out} \\ Q_{\rm out} \end{pmatrix} = \frac{3\sigma_{\rm T}}{16\pi} \begin{pmatrix} 1+\cos^2\theta & \sin^2\theta \\ \sin^2\theta & 1+\cos^2\theta \end{pmatrix} \begin{pmatrix} I_{\rm in} \\ Q_{\rm in} \end{pmatrix}. \tag{6.3.133}$$

Under an arbitrary rotation of the polarization basis, the pair (Q, U) transforms according to Eq. (6.3.96). To evaluate the scattering cross section for an incoming radiation state $(I_{\rm in}, Q_{\rm in}, U_{\rm in})$ along direction $\boldsymbol{n}$ redirected into an outgoing state $(I_{\rm out}, Q_{\rm out}, U_{\rm out})$ along $\boldsymbol{n}'$, we align the coordinates with a fixed reference plane (y, z) (see Fig. 6.11 and Ref. [30]):

1. Rotate about $\boldsymbol{n}$ to bring the scattering plane $(\boldsymbol{n}, \boldsymbol{n}')$ into the plane $(\boldsymbol{n}, \boldsymbol{z})$, corresponding to a rotation (6.3.96) with $\phi = \alpha$.

2. Rotate about the z-axis to bring this plane into coincidence with the reference (y, z) plane; this spatial rotation leaves the Stokes parameters invariant.

3. Apply the scattering transfer matrix of Eq. (6.3.133).

4. Rotate back about the z-axis to restore the scattering plane to $(\boldsymbol{z}, \boldsymbol{n}')$ (again leaving Stokes parameters unaffected).

5. Rotate about the outgoing direction $\boldsymbol{n}'$ through angle $\phi = \alpha'$ to align with the final polarization basis.

Because circular polarization decouples completely and remains zero if initially unexcited, we omit the Stokes parameter V. The preceding transformation sequence yields the linear relation between outgoing and incoming Stokes parameters:

$$\begin{pmatrix} I_1^{\rm out} \\ I_2^{\rm out} \\ U^{\rm out} \end{pmatrix} = \frac{3\sigma_{\rm T}}{4\pi} \boldsymbol{P} \begin{pmatrix} I_1^{\rm in} \\ I_2^{\rm in} \\ U^{\rm in} \end{pmatrix} = \frac{3\sigma_T}{8\pi} R(\alpha') M_{\rm I_1\, I_2\, U} R(\alpha) \begin{pmatrix} I_1^{\rm in} \\ I_2^{\rm in} \\ U^{\rm in} \end{pmatrix}. \tag{6.3.134}$$

Setting $\mu \equiv \cos\theta$ and $\mu' \equiv \cos\theta'$ such that

$$\begin{aligned} \hat{\boldsymbol{n}}' \cdot \hat{\boldsymbol{n}} &= (\sin\theta'\cos\phi', \sin\theta'\sin\phi', \cos\theta') \cdot (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta) \\ &= \sin\theta\sin\theta'\cos(\phi - \phi') + \cos\theta\cos\theta' \\ &= \sqrt{1-\mu^2}\sqrt{1-\mu'^2}\cos\delta + \mu\,\mu', \end{aligned} \tag{6.3.135}$$

the scattering matrix $\boldsymbol{P}$ decomposes into three azimuthal harmonic blocks:

$$\boldsymbol{P} = R(\alpha') M_{\rm I_1\, I_2\, U} R(\alpha) = \boldsymbol{P}^{(0)} + \sqrt{(1-\mu^2)(1-\mu'^2)}\,\boldsymbol{P}^{(1)} + \boldsymbol{P}^{(2)}, \tag{6.3.136}$$

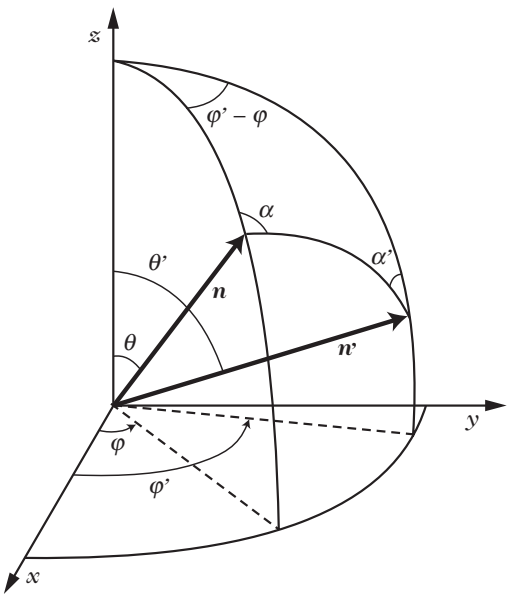


Figure 6.11: Definition of geometric vectors and spherical angles characterizing Thomson scattering between incident direction $\boldsymbol{n}$ and scattered direction $\boldsymbol{n}'$.

where

$$\boldsymbol{P}^{(0)} = \frac{3}{4}\begin{pmatrix} 2(1-\mu^2)(1-\mu'^2)+\mu^2\mu'^2 & \mu^2 & 0 \\ \mu'^2 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{6.3.137}$$

$$\boldsymbol{P}^{(1)} = \frac{3}{4}\begin{pmatrix} 4\mu\mu'\cos\delta & 0 & 2\mu\sin\delta \\ 0 & 0 & 0 \\ -4\mu'\sin\delta & 0 & 2\cos\delta \end{pmatrix}, \tag{6.3.138}$$

$$\boldsymbol{P}^{(2)} = \frac{3}{4}\begin{pmatrix} \mu^2\mu'^2\cos 2\delta & -\mu^2\cos 2\delta & \mu^2\mu'\sin 2\delta \\ -\mu'^2\cos 2\delta & \cos 2\delta & -\mu'\sin 2\delta \\ -2\mu\mu'^2\sin 2\delta & 2\mu\sin 2\delta & 2\mu\mu'\cos 2\delta \end{pmatrix}, \tag{6.3.139}$$

with the azimuthal difference $\delta \equiv \phi' - \phi$. The standard Stokes parameters (I, Q, U) are recovered via the algebraic transformation

$$\begin{pmatrix} I_1 \\ I_2 \\ U \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 2 \end{pmatrix}\begin{pmatrix} I \\ Q \\ U \end{pmatrix}. \tag{6.3.140}$$

In the rest frame of the baryon-electron fluid, the collision term governing each polarization state λ is given by

$$C^{(\lambda)} = \sigma_{\rm T} n_{\rm e}\left[\int f^{(\lambda')}(E', \boldsymbol{n}')P^{\lambda}{}_{\lambda'}(\boldsymbol{n}, \boldsymbol{n}')\frac{{\rm d}^2\boldsymbol{n}'}{4\pi} - f^{(\lambda)}(E', \boldsymbol{n})\right], \tag{6.3.141}$$

where $P^{\lambda}{}_{\lambda'}(\boldsymbol{n}, \boldsymbol{n}')$ represents the 2×2 submatrix corresponding to (I_1, I_2). This expression provides the full polarized generalization of Eq. (6.3.47).

Expanding the collision integral in the harmonic bases of Eqs. (6.3.124) and (6.3.125), we group the temperature and polarization perturbations into the triplet $\boldsymbol{\mathcal{T}} \equiv (\Theta, Q + {\rm i}U, Q - {\rm i}U)^T$.

One can then establish that the collision operator takes the compact form (see Ref. [31])

$$C[\boldsymbol{\mathcal{T}}] = \tau' \left\{ -\boldsymbol{\mathcal{T}} + \begin{pmatrix} \Theta_0 + n_i V_{\rm b}^i \\ 0 \\ 0 \end{pmatrix} + \frac{1}{10} \sum_{m=-2}^{2} \int \boldsymbol{\mathcal{P}}^{(m)}(\boldsymbol{n}, \boldsymbol{n}') \boldsymbol{\mathcal{T}}(\boldsymbol{n}') \mathrm{d}^2 \boldsymbol{n}' \right\}, \tag{6.3.142}$$

where the coupling matrix $\boldsymbol{\mathcal{P}}^{(m)}(\boldsymbol{n}, \boldsymbol{n}')$ follows from Eqs. (6.3.136)–(6.3.140) and reads explicitly:

$$\boldsymbol{\mathcal{P}}^{(m)} = \begin{pmatrix} Y_{2m}(\boldsymbol{n}')Y_{2m}(\boldsymbol{n}) & -\sqrt{\frac{3}{2}}{}_2Y_{2m}(\boldsymbol{n}')Y_{2m}(\boldsymbol{n}) & -\sqrt{\frac{3}{2}}{}_{-2}Y_{2m}(\boldsymbol{n}')Y_{2m}(\boldsymbol{n}) \\ -\sqrt{6}Y_{2m}(\boldsymbol{n}')_2Y_{2m}(\boldsymbol{n}) & 3_2Y_{2m}(\boldsymbol{n}')_2Y_{2m}(\boldsymbol{n}) & 3_{-2}Y_{2m}(\boldsymbol{n}')_2Y_{2m}(\boldsymbol{n}) \\ -\sqrt{6}Y_{2m}(\boldsymbol{n}')_{-2}Y_{2m}(\boldsymbol{n}) & 3_2Y_{2m}(\boldsymbol{n}')_{-2}Y_{2m}(\boldsymbol{n}) & 3_{-2}Y_{2m}(\boldsymbol{n}')_{-2}Y_{2m}(\boldsymbol{n}) \end{pmatrix}. \tag{6.3.143}$$

When restricted to the unpolarized temperature sector, this general result reduces directly to Eq. (6.3.57).

We begin with the process

$$\gamma(k) + e(p) \longleftrightarrow \gamma(k') + e(p'). \tag{1}$$

The invariant Boltzmann collision term for the photon distribution function is

$$\begin{aligned} C[f_\gamma(k)] = \frac{1}{2E_k} \int & \frac{\mathrm{d}^3\boldsymbol{p}}{(2\pi)^3\, 2E_p} \frac{\mathrm{d}^3\boldsymbol{k}'}{(2\pi)^3\, 2E_{k'}} \frac{\mathrm{d}^3\boldsymbol{p}'}{(2\pi)^3\, 2E_{p'}} \\ & \times (2\pi)^4 \delta^{(4)}(p + k - p' - k')\, |\mathcal{M}|^2 \\ & \times \Big[f_\gamma(k') f_e(p') \big(1 + f_\gamma(k)\big) \big(1 - f_e(p)\big) \\ & \qquad - f_\gamma(k) f_e(p) \big(1 - f_e(p')\big) \big(1 + f_\gamma(k')\big) \Big]. \end{aligned} \tag{2}$$

For the non-relativistic, non-degenerate electrons relevant for Thomson scattering,

$$f_e \ll 1, \qquad 1 - f_e \simeq 1. \tag{3}$$

In the Thomson limit,

$$E_k, E_{k'} \ll m_e, \tag{4}$$

so that electron recoil may be neglected,

$$E_{k'} = E_k, \tag{5}$$

and the electron distribution is unchanged by the scattering,

$$f_e(p') \simeq f_e(p). \tag{6}$$

The statistical factor in (2) then becomes

$$\begin{aligned} & f_\gamma(k') f_e(p') \big(1 + f_\gamma(k)\big) - f_\gamma(k) f_e(p) \big(1 + f_\gamma(k')\big) \\ & \quad \simeq f_e(p) \left[f_\gamma(k') + f_\gamma(k') f_\gamma(k) - f_\gamma(k) - f_\gamma(k) f_\gamma(k') \right] \\ & \quad = f_e(p) \left[f_\gamma(k') - f_\gamma(k) \right]. \end{aligned} \tag{7}$$

Thus, in the elastic Thomson limit, the Bose-enhancement terms cancel, and the collision integral takes the usual gain-minus-loss form.

From the invariant collision integral to the cross section. The invariant differential cross section for the process is

$$\mathrm{d}\sigma = \frac{(2\pi)^4\delta^{(4)}(p+k-p'-k')\,|\mathcal{M}|^2}{4p\cdot k}\,\frac{\mathrm{d}^3\boldsymbol{k}'}{(2\pi)^3\,2E_{k'}}\,\frac{\mathrm{d}^3\boldsymbol{p}'}{(2\pi)^3\,2E_{p'}}. \tag{8}$$

Introducing the relative velocity

$$v_{\mathrm{rel}} = \frac{p\cdot k}{E_pE_k}, \tag{9}$$

the collision term may be written as

$$C[f_\gamma(k)] = \int \frac{\mathrm{d}^3\boldsymbol{p}}{(2\pi)^3} f_e(p)\, v_{\mathrm{rel}} \int \mathrm{d}\sigma\,\left[f_\gamma(k') - f_\gamma(k)\right]. \tag{10}$$

In the electron rest frame,

$$p^\mu = (m_e, \mathbf{0}), \tag{11}$$

and hence

$$v_{\mathrm{rel}} = 1. \tag{12}$$

The electron number density is

$$n_e = \int \frac{\mathrm{d}^3\boldsymbol{p}}{(2\pi)^3} f_e(p). \tag{13}$$

Therefore,

$$C[f_\gamma(\boldsymbol{n})] = n_e \int \mathrm{d}\sigma\,\left[f_\gamma(\boldsymbol{n}') - f_\gamma(\boldsymbol{n})\right]. \tag{14}$$

Writing

$$\mathrm{d}\sigma = \frac{\mathrm{d}\sigma}{\mathrm{d}\Omega_{\boldsymbol{n}'}}\mathrm{d}\Omega_{\boldsymbol{n}'}, \tag{15}$$

we obtain

$$\begin{aligned} C[f_\gamma(\boldsymbol{n})] &= n_e \int \mathrm{d}\Omega_{\boldsymbol{n}'}\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega_{\boldsymbol{n}'}}\left[f_\gamma(\boldsymbol{n}') - f_\gamma(\boldsymbol{n})\right] = C[f_\gamma(\boldsymbol{n})]_{\mathrm{in}} + C[f_\gamma(\boldsymbol{n})]_{\mathrm{out}} \\ &= n_e \int \mathrm{d}\Omega_{\boldsymbol{n}'}\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega_{\boldsymbol{n}'}} f_\gamma(\boldsymbol{n}') - n_e \int \mathrm{d}\Omega_{\boldsymbol{n}'}\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega_{\boldsymbol{n}'}} f_\gamma(\boldsymbol{n}). \end{aligned} \tag{16}$$

This equation is the direct origin of the gain-minus-loss structure of the Thomson collision term.

Relation between the distribution function and Θ. The photon distribution is locally thermal,

$$f_\gamma(E,\boldsymbol{n}) = \frac{1}{e^{E/T(\boldsymbol{n})}-1}, \qquad T(\boldsymbol{n}) = \bar{T}(1+\Theta). \tag{17}$$

Expanding to first order in Θ,

$$f_\gamma = \bar{f} + \left.\frac{\partial f}{\partial T}\right|_{\bar{T}} \bar{T}\Theta, \tag{18}$$

where

$$\bar{f} = \frac{1}{e^{E/\bar{T}}-1}. \tag{19}$$

Since the background distribution depends on E and T only through $x = E/T$, the chain rule gives

$$\frac{\partial \bar{f}}{\partial T} = \frac{\mathrm{d}F}{\mathrm{d}x}\frac{\partial (E/T)}{\partial T} = -\frac{E}{T^2}\frac{\mathrm{d}F}{\mathrm{d}x}, \qquad \frac{\partial \bar{f}}{\partial E} = \frac{\mathrm{d}F}{\mathrm{d}x}\frac{\partial (E/T)}{\partial E} = \frac{1}{T}\frac{\mathrm{d}F}{\mathrm{d}x}. \tag{20}$$

Hence,

$$T\frac{\partial \bar{f}}{\partial T} = -E\frac{\partial \bar{f}}{\partial E}. \tag{21}$$

Therefore,

$$f_\gamma = \bar{f} + \delta f, \qquad \delta f = -E\frac{\partial \bar{f}}{\partial E}\Theta. \tag{22}$$

Since Thomson scattering is elastic,

$$E' = E, \tag{23}$$

the common spectral factor

$$-E\frac{\partial \bar{f}}{\partial E} \tag{24}$$

factors out of the collision term. Thus the angular collision term may be expressed directly in terms of Θ, together with the polarization variables Q and U.

Thomson scattering amplitude and polarization basis. Consider an incoming photon propagating in the direction $\boldsymbol{n}'$ and scattered into the direction $\boldsymbol{n}$. Define the scattering angle β by

$$\boldsymbol{n}\cdot\boldsymbol{n}' = \cos\beta. \tag{25}$$

In the electron rest frame and in the Thomson limit, the scattering amplitude is

$$\mathcal{A} = -\frac{e^2}{m_e}\,\boldsymbol{\epsilon}^*(\boldsymbol{n})\cdot\boldsymbol{\epsilon}(\boldsymbol{n}'). \tag{26}$$

Choose the polarization basis such that $\boldsymbol{\epsilon}^{(1)}$ lies in the scattering plane and $\boldsymbol{\epsilon}^{(2)}$ is perpendicular to it. Then

$$E_\parallel^{\mathrm{out}} = \frac{e^2}{m_e}\cos\beta\, E_\parallel^{\mathrm{in}}, \tag{27}$$

while

$$E_\perp^{\mathrm{out}} = \frac{e^2}{m_e}E_\perp^{\mathrm{in}}. \tag{28}$$

Using

$$\sigma_T = \frac{8\pi}{3}\left(\frac{e^2}{m_e}\right)^2, \tag{29}$$

the scattered field intensities may be written as

$$|E_\parallel^{\mathrm{out}}|^2 = \frac{3}{8\pi}n_e\sigma_T\cos^2\beta\,|E_\parallel^{\mathrm{in}}|^2, \tag{30}$$

and

$$|E_\perp^{\mathrm{out}}|^2 = \frac{3}{8\pi}n_e\sigma_T|E_\perp^{\mathrm{in}}|^2. \tag{31}$$

Stokes parameters. Define the Stokes parameters in the local scattering-plane basis by

$$\Theta \equiv |E_\parallel|^2 + |E_\perp|^2, \tag{32}$$

$$Q \equiv |E_\parallel|^2 - |E_\perp|^2, \tag{33}$$

and

$$U \equiv 2\,\mathrm{Re}\left(E_\parallel E_\perp^*\right). \tag{34}$$

The inverse relations are

$$|E_\parallel|^2 = \frac{1}{2}(\Theta + Q), \qquad |E_\perp|^2 = \frac{1}{2}(\Theta - Q). \tag{35}$$

For the scattered intensity,

$$\begin{aligned}\Theta^{\rm out} &= |E_\parallel^{\rm out}|^2 + |E_\perp^{\rm out}|^2 \\ &= \frac{3}{8\pi} n_e \sigma_T \left[\cos^2\beta\,|E_\parallel^{\rm in}|^2 + |E_\perp^{\rm in}|^2\right] \\ &= \frac{3}{16\pi} n_e \sigma_T \left[(1+\cos^2\beta)\Theta^{\rm in} - \sin^2\beta\, Q^{\rm in}\right].\end{aligned} \tag{36}$$

Similarly,

$$\begin{aligned}Q^{\rm out} &= |E_\parallel^{\rm out}|^2 - |E_\perp^{\rm out}|^2 \\ &= \frac{3}{16\pi} n_e \sigma_T \left[(1+\cos^2\beta) Q^{\rm in} - \sin^2\beta\,\Theta^{\rm in}\right],\end{aligned} \tag{37}$$

and

$$\begin{aligned}U^{\rm out} &= 2\,\mathrm{Re}\left(E_\parallel^{\rm out} E_\perp^{\rm out*}\right) \\ &= \frac{3}{8\pi} n_e \sigma_T \cos\beta\, U^{\rm in}.\end{aligned} \tag{38}$$

Complex polarization basis. Defining

$$\mathcal{T} = \begin{pmatrix} \Theta \\ Q + iU \\ Q - iU \end{pmatrix}. \tag{39}$$

and using

$$Q = \frac{1}{2}\left[(Q+iU) + (Q-iU)\right], \tag{40}$$

with

$$iU = \frac{1}{2}\left[(Q+iU) - (Q-iU)\right], \tag{41}$$

we find

$$\begin{aligned}Q^{\rm out} + iU^{\rm out} = \frac{3}{16\pi} n_e \sigma_T \Big[&-\sin^2\beta\,\Theta^{\rm in} \\ &+ \frac{1}{2}(1+\cos\beta)^2 (Q^{\rm in} + iU^{\rm in}) \\ &+ \frac{1}{2}(1-\cos\beta)^2 (Q^{\rm in} - iU^{\rm in})\Big].\end{aligned} \tag{42}$$

Therefore,

$$\mathcal{T}^{\rm out} = \frac{n_e \sigma_T}{4\pi} S(\beta) \mathcal{T}^{\rm in}, \tag{43}$$

with

$$S(\beta) = \frac{3}{4} \begin{pmatrix} 1+\cos^2\beta & -\frac{1}{2}\sin^2\beta & -\frac{1}{2}\sin^2\beta \\ -\sin^2\beta & \frac{1}{2}(1+\cos\beta)^2 & \frac{1}{2}(1-\cos\beta)^2 \\ -\sin^2\beta & \frac{1}{2}(1-\cos\beta)^2 & \frac{1}{2}(1+\cos\beta)^2 \end{pmatrix}. \tag{44}$$

Thus the polarized differential cross-section operator in the scattering-plane basis is

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega_{\boldsymbol{n}'}} = \frac{\sigma_T}{4\pi} S(\beta). \tag{45}$$

Rotation between the fixed basis and scattering-plane basis. The matrix $S(\beta)$ is written in the local scattering-plane basis. To express the result in the fixed polarization basis $(\boldsymbol{e}_\theta, \boldsymbol{e}_\varphi)$, introduce the angles γ' and γ.
Under a rotation of the polarization basis by an angle α,

$$Q \pm iU \longrightarrow e^{\mp 2i\alpha}(Q \pm iU). \tag{46}$$

Hence

$$R(\alpha) = \mathrm{diag}\left(1, e^{2i\alpha}, e^{-2i\alpha}\right). \tag{47}$$

The incoming state is first rotated into the scattering-plane basis:

$$\mathcal{T}^{\rm in}_{\rm sc} = R(\gamma')\mathcal{T}(\boldsymbol{n}'). \tag{48}$$

The scattering matrix is then applied:

$$\mathcal{T}^{\rm out}_{\rm sc} = S(\beta)R(\gamma')\mathcal{T}(\boldsymbol{n}'). \tag{49}$$

Finally, the outgoing state is rotated back to the fixed basis:

$$\mathcal{T}_{\rm out} = R(-\gamma)S(\beta)R(\gamma')\mathcal{T}(\boldsymbol{n}'). \tag{50}$$

Therefore the polarized differential cross-section operator in the fixed basis is

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega_{\boldsymbol{n}'}} = \frac{\sigma_T}{4\pi} R(-\gamma)S(\beta)R(\gamma'). \tag{51}$$

Collision term in the electron rest frame. Substituting (51) into the gain-minus-loss expression (16), the in-scattering term is

$$C[f_\gamma(\boldsymbol{n})]_{\rm in} = n_e \sigma_T \int \frac{\mathrm{d}\Omega_{\boldsymbol{n}'}}{4\pi} R(-\gamma)S(\beta)R(\gamma')\mathcal{T}(\boldsymbol{n}'). \tag{52}$$

The out-scattering term is

$$C[f_\gamma(\boldsymbol{n})]_{\rm out} = -n_e \int \mathrm{d}\Omega_{\boldsymbol{n}'} \frac{\mathrm{d}\sigma}{\mathrm{d}\Omega_{\boldsymbol{n}'}} \mathcal{T}(\boldsymbol{n}). \tag{53}$$

Using

$$\int \mathrm{d}\Omega_{\boldsymbol{n}'} \frac{\mathrm{d}\sigma}{\mathrm{d}\Omega_{\boldsymbol{n}'}} = \sigma_T, \tag{54}$$

we obtain

$$C[\mathcal{T}](\boldsymbol{n})|_{\rm out} = -n_e \sigma_T \mathcal{T}(\boldsymbol{n}). \tag{55}$$

Combining the gain and loss terms gives

$$\boxed{C[\mathcal{T}](\boldsymbol{n}) = n_e \sigma_T \int \frac{\mathrm{d}\Omega_{\boldsymbol{n}'}}{4\pi} R(-\gamma)S(\beta)R(\gamma')\mathcal{T}(\boldsymbol{n}') - n_e \sigma_T \mathcal{T}(\boldsymbol{n}).} \tag{56}$$

Multipole decomposition of the scattering kernel. The product

$$R(-\gamma)S(\beta)R(\gamma') \tag{57}$$

can be written explicitly as

$$R(-\gamma)SR(\gamma') = \frac{3}{4}\begin{pmatrix} 1+\cos^2\beta & -\frac{1}{2}\sin^2\beta\, e^{2i\gamma'} & -\frac{1}{2}\sin^2\beta\, e^{-2i\gamma'} \\ -\sin^2\beta\, e^{-2i\gamma} & \frac{1}{2}(1+\cos\beta)^2 e^{-2i\gamma}e^{2i\gamma'} & \frac{1}{2}(1-\cos\beta)^2 e^{-2i\gamma}e^{-2i\gamma'} \\ -\sin^2\beta\, e^{2i\gamma} & \frac{1}{2}(1-\cos\beta)^2 e^{2i\gamma}e^{2i\gamma'} & \frac{1}{2}(1+\cos\beta)^2 e^{2i\gamma}e^{-2i\gamma'} \end{pmatrix}. \tag{58}$$

Expressing the angular dependence in terms of spin-weighted spherical harmonics with $\ell \leq 2$, we obtain

$$\begin{aligned} &R(-\gamma)SR(\gamma') \\ &= \frac{1}{2}\sqrt{\frac{4\pi}{5}}\begin{pmatrix} Y_{20}(\beta,\gamma') + 2\sqrt{5}\, Y_{00}(\beta,\gamma') & -\sqrt{\frac{3}{2}}\, Y_{2,-2}(\beta,\gamma') & -\sqrt{6}\, Y_{22}(\beta,\gamma') \\ -\sqrt{6}\,{}_2Y_{20}(\beta,\gamma')e^{-2i\gamma} & 3\,{}_2Y_{2,-2}(\beta,\gamma')e^{-2i\gamma} & 3\,{}_2Y_{22}(\beta,\gamma')e^{-2i\gamma} \\ -\sqrt{\frac{3}{2}}\,{}_{-2}Y_{20}(\beta,\gamma')e^{2i\gamma} & 3\,{}_{-2}Y_{2,-2}(\beta,\gamma')e^{2i\gamma} & 3\,{}_{-2}Y_{22}(\beta,\gamma')e^{2i\gamma} \end{pmatrix}. \end{aligned} \tag{59}$$

Using the spin-weighted addition theorem,

$${}_sY_{2s'}(\beta,\gamma')e^{-si\gamma} = \sqrt{\frac{4\pi}{5}}\sum_{m=-2}^{2} {}_{-s}Y^*_{2m}(\boldsymbol{n}')\,{}_sY_{2m}(\boldsymbol{n}), \tag{60}$$

the kernel becomes

$$R(-\gamma)SR(\gamma') = \frac{4\pi}{10}P(\boldsymbol{n},\boldsymbol{n}') + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{61}$$

where

$$P(\boldsymbol{n},\boldsymbol{n}') = \sum_{m=-2}^{2} P_m(\boldsymbol{n},\boldsymbol{n}'). \tag{62}$$

The individual matrices are

$$P_m(\boldsymbol{n},\boldsymbol{n}') = \begin{pmatrix} Y_{2m}(\boldsymbol{n})Y^*_{2m}(\boldsymbol{n}') & -\sqrt{\frac{3}{2}}Y_{2m}(\boldsymbol{n})\,{}_2Y^*_{2m}(\boldsymbol{n}') & -\sqrt{\frac{3}{2}}Y_{2m}(\boldsymbol{n})\,{}_{-2}Y^*_{2m}(\boldsymbol{n}') \\ -\sqrt{6}\,{}_2Y_{2m}(\boldsymbol{n})Y^*_{2m}(\boldsymbol{n}') & 3\,{}_2Y_{2m}(\boldsymbol{n})\,{}_2Y^*_{2m}(\boldsymbol{n}') & 3\,{}_2Y_{2m}(\boldsymbol{n})\,{}_{-2}Y^*_{2m}(\boldsymbol{n}') \\ -\sqrt{6}\,{}_{-2}Y_{2m}(\boldsymbol{n})Y^*_{2m}(\boldsymbol{n}') & 3\,{}_{-2}Y_{2m}(\boldsymbol{n})\,{}_2Y^*_{2m}(\boldsymbol{n}') & 3\,{}_{-2}Y_{2m}(\boldsymbol{n})\,{}_{-2}Y^*_{2m}(\boldsymbol{n}') \end{pmatrix}. \tag{63}$$

Substituting this decomposition into the collision term gives

$$C[\mathcal{T}]_{\text{rest}} = an_e\sigma_T\Big[\frac{1}{10}\int \mathrm{d}\Omega_{\boldsymbol{n}'}\sum_{m=-2}^{2} P_m(\boldsymbol{n},\boldsymbol{n}')\mathcal{T}(\boldsymbol{n}') + \frac{1}{4\pi}\int \mathrm{d}\Omega_{\boldsymbol{n}'}\,\Theta(\boldsymbol{n}')\begin{pmatrix}1\\0\\0\end{pmatrix} - \mathcal{T}(\boldsymbol{n})\Big]. \tag{64}$$

Here the factor a appears when converting from physical time to conformal time,

$$\mathrm{d}t = a\,\mathrm{d}\eta. \tag{65}$$

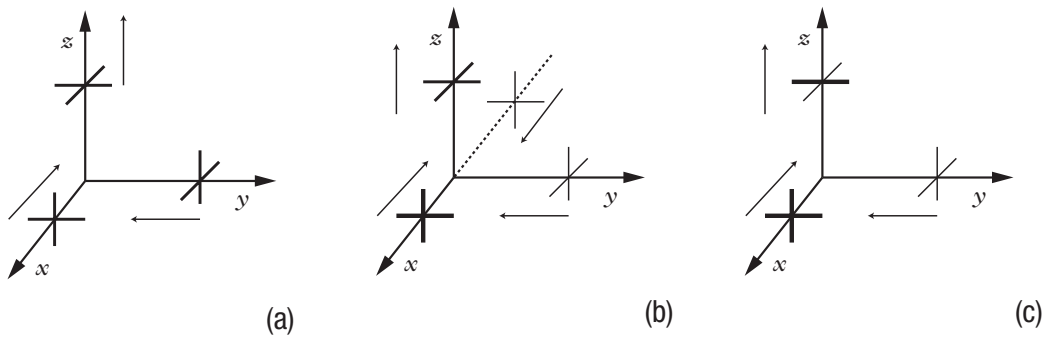


Figure 6.12: Heuristic generation of CMB polarization by Thomson scattering. (a) Under isotropic illumination, the scattered intensities along orthogonal directions balance exactly, producing an unpolarized scattered beam. (b) In the presence of a dipolar temperature anisotropy, opposing hot and cold regions cancel upon angular averaging, again producing zero net linear polarization. (c) A quadrupolar temperature anisotropy illuminates the electron with unequal orthogonal intensities, generating net linear polarization in the scattered radiation.

Transformation to the baryon frame. We now transform from the electron rest frame to the coordinate frame in which the baryon fluid has velocity $\boldsymbol{V}^{(b)}$.
To first order in the baryon velocity, the photon energy transforms as

$$E_{\rm rest} = E\left(1 - \boldsymbol{n}\cdot\boldsymbol{V}^{(b)}\right). \tag{66}$$

Consequently, the temperature perturbation in the electron rest frame is

$$\Theta_{\rm rest} = \Theta - \boldsymbol{n}\cdot\boldsymbol{V}^{(b)}. \tag{67}$$

Thus the loss term contributes

$$-\Theta_{\rm rest} = -\Theta + \boldsymbol{n}\cdot\boldsymbol{V}^{(b)}. \tag{68}$$

Defining the temperature monopole by

$$\Theta_0 = \frac{1}{4\pi}\int \mathrm{d}\Omega_{\boldsymbol{n}'}\Theta(\boldsymbol{n}'), \tag{69}$$

the collision term in the coordinate frame is

$$C[\mathcal{T}] = a n_e \sigma_T \left[\frac{1}{10}\int \mathrm{d}\Omega_{\boldsymbol{n}'}\sum_{m=-2}^{2} P_m(\boldsymbol{n},\boldsymbol{n}')\mathcal{T}(\boldsymbol{n}') - \mathcal{T}(\boldsymbol{n}) + \left(\Theta_0 + \boldsymbol{n}\cdot\boldsymbol{V}^{(b)}\right)\begin{pmatrix}1\\0\\0\end{pmatrix}\right]. \tag{70}$$

Origin of the cosmic microwave background polarization

The angular structure of the Thomson scattering cross section provides an intuitive, geometric explanation for the physical origin of CMB polarization (Fig. 6.12).

An electromagnetic wave incident along the x-axis has its electric field polarized in the (y, z) plane. If an unpolarized beam is deflected by 90° into the z-direction, the scattered photon becomes purely linearly polarized along the y-axis. When radiation arrives from all directions with equal intensity (isotropic illumination), statistical isotropy ensures that the y-polarization

sourced by waves arriving from the $\pm x$ directions exactly balances the x-polarization sourced by waves arriving from the $\pm y$ directions, resulting in zero net polarization on average (Fig. 6.12a).

If the incoming radiation field exhibits a dipolar temperature anisotropy, the y-directed polarization generated by photons arriving from $+x$ (hot) and $-x$ (cold) averages out to the same mean amplitude as the x-directed polarization from the orthogonal directions. Consequently, a dipole anisotropy also yields an unpolarized scattered beam (Fig. 6.12b).

To produce a net linear polarization in the scattered light, the incident radiation field must possess a non-zero local quadrupole moment (Fig. 6.12c). In this case, orthogonal directions subtend regions of unequal incident flux, so the scattered orthogonal polarizations fail to balance, yielding net linear polarization.

This geometric requirement has immediate cosmological implications:

- Because Thomson scattering can only operate effectively when free electrons are present, CMB polarization is generated almost exclusively during the recombination epoch (and later during reionization).
- Prior to decoupling, tight coupling between photons and baryons enforces near-complete isotropy in the baryon rest frame, keeping the radiation quadrupole very small ($\Theta_2 \ll \Theta_0, \Theta_1$). Consequently, the polarization signal is substantially weaker than the temperature anisotropies (by roughly an order of magnitude).
- For scalar cosmological perturbations, causal velocity gradients only build up quadrupoles inside the Hubble radius. Therefore, scalar-induced polarization is heavily suppressed on super-Hubble scales (at multipoles ℓ below the first acoustic peak). Sub-horizon acoustic oscillations generate polarization peaks that are systematically out of phase with the temperature acoustic peaks.
- In contrast, primordial tensor metric perturbations (gravitational waves) intrinsically produce a quadrupole tidal field even on super-Hubble scales, thereby generating large-angle polarization patterns in both E-modes and B-modes.

Figure 6.12 explains why a quadrupole illumination pattern is what Thomson scattering needs in order to polarize; it remains to see concretely where that quadrupole comes from. The mechanism invoked in the bullet above, causal velocity gradients inside the Hubble radius, is illustrated in Fig. 6.13: it is purely kinematic, generated entirely by the infall velocity field around an overdensity, with no quadrupole required in the density distribution itself. The $+$ and $\times$ polarizations form the linear polarization basis for gravitational waves and are not helicity eigenstates. The tensor-generated quadrupole has a different angular structure from the scalar-generated quadrupole, allowing tensor perturbations to generate both E- and B-mode polarization, whereas scalar perturbations being parity even generate only E-modes. In the linear polarization basis, the $+$ polarization has even parity and contributes to the E-type polarization pattern, while the $\times$ polarization has odd parity and contributes to the B-type polarization pattern.

Temperature and polarization hierarchy

Aligning the coordinate basis vector $\boldsymbol{e}_3$ with the wavevector $\boldsymbol{k}$, the spatial gradient term along the line of sight becomes

$$n^i \partial_i \to \mathrm{i} n_i k^i = \mathrm{i}\sqrt{\frac{4\pi}{3}} k Y_{10}(\boldsymbol{n}).$$

This directional operator multiplies the angular expansion of the temperature and polarization fields, generating products of spherical harmonics $Y_{10s} Y_{\ell m}$. By standard angular momentum

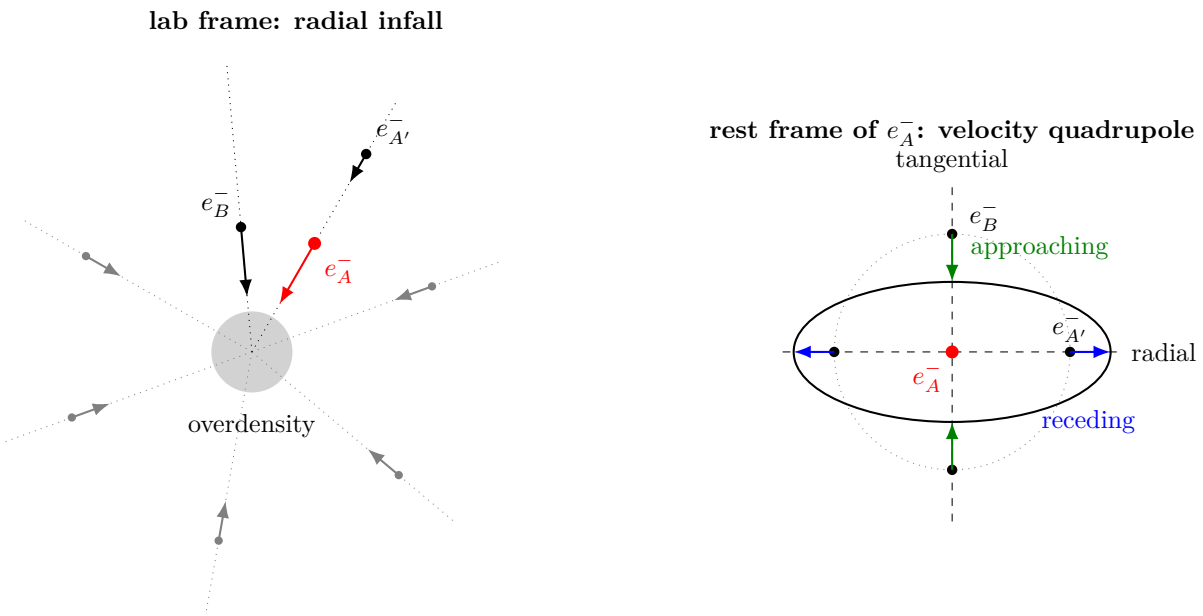


Figure 6.13: Radial infall onto an overdense region, and the velocity quadrupole it generates around a single infalling electron. *Left:* electrons stream radially inward, with infall speed increasing at smaller r. Electron e^-_A is singled out; $e^-_{A'}$ shares its radial line but sits farther out, while e^-_B sits at the same radius on a neighboring line. *Right:* boosting to the instantaneous rest frame of e^-_A, $e^-_{A'}$ recedes along the radial direction (it falls slower than A, so the separation grows), while e^-_B approaches along the tangential direction (both fall toward the same center, so their transverse separation shrinks as r decreases). The resulting distortion, stretched radially and squeezed tangentially, is a quadrupole ($\ell = 2$) pattern generated entirely by the velocity field, with no quadrupole present in the underlying density distribution.

coupling, these products expand as

$$\sqrt{\frac{4\pi}{3}}Y_{10}{}_sY_{\ell m} = \frac{{}_s\kappa^m_\ell}{\sqrt{(2\ell+1)(2\ell-1)}}{}_sY_{\ell-1,m} - \frac{ms}{\ell(\ell+1)}{}_sY_{\ell m} + \frac{{}_s\kappa^m_{\ell+1}}{\sqrt{(2\ell+1)(2\ell+3)}}{}_sY_{\ell+1,m}, \tag{6.3.144}$$

where the coupling coefficients are defined by

$${}_s\kappa^m_\ell \equiv \sqrt{\frac{(\ell^2-m^2)(\ell^2-s^2)}{\ell^2}}. \tag{6.3.145}$$

Projecting the linearized Boltzmann equations onto this harmonic basis yields the complete, coupled Boltzmann hierarchy for the multipole moments:

$$\Theta^{(m)\prime}_\ell = k\left[\frac{{}_0\kappa^m_\ell}{2\ell-1}\Theta^{(m)}_{\ell-1} - \frac{{}_0\kappa^m_{\ell+1}}{2\ell+3}\Theta^{(m)}_{\ell+1}\right] - \tau'\Theta^{(m)}_\ell + S^{(m)}_\ell, \tag{6.3.146}$$

$$E^{(m)\prime}_\ell = k\left[\frac{{}_2\kappa^m_\ell}{2\ell-1}E^{(m)}_{\ell-1} - \frac{2m}{\ell(\ell+1)}B^{(m)}_\ell - \frac{{}_2\kappa^m_{\ell+1}}{2\ell+3}E^{(m)}_{\ell+1}\right] - \tau'\left[E^{(m)}_\ell + \sqrt{6}P^{(m)}\delta_{\ell,2}\right], \tag{6.3.147}$$

$$B^{(m)\prime}_\ell = k\left[\frac{{}_2\kappa^m_\ell}{2\ell-1}B^{(m)}_{\ell-1} + \frac{2m}{\ell(\ell+1)}E^{(m)}_\ell - \frac{{}_2\kappa^m_{\ell+1}}{2\ell+3}B^{(m)}_{\ell+1}\right] - \tau'B^{(m)}_\ell. \tag{6.3.148}$$

The gravitational and scattering source terms $S^{(m)}_\ell$ are given by

$$S^{(0)}_0 = \tau'\Theta^{(0)}_0 + \Psi',$$
$$S^{(0)}_1 = -k\tau' V_{\rm b} + k\Phi,$$

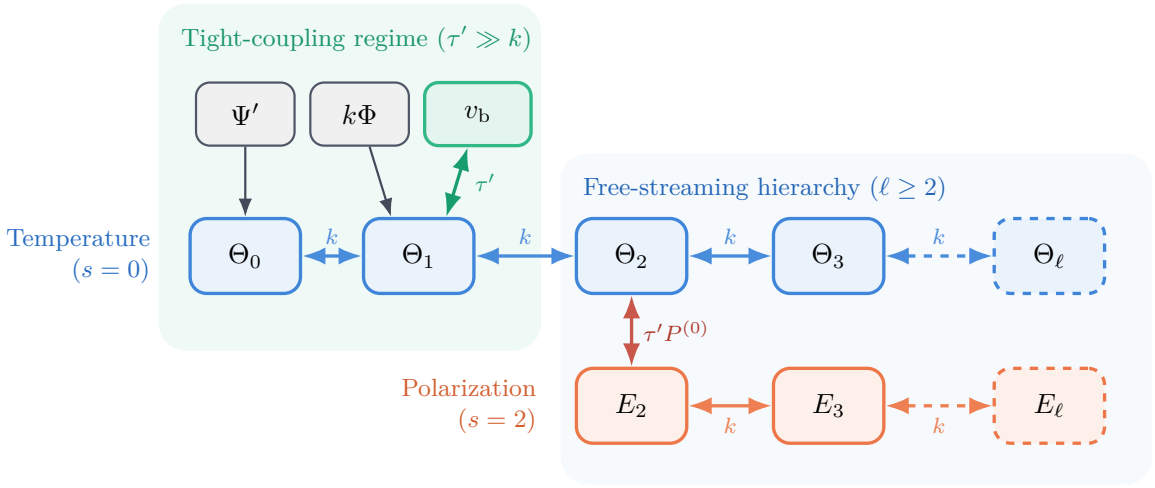


Figure 6.14: Structure of the coupled Boltzmann hierarchy for scalar perturbations ($m = 0$). Spatial gradients k link adjacent multipoles into a streaming ladder. In the tight-coupling regime ($\tau' \gg k$), the Bardeen potential time-derivative Ψ' sources the monopole Θ_0, gravitational acceleration $k\Phi$ drives the dipole Θ_1, and strong Thomson drag locks Θ_1 to the baryon velocity v_b, while higher multipoles ($\ell \geq 2$) are collision-damped ($-\tau'\Theta_\ell$). Residual quadrupole anisotropy Θ_2 sources E-mode polarization E_2 via Thomson scattering ($\tau' P^{(0)}$), which subsequently free-streams into higher polarization multipoles E_ℓ.

$$
\begin{aligned}
S_2^{(0)} &= \tau' P^{(0)}, \\
S_1^{(1)} &= -\tau' V_\mathrm{b}^{(1)}, \\
S_2^{(1)} &= \tau' P^{(1)} + \frac{k}{\sqrt{3}} \Phi^{(1)}, \\
S_2^{(2)} &= \tau' P^{(2)} - H^{(2)\prime},
\end{aligned}
\tag{6.3.149}
$$

where the vector and tensor metric perturbations are decomposed as

$$
\Phi_i = -\frac{\mathrm{i}}{\sqrt{2}} \sum_{m=\pm 1} e_i^{\pm} \Phi^{(m)}, \qquad E_{ij} = -\sqrt{\frac{3}{8}} \sum_{m=\pm 2} e_i^{\pm} \otimes e_j^{\pm} H^{(m)}. \tag{6.3.150}
$$

The polarization source term $P^{(m)}$, reflecting the anisotropic nature of Compton scattering, is determined by

$$
P^{(m)} = \frac{1}{10} \left[\Theta_2^{(m)} - \sqrt{6} E_2^{(m)} \right], \tag{6.3.151}
$$

which involves solely the quadrupolar moments ($\ell = 2$) of the temperature and of the E-mode polarization. The structural coupling of this multipole system is summarized schematically in Fig. 6.14.

We work in Fourier space for a fixed wavenumber vector $\mathbf{k}$ with direction $\hat{k}$. The direction of photon propagation is $\hat{n}$, with $\mu \equiv \hat{k} \cdot \hat{n} = \cos\theta$. The polarized radiation field vector is represented as:

$$
\mathbf{T}(\hat{n}) = \begin{pmatrix} \Theta(\hat{n}) \\ (Q + iU)(\hat{n}) \\ (Q - iU)(\hat{n}) \end{pmatrix}, \tag{1}
$$

where Θ corresponds to spin weight $s = 0$, and $Q \pm iU$ correspond to spin weights $s = \pm 2$.

To ensure that the resulting multipole differential equations remain purely real, we employ the standard phase convention with explicit $(-i)^\ell$ factors:

$$\Theta^{(m)}(\hat{n}) = \sum_{\ell=|m|}^{\infty} (-i)^\ell \Theta_\ell^{(m)} \, {}_0Y_\ell^m(\hat{n}), \tag{2}$$

$$(Q+iU)^{(m)}(\hat{n}) = -\sum_{\ell=\max(2,|m|)}^{\infty} (-i)^\ell \left(E_\ell^{(m)} + iB_\ell^{(m)}\right) {}_{+2}Y_\ell^m(\hat{n}), \tag{3}$$

$$(Q-iU)^{(m)}(\hat{n}) = -\sum_{\ell=\max(2,|m|)}^{\infty} (-i)^\ell \left(E_\ell^{(m)} - iB_\ell^{(m)}\right) {}_{-2}Y_\ell^m(\hat{n}). \tag{4}$$

Using the orthonormality of spin-weighted spherical harmonics,

$$\int d^2n \; {}_sY_\ell^{m*}(\hat{n}) \, {}_sY_{\ell'}^{m'}(\hat{n}) = \delta_{\ell\ell'}\delta_{mm'}, \tag{5}$$

the inverse projection relations are:

$$\Theta_\ell^{(m)} = i^\ell \int d^2n \; {}_0Y_\ell^{m*}(\hat{n}) \, \Theta^{(m)}(\hat{n}), \tag{6}$$

$$E_\ell^{(m)} + iB_\ell^{(m)} = -i^\ell \int d^2n \; {}_{+2}Y_\ell^{m*}(\hat{n}) \, (Q+iU)^{(m)}(\hat{n}), \tag{7}$$

$$E_\ell^{(m)} - iB_\ell^{(m)} = -i^\ell \int d^2n \; {}_{-2}Y_\ell^{m*}(\hat{n}) \, (Q-iU)^{(m)}(\hat{n}). \tag{8}$$

where the spin-weighted spherical harmonics couples adjacent multipoles according to the Wigner d-matrix recursion relation:

$$\mu \, {}_sY_\ell^m(\hat{n}) = \frac{{}_s\kappa_\ell^m}{2\ell-1} \, {}_sY_{\ell-1}^m(\hat{n}) + \frac{ms}{\ell(\ell+1)} \, {}_sY_\ell^m(\hat{n}) + \frac{{}_s\kappa_{\ell+1}^m}{2\ell+3} \, {}_sY_{\ell+1}^m(\hat{n}), \tag{9}$$

where the coupling coefficient is defined as:

$${}_s\kappa_\ell^m \equiv \sqrt{\frac{(\ell^2-m^2)(\ell^2-s^2)}{\ell^2}}. \tag{10}$$

The streaming operator acting on the radiation vector is:

$$\hat{L}[\mathbf{T}] = \left(\frac{\partial}{\partial\tau} + ik\mu\right)\mathbf{T}. \tag{11}$$

We now look at each spin one by one.
Intensity Channel ($s=0$): We project $\hat{L}[\Theta^{(m)}]$ using equation (6):

$$\left(\dot{\Theta}_\ell^{(m)}\right)_{\text{LHS}} = \dot{\Theta}_\ell^{(m)} + i^\ell \int d^2n \; {}_0Y_\ell^{m*}(\hat{n}) \left[ik\mu \sum_{\ell'} (-i)^{\ell'} \Theta_{\ell'}^{(m)} \, {}_0Y_{\ell'}^m(\hat{n})\right]. \tag{12}$$

Substitute the recursion relation (9) for $s=0$ (noting that the diagonal term vanishes for $s=0$):

$$= \dot{\Theta}_\ell^{(m)} + ik \sum_{\ell'} (-i)^{\ell'} i^\ell \Theta_{\ell'}^{(m)} \int d^2n \; {}_0Y_\ell^{m*}(\hat{n}) \left[\frac{{}_0\kappa_{\ell'}^m}{2\ell'-1} {}_0Y_{\ell'-1}^m + \frac{{}_0\kappa_{\ell'+1}^m}{2\ell'+3} {}_0Y_{\ell'+1}^m\right]. \tag{13}$$

Applying orthonormality yields two non-zero terms:

1. **Term 1 ($\ell' = \ell + 1$):**

$$ik\,(-i)^{\ell+1} i^{\ell}\,\frac{{}^{0}\kappa_{\ell+1}^{m}}{2(\ell+1)-1}\,\Theta_{\ell+1}^{(m)} = ik(-i)\frac{{}^{0}\kappa_{\ell+1}^{m}}{2\ell+1}\Theta_{\ell+1}^{(m)} = k\frac{{}^{0}\kappa_{\ell+1}^{m}}{2\ell+3}\Theta_{\ell+1}^{(m)}. \tag{14}$$

2. **Term 2 ($\ell' = \ell - 1$):**

$$ik\,(-i)^{\ell-1} i^{\ell}\,\frac{{}^{0}\kappa_{\ell}^{m}}{2(\ell-1)+3}\,\Theta_{\ell-1}^{(m)} = ik(+i)\frac{{}^{0}\kappa_{\ell}^{m}}{2\ell+1}\Theta_{\ell-1}^{(m)} = -k\frac{{}^{0}\kappa_{\ell}^{m}}{2\ell-1}\Theta_{\ell-1}^{(m)}. \tag{15}$$

Thus, the left-hand side is:

$$\left(\dot{\Theta}_{\ell}^{(m)}\right)_{\text{LHS}} = \dot{\Theta}_{\ell}^{(m)} - k\frac{{}^{0}\kappa_{\ell}^{m}}{2\ell-1}\Theta_{\ell-1}^{(m)} + k\frac{{}^{0}\kappa_{\ell+1}^{m}}{2\ell+3}\Theta_{\ell+1}^{(m)}. \tag{16}$$

Moving spatial gradient terms to the right-hand side gives the streaming contribution:

$$\boxed{\left(\dot{\Theta}_{\ell}^{(m)}\right)_{\text{stream}} = k\left[\frac{{}^{0}\kappa_{\ell}^{m}}{2\ell-1}\Theta_{\ell-1}^{(m)} - \frac{{}^{0}\kappa_{\ell+1}^{m}}{2\ell+3}\Theta_{\ell+1}^{(m)}\right].} \tag{17}$$

Polarization Channels ($s = \pm 2$): Let $P_+ \equiv Q + iU$ ($s = +2$) and $P_- \equiv Q - iU$ ($s = -2$).
Projecting $\hat{L}[P_+^{(m)}]$ using equation (7):

$$\begin{aligned}\left(\dot{E}_{\ell}^{(m)} + i\dot{B}_{\ell}^{(m)}\right)_{\text{LHS}} &= \dot{E}_{\ell}^{(m)} + i\dot{B}_{\ell}^{(m)} \\ &\quad - i^{\ell}\int d^2n\; {}_{+2}Y_{\ell}^{m*}(\hat{n})\left[ik\mu\left(-\sum_{\ell'}(-i)^{\ell'}(E_{\ell'}^{(m)} + iB_{\ell'}^{(m)})\,{}_{+2}Y_{\ell'}^{m}\right)\right] \\ &= (\dot{E}_{\ell} + i\dot{B}_{\ell}) - ik\sum_{\ell'}(-i)^{\ell'} i^{\ell}(E_{\ell'} + iB_{\ell'})\int d^2n\; {}_{+2}Y_{\ell}^{m*}\,(\mu\,{}_{+2}Y_{\ell'}^{m})\,.\end{aligned} \tag{18}$$

Substitute the recursion relation (9) with $s = +2$:

$$\mu\,{}_{+2}Y_{\ell'}^{m} = \frac{{}^{2}\kappa_{\ell'}^{m}}{2\ell'-1}\,{}_{+2}Y_{\ell'-1}^{m} + \frac{2m}{\ell'(\ell'+1)}\,{}_{+2}Y_{\ell'}^{m} + \frac{{}^{2}\kappa_{\ell'+1}^{m}}{2\ell'+3}\,{}_{+2}Y_{\ell'+1}^{m}. \tag{19}$$

Evaluating the three Kronecker delta contributions:

1. **Off-diagonal $\ell' = \ell - 1$:** Phase factor $-ik(-i)^{\ell-1}i^{\ell} = -k$.

$$\text{Contribution} = -k\frac{{}^{2}\kappa_{\ell}^{m}}{2\ell-1}(E_{\ell-1} + iB_{\ell-1}). \tag{20}$$

2. **Diagonal $\ell' = \ell$:** Phase factor $-ik(-i)^{\ell}i^{\ell} = -ik$.

$$\text{Contribution} = -ik\frac{2m}{\ell(\ell+1)}(E_{\ell} + iB_{\ell}). \tag{21}$$

3. **Off-diagonal $\ell' = \ell + 1$:** Phase factor $-ik(-i)^{\ell+1}i^{\ell} = +k$.

$$\text{Contribution} = +k\frac{{}^{2}\kappa_{\ell+1}^{m}}{2\ell+3}(E_{\ell+1} + iB_{\ell+1}). \tag{22}$$

Putting these together for $s = +2$:

$$\left(\dot{E}_\ell + i\dot{B}_\ell\right)_{\text{LHS}} = (\dot{E}_\ell + i\dot{B}_\ell) - k\frac{2\kappa_\ell^m}{2\ell-1}(E_{\ell-1} + iB_{\ell-1}) - ik\frac{2m}{\ell(\ell+1)}(E_\ell + iB_\ell) + k\frac{2\kappa_{\ell+1}^m}{2\ell+3}(E_{\ell+1} + iB_{\ell+1}). \quad (23)$$

Similarly, repeating for $s = -2$ (where the diagonal term flips sign due to $s = -2$):

$$\left(\dot{E}_\ell - i\dot{B}_\ell\right)_{\text{LHS}} = (\dot{E}_\ell - i\dot{B}_\ell) - k\frac{2\kappa_\ell^m}{2\ell-1}(E_{\ell-1} - iB_{\ell-1}) + ik\frac{2m}{\ell(\ell+1)}(E_\ell - iB_\ell) + k\frac{2\kappa_{\ell+1}^m}{2\ell+3}(E_{\ell+1} - iB_{\ell+1}). \quad (24)$$

To isolate $\dot{E}_\ell$, add equations (23) and (24):

$$\begin{aligned} 2\left(\dot{E}_\ell\right)_{\text{LHS}} &= 2\dot{E}_\ell - 2k\frac{2\kappa_\ell^m}{2\ell-1}E_{\ell-1} + 2k\frac{2\kappa_{\ell+1}^m}{2\ell+3}E_{\ell+1} - ik\frac{2m}{\ell(\ell+1)}\left[(E_\ell + iB_\ell) - (E_\ell - iB_\ell)\right] \\ &= 2\dot{E}_\ell - 2k\frac{2\kappa_\ell^m}{2\ell-1}E_{\ell-1} + 2k\frac{2\kappa_{\ell+1}^m}{2\ell+3}E_{\ell+1} + \frac{4km}{\ell(\ell+1)}B_\ell. \end{aligned} \quad (25)$$

Dividing by 2 and moving spatial terms to the right-hand side:

$$\boxed{\left(\dot{E}_\ell^{(m)}\right)_{\text{stream}} = k\left[\frac{2\kappa_\ell^m}{2\ell-1}E_{\ell-1}^{(m)} - \frac{2m}{\ell(\ell+1)}B_\ell^{(m)} - \frac{2\kappa_{\ell+1}^m}{2\ell+3}E_{\ell+1}^{(m)}\right].} \quad (26)$$

To isolate $\dot{B}_\ell$, subtract equation (24) from (23):

$$\begin{aligned} 2i\left(\dot{B}_\ell\right)_{\text{LHS}} &= 2i\dot{B}_\ell - 2ik\frac{2\kappa_\ell^m}{2\ell-1}B_{\ell-1} + 2ik\frac{2\kappa_{\ell+1}^m}{2\ell+3}B_{\ell+1} - ik\frac{2m}{\ell(\ell+1)}\left[(E_\ell + iB_\ell) + (E_\ell - iB_\ell)\right] \\ &= 2i\dot{B}_\ell - 2ik\frac{2\kappa_\ell^m}{2\ell-1}B_{\ell-1} + 2ik\frac{2\kappa_{\ell+1}^m}{2\ell+3}B_{\ell+1} - i\frac{4km}{\ell(\ell+1)}E_\ell. \end{aligned} \quad (27)$$

Dividing by $2i$ and moving spatial terms to the right-hand side:

$$\boxed{\left(\dot{B}_\ell^{(m)}\right)_{\text{stream}} = k\left[\frac{2\kappa_\ell^m}{2\ell-1}B_{\ell-1}^{(m)} + \frac{2m}{\ell(\ell+1)}E_\ell^{(m)} - \frac{2\kappa_{\ell+1}^m}{2\ell+3}B_{\ell+1}^{(m)}\right].} \quad (28)$$

The collision term in the Boltzmann equation is:

$$C[\mathbf{T}] = \tau'\left[-\mathbf{T}(\hat{n}) + \mathbf{T}_{\text{in, low}}(\hat{n}) + \mathbf{S}_{\text{quad}}^{(m)}(\hat{n})\right], \quad (29)$$

where $\tau' \equiv an_e\sigma_T$ is the differential optical depth.

1. **Scattering Out (Damping):** $-\tau'\mathbf{T}(\hat{n})$ projects directly onto $-\tau'\Theta_\ell$, $-\tau'E_\ell$, and $-\tau'B_\ell$.

2. **Monopole and Dipole In-Scattering:**

$$\mathbf{T}_{\text{in, low}}(\hat{n}) = \left(\Theta_0 - i\frac{V_b}{k}\cdot\hat{n}\right)\mathbf{e}_\Theta, \quad (30)$$

which projects exclusively onto $\ell = 0$ and $\ell = 1$ for intensity:

$$S_0^{(m)} = \tau'\Theta_0^{(m)}, \qquad S_1^{(m)} = -\tau'\frac{i}{k}V_b^{(m)}. \quad (31)$$

The quadrupole collision term is generated by integrating the Thomson phase matrix $P^{(m)}(\hat{n}, \hat{n}')$ over incoming directions $\hat{n}'$:

$$\mathbf{S}^{(m)}_{\rm quad}(\hat{n}) = \frac{\tau'}{10} \int d^2n'\; P^{(m)}(\hat{n}, \hat{n}')\, \mathbf{T}^{(m)}(\hat{n}'), \tag{32}$$

where the phase matrix is:

$$P^{(m)}(\hat{n}, \hat{n}') = \begin{pmatrix} Y_{2m}(\hat{n})Y^*_{2m}(\hat{n}') & \sqrt{\frac{3}{2}}Y_{2m}(\hat{n})_{+2}Y^*_{2m}(\hat{n}') & \sqrt{\frac{3}{2}}Y_{2m}(\hat{n})_{-2}Y^*_{2m}(\hat{n}') \\ \sqrt{6}_{+2}Y_{2m}(\hat{n})Y^*_{2m}(\hat{n}') & -3_{+2}Y_{2m}(\hat{n})_{+2}Y^*_{2m}(\hat{n}') & -3_{+2}Y_{2m}(\hat{n})_{-2}Y^*_{2m}(\hat{n}') \\ \sqrt{6}_{-2}Y_{2m}(\hat{n})Y^*_{2m}(\hat{n}') & -3_{-2}Y_{2m}(\hat{n})_{+2}Y^*_{2m}(\hat{n}') & -3_{-2}Y_{2m}(\hat{n})_{-2}Y^*_{2m}(\hat{n}') \end{pmatrix}. \tag{33}$$

Inserting the multipole expansion (2)–(4) into $\mathbf{T}^{(m)}(\hat{n}')$ and integrating:

$$\int d^2n'\; {}_sY^*_{2m}(\hat{n}')\, {}_sY^m_{\ell'}(\hat{n}') = \delta_{\ell' 2}. \tag{34}$$

Thus, **only quadrupole terms ($\ell' = 2$) survive the integration**. For $\ell' = 2$, the phase factor evaluates to $(-i)^2 = -1$. Evaluating matrix-vector multiplication row by row gives:

Row 1 (Θ Component)

$$\begin{aligned} \mathbf{I}_1 &= Y_{2m}(\hat{n}) \left[(-i)^2 \Theta^{(m)}_2\right] + \sqrt{\frac{3}{2}} Y_{2m}(\hat{n}) \left[-(-i)^2 (E^{(m)}_2 + iB^{(m)}_2)\right] \\ &\quad + \sqrt{\frac{3}{2}} Y_{2m}(\hat{n}) \left[-(-i)^2 (E^{(m)}_2 - iB^{(m)}_2)\right] \\ &= -Y_{2m}(\hat{n}) \Theta^{(m)}_2 + \sqrt{\frac{3}{2}} Y_{2m}(\hat{n}) \left(E^{(m)}_2 + iB^{(m)}_2\right) + \sqrt{\frac{3}{2}} Y_{2m}(\hat{n}) \left(E^{(m)}_2 - iB^{(m)}_2\right) \\ &= -Y_{2m}(\hat{n}) \left[\Theta^{(m)}_2 - \sqrt{6} E^{(m)}_2\right]. \end{aligned} \tag{35}$$

Row 2 ($Q + iU$ Component)

$$\begin{aligned} \mathbf{I}_2 &= \sqrt{6}_{+2}Y_{2m}(\hat{n}) \left[(-i)^2 \Theta^{(m)}_2\right] - 3_{+2}Y_{2m}(\hat{n}) \left[-(-i)^2 (E^{(m)}_2 + iB^{(m)}_2)\right] \\ &\quad - 3_{+2}Y_{2m}(\hat{n}) \left[-(-i)^2 (E^{(m)}_2 - iB^{(m)}_2)\right] \\ &= -\sqrt{6}_{+2}Y_{2m}(\hat{n}) \Theta^{(m)}_2 + 3_{+2}Y_{2m}(\hat{n}) \left(E^{(m)}_2 + iB^{(m)}_2\right) + 3_{+2}Y_{2m}(\hat{n}) \left(E^{(m)}_2 - iB^{(m)}_2\right) \\ &= -\sqrt{6}_{+2}Y_{2m}(\hat{n}) \left[\Theta^{(m)}_2 - \sqrt{6} E^{(m)}_2\right]. \end{aligned} \tag{36}$$

Row 3 ($Q - iU$ Component)

By identical calculation:

$$\mathbf{I}_3 = -\sqrt{6}_{-2}Y_{2m}(\hat{n}) \left[\Theta^{(m)}_2 - \sqrt{6} E^{(m)}_2\right]. \tag{37}$$

Quadrupole Source Vector

Defining the scalar perturbation $\Pi^{(m)} \equiv \Theta^{(m)}_2 - \sqrt{6} E^{(m)}_2$, the integrated vector is:

$$\mathbf{S}^{(m)}_{\rm quad}(\hat{n}) = \frac{\tau'}{10} \Pi^{(m)} \begin{pmatrix} -{}_0Y_{2m}(\hat{n}) \\ -\sqrt{6}_{+2}Y_{2m}(\hat{n}) \\ -\sqrt{6}_{-2}Y_{2m}(\hat{n}) \end{pmatrix}. \tag{38}$$

Projecting $\mathbf{S}^{(m)}_{\rm quad}(\hat{n})$ onto the harmonic basis at $\ell = 2$:

1. **Intensity (Θ_2):**

$$\dot{\Theta}_2\Big|_{\rm quad} = i^2 \int d^2n\, {}_0Y^*_{2m} \left(-\frac{\tau'}{10}\Pi^{(m)}\, {}_0Y_{2m}\right) = (-1)\left(-\frac{\tau'}{10}\Pi^{(m)}\right) = +\frac{\tau'}{10}\Pi^{(m)}. \quad (39)$$

2. **E-mode Polarization (E_2):**

$$\dot{E}_2\Big|_{\rm quad} = -i^2 \int d^2n\, {}_{+2}Y^*_{2m} \left(-\frac{\tau'}{10}\sqrt{6}\Pi^{(m)}\, {}_{+2}Y_{2m}\right) = -(-1)\left(-\frac{\tau'}{10}\sqrt{6}\Pi^{(m)}\right) \quad (40)$$

$$= -\tau'\sqrt{6}P^{(m)}, \quad (41)$$

where $P^{(m)} \equiv \frac{1}{10}\Pi^{(m)} = \frac{1}{10}\left(\Theta^{(m)}_2 - \sqrt{6}E^{(m)}_2\right)$.

3. **B-mode Polarization (B_2):** Parity invariance of Thomson scattering guarantees that no B-mode source is generated:

$$\dot{B}_\ell\Big|_{\rm quad} = 0. \quad (42)$$

Combining streaming (LHS) and collision (RHS) terms yields the complete set of equations.

Intensity Hierarchy (6.3.146)

$$\boxed{\dot{\Theta}^{(m)}_\ell = k\left[\frac{{}_0\kappa^m_\ell}{2\ell-1}\Theta^{(m)}_{\ell-1} - \frac{{}_0\kappa^m_{\ell+1}}{2\ell+3}\Theta^{(m)}_{\ell+1}\right] - \tau'\Theta^{(m)}_\ell + S^{(m)}_\ell,} \quad (43)$$

where the source term is defined as:

$$S^{(m)}_\ell = \tau'\Theta^{(m)}_0\delta_{\ell 0} - \tau'\frac{i}{k}V^{(m)}_{\rm b}\delta_{\ell 1} + \frac{\tau'}{10}\Pi^{(m)}\delta_{\ell 2}. \quad (44)$$

E-mode Hierarchy (6.3.147)

$$\boxed{\dot{E}^{(m)}_\ell = k\left[\frac{{}_2\kappa^m_\ell}{2\ell-1}E^{(m)}_{\ell-1} - \frac{2m}{\ell(\ell+1)}B^{(m)}_\ell - \frac{{}_2\kappa^m_{\ell+1}}{2\ell+3}E^{(m)}_{\ell+1}\right] - \tau'\left[E^{(m)}_\ell + \sqrt{6}\,P^{(m)}\delta_{\ell 2}\right],} \quad (45)$$

where $P^{(m)} \equiv \frac{1}{10}\left(\Theta^{(m)}_2 - \sqrt{6}E^{(m)}_2\right)$.

B-mode Hierarchy (6.3.148)

$$\boxed{\dot{B}^{(m)}_\ell = k\left[\frac{{}_2\kappa^m_\ell}{2\ell-1}B^{(m)}_{\ell-1} + \frac{2m}{\ell(\ell+1)}E^{(m)}_\ell - \frac{{}_2\kappa^m_{\ell+1}}{2\ell+3}B^{(m)}_{\ell+1}\right] - \tau'B^{(m)}_\ell.} \quad (46)$$

Significantly, the polarization source $P^{(m)}$ enters only into the evolution equation for the quadrupole of E. The temperature evolution equations should be compared with the unpolarized equations (6.3.69) and (6.3.72) derived previously. Because the system is invariant under azimuthal reflection $m \to -m$, one only needs to solve the hierarchy for positive m. Parity symmetry strictly demands that $B^{(0)}_\ell = 0$ for any scalar perturbation source.

This hierarchy replaces the hydrodynamic fluid equations for the radiation component, while

leaving the governing equations for cold dark matter and baryons unaltered. For massless neutrinos, an identical hierarchy applies with $\tau' = 0$ (since neutrinos do not undergo Thomson scattering).

In the tight-coupling regime ($\tau' \gg k$), expanding the scalar evolution equations to leading order yields

$$E_2^{(0)} = -\sqrt{6}P^{(0)}, \qquad P^{(0)} = \frac{1}{4}\Theta_2^{(0)}, \qquad \sqrt{6}\Theta_2^{(0)} = -4E_2^{(0)},$$

which implies the tight-coupling quadrupole relation

$$\Theta_2^{(0)} = \frac{8}{9}\frac{k}{\tau'}\Theta_1^{(0)}. \tag{6.3.152}$$

Recalling the shear stress relation of Eq. (6.3.73), we recover

$$\frac{k^2}{12}\pi_\gamma = \frac{1}{5}\Theta_2^{(0)} = -\frac{8}{45}\frac{k^2}{\tau'}V_\gamma, \tag{6.3.153}$$

which confirms the shear viscosity relation employed in our earlier analysis of Silk damping in Section 6.2.3.

Power spectra

Given the modal decompositions above, the angular power spectra between two observable fields $X, Z \in \{\Theta, E, B\}$ are given by

$$(2\ell+1)^2 C_\ell^{XZ} = \frac{2}{\pi}\int k^2\,\mathrm{d}k \sum_{m=-2}^{2} X_\ell^{(m)*}(\eta_0, k) Z_\ell^{(m)}(\eta_0, k). \tag{6.3.154}$$

Figure 6.15 illustrates the four non-vanishing spectra generated by scalar and tensor perturbations. While scalar perturbations generate Θ–Θ, E–E, and Θ–E correlations, they cannot source B-modes. Parity symmetry requires that $C_\ell^{\Theta B} = C_\ell^{EB} = 0$ in the absence of parity-violating physics. Primordial tensor perturbations, however, generate all four non-vanishing power spectra.

Cosmic birefringence and parity violation. As demonstrated in Section 6.3.3, the E-mode is an authentic scalar under spatial parity ($+1$), whereas the B-mode is a pseudoscalar (-1). Under parity reflection, the cross-correlations transform as

$$C_\ell^{TB} \to (+1)(-1)C_\ell^{TB} = -C_\ell^{TB}, \qquad C_\ell^{EB} \to (+1)(-1)C_\ell^{EB} = -C_\ell^{EB}, \tag{6.3.155}$$

which identically forces $C_\ell^{TB} = C_\ell^{EB} = 0$ in any parity-conserving universe.

However, if physical interactions violate parity during cosmic propagation—such as through a rolling pseudoscalar axion-like field coupled to electromagnetism via a Chern–Simons interaction $\mathcal{L}_{\rm int} = -\frac{1}{4}g_{\phi\gamma}\phi F_{\mu\nu}\tilde{F}^{\mu\nu}$—the phase velocities of right- and left-circularly polarized photons differ. For a homogeneous background $\phi(t)$, the dispersion relations for the two circular helicity states $\boldsymbol{e}_\pm = (\boldsymbol{e}_1 \pm \mathrm{i}\boldsymbol{e}_2)/\sqrt{2}$ become

$$\omega_\pm^2 = k^2 \mp g_{\phi\gamma}\dot{\phi}k \implies k_\pm \approx \omega \pm \frac{1}{2}g_{\phi\gamma}\dot{\phi}. \tag{6.3.156}$$

The accumulated phase difference between opposite circular helicities during the journey from the last scattering surface ($\eta_{\rm LSS}$) to today (η_0) rotates the plane of linear polarization by an in-flight rotation angle β:

$$\beta = \frac{1}{2}\int_{\eta_{\rm LSS}}^{\eta_0}(k_+ - k_-)\,\mathrm{d}\eta = \frac{1}{2}g_{\phi\gamma}(\phi_{\rm obs} - \phi_{\rm LSS}). \tag{6.3.157}$$

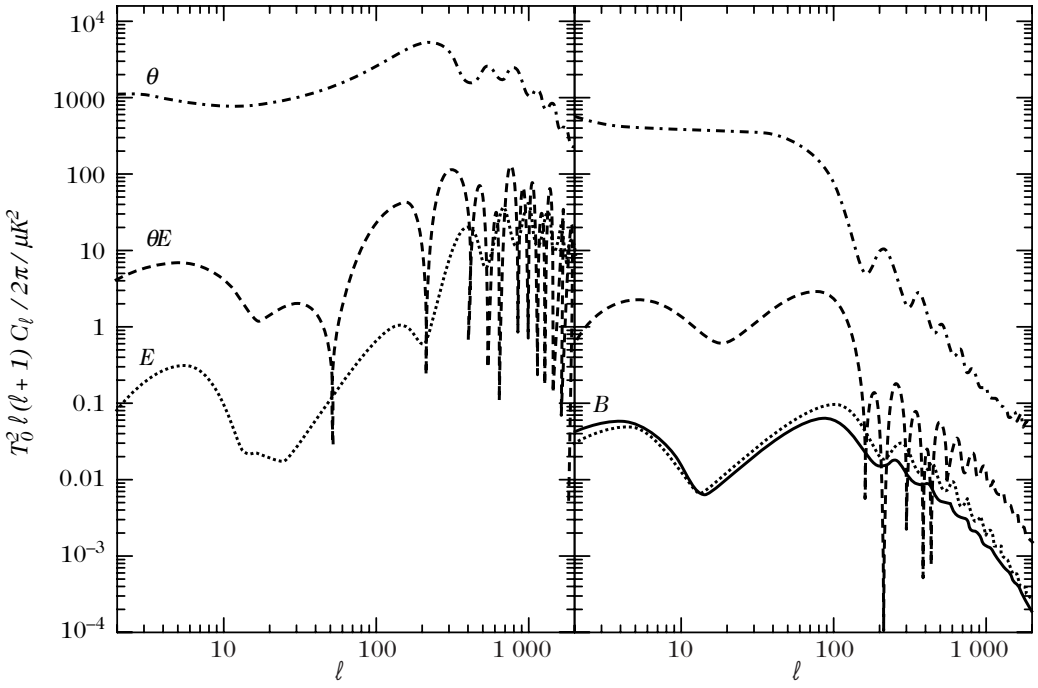


Figure 6.15: The four primary angular power spectra Θ–Θ, Θ–E, E–E, and B–B generated by scalar modes (left panel) and tensor modes (right panel) in an inflationary scenario. Note that the B-mode power spectrum B–B is sourced exclusively by primordial tensor perturbations, while the cross-correlation E–B vanishes identically due to parity conservation.

Because the wavenumber splitting $k_+ - k_- = g_{\phi\gamma}\dot{\phi}$ is constant, this rotation is strictly achromatic (frequency-independent), distinguishing it fundamentally from chromatic astrophysical Faraday rotation ($\beta_{\rm Faraday} \propto \lambda^2$).

At each line of sight, this rotation transforms the linear polarization Stokes parameters as $(Q \pm \mathrm{i}U)_{\rm obs} = \mathrm{e}^{\pm 2\mathrm{i}\beta}(Q \pm \mathrm{i}U)_{\rm intrinsic}$, rotating the harmonic multipoles according to

$$E_{\ell m}^{(\rm obs)} = E_{\ell m}\cos 2\beta - B_{\ell m}\sin 2\beta, \qquad B_{\ell m}^{(\rm obs)} = E_{\ell m}\sin 2\beta + B_{\ell m}\cos 2\beta. \tag{6.3.158}$$

Assuming parity-symmetric primordial perturbations ($C_\ell^{EB} = C_\ell^{TB} = 0$), the observed angular power spectra become

$$C_\ell^{TE,\rm obs} = C_\ell^{TE}\cos 2\beta, \tag{6.3.159}$$

$$C_\ell^{TB,\rm obs} = C_\ell^{TE}\sin 2\beta, \tag{6.3.160}$$

$$C_\ell^{EE,\rm obs} = C_\ell^{EE}\cos^2 2\beta + C_\ell^{BB}\sin^2 2\beta, \tag{6.3.161}$$

$$C_\ell^{BB,\rm obs} = C_\ell^{BB}\cos^2 2\beta + C_\ell^{EE}\sin^2 2\beta, \tag{6.3.162}$$

$$C_\ell^{EB,\rm obs} = \frac{1}{2}\sin 4\beta\left(C_\ell^{EE} - C_\ell^{BB}\right). \tag{6.3.163}$$

Because primordial E-mode power dominates over B-mode power ($C_\ell^{EE} \gg C_\ell^{BB}$), even a sub-degree cosmic rotation angle $\beta \sim 0.3^\circ$ produces detectable non-vanishing signatures $C_\ell^{EB,\rm obs} \approx 2\beta C_\ell^{EE}$ and $C_\ell^{TB,\rm obs} \approx 2\beta C_\ell^{TE}$ tracing the acoustic peaks. Recent analyses of Planck and WMAP polarization data utilizing Galactic dust to break instrumental calibration degeneracies have reported tentative hints of $\beta \approx 0.30^\circ \pm 0.11^\circ$, providing an active observational window into cosmic parity violation.

Polarization propagation, tensor birefringence, and the Lorentz oscillator

In vacuum, light obeys the wave equation $\Box A_\mu = 0$, where spatial symmetry ensures that both transverse polarization states travel along identical null geodesics at the speed of light c. Because empty space treats every transverse direction equally, the orientation of a polarization stick emitted at the surface of last scattering remains frozen in orientation throughout its journey to our detectors.

The situation changes fundamentally when photons propagate through a background medium, an astrophysical plasma, or a cosmological field condensate. The interaction with background charges or fields dresses the electromagnetic wave, breaking the symmetry between different polarization states. The medium responds to the wave through a tensor of response coefficients. The modes that travel through the medium without distortion are simply the eigenvectors of this tensor, and their propagation speeds are given by the corresponding eigenvalues. If the eigenvalues differ, the medium is birefringent: an arbitrary initial polarization decomposes into these eigenmodes, which drift out of phase as they travel, altering the net polarization pattern that emerges.

To trace this behavior from first principles, consider a charge q of mass m displaced by $\boldsymbol{r}$ from its rest position in the presence of an anisotropic restoring force $-m\boldsymbol{\Omega}_0^2\boldsymbol{r}$, frictional damping $-m\gamma\dot{\boldsymbol{r}}$, the driving electric field $\boldsymbol{E}(t)$ of the wave, and a steady background magnetic field $\boldsymbol{B}_0$:

$$m\ddot{\boldsymbol{r}} = -m\boldsymbol{\Omega}_0^2\boldsymbol{r} - m\gamma\dot{\boldsymbol{r}} + q\boldsymbol{E} + q\,\dot{\boldsymbol{r}} \times \boldsymbol{B}_0. \tag{1}$$

For a monochromatic wave $\boldsymbol{E}\mathrm{e}^{-\mathrm{i}\omega t}$, the charge exhibits steady-state harmonic response $\boldsymbol{r}\mathrm{e}^{-\mathrm{i}\omega t}$, with velocities $\dot{\boldsymbol{r}} = -\mathrm{i}\omega\boldsymbol{r}$ and accelerations $\ddot{\boldsymbol{r}} = -\omega^2\boldsymbol{r}$. In the absence of $\boldsymbol{B}_0$:

$$\begin{gathered} -m\omega^2\boldsymbol{r} \overset{(1)}{=} -m\boldsymbol{\Omega}_0^2\boldsymbol{r} + \mathrm{i}m\gamma\omega\,\boldsymbol{r} + q\boldsymbol{E}, \\ \left[\boldsymbol{\Omega}_0^2 - \omega^2\mathbf{1} - \mathrm{i}\omega\gamma\mathbf{1}\right]\boldsymbol{r} \overset{(2)}{=} \frac{q}{m}\boldsymbol{E}, \qquad \boldsymbol{r} \overset{(3)}{=} \frac{q}{m}\left[\boldsymbol{\Omega}_0^2 - \omega^2\mathbf{1} - \mathrm{i}\omega\gamma\mathbf{1}\right]^{-1}\boldsymbol{E}. \end{gathered} \tag{2}$$

(1) Insert the harmonic forms into Newton's law with $\boldsymbol{B}_0 = 0$; the common factor $\mathrm{e}^{-\mathrm{i}\omega t}$ cancels, and $-m\gamma(-\mathrm{i}\omega)\boldsymbol{r} = \mathrm{i}m\gamma\omega\boldsymbol{r}$.

(2) Collect every displacement term with $\boldsymbol{r}$ to the left-hand side and divide through by m.

(3) Multiply by the inverse matrix, which exists whenever the driving frequency ω is off-resonance.

For a medium containing N such charges per unit volume, the macroscopic dipole moment density is $\boldsymbol{P} = Nq\boldsymbol{r} = \varepsilon_0\boldsymbol{\chi}\boldsymbol{E}$. This defines the susceptibility matrix $\boldsymbol{\chi}(\omega)$ and relative permittivity tensor $\boldsymbol{\varepsilon}_r(\omega)$:

$$\boldsymbol{\chi}(\omega) = \omega_p^2\left[\boldsymbol{\Omega}_0^2 - \omega^2\mathbf{1} - \mathrm{i}\omega\gamma\mathbf{1}\right]^{-1}, \qquad \boldsymbol{\varepsilon}_r(\omega) = \mathbf{1} + \boldsymbol{\chi}(\omega), \qquad \omega_p^2 \equiv \frac{Nq^2}{\varepsilon_0 m}. \tag{3}$$

The plasma frequency ω_p measures the density and mobility of the medium's charges.

The propagation of electromagnetic waves follows from Maxwell's curl equations, $\nabla \times \boldsymbol{E} = -\partial_t\boldsymbol{B}$ and $\nabla \times \boldsymbol{B} = \mu_0\varepsilon_0\boldsymbol{\varepsilon}_r\partial_t\boldsymbol{E}$. Testing a plane wave travelling along $\hat{\boldsymbol{z}}$ with spatial form $\mathrm{e}^{\mathrm{i}(kz-\omega t)}$:

$$\begin{gathered} \boldsymbol{B} \overset{(1)}{=} \frac{k}{\omega}\,\hat{\boldsymbol{z}} \times \boldsymbol{E}, \\ \mathrm{i}k\,\frac{k}{\omega}\,\hat{\boldsymbol{z}} \times (\hat{\boldsymbol{z}} \times \boldsymbol{E}) \overset{(2)}{=} -\frac{\mathrm{i}\omega}{c^2}\boldsymbol{\varepsilon}_r\boldsymbol{E}, \\ \frac{c^2k^2}{\omega^2}\boldsymbol{E}_\perp \overset{(3)}{=} \boldsymbol{\varepsilon}_r\boldsymbol{E}, \qquad \left[\boldsymbol{\varepsilon}_{r\perp} - n^2\mathbf{1}\right]\boldsymbol{E}_\perp \overset{(4)}{=} 0, \qquad n \equiv \frac{ck}{\omega}. \end{gathered} \tag{4}$$

(1) The first Maxwell curl equation: $\mathrm{i}k\,\hat{\boldsymbol{z}} \times \boldsymbol{E} = \mathrm{i}\omega\boldsymbol{B}$.

(2) The second curl equation, substituting $\boldsymbol{B}$ from the first line and using $\mu_0\varepsilon_0 = 1/c^2$.

(3) The vector triple-product identity $\hat{\boldsymbol{z}} \times (\hat{\boldsymbol{z}} \times \boldsymbol{E}) = \hat{\boldsymbol{z}}(\hat{\boldsymbol{z}} \cdot \boldsymbol{E}) - \boldsymbol{E} = -\boldsymbol{E}_\perp$, the part of $\boldsymbol{E}$ transverse to $\hat{\boldsymbol{z}}$; then multiply by $\mathrm{i}c^2/\omega$.

(4) Because $\boldsymbol{\varepsilon}_r$ does not couple the longitudinal axis z to transverse coordinates, $\varepsilon_{r,zz}E_z = 0 \implies E_z = 0$, leaving a 2×2 eigenvalue problem for the transverse block $\boldsymbol{\varepsilon}_{r\perp}$.

This polarimetric eigenvalue problem shows that *the polarizations that travel through the medium unchanged are the eigenvectors of $\boldsymbol{\varepsilon}_{r\perp}$, and their refractive indices squared n^2 are its eigenvalues.*

When the restoring springs are anisotropic with different stiffnesses along x and y, $\boldsymbol{\Omega}_0^2 = \mathrm{diag}(\omega_{0x}^2, \omega_{0y}^2, \omega_{0z}^2)$ with $\boldsymbol{B}_0 = 0$ and $\gamma = 0$, the transverse dielectric tensor is purely diagonal:

$$\boldsymbol{\varepsilon}_{r\perp} = \begin{pmatrix} 1 + \dfrac{\omega_p^2}{\omega_{0x}^2 - \omega^2} & 0 \\ 0 & 1 + \dfrac{\omega_p^2}{\omega_{0y}^2 - \omega^2} \end{pmatrix}. \tag{5}$$

Its eigenvectors are the linear states along $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$, possessing distinct refractive indices $n_x \neq n_y$ when $\omega_{0x} \neq \omega_{0y}$ (linear birefringence). A wave linearly polarized at 45°, $\boldsymbol{E} = \frac{E_0}{\sqrt{2}}(\hat{\boldsymbol{x}} + \hat{\boldsymbol{y}})$, splits into these two eigen-components. After traversing path length L:

$$\begin{aligned} \boldsymbol{E}(L) &\overset{(1)}{=} \frac{E_0}{\sqrt{2}}(\hat{\boldsymbol{x}}\mathrm{e}^{\mathrm{i}n_x\omega L/c} + \hat{\boldsymbol{y}}\mathrm{e}^{\mathrm{i}n_y\omega L/c}) \overset{(2)}{=} \mathrm{e}^{\mathrm{i}n_x\omega L/c}\frac{E_0}{\sqrt{2}}(\hat{\boldsymbol{x}} + \hat{\boldsymbol{y}}\mathrm{e}^{\mathrm{i}\delta}), \qquad \delta \equiv (n_y - n_x)\frac{\omega L}{c}, \\ Q &\overset{(3)}{=} \tfrac{1}{2}E_0^2 - \tfrac{1}{2}E_0^2 = 0, \qquad U \overset{(4)}{=} 2 \cdot \tfrac{1}{2}E_0^2\cos(-\delta) = E_0^2\cos\delta. \end{aligned} \tag{6}$$

(1) Each eigenvector accumulates its own optical phase $kL = n\omega L/c$ while preserving its direction.

(2) Factor out the overall phase of the x component, which is common to the wave and unobservable in Stokes parameters.

(3) Evaluate Stokes parameter Q with $a_1 = a_2 = E_0/\sqrt{2}$.

(4) Evaluate Stokes parameter U with phase difference $\theta_1 - \theta_2 = -\delta$.

The polarization stick is never turned. Rather, the accumulated phase lag δ deforms the linear state into an ellipse whose principal axes remain locked to the crystal axes, generating circular polarization $V = E_0^2 \sin\delta$.

To produce a genuine spatial rotation of the polarization plane, the medium's eigenvectors must be circular helicities rather than straight lines. This occurs when an isotropic electron gas ($\boldsymbol{\Omega}_0^2 = \omega_0^2\mathbf{1}$) is immersed in a longitudinal magnetic field $\boldsymbol{B}_0 = B_0\hat{\boldsymbol{z}}$. The magnetic Lorentz force on velocity $\dot{\boldsymbol{r}} = -\mathrm{i}\omega\boldsymbol{r}$ is $q\dot{\boldsymbol{r}} \times \boldsymbol{B}_0 = -\mathrm{i}\omega qB_0(y, -x, 0)$, coupling transverse coordinates antisymmetrically:

$$M\begin{pmatrix} x \\ y \end{pmatrix} = \frac{q}{m}\begin{pmatrix} E_x \\ E_y \end{pmatrix}, \qquad M = \begin{pmatrix} a & \mathrm{i}\omega\omega_c \\ -\mathrm{i}\omega\omega_c & a \end{pmatrix}, \qquad a \equiv \omega_0^2 - \omega^2 - \mathrm{i}\omega\gamma, \qquad \omega_c \equiv \frac{qB_0}{m}. \tag{7}$$

The cyclotron frequency ω_c introduces antisymmetric off-diagonal entries that break spatial reflection symmetry. Evaluating M on the circular helicity basis vectors $\boldsymbol{e}_\pm = (\hat{\boldsymbol{x}} \pm \mathrm{i}\hat{\boldsymbol{y}})/\sqrt{2}$:

$$\begin{aligned} M\boldsymbol{e}_\pm &\overset{(1)}{=} \frac{1}{\sqrt{2}}\begin{pmatrix} a + \mathrm{i}\omega\omega_c(\pm\mathrm{i}) \\ -\mathrm{i}\omega\omega_c + a(\pm\mathrm{i}) \end{pmatrix} \overset{(2)}{=} \frac{1}{\sqrt{2}}\begin{pmatrix} a \mp \omega\omega_c \\ \pm\mathrm{i}(a \mp \omega\omega_c) \end{pmatrix} \overset{(3)}{=} (a \mp \omega\omega_c)\boldsymbol{e}_\pm, \\ n_\pm^2 &\overset{(4)}{=} 1 + \frac{\omega_p^2}{\omega_0^2 - \omega^2 \mp \omega\omega_c} \qquad (\gamma = 0). \end{aligned} \tag{8}$$

(1) Multiply the matrix M into the column vector $(1, \pm\mathrm{i})/\sqrt{2}$.

(2) In the first component $\mathrm{i}\cdot(\pm\mathrm{i}) = \mp 1$; in the second component $-\mathrm{i}\omega\omega_c = \pm\mathrm{i}(\mp\omega\omega_c)$ because $(\pm\mathrm{i})(\mp 1) = -\mathrm{i}$.

(3) Both components share the common factor $(a \mp \omega\omega_c)$, recovering the original helicity vector $\boldsymbol{e}_\pm$.

(4) Because $\boldsymbol{\chi} = \omega_p^2 M^{-1}$, the inverse matrix shares the same eigenvectors with inverted eigenvalues, giving $\boldsymbol{\varepsilon}_{r\perp} = \mathbf{1} + \boldsymbol{\chi}$ the eigenvalues $1 + \omega_p^2/(a \mp \omega\omega_c)$.

The two propagating eigenmodes are the circular helicities, with split refractive indices $n_+ \neq n_-$. Because an arbitrary linear polarization stick at angle ψ is an equal mixture of both helicities, $\boldsymbol{\epsilon}(\psi) = 2^{-1/2}(\mathrm{e}^{-\mathrm{i}\psi}\boldsymbol{e}_+ + \mathrm{e}^{\mathrm{i}\psi}\boldsymbol{e}_-)$, after traveling distance L with optical phases $\phi_\pm = n_\pm \omega L/c$:

$$\begin{aligned}\frac{\mathrm{e}^{-\mathrm{i}\psi}\mathrm{e}^{\mathrm{i}\phi_+}\boldsymbol{e}_+ + \mathrm{e}^{\mathrm{i}\psi}\mathrm{e}^{\mathrm{i}\phi_-}\boldsymbol{e}_-}{\sqrt{2}} &\overset{(1)}{=} \mathrm{e}^{\mathrm{i}\bar{\phi}}\frac{\mathrm{e}^{-\mathrm{i}(\psi-\Delta/2)}\boldsymbol{e}_+ + \mathrm{e}^{\mathrm{i}(\psi-\Delta/2)}\boldsymbol{e}_-}{\sqrt{2}} \\ &\overset{(2)}{=} \mathrm{e}^{\mathrm{i}\bar{\phi}}\left[\cos(\psi-\Delta/2)\hat{\boldsymbol{x}} + \sin(\psi-\Delta/2)\hat{\boldsymbol{y}}\right], \end{aligned} \tag{9}$$

(1) Define the mean phase $\bar{\phi} \equiv (\phi_+ + \phi_-)/2$ and phase difference $\Delta \equiv \phi_+ - \phi_-$, writing $\phi_\pm = \bar{\phi} \pm \Delta/2$.

(2) Apply the helicity identity at the shifted orientation angle $\psi - \Delta/2$.

The transmitted wave remains purely linearly polarized, but its polarization plane rotates uniformly by $\beta = -\Delta/2$ for every initial angle ψ.

For cosmological photons crossing ionized plasma ($\omega_0 = 0$, $\omega \gg \omega_p, \omega_c$):

$$\begin{aligned} n_\pm &\overset{(1)}{=} \left[1 - \frac{\omega_p^2}{\omega^2}\left(1 \pm \frac{\omega_c}{\omega}\right)^{-1}\right]^{1/2} \overset{(2)}{=} 1 - \frac{\omega_p^2}{2\omega^2}\left(1 \mp \frac{\omega_c}{\omega}\right), \\ \Delta &\overset{(3)}{=} (n_+ - n_-)\frac{\omega L}{c} \overset{(4)}{=} \frac{\omega_p^2\omega_c}{\omega^3}\frac{\omega L}{c} = \frac{\omega_p^2\omega_c L}{c\omega^2} \overset{(5)}{=} \frac{\omega_p^2\omega_c L}{4\pi^2 c^3}\lambda^2. \end{aligned} \tag{10}$$

(1) Factor ω^2 out of $\omega(\omega \pm \omega_c)$ and take the square root of $n_\pm^2$.

(2) Expand to first order: $(1 \pm x)^{-1} \approx 1 \mp x$ and $(1 - y)^{1/2} \approx 1 - y/2$, neglecting higher-order terms.

(3) The definition of phase difference Δ with $\phi_\pm = n_\pm \omega L/c$.

(4) The isotropic terms $1 - \omega_p^2/(2\omega^2)$ cancel in the difference $n_+ - n_-$, leaving $\omega_p^2\omega_c/\omega^3$.

(5) Substitute $\omega = 2\pi c/\lambda$, so that $1/\omega^2 = \lambda^2/(4\pi^2 c^2)$.

The index difference falls as ω^{-3} while path length in wavelengths grows as ω, producing the classical λ^2 scaling of Faraday rotation: $|\beta| = \mathrm{RM}\,\lambda^2$, where $\mathrm{RM} = \frac{e^3}{8\pi^2\varepsilon_0 m_e^2 c^3}\int n_e B_\parallel \,\mathrm{d}\ell$. This microscopic derivation underscores the profound physical distinction between astrophysical plasma rotation and cosmic birefringence: in Faraday rotation, the index splitting $n_+ - n_- \propto \omega^{-3}$ yields a chromatic λ^2 rotation; in cosmic birefringence induced by a rolling pseudoscalar axion field $\phi(t)$ coupled via $g_{\phi\gamma}\phi\boldsymbol{E}\cdot\boldsymbol{B}$, the circular helicity splitting $k_+ - k_- = g_{\phi\gamma}\dot{\phi}$ is constant across all wavelengths. The resulting rotation $\beta = \frac{1}{2}g_{\phi\gamma}\Delta\phi$ is strictly achromatic, turning every CMB polarization stick by the same angle regardless of observing frequency.

In principle, direct numerical integration of the coupled Boltzmann hierarchy can be carried out to obtain these power spectra. Historically, this full-hierarchy approach was the standard methodology, following the evolution of the photon distribution function through the entire cosmic expansion history. However, this brute-force method is computationally demanding: it

requires integrating an extensive set of coupled differential equations (typically on the order of $5\ell_{\rm max}$ equations) for hundreds or thousands of wavenumbers k spanning $k\eta_0 \in [\ell_{\rm min}/10, 2\ell_{\rm max}]$. Moreover, careful boundary truncation algorithms are necessary to prevent artificial reflection of power from high to low multipoles.

6.3.4 Numerical integration

Radial modes

A significant computational breakthrough is achieved by separating the radial and angular components of the basis functions ${}_sG_{\ell m}$. Using the Rayleigh expansion for plane waves into spherical harmonics and spherical Bessel functions [Eq. (B.20)],

$$\exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}) = \sum_\ell (-\mathrm{i})^\ell \sqrt{4\pi(2\ell+1)}\, j_\ell(kr) Y_{\ell 0}(\boldsymbol{n}),$$

where $\boldsymbol{e}_3 = \boldsymbol{k}/k$ and $\boldsymbol{x} = -r\boldsymbol{n}$ (with $\boldsymbol{n}$ pointing along the photon propagation direction toward the observer), and applying the recurrence relations for spherical Bessel functions [Eq. (B.46)], we can recast the basis functions as

$$G_{\ell' m} = \sum_\ell (-\mathrm{i})^\ell \sqrt{4\pi(2\ell+1)}\, j_\ell^{(\ell' m)}(kr) Y_{\ell m}(\boldsymbol{n}), \tag{6.3.164}$$

$${}_{\pm2}G_{2m} = \sum_\ell (-\mathrm{i})^\ell \sqrt{4\pi(2\ell+1)}\left[\epsilon_\ell^{(m)}(kr) \pm \mathrm{i}\beta_\ell^{(m)}(kr)\right] {}_{\pm2}Y_{\ell m}(\boldsymbol{n}). \tag{6.3.165}$$

The radial transfer functions $j_\ell^{(\ell' m)}(x)$, $\epsilon_\ell^{(m)}(x)$, and $\beta_\ell^{(m)}(x)$ (with argument $x \equiv kr$) are given explicitly by

$$\begin{aligned}
j_\ell^{(00)} &= j_\ell, \\
j_\ell^{(10)} &= j_\ell', \\
j_\ell^{(20)} &= \frac{1}{2}\left(3j_\ell'' + j_\ell\right), \\
j_\ell^{(11)} &= \sqrt{\frac{\ell(\ell+1)}{2}}\frac{j_\ell}{x}, \\
j_\ell^{(21)} &= \sqrt{\frac{3\ell(\ell+1)}{2}}\left(\frac{j_\ell}{x}\right)', \\
j_\ell^{(22)} &= \sqrt{\frac{3}{8}\frac{(\ell+2)!}{(\ell-2)!}}\frac{j_\ell}{x^2},
\end{aligned} \tag{6.3.166}$$

alongside the polarization radial functions:

$$\epsilon_\ell^{(0)} = j_\ell^{(22)}, \tag{6.3.167}$$

$$\epsilon_\ell^{(1)} = \sqrt{\frac{(\ell-1)(\ell+2)}{2}}\left[\frac{j_\ell}{x^2} + \frac{j_\ell'}{x}\right], \tag{6.3.168}$$

$$\epsilon_\ell^{(2)} = \frac{1}{4}\left[-j_\ell + j_\ell'' + 2\frac{j_\ell}{x^2} + 4\frac{j_\ell'}{x}\right], \tag{6.3.169}$$

$$\beta_\ell^{(0)} = 0, \tag{6.3.170}$$

$$\beta_\ell^{(1)} = \sqrt{\frac{(\ell-1)(\ell+2)}{2}}\frac{j_\ell}{x}, \tag{6.3.171}$$

$$\beta_\ell^{(2)} = \frac{1}{2}\left[j_\ell' + 2\frac{j_\ell}{x}\right]. \tag{6.3.172}$$

Furthermore, these functions satisfy the reflection relations $\epsilon_\ell^{(-m)} = \epsilon_\ell^{(m)}$ and $\beta_\ell^{(-m)} = -\beta_\ell^{(m)}$.

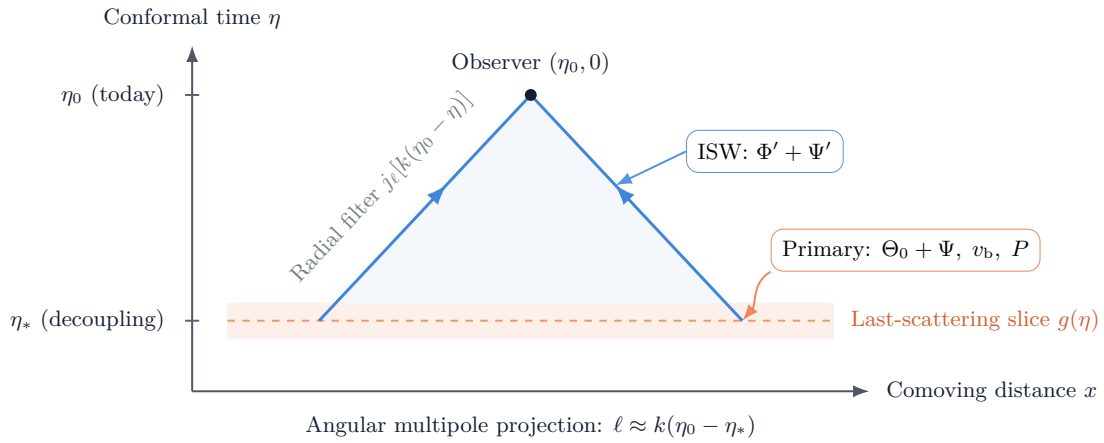


Figure 6.16: Conformal spacetime geometry of the line-of-sight integration method. Photons propagate along the past light cone $\chi = \eta_0 - \eta$ to the observer at $(\eta_0, 0)$. Primary acoustic sources—effective temperature monopole $(\Theta_0 + \Psi)$, Doppler dipole v_b, and polarization quadrupole P—originate at the last-scattering surface η_*, weighted by the visibility window $g(\eta)$. Along the intervening line-of-sight corridor $(\eta_* < \eta < \eta_0)$, time-varying gravitational potentials accumulate as the Integrated Sachs–Wolfe (ISW) contribution $\Phi' + \Psi'$. The radial spherical Bessel filter $j_\ell[k(\eta_0 - \eta)]$ geometrically maps spatial wavenumber k into observed angular multipoles $\ell \approx k(\eta_0 - \eta_*)$.

Integral solution

The Boltzmann hierarchy of Eqs. (6.3.146)–(6.3.148) admits an exact formal solution in integral form. Projecting onto the radial mode functions of Eqs. (6.3.164) and (6.3.165), the present-day temperature multipoles evaluate to

$$\Theta_\ell^{(m)}(\eta_0, k) = (2\ell + 1) \int_0^{\eta_0} \mathrm{d}\eta \, \mathrm{e}^{-\tau} \sum_{\ell'} S_{\ell'}^{(m)}(\eta) j_\ell^{(\ell' m)}[k(\eta_0 - \eta)]. \tag{6.3.173}$$

Similarly, the formal solution for the polarization multipoles is given by

$$\left[E_\ell^{(m)} \pm \mathrm{i} B_\ell^{(m)}\right](\eta_0, k) = -(2\ell + 1)\sqrt{6} \int_0^{\eta_0} \mathrm{d}\eta \, \tau' \mathrm{e}^{-\tau} P^{(m)}(\eta) \left[\epsilon_\ell^{(m)} \pm \mathrm{i}\beta_\ell^{(m)}\right][k(\eta_0 - \eta)]. \tag{6.3.174}$$

This line-of-sight integration method, developed by Seljak and Zaldarriaga (Refs. [32, 33]), represents a profound generalization of the Sachs–Wolfe formula: it expresses the observed multipoles as line-of-sight integrals over the source functions evaluated throughout cosmic history. The conformal spacetime geometry and physical source decomposition underlying this integral projection are depicted in Fig. 6.16.

Line of Sight integral

We now derive the line-of-sight solution directly from the Boltzmann hierarchy. The useful observation is that, after summing the harmonic expansions, the free-streaming terms can be written simply as $-ik\mu$, where

$$\mu \equiv \hat{\mathbf{k}} \cdot \mathbf{n}. \tag{1}$$

For a fixed Fourier mode $\mathbf{k}$, the temperature and polarization perturbations are expanded in the total-angular-momentum basis as

$$\Theta(\eta, \mathbf{x}, \mathbf{n}) = \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^{3/2}} \sum_{m,\ell} \Theta_\ell^{(m)}(\eta, \mathbf{k}) \, {}_0G_{\ell m}(\mathbf{x}, \mathbf{n}), \tag{2}$$

$$E(\eta,\mathbf{x},\mathbf{n})+iB(\eta,\mathbf{x},\mathbf{n})=\int\frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^{3/2}}\sum_{m,\ell}\left[E_\ell^{(m)}(\eta,\mathbf{k})+iB_\ell^{(m)}(\eta,\mathbf{k})\right]{}_2G_{\ell m}(\mathbf{x},\mathbf{n}),\tag{3}$$

$$E(\eta,\mathbf{x},\mathbf{n})-iB(\eta,\mathbf{x},\mathbf{n})=\int\frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^{3/2}}\sum_{m,\ell}\left[E_\ell^{(m)}(\eta,\mathbf{k})-iB_\ell^{(m)}(\eta,\mathbf{k})\right]{}_{-2}G_{\ell m}(\mathbf{x},\mathbf{n}).\tag{4}$$

The total-angular-momentum functions are defined by

$${}_sG_{\ell m}(\mathbf{x},\mathbf{n})\equiv(-i)^\ell\sqrt{\frac{4\pi}{2\ell+1}}\,{}_sY_{\ell m}(\mathbf{n})\,e^{i\mathbf{k}\cdot\mathbf{x}}.\tag{5}$$

Here and below we suppress the explicit $\mathbf{k}$ dependence whenever there is no ambiguity.

Temperature.

The temperature hierarchy is

$$\Theta_\ell^{(m)\prime}=k\left[\frac{{}_0\kappa_\ell^m}{2\ell-1}\Theta_{\ell-1}^{(m)}-\frac{{}_0\kappa_{\ell+1}^m}{2\ell+3}\Theta_{\ell+1}^{(m)}\right]-\tau'\Theta_\ell^{(m)}+S_\ell^{(m)},\tag{6}$$

where

$${}_s\kappa_\ell^m=\sqrt{\frac{(\ell^2-m^2)(\ell^2-s^2)}{\ell^2}}.\tag{7}$$

Using the recurrence relation for the total-angular-momentum functions, the free-streaming terms in Eq. (6) can be recombined into multiplication by μ. Therefore the summed temperature perturbation satisfies

$$\left(\partial_\eta-ik\mu+\tau'\right)\Theta^{(m)}(\eta,\mathbf{n},\mathbf{k})=S^{(m)}(\eta,\mathbf{n},\mathbf{k}).\tag{8}$$

The source appearing here is obtained by summing the source multipoles in the same basis:

$$S^{(m)}(\eta,\mathbf{n},\mathbf{k})=\sum_{\ell'=0}^{2}S_{\ell'}^{(m)}(\eta,\mathbf{k})\,{}_0G_{\ell' m}(\mathbf{0},\mathbf{n}).\tag{9}$$

The sum terminates at $\ell'=2$ because the Thomson-scattering source contains only the monopole, dipole and quadrupole pieces.

Explicit source terms for the three helicity sectors We now make the source term explicit in the scalar, vector and tensor sectors. This will allow the final line-of-sight solution to be written directly in terms of the physical source functions.

Scalar sector: $m=0$**.**

For scalar perturbations, the non-vanishing source multipoles are

$$S_0^{(0)}=\tau'\Theta_0^{(0)},\tag{10}$$

$$S_1^{(0)}=ik\left(\Psi+\Phi+\tau' v_b\right),\tag{11}$$

$$S_2^{(0)}=\tau' P^{(0)}=\frac{\tau'}{10}\left[\Theta_2^{(0)}-\sqrt{6}\,E_2^{(0)}\right].\tag{12}$$

Hence

$$\begin{aligned}S^{(0)}(\eta,\mathbf{n},\mathbf{k})=&\,S_0^{(0)}(\eta,\mathbf{k})\,{}_0G_{00}(\mathbf{0},\mathbf{n})+S_1^{(0)}(\eta,\mathbf{k})\,{}_0G_{10}(\mathbf{0},\mathbf{n})\\&+S_2^{(0)}(\eta,\mathbf{k})\,{}_0G_{20}(\mathbf{0},\mathbf{n}).\end{aligned}\tag{13}$$

Vector sector: $m=\pm1$**.**

For vector perturbations, the source starts at the dipole:

$$S_1^{(\pm1)} = \tau' v_{b,\pm1}, \tag{14}$$

$$\begin{aligned} S_2^{(\pm1)} &= -\frac{ik}{\sqrt{3}}\sigma_{\pm1} + \tau' P^{(\pm1)} \\ &= -\frac{ik}{\sqrt{3}}\sigma_{\pm1} + \frac{\tau'}{10}\left[\Theta_2^{(\pm1)} - \sqrt{6}\, E_2^{(\pm1)}\right]. \end{aligned} \tag{15}$$

Thus

$$\begin{aligned} S^{(\pm1)}(\eta, \mathbf{n}, \mathbf{k}) = S_1^{(\pm1)}(\eta, \mathbf{k})\, {}_0G_{1,\pm1}(\mathbf{0}, \mathbf{n}) \\ + S_2^{(\pm1)}(\eta, \mathbf{k})\, {}_0G_{2,\pm1}(\mathbf{0}, \mathbf{n}). \end{aligned} \tag{16}$$

Tensor sector: $m = \pm 2$**.**
For tensor perturbations, only the quadrupole source is present:

$$S_2^{(\pm2)} = \frac{1}{\sqrt{3}}\dot{h}_{\pm2} + \tau' P^{(\pm2)} = \frac{1}{\sqrt{3}}\dot{h}_{\pm2} + \frac{\tau'}{10}\left[\Theta_2^{(\pm2)} - \sqrt{6}\, E_2^{(\pm2)}\right]. \tag{17}$$

Hence

$$S^{(\pm2)}(\eta, \mathbf{n}, \mathbf{k}) = S_2^{(\pm2)}(\eta, \mathbf{k})\, {}_0G_{2,\pm2}(\mathbf{0}, \mathbf{n}). \tag{18}$$

Formal solution for the temperature Define the optical depth between η and η_0 by

$$\tau(\eta_0, \eta) \equiv \int_\eta^{\eta_0} d\tilde{\eta}\, \tau'(\tilde{\eta}). \tag{19}$$

The integrating factor for Eq. (8) is

$$\exp\left[-ik\mu\eta + \int^\eta d\tilde{\eta}\, \tau'(\tilde{\eta})\right]. \tag{20}$$

Multiplying Eq. (8) by this factor and integrating from $\eta_{\rm in}$ to η gives

$$\begin{aligned} \Theta^{(m)}(\eta, \mathbf{n}, \mathbf{k}) = \Theta^{(m)}(\eta_{\rm in}, \mathbf{n}, \mathbf{k}) e^{-ik\mu(\eta-\eta_{\rm in}) - \tau(\eta,\eta_{\rm in})} \\ + \int_{\eta_{\rm in}}^{\eta} d\tilde{\eta}\, e^{-ik\mu(\eta-\tilde{\eta}) - \tau(\eta,\tilde{\eta})} S^{(m)}(\tilde{\eta}, \mathbf{n}, \mathbf{k}). \end{aligned} \tag{21}$$

We are interested in the perturbation observed at η_0. Choosing $\eta_{\rm in}$ sufficiently early, the optical depth to the present is very large,

$$e^{-\tau(\eta_0,\eta_{\rm in})} \ll 1, \tag{22}$$

and hence the initial-value term can be neglected. Thus

$$\Theta^{(m)}(\eta_0, \mathbf{n}, \mathbf{k}) = \int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-ik\mu(\eta_0-\eta)} e^{-\tau(\eta_0,\eta)} S^{(m)}(\eta, \mathbf{n}, \mathbf{k}). \tag{23}$$

The combination

$$g(\eta) \equiv \tau'(\eta) e^{-\tau(\eta_0,\eta)} \tag{24}$$

is the visibility function. It is strongly peaked around the epoch of last scattering, while at late times the small scattering rate leaves the photons to evolve predominantly by free streaming.

Expansion of the free-streaming term Let

$$\chi \equiv \eta_0 - \eta, \qquad x \equiv k\chi. \tag{25}$$

Along the line of sight, $\mathbf{x} = -\mathbf{n}\chi$, and therefore

$${}_sG_{\ell m}(-\mathbf{n}\chi, \mathbf{n}) = (-i)^\ell \sqrt{\frac{4\pi}{2\ell+1}}\, {}_sY_{\ell m}(\mathbf{n}) e^{-ik\mu\chi}. \tag{26}$$

For the temperature sector, the total-angular-momentum expansion is

$${}_0G_{\ell m}(-\mathbf{n}\chi, \mathbf{n}) = \sum_L (-i)^L \sqrt{4\pi(2L+1)}\, j_L^{(\ell m)}(x) Y_{Lm}(\mathbf{n}). \tag{27}$$

Using

$${}_0G_{Lm}(\mathbf{0}, \mathbf{n}) = (-i)^L \sqrt{\frac{4\pi}{2L+1}} Y_{Lm}(\mathbf{n}), \tag{28}$$

the preceding equation can equivalently be written as

$${}_0G_{\ell m}(-\mathbf{n}\chi, \mathbf{n}) = \sum_L (2L+1) j_L^{(\ell m)}(x) {}_0G_{Lm}(\mathbf{0}, \mathbf{n}). \tag{29}$$

This relation shows explicitly where the factor $2L+1$ in the observed multipoles originates. Substituting Eq. (9) into Eq. (23), we obtain

$$\Theta^{(m)}(\eta_0, \mathbf{n}, \mathbf{k}) = \sum_{\ell'=0}^{2} \int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)} S_{\ell'}^{(m)}(\eta, \mathbf{k}) \times {}_0G_{\ell' m}(-\mathbf{n}\chi, \mathbf{n}). \tag{30}$$

Using Eq. (27),

$$\Theta^{(m)}(\eta_0, \mathbf{n}, \mathbf{k}) = \sum_L (-i)^L \sqrt{4\pi(2L+1)} Y_{Lm}(\mathbf{n}) \times \int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)} \sum_{\ell'=0}^{2} S_{\ell'}^{(m)}(\eta, \mathbf{k}) j_L^{(\ell' m)}(x). \tag{31}$$

At the observer, Eq. (2) gives

$$\Theta^{(m)}(\eta_0, \mathbf{n}, \mathbf{k}) = \sum_L (-i)^L \sqrt{\frac{4\pi}{2L+1}} \Theta_L^{(m)}(\eta_0, \mathbf{k}) Y_{Lm}(\mathbf{n}). \tag{32}$$

Using

$$\sqrt{4\pi(2L+1)} = (2L+1)\sqrt{\frac{4\pi}{2L+1}}, \tag{33}$$

comparison of the coefficients of Y_{Lm} gives

$$\Theta_L^{(m)}(\eta_0, k) = (2L+1) \int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)} \sum_{\ell'=0}^{2} S_{\ell'}^{(m)}(\eta, k) j_L^{(\ell' m)}[k(\eta_0 - \eta)]\,. \tag{34}$$

This is the desired numerical line-of-sight solution for the temperature hierarchy.

Explicit temperature line-of-sight solutions Using the source terms above, we can now write the temperature line-of-sight solution separately for the three helicity sectors.
Scalar sector: $m = 0$**.**
Substituting Eqs. (10)–(12) into Eq. (34), we obtain

$$\Theta_L^{(0)}(\eta_0, k) = (2L+1)\int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)}\Big[\tau'\Theta_0^{(0)} j_L^{(00)}(x) + ik\left(\Psi+\Phi+\tau' v_b\right) j_L^{(10)}(x) + \frac{\tau'}{10}\left(\Theta_2^{(0)} - \sqrt{6}\, E_2^{(0)}\right) j_L^{(20)}(x)\Big]. \quad (35)$$

Vector sector: $m = \pm 1$**.**
Similarly,

$$\Theta_L^{(\pm1)}(\eta_0, k) = (2L+1)\int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)}\Big[\tau' v_{b,\pm1} j_L^{(1,\pm1)}(x) - \frac{ik}{\sqrt{3}}\sigma_{\pm1} j_L^{(2,\pm1)}(x) + \frac{\tau'}{10}\left(\Theta_2^{(\pm1)} - \sqrt{6}\, E_2^{(\pm1)}\right) j_L^{(2,\pm1)}(x)\Big]. \quad (36)$$

Tensor sector: $m = \pm 2$**.**
Finally,

$$\Theta_L^{(\pm2)}(\eta_0, k) = (2L+1)\int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)}\Big[\frac{1}{\sqrt{3}}\dot{h}_{\pm2} j_L^{(2,\pm2)}(x) + \frac{\tau'}{10}\left(\Theta_2^{(\pm2)} - \sqrt{6}\, E_2^{(\pm2)}\right) j_L^{(2,\pm2)}(x)\Big]. \quad (37)$$

In all three expressions,

$$x = k(\eta_0 - \eta). \quad (38)$$

Polarization The E- and B-mode hierarchies can be combined into the helicity variables

$$\mathcal{P}_\pm^{(m)} \equiv E^{(m)} \pm iB^{(m)}. \quad (39)$$

Using the harmonic expansion in terms of ${}_{\pm2}G_{\ell m}$, the E- and B-mode hierarchies can likewise be written as the angular equation

$$\left(\partial_\eta - ik\mu + \tau'\right)\mathcal{P}_\pm^{(m)} = S_\pm^{(m)}. \quad (40)$$

The source term follows directly from the $\ell = 2$ term in the polarization hierarchy. With

$$P^{(m)} = \frac{1}{10}\left[\Theta_2^{(m)} - \sqrt{6}\, E_2^{(m)}\right], \quad (41)$$

we have

$$S_\pm^{(m)}(\eta, \mathbf{n}, \mathbf{k}) = -\sqrt{6}\,\tau'(\eta) P^{(m)}(\eta, \mathbf{k})\, {}_{\pm2}G_{2m}(\mathbf{0}, \mathbf{n}). \quad (42)$$

The formal solution is therefore

$$\mathcal{P}_\pm^{(m)}(\eta_0, \mathbf{n}, \mathbf{k}) = \int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-ik\mu(\eta_0-\eta)} e^{-\tau(\eta_0,\eta)} S_\pm^{(m)}(\eta, \mathbf{n}, \mathbf{k}). \quad (43)$$

Equivalently, since the $\ell = 2$ source is proportional to ${}_{\pm2}G_{2m}$,

$$\mathcal{P}_\pm^{(m)}(\eta_0, \mathbf{n}, \mathbf{k}) = -\sqrt{6}\int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)}\tau'(\eta) P^{(m)}(\eta, \mathbf{k})\, {}_{\pm2}G_{2m}(-\mathbf{n}\chi, \mathbf{n}). \quad (44)$$

Expansion of the polarization free-streaming term For the polarization sector, the total-angular-momentum expansion is

$$ {}_{\pm2}G_{\ell m}(-\mathbf{n}\chi,\mathbf{n}) = \sum_L (-i)^L \sqrt{4\pi(2L+1)}\left[\epsilon_L^{(\ell m)}(x) \pm i\beta_L^{(\ell m)}(x)\right] Y_{Lm}(\mathbf{n}). \tag{45}$$

For the quadrupole source, $\ell = 2$, and therefore

$$\begin{aligned}\mathcal{P}_\pm^{(m)}(\eta_0,\mathbf{n},\mathbf{k}) = & -\sqrt{6}\sum_L (-i)^L \sqrt{4\pi(2L+1)}\left[\epsilon_L^{(2m)}(x) \pm i\beta_L^{(2m)}(x)\right] Y_{Lm}(\mathbf{n}) \\ & \times \int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)}\tau'(\eta) P^{(m)}(\eta,\mathbf{k}).\end{aligned} \tag{46}$$

At the observer,

$$\mathcal{P}_\pm^{(m)}(\eta_0,\mathbf{n},\mathbf{k}) = \sum_L (-i)^L \sqrt{\frac{4\pi}{2L+1}}\left[E_L^{(m)}(\eta_0,\mathbf{k}) \pm iB_L^{(m)}(\eta_0,\mathbf{k})\right] {}_{\pm2}Y_{Lm}(\mathbf{n}). \tag{47}$$

Thus, using

$$\sqrt{4\pi(2L+1)} = (2L+1)\sqrt{\frac{4\pi}{2L+1}}, \tag{48}$$

we obtain

$$\begin{aligned}E_L^{(m)}(\eta_0,k) \pm iB_L^{(m)}(\eta_0,k) = & -(2L+1)\sqrt{6}\int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)}\tau'(\eta)P^{(m)}(\eta,k) \\ & \times \left[\epsilon_L^{(2m)}(x) \pm i\beta_L^{(2m)}(x)\right].\end{aligned} \tag{49}$$

Using

$$P^{(m)} = \frac{1}{10}\left[\Theta_2^{(m)} - \sqrt{6}\,E_2^{(m)}\right], \tag{50}$$

this becomes

$$\begin{aligned}E_L^{(m)}(\eta_0,k) \pm iB_L^{(m)}(\eta_0,k) = & -\frac{(2L+1)\sqrt{6}}{10}\int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)}\tau'(\eta)\left[\Theta_2^{(m)} - \sqrt{6}\,E_2^{(m)}\right] \\ & \times \left[\epsilon_L^{(2m)}(x) \pm i\beta_L^{(2m)}(x)\right].\end{aligned} \tag{51}$$

Separating the real and imaginary parts gives

$$E_L^{(m)}(\eta_0,k) = -(2L+1)\sqrt{6}\int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)}\tau'(\eta)P^{(m)}(\eta,k)\epsilon_L^{(2m)}[k(\eta_0-\eta)], \tag{52}$$

$$B_L^{(m)}(\eta_0,k) = -(2L+1)\sqrt{6}\int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)}\tau'(\eta)P^{(m)}(\eta,k)beta_L^{(2m)}[k(\eta_0-\eta)]. \tag{53}$$

For scalar perturbations,

$$\beta_L^{(20)}(x) = 0, \tag{54}$$

and therefore

$$B_L^{(0)} = 0. \tag{55}$$

Numerical form The line-of-sight method separates the calculation into a low-order Boltzmann evolution and a subsequent free-streaming projection.
The low-order hierarchy is solved numerically to obtain

$$S_{\ell'}^{(m)}(\eta,k), \qquad \Theta_2^{(m)}(\eta,k), \qquad E_2^{(m)}(\eta,k), \tag{56}$$

from which

$$P^{(m)}(\eta,k) = \frac{1}{10}\left[\Theta_2^{(m)}(\eta,k) - \sqrt{6}\,E_2^{(m)}(\eta,k)\right] \tag{57}$$

is constructed.
The high temperature multipoles are then obtained without evolving the full hierarchy:

$$\Theta_L^{(m)}(\eta_0,k) = (2L+1)\int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)} \sum_{\ell'=0}^{2} S_{\ell'}^{(m)}(\eta,k) j_L^{(\ell' m)}[k(\eta_0-\eta)]\,. \tag{58}$$

The polarization multipoles are obtained from

$$E_L^{(m)}(\eta_0,k) \pm iB_L^{(m)}(\eta_0,k) = -(2L+1)\sqrt{6}\int_{\eta_{\rm in}}^{\eta_0} d\eta\, e^{-\tau(\eta_0,\eta)}\tau'(\eta)P^{(m)}(\eta,k) \times \left[\epsilon_L^{(2m)} \pm i\beta_L^{(2m)}\right][k(\eta_0-\eta)]\,. \tag{59}$$

Thus the source functions contain the local physics of the perturbations and Thomson scattering, while $j_L^{(\ell' m)}$, $\epsilon_L^{(2m)}$ and $\beta_L^{(2m)}$ describe the subsequent free streaming from the source to the observer.

The primary computational advantage of this formulation is that the source terms $S_\ell^{(m)}$ and $P^{(m)}$ depend exclusively on moments of order $\ell \le 2$. As a result, one does not need to solve the full Boltzmann hierarchy up to high $\ell_{\rm max}$; it is sufficient to solve the coupled differential system truncated at $\ell = 2$ (or conservatively $\ell \sim 8$–10) to determine the source perturbations. The high-ℓ multipoles observed today are then obtained by evaluating the one-dimensional line-of-sight integrals (6.3.173) and (6.3.174), a step that is computationally rapid because the spherical Bessel functions can be precomputed and tabulated on a fine grid.

The structure of Eq. (6.3.173) closely parallels the expressions obtained in Section 6.1.1, which omitted the anisotropic stress of radiation prior to the introduction of Thomson scattering. In particular, the functions $j_\ell^{(00)}$ and $j_\ell^{(10)}$ for scalar modes and $j_\ell^{(11)}$ for vector modes reappear naturally in this harmonic framework. The slight difference in normalization between $j_\ell^{(22)}$ and $\tilde{j}_\ell^{(22)}$ is attributed to the tensor basis conventions noted in Eq. (6.3.46).

Freely available codes

Several publicly available, highly optimized numerical packages implement the line-of-sight integration method to integrate the coupled Einstein–Boltzmann hierarchy for both temperature and polarization:

- `CMBFAST` (Seljak & Zaldarriaga, Ref. [34]), the pioneering code that established the line-of-sight method;
- `CAMB` (Lewis, Challinor & Lasenby, Ref. [35]), based on a covariant approach and written in modern Fortran;
- `CMBEASY` (Doran, Ref. [36]), a C++ object-oriented implementation;

- `COSMICS` (Bertschinger, Ref. [37]), which integrates the complete Boltzmann hierarchy in synchronous or conformal Newtonian gauge.

These packages integrate the perturbation equations within the standard ΛCDM cosmological model, and can be readily extended to encompass dynamical dark energy, non-standard neutrino properties, modified gravity, spatial curvature, and primordial non-Gaussianities.

6.4 Anisotropies of the cosmic microwave background

6.4.1 Observations of the cosmic microwave background anisotropies

We observe the cosmic microwave background on the celestial sphere surrounding our vantage point. Any full-sky temperature anisotropy map $\Theta^{\rm obs}(\boldsymbol{e}) \equiv \Delta T(\boldsymbol{e})/T_0$ measured along line-of-sight directions $\boldsymbol{e}$ can be decomposed in spherical harmonics:

$$\Theta^{\rm obs}(\boldsymbol{e}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m}^{\rm obs} Y_{\ell m}(\boldsymbol{e}). \tag{6.4.1}$$

Under the hypothesis of statistical isotropy, the angular two-point correlation function of the observed temperature field,

$$C^{\rm obs}(\vartheta) \equiv \left\langle \Theta^{\rm obs}(\boldsymbol{e}) \Theta^{\rm obs}(\boldsymbol{e}') \right\rangle_{\boldsymbol{e}\cdot\boldsymbol{e}'=\cos\vartheta}, \tag{6.4.2}$$

depends exclusively on the angular separation $\cos\vartheta = \boldsymbol{e}\cdot\boldsymbol{e}'$ and can be obtained by spatial averaging across the sky. Expanding this correlation function in Legendre polynomials,

$$C^{\rm obs}(\vartheta) = \frac{1}{4\pi} \sum_{\ell=0}^{\infty} (2\ell+1) C_\ell^{\rm obs} P_\ell(\cos\vartheta), \tag{6.4.3}$$

the observed angular power spectrum multipoles are extracted as

$$C_\ell^{\rm obs} = \frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} \left| a_{\ell m}^{\rm obs} \right|^2. \tag{6.4.4}$$

Cosmological theoretical models predict only the ensemble statistical properties of primordial perturbations, not the exact fluctuation pattern in our particular Universe. Concretely: inflation predicts a definite value for C_ℓ, computable from the primordial power spectrum and the radiation transfer function, but it never predicts any individual coefficient $a_{\ell m}$ – each $a_{\ell m}$ is a single random draw from a distribution of variance C_ℓ, and no theory of inflation makes, or could make, a prediction for the outcome of that draw. The comparison between theory and observation therefore requires invoking an ergodic hypothesis: we equate the spatial average over the celestial sphere of our unique observable Universe, $C_\ell^{\rm obs}$, with the theoretical ensemble average $C_\ell = \langle |a_{\ell m}|^2 \rangle$ evaluated over all possible realizations of the cosmological model. Because our sky provides only a finite statistical sample of $2\ell+1$ independent azimuthal modes for a given multipole ℓ, this comparison is fundamentally limited by cosmic variance.

Estimator of C_ℓ and cosmic variance

In linear cosmological perturbation theory, the temperature fluctuation field is modeled as a homogeneous and isotropic random Gaussian field. The multipole coefficients $a_{\ell m}$ are independent complex random variables of vanishing mean, satisfying

$$\langle a_{\ell m} \rangle = 0, \qquad \langle a_{\ell m} a_{\ell' m'}^* \rangle = \delta_{\ell\ell'} \delta_{mm'} C_\ell.$$

Because the true ensemble power spectrum C_ℓ cannot be measured directly, we construct an unbiased estimator $\widehat{C}_\ell$ by summing over all azimuthal modes m:

$$\widehat{C}_\ell = \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell} a_{\ell m}a_{\ell m}^* = \frac{C_\ell}{2\ell+1}V_\ell, \tag{6.4.5}$$

where the normalized variable V_ℓ is defined by

$$V_\ell \equiv \sum_m \frac{(a_{\ell m}-\langle a_{\ell m}\rangle)(a_{\ell m}^*-\langle a_{\ell m}^*\rangle)}{\langle a_{\ell m}a_{\ell m}^*\rangle} = \sum_{m=-\ell}^{\ell}\frac{|a_{\ell m}|^2}{C_\ell}.$$

Since $a_{\ell m}$ obeys Normal distribution, therefore C_ℓ obeys χ^2 distribution. The observed spectrum $C_\ell^{\rm obs}$ defined in Eq. (6.4.4) is precisely an empirical realization of this estimator $\widehat{C}_\ell$.

In standard single-field slow-roll inflation, the primordial perturbations follow nearly Gaussian statistics, so that each multipole amplitude $a_{\ell m}$ obeys a one-point Gaussian probability density function:

$$\mathcal{P}(a_{\ell m}) = \frac{1}{\sqrt{2\pi C_\ell}}\exp\left(-\frac{a_{\ell m}^2}{2C_\ell}\right). \tag{6.4.6}$$

Equation (6.4.6) is, as written, a probability distribution over the data $a_{\ell m}$ at some fixed, given value of C_ℓ: it integrates to 1 over all possible outcomes $a_{\ell m}$. Now hold the data fixed instead, at whatever value was actually measured on the sky, and read the very same expression as a function of C_ℓ. Under this second reading it is no longer a distribution over outcomes (it does not integrate to 1 over C_ℓ); it is called the **likelihood** of C_ℓ, and the **maximum-likelihood principle** says: the best estimate of C_ℓ is whichever value makes the data we actually observed the most probable.

It is worth being precise about what C_ℓ is standing for at this point. Inflation does predict a specific theoretical value for C_ℓ, computable from $P_\zeta(k)$ and the radiation transfer function once a cosmological model is chosen (§10.9 of Chapter 10). But here C_ℓ is being treated purely as an unknown parameter to be read off from the map, with no reference yet to that theoretical prediction. We are not, at this stage, testing inflation against the data; we are only turning a set of noisy numbers into our best single estimate of the sky's own variance at that ℓ. Comparing that estimate to a theoretical $C_\ell(\boldsymbol{\lambda})$, to fit cosmological parameters, is a separate step, taken up in Chapter 10, §10.9.

Applying the principle: the likelihood of the $2\ell+1$ modes at fixed ℓ is the product of Eq. (6.4.6) over m,

$$L(C_\ell) = \prod_{m=-\ell}^{\ell}\frac{1}{\sqrt{2\pi C_\ell}}\exp\left(-\frac{a_{\ell m}^2}{2C_\ell}\right),$$

and maximizing $L(C_\ell)$ is equivalent to minimizing $\chi^2(C_\ell) \equiv -2\ln L(C_\ell) = (2\ell+1)\ln(2\pi C_\ell) + \frac{1}{C_\ell}\sum_m a_{\ell m}^2$. Differentiating and setting the result to zero,

$$\frac{\mathrm{d}\chi^2}{\mathrm{d}C_\ell} = \frac{2\ell+1}{C_\ell} - \frac{1}{C_\ell^2}\sum_{m=-\ell}^{\ell}a_{\ell m}^2 = 0 \implies \widehat{C}_\ell = \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell}a_{\ell m}^2,$$

recovering Eq. (6.4.5) exactly: the maximum-likelihood estimator is nothing more exotic than the sample variance of the $2\ell+1$ measured modes, since their mean is already known to be zero.

Because V_ℓ represents the sum of the squares of $2\ell+1$ independent standard Gaussian random variables, it obeys a chi-square (χ^2) distribution with $2\ell+1$ degrees of freedom:

$$\mathcal{P}_{\chi^2_\ell}(V_\ell) = \frac{V_\ell^{(2\ell+1)/2-1}}{2^{(2\ell+1)/2}\Gamma\left(\frac{2\ell+1}{2}\right)}\exp\left(-\frac{V_\ell}{2}\right). \tag{6.4.7}$$

From this result, the probability distribution function of the estimator $\widehat{C}_\ell$ follows directly:

$$\mathcal{P}(\widehat{C}_\ell) = \frac{2\ell+1}{C_\ell}\mathcal{P}_{\chi^2_\ell}\left(\frac{(2\ell+1)\widehat{C}_\ell}{C_\ell}\right) = \frac{2\ell+1}{C_\ell}\frac{\left[(2\ell+1)\frac{\hat{C}_\ell}{C_\ell}\right]^{\ell-\frac{1}{2}}}{2^{\ell+\frac{1}{2}}\Gamma\left(\ell+\frac{1}{2}\right)}e^{-\frac{(2\ell+1)}{2C_\ell}\hat{C}_\ell}. \tag{6.4.8}$$

In the high-multipole limit $\ell \to \infty$, the central limit theorem ensures that this distribution approaches a Gaussian, and the estimator converges strictly to the true value:

$$\lim_{\ell\to\infty} \widehat{C}_\ell = C_\ell. \tag{6.4.9}$$

The estimator is therefore consistent. Furthermore, since $\langle \widehat{C}_\ell \rangle = \frac{C_\ell}{2\ell+1}\langle V_\ell \rangle = \frac{C_\ell}{2\ell+1}(2\ell+1) = C_\ell$, it is strictly unbiased. The intrinsic variance of the estimator evaluates to (from (6.4.7))

$$\mathrm{var}(C_\ell) = \langle \widehat{C}^2_\ell \rangle - \langle \widehat{C}_\ell \rangle^2 = \left(\frac{C_\ell}{2\ell+1}\right)^2 \times 2(2\ell+1) = \frac{2}{2\ell+1}C^2_\ell.$$

By the Cramér–Rao bound, $\widehat{C}_\ell$ is the minimum-variance unbiased estimator of C_ℓ. The relative uncertainty on the power spectrum measurement is fundamentally bounded by the finite number of independent orientations on the sky:

$$\frac{\langle \widehat{C}^2_\ell \rangle - \langle \widehat{C}_\ell \rangle^2}{C^2_\ell} = \frac{2}{2\ell+1}. \tag{6.4.10}$$

This statistical uncertainty constitutes the cosmic variance. It represents an irreducible physical limitation that dominates the error budget at low multipoles ($\ell \lesssim 10$).

Consequently, no single $a_{\ell m}$ can directly test cosmological models. Scientific inference requires constructing statistical estimators. If primordial fluctuations deviate from Gaussianity (e.g., through non-Gaussian initial conditions or non-linear gravitational evolution), optimal estimator construction becomes significantly more intricate, as two-point correlation functions no longer exhaust the information content of the field.

Status of the observations

Observational cosmology of the CMB has experienced unprecedented progress over recent decades (see Refs. [38–40] for detailed historical surveys).

The foundational breakthrough was achieved in 1992 by the Differential Microwave Radiometer (DMR) instrument aboard NASA's COBE satellite, which detected primordial CMB anisotropies for the first time on large angular scales ($\vartheta \gtrsim 7°$, corresponding to multipoles $\ell \lesssim 20$). As illustrated in Fig. 6.17, observational precision advanced rapidly over the subsequent decade. In 1997, large experimental uncertainties could not decisively distinguish between competing cosmological paradigms. Beginning in 1998, high-altitude balloon-borne experiments—notably BOOMERanG and MAXIMA, followed by Archeops—resolved the first acoustic peak with high significance and provided initial detections of the second peak. By 2003, the Wilkinson Microwave Anisotropy Probe (WMAP) mapped the full sky with sub-degree resolution, measuring the angular power spectrum up to $\ell \sim 1000$.

CMB experimental platforms fall into three primary categories:

1. **Ground-based experiments**: These instruments target small angular scales (high multipoles) using coherent interferometric arrays or focal-plane bolometers (e.g., DASI [41], CAT [42], VLA [43], and CBI). Located at high altitudes or exceptionally dry sites (such as the Atacama plateau or the South Pole), their sensitivity is ultimately limited by atmospheric water vapor fluctuations.

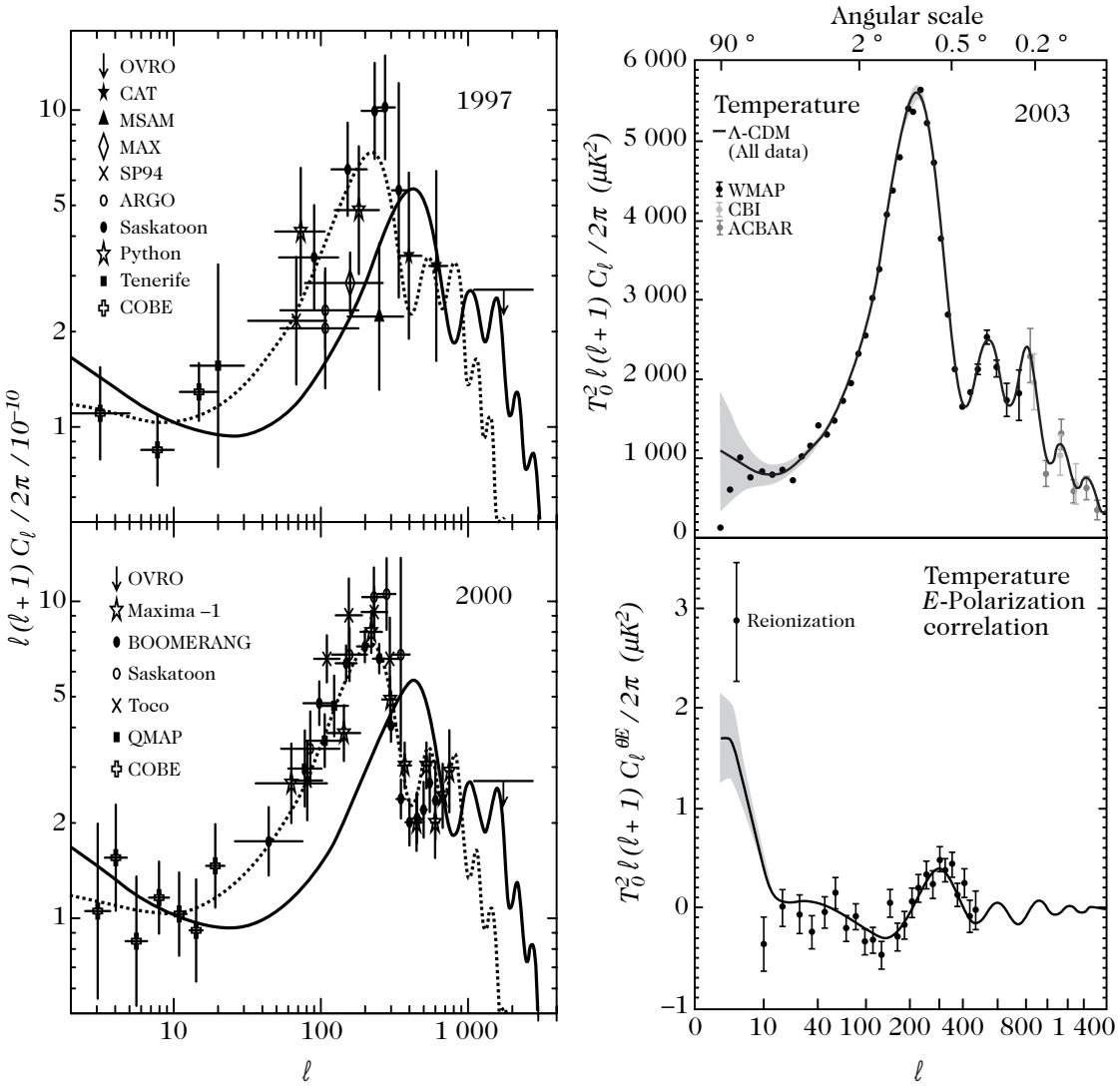


Figure 6.17: Evolution of CMB observational status between 1997 and 2003 (left panel). The solid curve represents the best-fit flat ΛCDM concordance model, while an open model ($\Omega_0 = 0.3$, dotted line) is firmly ruled out by the data. (Right panel): Full-sky angular power spectra for temperature (Θ–Θ) and the temperature–polarization cross-correlation (Θ–E) determined by the WMAP satellite.

2. **Stratospheric balloon experiments**: Operating at altitudes near 40 km, balloon-borne payloads fly above 99.5% of the Earth's atmosphere, substantially eliminating atmospheric emission and absorption. Flight durations range from overnight flights to circum-polar long-duration ballooning of ten days or more (e.g., BOOMERanG [44], MAXIMA [45], and Archeops [46]).

3. **Space missions**: Space platforms provide full-sky coverage in a thermally stable environment without atmospheric attenuation. Three dedicated space missions have surveyed the CMB: COBE [47], WMAP [48], and Planck [49].

The WMAP satellite mapped the cosmic microwave background in five frequency channels spanning 22 to 90 GHz. Key findings from its multi-year data releases (Ref. [50]) include:

- Full-sky temperature maps with an angular resolution of 0.3° (Fig. 6.18), measuring the angular power spectrum with high signal-to-noise ratio up to $\ell \sim 1000$.

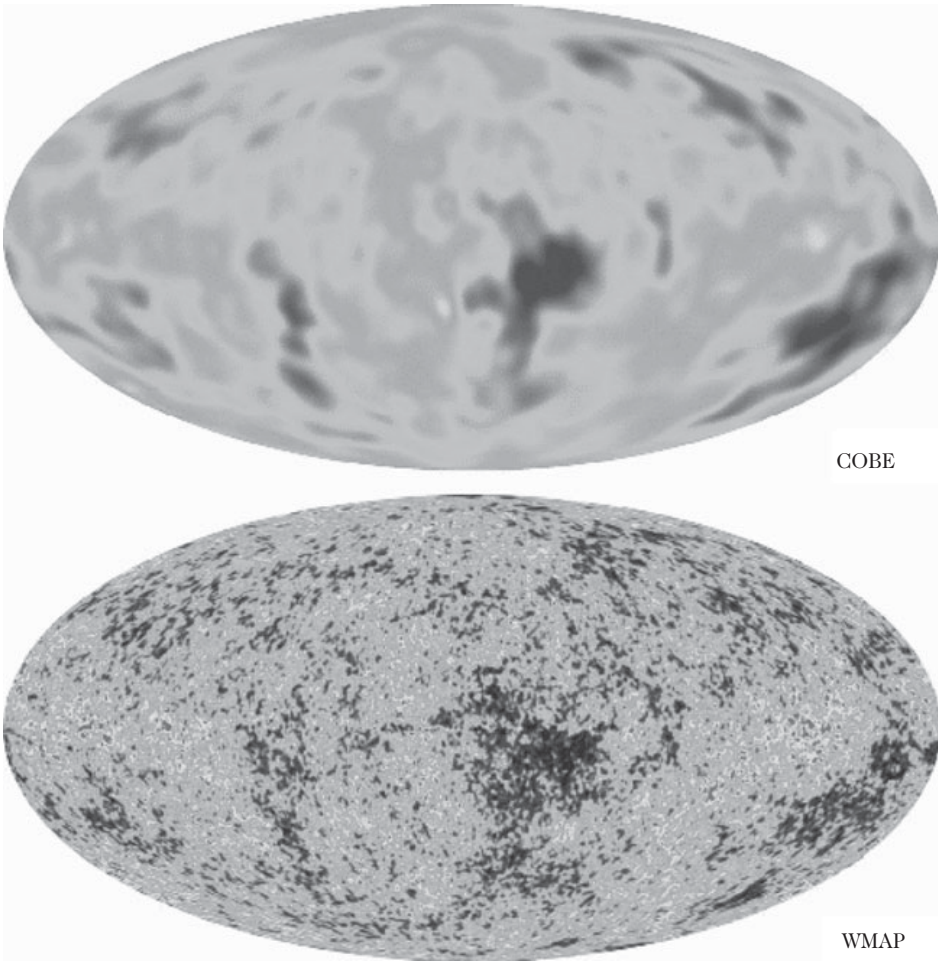


Figure 6.18: Comparison of the angular resolution achieved by the COBE-DMR instrument (top) and the WMAP satellite (bottom) across the microwave sky (WMAP Science Team, 2003).

- High-precision determination of the first acoustic peak position (Ref. [51]):

$$\ell^{(1)} = 220.8 \pm 0.7, \tag{6.4.11}$$

  with an amplitude of 74.7 ± 0.5 μK. The second acoustic peak is measured at $\ell^{(2)} = 530.9 \pm 3.8$ with amplitude 48.8 ± 0.9 μK. The first and second troughs occur at $\ell^{(-1)} = 412.4 \pm 1.9$ and $\ell^{(-2)} = 675.2 \pm 11.1$, respectively.

- Measurement of the temperature–polarization cross-correlation spectrum $C_\ell^{\Theta E}$ (Fig. 6.17), exhibiting an anti-peak at $\ell = 137 \pm 9$ and a positive peak at $\ell = 329 \pm 19$. The existence of this anti-correlation on angular scales larger than the sound horizon provides definitive confirmation of super-Hubble perturbations at decoupling, ruling out causal post-recombination mechanisms (such as cosmic strings and textures). The relative phase shift between temperature and polarization spectra provides compelling evidence in favor of adiabatic initial conditions.

- An excess in the large-scale Θ–E cross-correlation at $\ell < 20$, indicating an early reionization of the intergalactic medium at redshift $z_{\rm re} = 11.0^{+2.6}_{-2.5}$, with an integrated optical depth $\tau_{\rm re} = 0.089 \pm 0.03$.

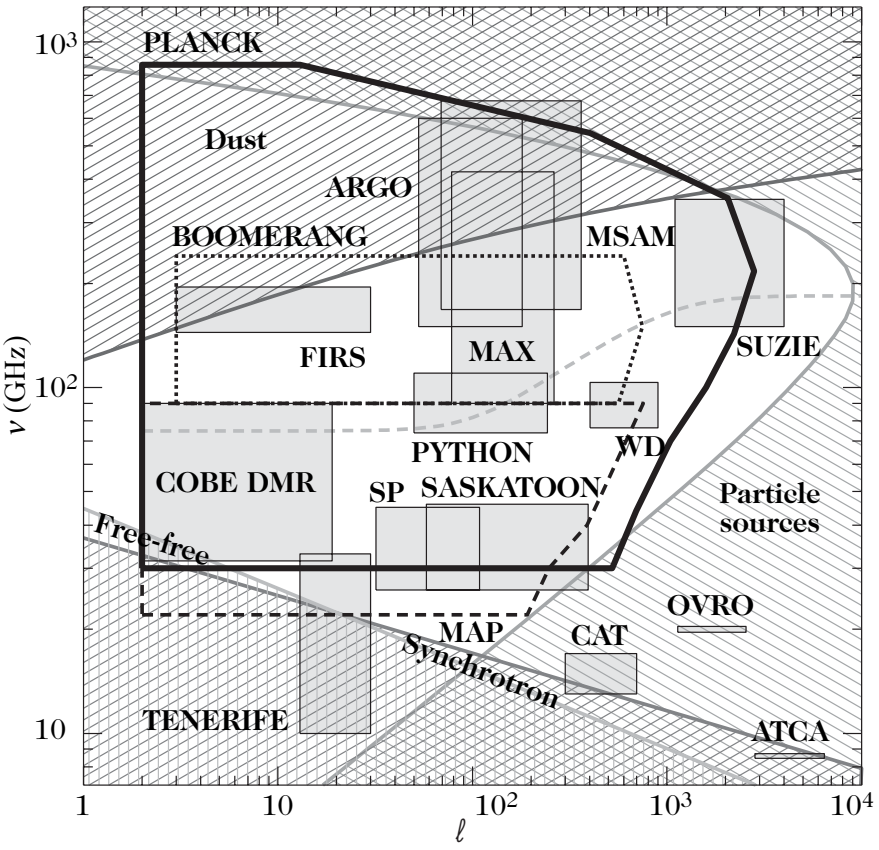


Figure 6.19: Angular and frequency sensitivity ranges for major ground-based, sub-orbital, and space CMB experiments. The shaded regions denote frequency regimes where Galactic foreground emissions (synchrotron, free–free, and thermal dust) exceed 20 μK in the cleanest 20% of the sky. From Ref. [53].

- Detection of primordial E-mode polarization:

$$\frac{\ell(\ell+1)}{2\pi} C_\ell^{EE} = (0.086 \pm 0.029)\ \mu\mathrm{K}^2 \qquad (\ell = 2\text{–}6),$$

  interpreted as the Thomson rescattering of CMB photons during reionization.

- Non-detection of primordial tensor B-modes, establishing the upper bound:

$$\frac{\ell(\ell+1)}{2\pi} C_\ell^{BB} = (-0.04 \pm 0.03)\ \mu\mathrm{K}^2 \qquad (\ell = 2\text{–}6).$$

- Strict constraints establishing that the temperature field is consistent with Gaussianity to high precision.

The WMAP observations firmly established the flat ΛCDM concordance model with $\Omega_{\Lambda 0} \sim 0.7$ as the standard framework of modern cosmology. Nevertheless, several subtle anomalies persist at low multipoles:

- A lower quadrupole ($\ell = 2$) amplitude than predicted by the concordance model, first identified by COBE;

- An unexpected planar alignment between the quadrupole and octopole ($\ell = 3$) modes;

- Hemispherical power asymmetries between the northern and southern ecliptic skies.

CMB polarization was first detected experimentally in September 2002 by the DASI interferometer operating at 30 GHz at the South Pole (Ref. [52]), confirming the predicted E-mode polarization at the 5σ confidence level.

The ESA Planck satellite was designed to survey the sky across nine frequency channels (20 GHz to 850 GHz) with angular resolution down to $4'$ ($\ell \sim 2000$), deploying low-frequency radiometers and high-frequency bolometers to map temperature and polarization with cosmic-variance-limited sensitivity.

Statistical isotropy and observation on an incomplete sky

Realistic observations cannot measure the primordial CMB across the entire celestial sphere because astrophysical foreground emissions—predominantly Galactic synchrotron, free–free, and thermal dust emission, as well as extragalactic point sources—obscure the cosmological signal. In practice, the Galactic plane and point sources are excised by applying an observational window or mask $W(\boldsymbol{e})$:

$$\widetilde{\Theta}(\boldsymbol{e}) = \Theta(\boldsymbol{e}) W(\boldsymbol{e}), \tag{6.4.12}$$

where $W(\boldsymbol{e})$ is a spatial mask equal to zero in masked regions and unity across clean sky areas (or smoothly apodized). Expanding the mask in spherical harmonics,

$$W(\boldsymbol{e}) = \sum_{\ell m} w_{\ell m} Y_{\ell m}(\boldsymbol{e}), \tag{6.4.13}$$

with the reality condition $w_{\ell m}^{*} = (-1)^{m} w_{\ell,-m}$, the pseudo-multipole coefficients $\widetilde{a}_{\ell m}$ of the masked map relate to the primordial full-sky coefficients $a_{\ell m}$ via

$$\widetilde{a}_{\ell m} = \sum_{\ell_1 m_1} a_{\ell_1 m_1} \sum_{\ell_2 m_2} w_{\ell_2 m_2} \int \mathrm{d}^2\boldsymbol{e}\, Y_{\ell_1 m_1}(\boldsymbol{e}) Y_{\ell_2 m_2}(\boldsymbol{e}) Y_{\ell m}^{*}(\boldsymbol{e}). \tag{6.4.14}$$

Using the angular momentum coupling formula for the integral of three spherical harmonics [Eq. (B.23)], this expression evaluates to

$$\widetilde{a}_{\ell m} = \sum_{\ell_1 m_1} a_{\ell_1 m_1} K_{\ell m}^{\ell_1 m_1}, \tag{6.4.15}$$

where the coupling kernel $K_{\ell m}^{\ell_1 m_1}$ is given explicitly in terms of Wigner $3j$-symbols by

$$K_{\ell m}^{\ell_1 m_1} = (-1)^{m} \sum_{\ell_2 m_2} w_{\ell_2 m_2} \sqrt{\frac{(2\ell_1+1)(2\ell_2+1)(2\ell+1)}{4\pi}} \begin{pmatrix} \ell_1 & \ell_2 & \ell \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \ell_1 & \ell_2 & \ell \\ m_1 & m_2 & -m \end{pmatrix}. \tag{6.4.16}$$

For an unmasked full sky, $W(\boldsymbol{e}) = 1$, corresponding to $w_{\ell m} = \sqrt{4\pi}\,\delta_{\ell 0}\delta_{m0}$, and the kernel collapses to the identity $K_{\ell m}^{\ell_1 m_1} = \delta_{\ell \ell_1}\delta_{m m_1}$. In the presence of a mask, the masked multipoles no longer satisfy the orthogonality condition $\langle \widetilde{a}_{\ell m} \widetilde{a}_{\ell' m'}^{*} \rangle = \delta_{\ell\ell'}\delta_{mm'} C_\ell$; instead, the covariance matrix becomes

$$\langle \widetilde{a}_{\ell m} \widetilde{a}_{\ell' m'}^{*} \rangle = \sum_{\ell_1 m_1} C_{\ell_1} K_{\ell m}^{\ell_1 m_1} \left(K_{\ell' m'}^{\ell_1 m_1} \right)^{*},$$

reflecting the fact that the cut sky breaks statistical isotropy and couples different multipoles (ℓ, m). Constructing unbiased, minimum-variance pseudo-C_ℓ estimators on partial skies therefore requires inverting the mode-coupling matrix.

6.5 Effects of the parameters on the angular power spectrum

Scalar, vector, and tensor perturbations impart distinct physical signatures onto the temperature anisotropies and polarization fields of the cosmic microwave background. Parity conservation ensures that primordial scalar and tensor perturbations produce only four non-vanishing angular auto- and cross-power spectra: Θ–Θ, E–E, Θ–E, and B–B. The cross-correlations Θ–B and E–B vanish identically by reflection symmetry. Primordial scalar density fluctuations source only the Θ–Θ, E–E, and Θ–E spectra, while tensor gravitational waves contribute to all four observables and provide the unique source for the primordial B–B auto-correlation.

The morphology of these angular power spectra depends sensitively on the cosmological parameters governing the background FLRW dynamics and the perturbation transfer functions, as well as on the primordial initial conditions.

6.5.1 Cosmological parameters

Baryon density: $\Omega_{\mathrm{b}0}h^2$

As established in Section 6.2.2, the principal physical quantity governing acoustic oscillations in the pre-recombination plasma is the baryon-to-photon momentum density ratio $R \equiv 3\rho_{\mathrm{b}}/(4\rho_\gamma)$. At the last-scattering surface, its value evaluates to

$$R_{\mathrm{LSS}} \simeq 27.37\,\Omega_{\mathrm{b}0}h^2.$$

The physical baryon density $\Omega_{\mathrm{b}0}h^2$ exerts four primary effects on the temperature angular power spectrum:

1. **Sound speed and acoustic scale**: Increasing $\Omega_{\mathrm{b}0}$ increases the plasma inertia and reduces the sound speed $c_{\mathrm{s}} = [3(1+R)]^{-1/2}$, decreasing the sound horizon $r_{\mathrm{s}}(\eta_{\mathrm{LSS}})$. This shifts the acoustic peaks toward slightly higher multipoles and compresses peak spacing.

2. **Odd-to-even peak contrast**: Baryons shift the zero-point of acoustic oscillations in the gravitational potential wells. In the adiabatic mode, the oscillation amplitude of odd peaks (compressions) is enhanced by $(1+6R)$ relative to even peaks (rarefactions). A higher baryon density therefore magnifies the first and third peaks while suppressing the second peak.

3. **Acoustic contrast**: The baryon density modulates the relative prominence of the acoustic Doppler velocity term. In the limit of vanishing baryon content ($\Omega_{\mathrm{b}0} \to 0$), the acoustic peaks are heavily smoothed, whereas increasing $\Omega_{\mathrm{b}0}$ sharpens peak contrast.

4. **Silk damping**: Increasing the baryon density increases the free electron density n_{e}, thereby shortening the photon mean free path $\lambda_{\mathrm{mfp}} = (\sigma_{\mathrm{T}} n_{\mathrm{e}})^{-1}$ and shifting Silk diffusion damping toward smaller spatial scales (higher ℓ).

Figure 6.20 demonstrates the systematic deformation of the temperature power spectrum induced by varying the baryon density.

Total matter density: $\Omega_{\mathrm{m}0}h^2$

In the concordance cosmological model, cold dark matter dominates the total matter density ($\Omega_{\mathrm{c}0} \gg \Omega_{\mathrm{b}0}$). Cold dark matter interacts with photons solely through the gravitational potential. The physical matter density $\Omega_{\mathrm{m}0}h^2$ dictates the redshift of matter–radiation equality:

$$1 + z_{\mathrm{eq}} = \frac{\Omega_{\mathrm{m}0}}{\Omega_{\mathrm{r}0}} \simeq 2.4 \times 10^4\,\Omega_{\mathrm{m}0}h^2.$$

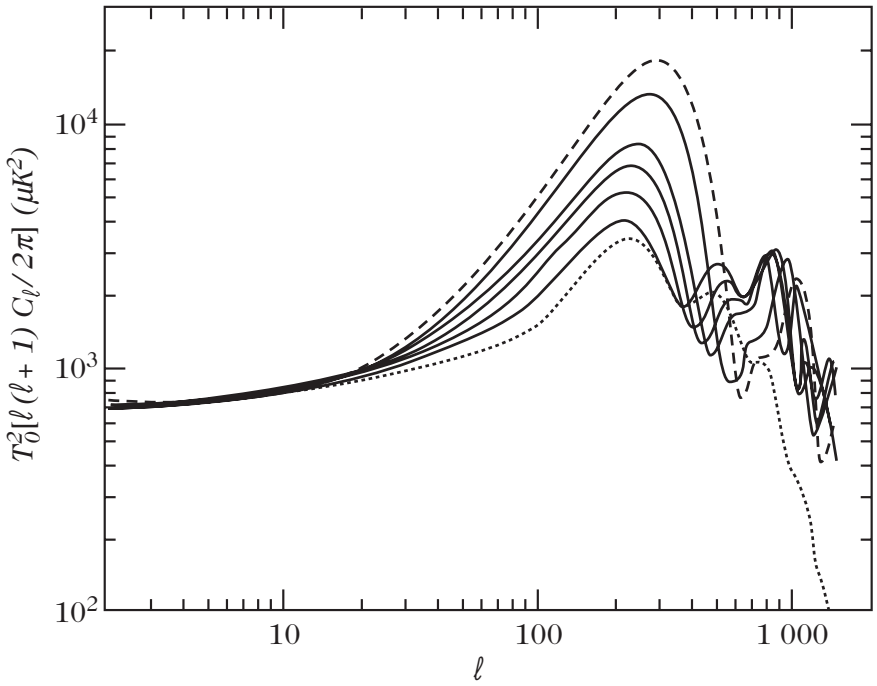


Figure 6.20: Influence of the physical baryon density $\Omega_{\mathrm{b}0}h^2$ on the CMB temperature angular power spectrum. Solid curves depict $\Omega_{\mathrm{b}0}h^2 = 0.0125, 0.0250, 0.0375, 0.0500$, and 0.0875. The extreme cases correspond to $\Omega_{\mathrm{b}0}h^2 = 0.0025$ (dotted line) and 0.1250 (dashed line). From Ref. [3].

Increasing $\Omega_{\mathrm{m}0}h^2$ causes equality to occur earlier (at higher redshift). In radiation-dominated eras, the decay of the Bardeen potentials driving acoustic oscillations induces a substantial early integrated Sachs–Wolfe (eISW) effect, which boosts the amplitude of the first acoustic peak. An earlier equality transition shortens the radiation-dominated era preceding decoupling, thereby attenuating the eISW enhancement and lowering the first peak amplitude relative to the higher multipoles.

Dark-matter–baryon ratio

The baryon-to-dark-matter ratio $\alpha \equiv \Omega_{\mathrm{b}0}/\Omega_{\mathrm{c}0}$ governs the relative source contributions in the perturbed Poisson equation. Because baryonic density perturbations can grow only after decoupling, increasing α diminishes the early gravitational potential well depths established by cold dark matter, lowering the overall amplitude of the acoustic peaks. Furthermore, the acoustic oscillations of the photon–baryon fluid leave a permanent characteristic imprint in the matter power spectrum, known as baryon acoustic oscillations (BAO).

Hubble constant: h

The dimensionless Hubble parameter h is constrained by the Friedmann constraint $\Omega_{\mathrm{m}0} + \Omega_{\Lambda 0} + \Omega_{K0} = 1$. Its impact on the angular power spectrum depends on which parameters are held fixed during parameter variation. For fixed physical matter density $\omega_{\mathrm{m}} \equiv \Omega_{\mathrm{m}0}h^2$, reducing h increases the comoving distance to the last-scattering surface, shifting the entire acoustic peak pattern toward higher multipoles ℓ. Conversely, if $\Omega_{\mathrm{m}0}$ is fixed while h varies, decreasing h lowers ω_{m}, delaying matter–radiation equality and enhancing the acoustic peak amplitudes via the early ISW effect (as shown in Fig. 6.21).

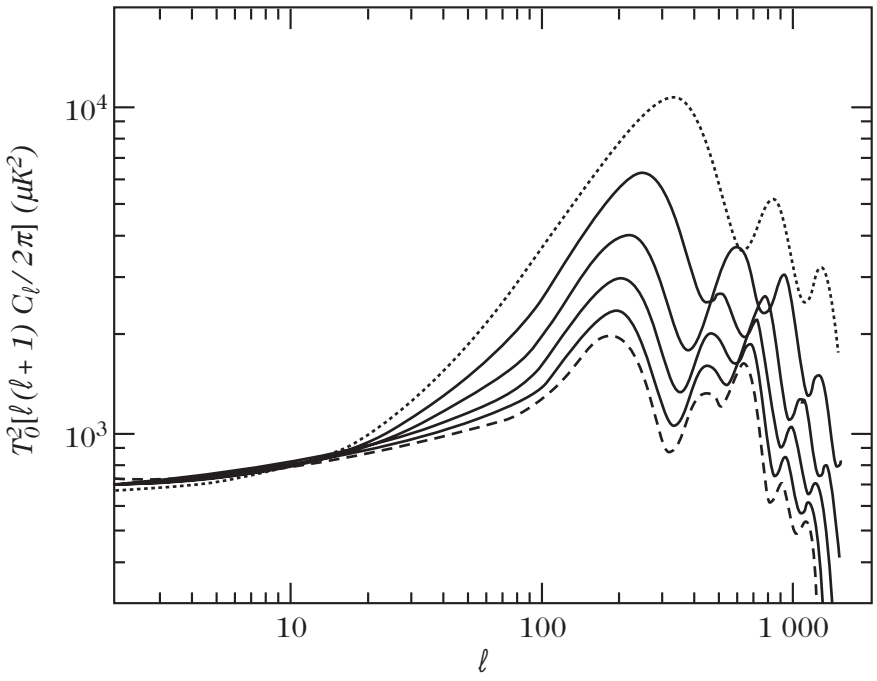


Figure 6.21: Influence of the Hubble constant h on the temperature power spectrum for a fixed baryon fraction $\Omega_{\rm b0} = 0.0125$. The curves correspond to $h = 0.20$ (dotted line), $0.35, 0.50, 0.65, 0.80$ (solid lines), and 0.95 (dashed line). From Ref. [3].

Cosmological constant: $\Omega_{\Lambda 0}$

Varying the vacuum energy density $\Omega_{\Lambda 0}$ in a spatially flat universe alters the expansion rate at late times, modifying the comoving angular diameter distance to recombination $D_{\rm A}(\eta_{\rm LSS})$:

$$D_{\rm A} = \int_0^{z_{\rm LSS}} \frac{{\rm d}z}{H(z)}.$$

An increase in $\Omega_{\Lambda 0}$ increases $D_{\rm A}$, which shifts the angular projection of the sound horizon toward smaller angular scales (larger multipoles ℓ). Furthermore, dark energy domination at low redshifts causes the gravitational potentials Φ and Ψ to decay at $z \lesssim 1$, producing a late integrated Sachs–Wolfe effect that boosts the temperature power spectrum at low multipoles ($\ell \lesssim 20$).

Curvature: $\Omega_{K0}h^2$

Spatial curvature fundamentally modifies the spatial geometry and the angular diameter distance relation:

$$D_{\rm A} = \frac{1}{\sqrt{|K|}} \chi_K\left(\sqrt{|K|}\chi\right),$$

where $\chi_K(x) = \sin x$ for closed spaces ($K > 0$) and $\sinh x$ for open spaces ($K < 0$). In an open universe ($\Omega_{K0} > 0$), photon geodesics diverge, reducing the apparent angular size of the sound horizon and shifting the entire acoustic peak spectrum toward substantially higher multipoles ($\ell \propto \Omega_0^{-1/2}$). In a closed universe, geodesics converge, shifting the peaks toward lower multipoles. Additionally, the transition to curvature domination triggers a late decay of the Bardeen potentials, enhancing the low-ℓ power via the late ISW effect.

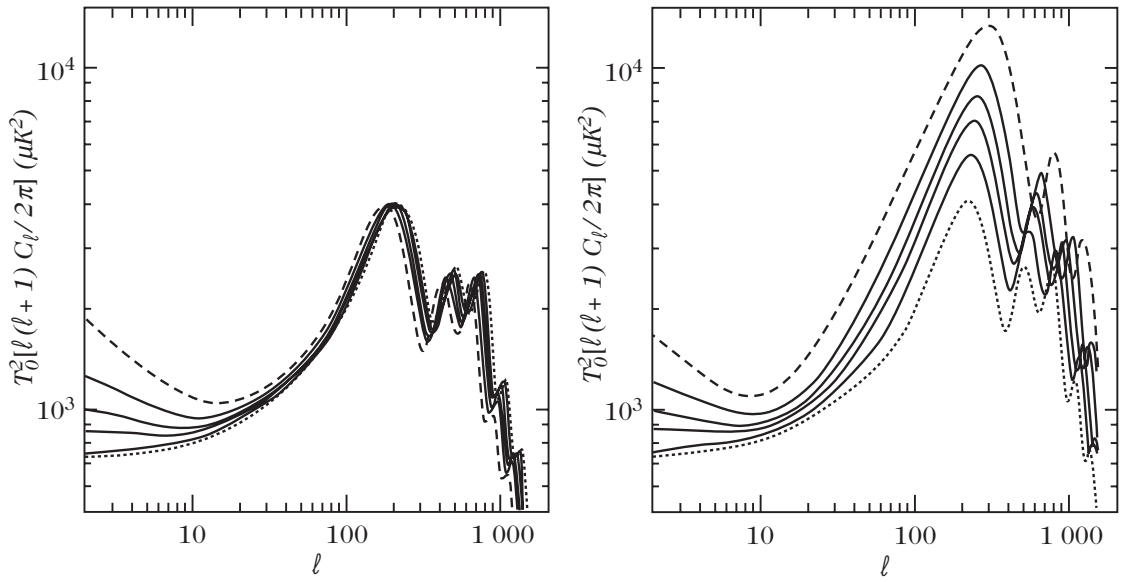


Figure 6.22: Influence of the cosmological constant $\Omega_{\Lambda 0}$ on the temperature angular power spectrum. (Left panel): $\Omega_{\Lambda 0}$ varies while keeping the critical density $\rho_{\mathrm{crit},0}$ and the physical matter density $\bar{\rho}_0$ fixed, which implies that h also varies; $\Omega_{\Lambda 0}$ takes the values 0 (dotted line), $0.4, 0.6, 0.7, 0.8$ (solid lines), and 0.9 (dashed line). (Right panel): $\rho_{\mathrm{b}0}$ and h are held constant so that $\Omega_{\mathrm{m}0}$ varies as $1 - \Omega_{\Lambda 0}$; $\Omega_{\Lambda 0}$ takes the values 0 (dashed line), $0.4, 0.6, 0.7, 0.8$ (solid lines), and 0.9 (dotted line). From Ref. [3].

Notice that we can simultaneously adjust the spatial curvature and the cosmological constant while keeping the resulting angular power spectrum almost unchanged. These compensating effects are illustrated in Fig. 6.23.

Parameter degeneracy

As demonstrated in the preceding figures, varying cosmological parameters simultaneously alters both the positions and relative amplitudes of the acoustic peaks, as well as the shape of the large-scale Sachs–Wolfe plateau. On large angular scales, cosmic variance imposes an unavoidable statistical limitation, allowing substantially different sets of cosmological parameters to produce angular power spectra that are observationally indistinguishable.

These geometric degeneracies arise primarily because the angular locations of the acoustic peaks are dictated by the acoustic angular scale:

$$\ell_{\mathrm{A}} = \frac{f_K(z_{\mathrm{LSS}})}{r_{\mathrm{s}}(z_{\mathrm{LSS}})},$$

introduced in Eq. (6.74). Linearizing this angular scale around a flat fiducial concordance model with $(\Omega_0, \Omega_{\mathrm{m}0}h^2, \Omega_{\Lambda 0}, w_Q, \Omega_{\mathrm{b}0}h^2) = (1, 0.15, 0.65, -1, 0.02)$, where w_Q denotes the dark energy equation of state parameter, the fractional shift of ℓ_{A} satisfies

$$\frac{\Delta\ell_{\mathrm{A}}}{\ell_{\mathrm{A}}} \simeq -1.1\frac{\Delta\Omega_0}{\Omega_0} - 0.24\frac{\Delta(\Omega_{\mathrm{m}0}h^2)}{\Omega_{\mathrm{m}0}h^2} - 0.17\frac{\Delta\Omega_{\Lambda 0}}{\Omega_{\Lambda 0}} - 0.11\Delta w_Q - 0.07\frac{\Delta(\Omega_{\mathrm{b}0}h^2)}{\Omega_{\mathrm{b}0}h^2}.$$

An explicit realization of this cosmological parameter degeneracy is illustrated in Fig. 6.24.

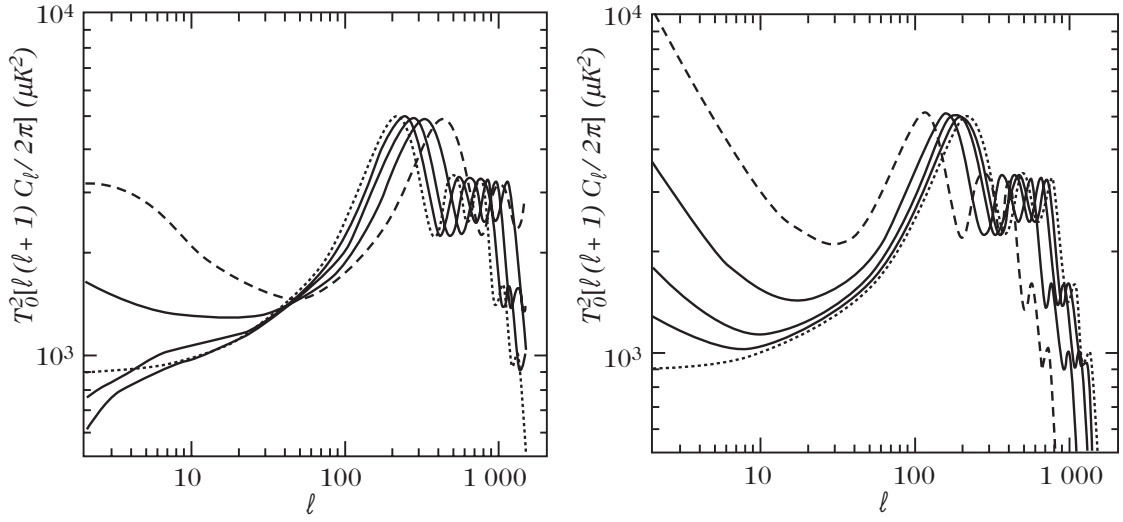


Figure 6.23: Influence of spatial curvature on the temperature angular power spectrum. (Left panel): Spatially hyperbolic (open) universes. The matter density is held constant while Ω_0 takes the values 1 (dotted line), $0.8, 0.6, 0.4$ (solid lines), and 0.2 (dashed line); the acoustic peaks shift toward smaller angular scales (higher ℓ) and potential decay produces a late integrated Sachs–Wolfe effect of opposite sign to the ordinary Sachs–Wolfe term. (Right panel): Spatially spherical (closed) universes. The matter density is held constant while Ω_0 takes the values 1 (dotted line), $1.2, 1.4, 2.0$ (solid lines), and 4.0 (dashed line); the Bardeen potentials grow during the pre-turnaround era, reinforcing the large-scale Sachs–Wolfe plateau. From Ref. [3].

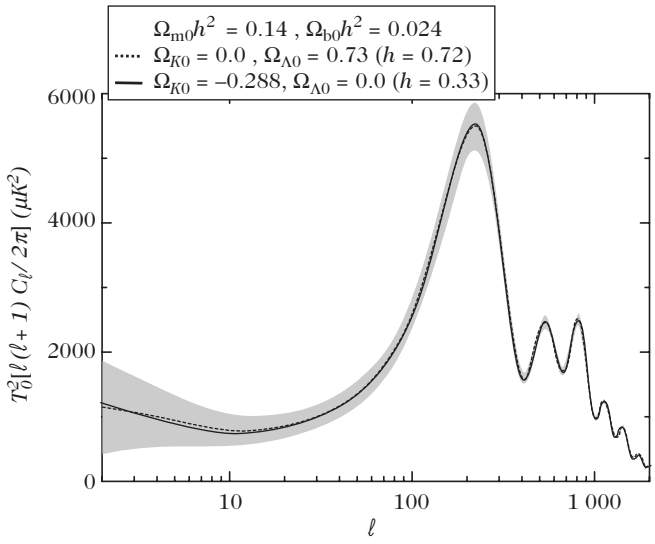


Figure 6.24: Illustration of cosmological parameter degeneracy in the temperature power spectrum. Both cosmological models produce nearly identical angular power spectra within the cosmic variance limit. Both models share identical physical densities $(\Omega_{m0}h^2, \Omega_{b0}h^2) = (0.14, 0.024)$. The first model (dotted line) is the concordance flat ΛCDM model with $(\Omega_{K0}, \Omega_{\Lambda0}) = (0, 0.73)$ and $h = 0.72$, while the second (solid line) is an open model without dark energy, having $(\Omega_{K0}, \Omega_{\Lambda0}) = (-0.288, 0)$ and a low Hubble constant $h = 0.33$. From Ref. [3].

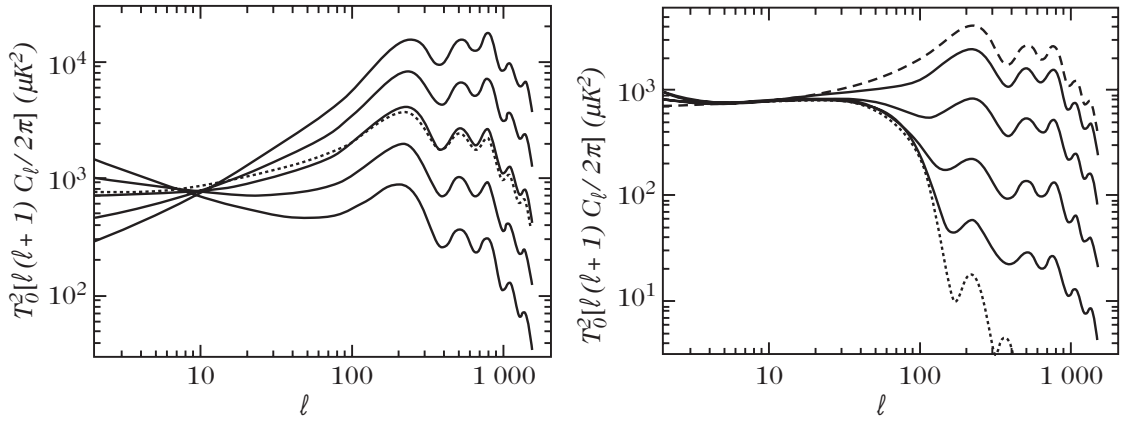


Figure 6.25: (Left panel): Influence of the scalar spectral index $n_{\rm s}$ on the temperature power spectrum. Curves display $n_{\rm s} = 0.5, 0.75, 1.0, 1.25$, and 1.5 (solid lines from bottom to top across the acoustic region). The dotted curve denotes $n_{\rm s} = 1.5$ rescaled by $\ell^{-0.5}$, matching the scale-invariant spectrum $n_{\rm s} = 1$. (Right panel): Influence of the tensor-to-scalar ratio T/S. The dotted and dashed curves show pure scalar and pure tensor contributions, respectively, with intermediate curves interpolating between $T/S = 0$ and ∞. From Ref. [3].

6.5.2 Parameters describing the primordial physics

Primordial physics governs the initial power spectra for both scalar and tensor fluctuations, which are parameterized by their amplitudes, spectral indices, and running of the spectral tilt (Chapter 8). It also determines whether primordial perturbations originate from adiabatic or isocurvature initial conditions.

Spectral index

As established in Eqs. (6.50) and (6.96), the angular power spectrum multipoles on large angular scales scale as $\ell(\ell+1)C_\ell \propto \ell^{n_{\rm s}-1}$ for scalar modes and $\ell(\ell+1)C_\ell \propto \ell^{n_{\rm T}}$ for tensor modes.

A red-tilted primordial spectrum ($n_{\rm s} < 1, n_{\rm T} < 0$) produces an excess of power at large angular scales (low ℓ), whereas a blue-tilted spectrum ($n_{\rm s} > 1$) enhances power at small angular scales. A deviation from exact scale invariance ($n_{\rm s} = 1$) therefore tilts the Sachs–Wolfe plateau and modifies the relative amplitude of the acoustic peaks compared with the low-ℓ plateau. This tilting is illustrated in the left panel of Fig. 6.25.

Primordial gravitational waves

Primordial gravitational waves contribute predominantly to temperature anisotropies at large angular scales ($\ell \lesssim 100$), decaying rapidly once modes enter the horizon during the radiation- or matter-dominated eras. Consequently, tensor perturbations tend to decrease the relative height of the acoustic peaks (which are of purely scalar origin) compared with the Sachs–Wolfe plateau.

When the total power spectrum is normalized to match observational data on large angular scales, a non-zero tensor-to-scalar ratio $r \equiv T/S$ lowers the required scalar amplitude, reducing the overall height of the acoustic peaks (as shown in the right panel of Fig. 6.25).

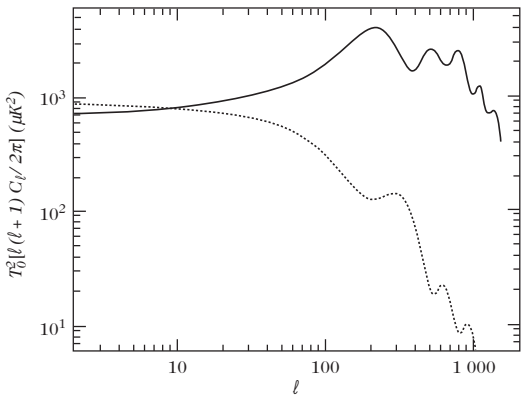


Figure 6.26: Comparison of the temperature angular power spectrum for adiabatic (solid line) and isocurvature (dashed line) initial conditions. The acoustic peaks in the isocurvature spectrum are shifted toward higher multipoles, clearly demonstrating the observational preference for adiabatic perturbations. From Ref. [3].

Adiabatic and isocurvature initial conditions

The choice of initial conditions alters the acoustic peak positions and their relative harmonic spacing. This structural difference stems directly from the oscillation phases described in Eq. (6.74). For isocurvature perturbations, the initial condition is characterized by a vanishing density perturbation at horizon entry, resulting in cosine acoustic oscillations instead of the sine oscillations characteristic of adiabatic modes. Consequently, the position of the first acoustic peak is shifted by a factor of roughly 3/2 toward higher multipoles ($\ell \sim 330$ for isocurvature versus $\ell \sim 220$ for adiabatic perturbations).

A second fundamental distinction appears in the Sachs–Wolfe normalization on super-Hubble scales:

$$\Theta_{\rm SW} = \frac{1}{3}\Phi \quad \text{(adiabatic)}, \qquad \Theta_{\rm SW} = 2\Phi \quad \text{(isocurvature)}.$$

As demonstrated in Fig. 6.26, the observed acoustic peak positions decisively exclude pure isocurvature models, confirming that primordial cosmic perturbations are predominantly adiabatic.

Chapter 7

Cosmological parameter estimation

7.1 Introduction

In the preceding chapters, we developed the relativistic perturbation theory and the kinetic Boltzmann framework required to compute the temperature anisotropies and polarization of the Cosmic Microwave Background (CMB). For any given cosmological model equipped with a specified set of initial conditions, we can calculate the predicted angular power spectrum C_ℓ with remarkable precision.

However, in observational cosmology, we are confronted with the inverse problem: given observed data from the sky, how do we deduce the underlying parameters of our Universe?

The shape and amplitude of the CMB angular power spectrum are governed by two conceptually distinct physical ingredients:

1. **The Initial Fluctuations:** The primordial power spectra of scalar, vector, and tensor perturbations generated in the very early Universe (e.g., during inflation).

2. **The Background Cosmological Parameters:** The parameters ($\Omega_\mathrm{b}, \Omega_\mathrm{cdm}, \Omega_\Lambda, \Omega_K, H_0$, etc.) that govern the background Friedmann–Lemaître expansion and dictate how perturbations evolve and project onto the sky from the last scattering surface to the observer today.

7.1.1 The Transfer Function Formalism

If we make no prior assumptions or theoretical hypotheses about the primordial initial fluctuations, an observed CMB spectrum could be reproduced by an almost arbitrary choice of cosmological parameters.

Under the assumptions of statistical homogeneity and spatial isotropy, linear perturbation theory ensures that each spatial Fourier mode $\boldsymbol{k}$ evolves independently. For an arbitrary initial power spectrum $P_m(k)$ of scalar ($m = 0$), vector ($m = \pm 1$), and tensor ($m = \pm 2$) perturbations, the resulting CMB angular power spectrum takes the generic form:

$$C_\ell = \sum_{m=-2}^{2} \int_0^\infty \mathrm{d}k \, T_m(\ell, k) \, P_m(k), \tag{7.1.1}$$

where $T_m(\ell, k)$ is the CMB **transfer function**, which is entirely determined by the background cosmological parameters.

Consequently, for a nearly arbitrary transfer function $T_m(\ell, k)$ and arbitrary observed multipoles C_ℓ, one could in principle find a custom primordial spectrum $P_m(k)$ such that Eq. (7.1.1) is satisfied.

Derivation: Line-of-Sight Transfer Function Formulation
To trace how Eq. (7.1.1) emerges from the underlying relativistic Boltzmann hierarchy, recall that in the line-of-sight integration formulation, the harmonic expansion coefficients $a_{\ell m}$ of the temperature anisotropy field are expressed in terms of the primordial curvature perturbation $\mathcal{R}(\boldsymbol{k})$ as:

$$a_{\ell m} = 4\pi(-i)^\ell \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3}\, \mathcal{R}(\boldsymbol{k})\, \Delta_\ell(k)\, Y^*_{\ell m}(\hat{\boldsymbol{k}}), \tag{1}$$

where $\Delta_\ell(k)$ is the line-of-sight radiation transfer function:

$$\Delta_\ell(k) = \int_0^{\eta_0} \mathrm{d}\eta\, S(k,\eta)\, j_\ell(k(\eta_0 - \eta)), \tag{2}$$

with $S(k,\eta)$ denoting the photon source term (comprising the Sachs–Wolfe potential, Doppler velocities, and the Integrated Sachs–Wolfe effect) and $j_\ell(x)$ the spherical Bessel function.
Assuming statistical isotropy for the primordial field,

$$\langle \mathcal{R}(\boldsymbol{k})\mathcal{R}^*(\boldsymbol{k}')\rangle = (2\pi)^3 \frac{P_\mathcal{R}(k)}{4\pi k^2}\, \delta^{(3)}(\boldsymbol{k} - \boldsymbol{k}'), \tag{3}$$

the ensemble average of the multipole moments evaluates to:

$$\begin{aligned}
\langle a_{\ell m} a^*_{\ell' m'}\rangle &= (4\pi)^2(-i)^\ell(i)^{\ell'} \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} \int \frac{\mathrm{d}^3\boldsymbol{k}'}{(2\pi)^3} \langle \mathcal{R}(\boldsymbol{k})\mathcal{R}^*(\boldsymbol{k}')\rangle\, \Delta_\ell(k)\Delta^*_{\ell'}(k')\, Y^*_{\ell m}(\hat{\boldsymbol{k}}) Y_{\ell' m'}(\hat{\boldsymbol{k}}') \\
&= (4\pi)^2 \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} \frac{P_\mathcal{R}(k)}{4\pi k^2} |\Delta_\ell(k)|^2\, Y^*_{\ell m}(\hat{\boldsymbol{k}}) Y_{\ell' m'}(\hat{\boldsymbol{k}}) \\
&= \frac{2}{\pi} \int_0^\infty \mathrm{d}k\, P_\mathcal{R}(k)\, |\Delta_\ell(k)|^2 \underbrace{\int_{S^2} Y^*_{\ell m}(\hat{\boldsymbol{k}}) Y_{\ell' m'}(\hat{\boldsymbol{k}})\, \mathrm{d}\Omega_{\hat{\boldsymbol{k}}}}_{\delta_{\ell\ell'}\delta_{mm'}} \\
&= \left(\frac{2}{\pi} \int_0^\infty \mathrm{d}k\, P_\mathcal{R}(k)\, |\Delta_\ell(k)|^2\right) \delta_{\ell\ell'}\delta_{mm'}.
\end{aligned} \tag{4}$$

Comparing with $\langle a_{\ell m} a^*_{\ell' m'}\rangle = C_\ell \delta_{\ell\ell'}\delta_{mm'}$, we identify the scalar transfer function:

$$T_0(\ell, k) = \frac{2}{\pi} |\Delta_\ell(k)|^2. \tag{5}$$

Since primordial vector perturbations decay as the Universe expands ($v_{\mathrm{vec}} \propto a^{-1}$ due to angular momentum conservation), the vector transfer function $T_{\pm 1}(\ell, k)$ is negligible in linear perturbation theory. Unless seeded continuously by external sources, vector perturbations do not contribute significantly to the CMB spectrum and are neglected in standard analyses.

An exception arises in the presence of **sources** (active seeds such as cosmic strings or topological defects), where inhomogeneous and anisotropic energy–momentum components continuously generate perturbations at late times. In that setting, vector perturbations can play an important role (see Section 7.8).

7.1.2 Parameter Measurement vs. Estimation

When determining cosmological parameters from the CMB, we are not "measuring" them in the classical laboratory sense, but rather **estimating** them under restrictive, physically motivated hypotheses regarding the initial conditions.

In a laboratory experiment, prior knowledge is routinely invoked (such as assuming the measuring apparatus obeys Maxwell's equations), but the apparatus has been independently calibrated. In cosmology, however, the CMB and large-scale structure surveys represent our primary empirical window into both the background expansion and the primordial initial conditions.

In an idealized setting, one might measure cosmological parameters via independent astrophysical means, and then, using the well-determined transfer functions $T_m(\ell, k)$, invert Eq. (7.1.1) to reconstruct the primordial power spectra $P_m(k)$, thereby directly probing high-energy physics and inflation near the Planck scale.

In practice, because the CMB provides the cleanest and most accurately modeled dataset, we use it simultaneously to constrain both the background parameters and the primordial power spectrum. To make this tractable, we parameterize the primordial initial spectrum using a small number of parameters—typically a power-law form predicted by slow-roll inflation:

$$k^3 P_0(k) = A_\mathrm{s} \left(\frac{k}{H_0}\right)^{n_\mathrm{s}-1}, \qquad P_{\pm 2}(k) = 0 \quad \text{or} \quad A_\mathrm{T} \left(\frac{k}{H_0}\right)^{n_\mathrm{T}}, \tag{7.1.2}$$

specified by the scalar amplitude A_s, spectral tilt n_s, and tensor-to-scalar ratio r.

Although an observed power spectrum provides hundreds of measured multipoles C_ℓ, the spectrum is a smooth function determined by a modest number of physical mechanisms. Consequently, only a small set of parameter combinations can be constrained independently, giving rise to **parameter degeneracies** (discussed in Section 7.6).

7.2 The physics of parameter dependence

To understand which parameters can be estimated with high precision and which suffer from degeneracies, we examine how cosmological parameters physically alter the acoustic peaks, damping tails, and large-angle plateaus of the CMB power spectrum.

7.2.1 The acoustic peaks

Before recombination, photons and baryons are tightly coupled via Thomson scattering (low energy limit of Compton scattering), forming a single relativistic fluid that oscillates acoustically inside gravitational potential wells.

The first acoustic peak corresponds to a perturbation mode with comoving wavelength $\lambda_1 = \pi/k_1$ that has undergone exactly one maximal compression since entering the horizon, reaching maximum amplitude at decoupling. For adiabatic initial perturbations, this scale satisfies:

$$c_\mathrm{s} k_1 \eta_\mathrm{dec} = \pi \qquad \text{(adiabatic perturbations)}, \tag{7.2.1}$$

whereas for isocurvature perturbations, the initial condition starts with vanishing density fluctuation, requiring:

$$c_\mathrm{s} k_1 \eta_\mathrm{dec} = \frac{3\pi}{2} \qquad \text{(isocurvature perturbations)}. \tag{7.2.2}$$

Subsequent acoustic peaks correspond to higher harmonic modes undergoing successive compressions and rarefactions:

$$c_\mathrm{s} k_n \eta_\mathrm{dec} = n\pi \quad \text{(adiabatic)}, \qquad c_\mathrm{s} k_n \eta_\mathrm{dec} = \frac{(2n+1)\pi}{2} \quad \text{(isocurvature)}. \tag{7.2.3}$$

Angular Diameter Distance and Peak Multipole Positions

The angle θ_n subtended on the sky by the physical wavelength λ_n on the last-scattering surface is given by:

$$\theta_n \simeq \frac{\lambda_n}{D_{\rm A}(z_{\rm dec})}, \qquad \ell_n \simeq \frac{\pi}{\theta_n}, \tag{7.2.4}$$

where $D_{\rm A}(z)$ is the angular diameter distance to redshift z, and $\ell(\theta) \approx \pi/\theta$ relates angular scales to spherical harmonic multipoles.

The physical sound horizon at decoupling, $r_{\rm s}(\eta_{\rm dec})$, is given by:

$$r_{\rm s}(\eta_{\rm dec}) = \frac{1}{H_0}\int_{z_{\rm dec}}^{\infty} \frac{c_{\rm s}(z)\,{\rm d}z}{[\Omega_{\rm r}(z+1)^4 + \Omega_{\rm m}(z+1)^3 + \Omega_\Lambda + \Omega_K(z+1)^2]^{1/2}}, \tag{7.2.5}$$

and the angular diameter distance to the decoupling epoch is:

$$D_{\rm A}(z_{\rm dec}) = \frac{a_0}{z_{\rm dec}+1}\,\chi_K\left(\frac{1}{H_0 a_0}\int_0^{z_{\rm dec}} \frac{{\rm d}z}{[\Omega_{\rm r}(z+1)^4 + \Omega_{\rm m}(z+1)^3 + \Omega_\Lambda + \Omega_K(z+1)^2]^{1/2}}\right). \tag{7.2.6}$$

Combining these relations with the comoving wavelength $\lambda_n = \frac{\pi}{k_n} = \frac{r_{\rm s}(\eta_{\rm dec})}{n}$, we find:

$$\ell_n \simeq \frac{\pi D_{\rm A}(z_{\rm dec})}{\lambda_n} = \frac{n\pi\,\chi_K\left(\frac{1}{H_0 a_0}\int_0^{z_{\rm dec}} \frac{{\rm d}z}{[\Omega_{\rm r}(z+1)^4+\Omega_{\rm m}(z+1)^3+\Omega_\Lambda+\Omega_K(z+1)^2]^{1/2}}\right)}{\frac{1}{z_{\rm dec}+1}\frac{1}{H_0}\int_{z_{\rm dec}}^{\infty} \frac{c_{\rm s}(z)\,{\rm d}z}{[\Omega_{\rm r}(z+1)^4+\Omega_{\rm m}(z+1)^3+\Omega_\Lambda+\Omega_K(z+1)^2]^{1/2}}}. \tag{7.2.7}$$

To understand the sign flip between the dashed and dotted curves in Fig. 7.1, consider the logarithmic derivative of $\ell_1 \approx \pi D_{\rm A}(z_*)/r_{\rm s}(z_*)$:

$$\frac{\partial \ln \ell_1}{\partial \Omega_\Lambda} = \frac{\partial \ln(H_0 D_{\rm A})}{\partial \Omega_\Lambda} - \frac{\partial \ln(H_0 r_{\rm s})}{\partial \Omega_\Lambda}. \tag{1}$$

1. Fixed h Case ($\Omega_{\rm m} = 1 - \Omega_\Lambda$): The sound horizon is set by high-z physics ($z > 1000$) where dark energy is suppressed by $(1+z)^{-3} \sim 10^{-9}$, so $\partial(H_0 r_{\rm s})/\partial\Omega_\Lambda \approx 0$. The dimensionless angular diameter distance is:

$$H_0 D_{\rm A} = \frac{1}{1+z_*}\int_0^{z_*} \frac{{\rm d}z}{\sqrt{(1-\Omega_\Lambda)(1+z)^3 + \Omega_\Lambda}}. \tag{2}$$

Differentiating with respect to Ω_Λ:

$$\frac{\partial(H_0 D_{\rm A})}{\partial \Omega_\Lambda} = \frac{1}{2(1+z_*)}\int_0^{z_*} \frac{(1+z)^3 - 1}{[(1-\Omega_\Lambda)(1+z)^3 + \Omega_\Lambda]^{3/2}}\,{\rm d}z > 0. \tag{3}$$

Since $(1+z)^3 > 1$ for all $z > 0$, the integrand is strictly positive, giving $\frac{\partial \ln \ell_1}{\partial \Omega_\Lambda}\big|_h > 0 \implies \ell_1$ shifts to higher multipoles.

2. Fixed $\omega_{\rm m} \equiv \Omega_{\rm m} h^2$ Case ($h = \sqrt{\omega_{\rm m}/(1-\Omega_\Lambda)}$): As Ω_Λ increases, h grows as $(1-\Omega_\Lambda)^{-1/2}$. In dimensionless units, the sound horizon is $H_0 r_{\rm s} = (100\,h\,{\rm km/s/Mpc}) r_{\rm s} \propto h \propto (1-\Omega_\Lambda)^{-1/2}$. Therefore:

$$\frac{\partial \ln(H_0 r_{\rm s})}{\partial \Omega_\Lambda} \approx \frac{1}{2(1-\Omega_\Lambda)}. \tag{4}$$

For standard concordance values ($\Omega_\Lambda \approx 0.7$), $\frac{1}{2(1-\Omega_\Lambda)} \approx 1.67$, which exceeds $\frac{\partial \ln(H_0 D_{\rm A})}{\partial \Omega_\Lambda} \approx 0.72$. Thus:

$$\left.\frac{\partial \ln \ell_1}{\partial \Omega_\Lambda}\right|_{\omega_{\rm m}} \approx 0.72 - 1.67 = -0.95 < 0, \tag{5}$$

analytically proving why ℓ_1 shifts to lower multipoles with increasing Ω_Λ when ω_m is held fixed!

Here, the spatial curvature function $\chi_K(r)$ is defined by:

$$\chi_K(r) = \begin{cases} r, & K = 0 \quad \text{(flat space)}, \\ \frac{1}{\sqrt{K}} \sin\!\left(\sqrt{K}\, r\right), & K > 0 \quad \text{(closed space)}, \\ \frac{1}{\sqrt{|K|}} \sinh\!\left(\sqrt{|K|}\, r\right), & K < 0 \quad \text{(open space)}. \end{cases} \tag{7.2.8}$$

Via the radial function χ_K, the multipole positions ℓ_n depend very sensitively on the spatial curvature Ω_K. Because curvature and dark energy are negligible at high redshifts ($z > z_\mathrm{dec} \approx 1090$), the denominator of Eq. (7.2.7) is almost completely independent of Ω_K and Ω_Λ; it depends predominantly on $\Omega_\mathrm{r}, \Omega_\mathrm{m}$, and the baryon density through the sound speed $c_\mathrm{s}(z)$:

$$c_\mathrm{s}^2 = \frac{1}{3} \frac{4\Omega_\gamma h^2}{4\Omega_\gamma h^2 + 3\Omega_\mathrm{b} h^2 (1+z)^{-1}}, \tag{7.2.9}$$

where $h = H_0/(100\,\mathrm{km\,s^{-1}\,Mpc^{-1}})$. Because the physical photon density $\Omega_\gamma h^2 \propto T_0^4$ is determined with extreme precision from the FIRAS measurement of the CMB monopole ($T_0 = 2.7255\,\mathrm{K}$), measuring the sound speed directly measures the physical baryon density $\Omega_\mathrm{b} h^2$.

In the tightly coupled photon–baryon fluid, photon radiation pressure provides the restoring force ($p = \frac{1}{3}\rho_\gamma c^2$), while both photons and non-relativistic baryons contribute to fluid inertia ($\rho = \rho_\gamma + \rho_\mathrm{b}$).
Differentiating pressure and density with respect to the scale factor a:

$$\frac{\mathrm{d}p}{\mathrm{d}a} = \frac{1}{3} c^2 \frac{\mathrm{d}\rho_\gamma}{\mathrm{d}a} = -\frac{4}{3} c^2 \frac{\rho_\gamma}{a}, \qquad \frac{\mathrm{d}\rho}{\mathrm{d}a} = \frac{\mathrm{d}\rho_\gamma}{\mathrm{d}a} + \frac{\mathrm{d}\rho_\mathrm{b}}{\mathrm{d}a} = -\frac{4\rho_\gamma + 3\rho_\mathrm{b}}{a}. \tag{1}$$

Taking the ratio $\mathrm{d}p/\mathrm{d}\rho$:

$$c_\mathrm{s}^2 = \frac{\mathrm{d}p/\mathrm{d}a}{\mathrm{d}\rho/\mathrm{d}a} = \frac{c^2}{3\left(1 + \frac{3\rho_\mathrm{b}}{4\rho_\gamma}\right)} = \frac{c^2}{3\left(1 + \frac{3\Omega_\mathrm{b} h^2}{4\Omega_\gamma h^2 (1+z)}\right)} = \frac{1}{3} \frac{4\Omega_\gamma h^2}{4\Omega_\gamma h^2 + 3\Omega_\mathrm{b} h^2 (1+z)^{-1}}. \tag{2}$$

In Fig. 7.1, we show how the position of the first peak ℓ_1 varies with cosmological parameters. Crucially, the response depends on which parameters are held constant:

- **Varying Ω_Λ at fixed h (dashed line):** As Ω_Λ increases, $\Omega_\mathrm{m} = 1 - \Omega_\Lambda$ decreases. The reduced matter density accelerates late-time expansion, increasing the dimensionless angular diameter distance $H_0 D_\mathrm{A}$, while the early-universe sound horizon $H_0 \lambda_1$ is unaffected. Hence, ℓ_1 shifts to higher multipoles (smaller angular scales).
- **Varying Ω_Λ at fixed $\Omega_\mathrm{m} h^2$ (dotted line):** Here $h = \sqrt{\Omega_\mathrm{m} h^2/(1-\Omega_\Lambda)}$ must increase with Ω_Λ. The radiation density parameter $\Omega_\mathrm{r} = \Omega_\mathrm{r} h^2/h^2$ decreases, increasing the comoving sound horizon $H_0 \lambda_1$. This increase overcomes the growth in $H_0 D_\mathrm{A}$, shifting ℓ_1 to *lower* multipoles.

Derivation: Logarithmic Sensitivity of Peak Position $\partial \ln \ell_1 / \partial \Omega_\Lambda$
The acoustic scale multipole is $\ell_1 \approx \pi D_\mathrm{A}(z_*)/r_\mathrm{s}(z_*)$. Taking the logarithmic derivative

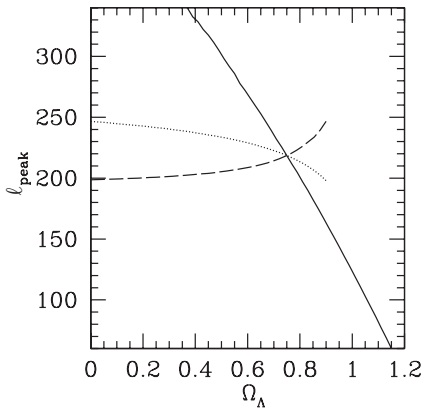


Figure 7.1: The position of the first acoustic peak ℓ_1 as a function of Ω_Λ. Solid line: varying Ω_K while leaving all other parameters fixed. Dashed line: varying Ω_Λ at fixed $h = 0.72$. Dotted line: varying Ω_Λ at fixed $\Omega_{\rm m}h^2$. The fixed baseline parameters are $\Omega_K = 0, h = 0.72, \Omega_{\rm b}h^2 = 0.022, \Omega_{\rm m} = 0.25, n_{\rm s} = 0.96, \tau_{\rm ri} = 0.085$, so that all curves intersect at $\Omega_\Lambda = 0.75$. Notice the pronounced sensitivity of ℓ_1 to spatial curvature.

with respect to Ω_Λ:

$$\frac{\partial \ln \ell_1}{\partial \Omega_\Lambda} = \frac{\partial \ln(H_0 D_{\rm A})}{\partial \Omega_\Lambda} - \frac{\partial \ln(H_0 r_{\rm s})}{\partial \Omega_\Lambda}. \tag{1}$$

1. **Fixed h:** The sound horizon $r_{\rm s}$ is established at $z \sim 1090$, where $\Omega_\Lambda(1+z)^{-3} \ll 1$, so $\partial \ln(H_0 r_{\rm s})/\partial\Omega_\Lambda \approx 0$. For a flat universe with $\Omega_{\rm m} = 1 - \Omega_\Lambda$:

$$H_0 D_{\rm A} = \frac{1}{1+z_*} \int_0^{z_*} \frac{{\rm d}z}{\sqrt{(1-\Omega_\Lambda)(1+z)^3 + \Omega_\Lambda}}. \tag{2}$$

   Differentiating the integrand:

$$\frac{\partial (H_0 D_{\rm A})}{\partial \Omega_\Lambda} = \frac{1}{2(1+z_*)} \int_0^{z_*} \frac{(1+z)^3 - 1}{\left[(1-\Omega_\Lambda)(1+z)^3 + \Omega_\Lambda\right]^{3/2}} \, {\rm d}z > 0, \tag{3}$$

   since $(1+z)^3 > 1$ for all $z > 0$. Thus $\partial \ln \ell_1/\partial\Omega_\Lambda > 0$, shifting ℓ_1 upward.

2. **Fixed $\omega_{\rm m} \equiv \Omega_{\rm m}h^2$:** Since $h = \omega_{\rm m}^{1/2}(1-\Omega_\Lambda)^{-1/2}$, the dimensionless quantity $H_0 r_{\rm s} \propto h \propto (1-\Omega_\Lambda)^{-1/2}$, giving:

$$\frac{\partial \ln(H_0 r_{\rm s})}{\partial \Omega_\Lambda} = \frac{1}{2(1-\Omega_\Lambda)}. \tag{4}$$

   For $\Omega_\Lambda \approx 0.7$, $\frac{1}{2(1-\Omega_\Lambda)} \approx 1.67$, which is larger than $\frac{\partial \ln(H_0 D_{\rm A})}{\partial \Omega_\Lambda} \approx 0.72$. Therefore:

$$\left.\frac{\partial \ln \ell_1}{\partial \Omega_\Lambda}\right|_{\omega_{\rm m}} \approx 0.72 - 1.67 = -0.95 < 0, \tag{5}$$

   explaining analytically why ℓ_1 decreases when $\Omega_{\rm m}h^2$ is kept fixed.

Matter Density, Odd/Even Peak Asymmetry, and Damping

During the radiation-dominated era, dark matter perturbations grow only logarithmically ($\delta_{\rm cdm} \propto \ln a$) due to the Meszaros effect. Once the Universe enters the matter-dominated epoch,

they grow linearly with the scale factor ($\delta_{\rm cdm} \propto a$). Consequently, the gravitational potential Φ at decoupling is sensitive to the epoch of matter-radiation equality, determined by $\Omega_{\rm m}h^2$. In particular, the ratio of the height of the first acoustic peak to the low-ℓ Sachs–Wolfe plateau provides a direct measure of $\Omega_{\rm m}h^2$.

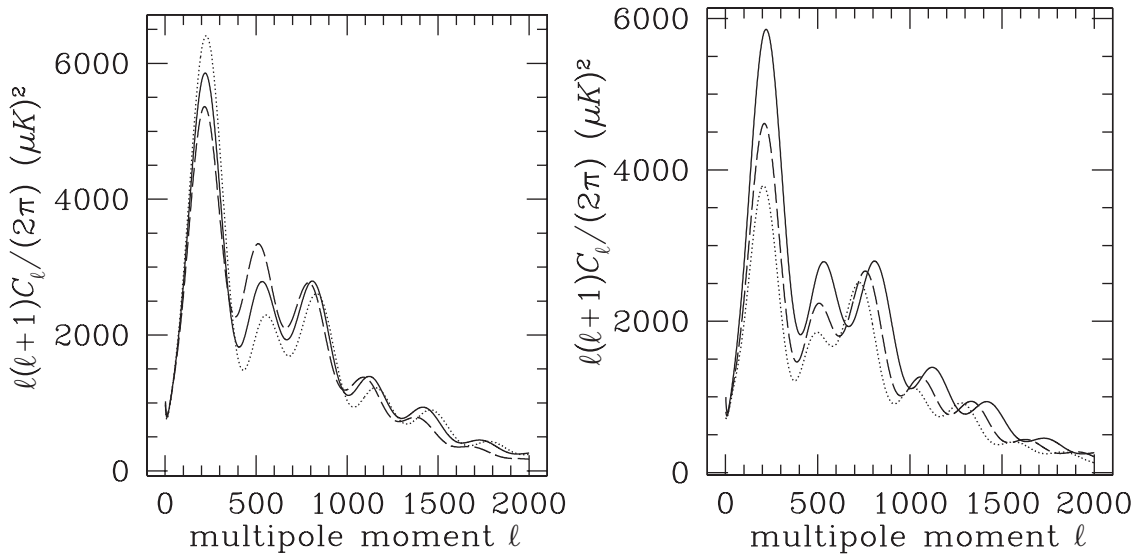


Figure 7.2: Left: The asymmetry of even and odd acoustic peaks and its dependence on $\Omega_{\rm b}h^2$. The temperature power spectrum is plotted for $\Omega_{\rm b}h^2 = 0.02$ (solid line), $\Omega_{\rm b}h^2 = 0.03$ (dotted line), and $\Omega_{\rm b}h^2 = 0.01$ (dashed line). Right: Fixing $\Omega_{\rm b}h^2 = 0.02$ and varying the physical matter density: $\Omega_{\rm m}h^2 = 0.12$ (solid), $\Omega_{\rm m}h^2 = 0.20$ (dashed), and $\Omega_{\rm m}h^2 = 0.30$ (dotted). Higher $\Omega_{\rm m}h^2$ increases peak asymmetry, while smaller $\Omega_{\rm m}h^2$ boosts the first peak via the early integrated Sachs–Wolfe effect.

The baryon density enters the CMB spectrum through the characteristic height asymmetry of odd and even acoustic peaks (Fig. 7.2):

- **Odd peaks (1st, 3rd, 5th):** Represent maximal gravitational *compressions* (overdensities) inside potential wells.
- **Even peaks (2nd, 4th, 6th):** Represent maximal *rarefactions* (underdensities).

If the oscillating fluid consisted purely of massless photons, the oscillations would be symmetric about zero, producing equal amplitudes for compressions and rarefactions. The inclusion of massive baryons provides extra gravitational attraction (baryon loading) without contributing pressure. This shifts the zero-point of the oscillation, enhancing compressions (odd peaks) while suppressing rarefactions (even peaks).

On small angular scales ($\ell \gtrsim 1000$), fluctuation amplitudes decay exponentially due to **Silk damping** (photon diffusion during recombination). The damping scale depends inversely on the electron density, making the damping tail another sensitive probe of $\Omega_{\rm b}h^2$.

7.2.2 Neutrinos

Neutrinos decouple from thermal equilibrium when the temperature drops to $T_\nu \sim 1.4\,\mathrm{MeV}$ ($z_\nu \sim 0.6 \times 10^{10}$), after which weak interactions become too slow compared to the cosmic expansion rate.

After decoupling, neutrinos propagate freely, described by the relativistic Liouville equation.

- **Massless Neutrinos:** Free stream across horizons, developing anisotropic stress and diluting as radiation ($\rho_\nu \propto a^{-4}$).

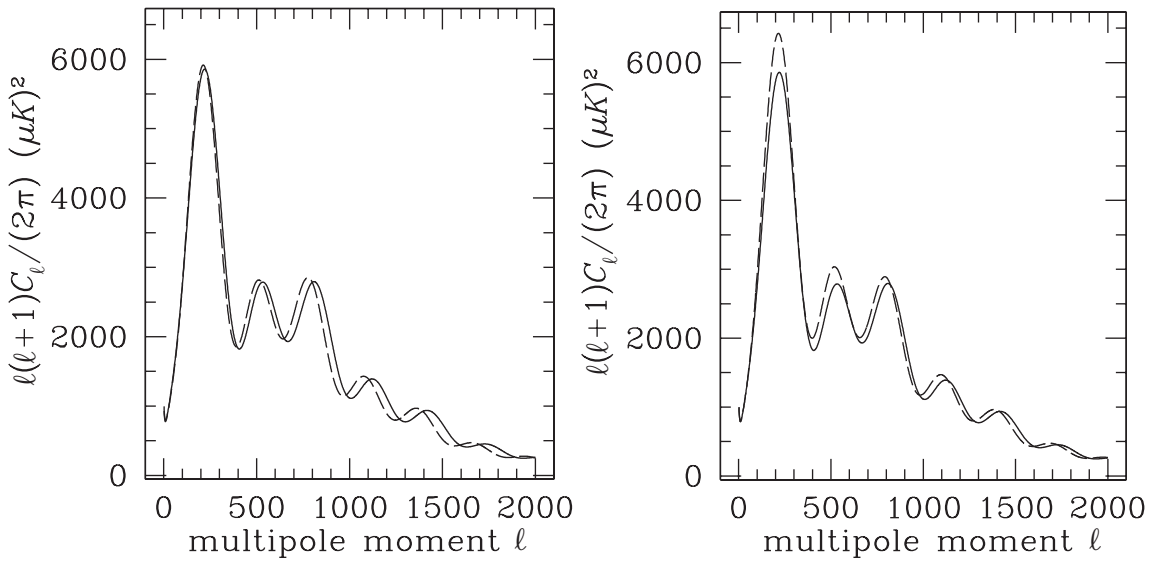


Figure 7.3: The CMB power spectrum for massless neutrinos and massive neutrinos with mass $m_\nu = 2\,\mathrm{eV}$ (dashed lines). Left: Cold dark matter density $\Omega_{\rm cdm}h^2 = 0.12$ is fixed. Right: Total matter density $\Omega_{\rm m}h^2 = 0.144$ is fixed. In all curves, $\Omega_{\rm tot} = 1, \Omega_{\rm b}h^2 = 0.022$, and $h = 0.7$. At fixed $\Omega_{\rm m}h^2$, massive neutrinos mimic a lower matter density because they are not yet fully non-relativistic at decoupling.

- **Massive Neutrinos:** As the Universe cools below their rest mass ($T_\nu < m_\nu$), their thermal velocities drop, anisotropic stress decays, and they transition to behaving as non-relativistic matter ($\rho_\nu \propto a^{-3}$).

This transition alters the CMB power spectrum (Fig. 7.3). However, because neutrino anisotropic stresses contribute only $\sim 5\%$ to CMB temperature fluctuations, and massive neutrinos mimic cold dark matter on large scales, the sensitivity of primary temperature anisotropies to neutrino mass is moderate.

Derivation: Massive Neutrino Density Parameter and Power Suppression
The present-day number density of a relativistic Fermi–Dirac species with $g_\nu = 2$ internal degrees of freedom is:

$$n_{\nu,0} = \frac{g_\nu}{(2\pi)^3}\int \frac{\mathrm{d}^3\boldsymbol{p}}{\mathrm{e}^{p/T_{\nu,0}}+1} = \frac{3\zeta(3)}{4\pi^2}T_{\nu,0}^3 \approx 112\,\mathrm{cm}^{-3} \quad \text{(per flavor)}, \tag{1}$$

where $T_{\nu,0} = (4/11)^{1/3}T_{\gamma,0} \approx 1.945\,\mathrm{K}$. If the neutrinos are non-relativistic today ($m_\nu \gg T_{\nu,0} \approx 1.68\times10^{-4}\,\mathrm{eV}$), their mass density is $\rho_\nu = \sum m_\nu n_{\nu,0}$, giving the density parameter:

$$\Omega_\nu h^2 = \frac{\sum m_\nu}{93.14\,\mathrm{eV}}. \tag{2}$$

On scales smaller than their free-streaming wavenumber $k_{\rm fs} \approx 0.018\sqrt{\Omega_{\rm m}}(m_\nu/1\,\mathrm{eV})^{1/2}h\,\mathrm{Mpc}^{-1}$, neutrino perturbations do not cluster, slowing the growth rate of matter fluctuations from $\delta_{\rm m} \propto a$ to $\delta_{\rm m} \propto a^{1-\frac{3}{5}f_\nu}$ with $f_\nu \equiv \Omega_\nu/\Omega_{\rm m}$. Over the matter era, this leads to the well-known suppression in the matter power spectrum:

$$\frac{\Delta P_{\rm m}(k)}{P_{\rm m}(k)} \approx -8\,\frac{\Omega_\nu}{\Omega_{\rm m}}. \tag{3}$$

7.2.3 Gravitational waves

CMB anisotropies generated by primordial tensor perturbations (gravitational waves) are significant only on angular scales that were super-Hubble before decoupling, corresponding to low multipoles:

$$\ell \lesssim 80. \tag{7.2.10}$$

On smaller scales ($\ell \gg 80$), tensor modes enter the horizon during the radiation- or early matter-dominated era and decay rapidly.

On the Sachs–Wolfe plateau, a small tensor contribution can be difficult to distinguish from a slightly red scalar spectral index ($n_\mathrm{s} < 1$), which also increases low-ℓ power relative to the acoustic peaks. However, scalar perturbations produce only gradient-type (E-mode) polarization due to parity symmetry, whereas tensor perturbations generate both E-mode and curl-type (B-mode) polarization. The detection of primordial B-mode polarization at $\ell \sim 80$ is therefore a definitive signature of inflationary gravitational waves.

7.3 Reionization

Although the Universe recombined and became neutral at $z_\mathrm{rec} \approx 1400$ ($z_* \approx 1090$), observations reveal that the diffuse intergalactic medium (IGM) today contains virtually no neutral hydrogen; it is almost completely ionized.

7.3.1 The Gunn–Peterson Trough

This conclusion is established by the absence of the **Gunn–Peterson trough** in quasar spectra. Quasars are highly luminous active galactic nuclei observable at high redshifts ($z \lesssim 7$). Gunn and Peterson (1965) demonstrated that even a tiny trace of diffuse neutral hydrogen would produce resonant Lyman-α absorption ($\lambda_\alpha = 121.6\,\mathrm{nm}$) for all emitted photons redshifting through the Ly-α resonance along the line of sight.

Evaluating the Ly-α optical depth implies a neutral hydrogen density in the diffuse IGM of:

$$\Omega_\mathrm{HI} h \lesssim 10^{-8}. \tag{7.3.1}$$

While discrete neutral clouds produce the "Lyman-α forest" of absorption lines, integrating the forest's total optical depth yields an average neutral fraction of only $\Omega_\mathrm{HI} h \simeq 10^{-5}$. Since galaxy formation could not have swept up 99.9% of all cosmic gas into stars, the intergalactic hydrogen must have been reionized.

In very high-redshift quasars ($z > 6$), complete Gunn–Peterson absorption troughs are observed, demonstrating that reionization was completing around $z_\mathrm{ri} \approx 6$, driven by ionizing UV photons from the earliest stars and galaxies.

7.3.2 Optical Depth to Reionization

Following reionization, CMB photons can Thomson-scatter once again off the liberated free electrons. Because the expanding Universe has diluted the electron density, the total scattering optical depth τ_ri is small (~ 5–10%):

$$\tau_\mathrm{ri}(z_\mathrm{ri}) = \sigma_\mathrm{T} \int_{t(z_\mathrm{ri})}^{t_0} c\, n_\mathrm{e}(t)\, \mathrm{d}t = \sigma_\mathrm{T} \int_0^{z_\mathrm{ri}} \frac{c\,(1+z)^2\, n_\mathrm{e}(z)}{H(z)}\, \mathrm{d}z. \tag{7.3.2}$$

Derivation: Suppression Factor $\mathrm{e}^{-2\tau_\mathrm{ri}}$ and Large-Scale Polarization
The probability that a CMB photon travels from the last scattering surface to the observer without undergoing a secondary scattering is $\mathrm{e}^{-\tau_\mathrm{ri}}$. The unscattered temperature

fluctuation is therefore:

$$\Delta T_{\rm unscattered}(\hat{\boldsymbol{n}}) = \mathrm{e}^{-\tau_{\rm ri}} \Delta T_{\rm prim}(\hat{\boldsymbol{n}}). \tag{1}$$

The scattered fraction $(1 - \mathrm{e}^{-\tau_{\rm ri}})$ has its direction randomized over the horizon scale at reionization ($\theta_{\rm ri} \sim 10^\circ \implies \ell_{\rm ri} \sim 10$). On small scales ($\ell \gg 10$), these scattered photons become completely uncorrelated with the primary field, washing out small-scale acoustic peak power by:

$$C_\ell = \mathrm{e}^{-2\tau_{\rm ri}} C_\ell^{\rm prim} \qquad (\ell \gg 10). \tag{2}$$

On large scales ($\ell \leq 10$), rescattering of the quadrupole anisotropy generates a distinctive bump in the E-mode polarization power spectrum with amplitude $C_\ell^{EE} \propto \tau_{\rm ri}^2$, breaking the degeneracy between the primordial amplitude $A_{\rm s}$ and the optical depth $\tau_{\rm ri}$ (the $A_{\rm s}\mathrm{e}^{-2\tau_{\rm ri}}$ degeneracy).

7.4 CMB data

Having established the theoretical framework, we now examine how observational data are modeled and analyzed to constrain cosmological parameters.

7.4.1 The Data Modeling Paradigm: Signal, Noise, and Likelihood

Parameter estimation in observational cosmology begins with a clear model of how measurements are produced. Any measured data vector $\boldsymbol{d}$ can be decomposed into an astrophysical signal $\boldsymbol{s}$ and observational noise $\boldsymbol{n}$:

$$\boldsymbol{d} = \boldsymbol{s} + \boldsymbol{n}. \tag{7.4.1}$$

Here, the noise $\boldsymbol{n}$ accounts for instrumental and environmental uncertainties (such as detector thermal noise, cryogenic fluctuations, and atmospheric emission). In well-characterized experiments, the noise has zero mean, $\langle \boldsymbol{n} \rangle = 0$, and is modeled as a multivariate Gaussian with noise covariance matrix $\boldsymbol{N} \equiv \langle \boldsymbol{n}\boldsymbol{n}^\top \rangle$:

$$P(\boldsymbol{n}) = \frac{1}{|2\pi \boldsymbol{N}|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{n}^\top \boldsymbol{N}^{-1} \boldsymbol{n}\right). \tag{7.4.2}$$

The mathematical form of the likelihood function $P(\boldsymbol{d} \mid \boldsymbol{\theta})$ depends fundamentally on the physical nature of the signal $\boldsymbol{s}$, which falls into two distinct regimes across cosmology:

1. Deterministic signals (mean-shifting likelihood). When observing cosmological background evolution—such as Type Ia supernova distance moduli $\mu(z)$ or cosmic chronometer expansion rates $H(z)$—cosmological theory predicts a smooth, deterministic curve $\boldsymbol{\mu}(\boldsymbol{\theta})$ for a chosen set of parameters $\boldsymbol{\theta}$. The underlying signal is not a random variable; rather, $\boldsymbol{s} = \boldsymbol{\mu}(\boldsymbol{\theta})$ is completely determined by the model.

In this regime, the quantity that fluctuates stochastically from measurement to measurement is the residual between the data and the theoretical prediction:

$$\boldsymbol{d} - \boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{n}. \tag{7.4.3}$$

Because the noise $\boldsymbol{n}$ follows Eq. (7.4.2), the residual is zero-mean Gaussian distributed. Equivalently, the data vector $\boldsymbol{d}$ is Gaussian distributed about the theoretical mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ with fixed noise covariance $\boldsymbol{N}$:

$$P(\boldsymbol{d} \mid \boldsymbol{\theta}) = \frac{1}{|2\pi \boldsymbol{N}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{d} - \boldsymbol{\mu}(\boldsymbol{\theta}))^\top \boldsymbol{N}^{-1} (\boldsymbol{d} - \boldsymbol{\mu}(\boldsymbol{\theta}))\right). \tag{7.4.4}$$

Here, the cosmological parameters $\boldsymbol{\theta}$ shift the expected *mean* of the distribution, while the covariance $\boldsymbol{N}$ is fixed by the instrument.

2. Stochastic signals (variance-shifting likelihood, e.g., the CMB). When mapping the Cosmic Microwave Background, our telescopes record primordial temperature and polarization fluctuations across the sky. Standard inflationary cosmology does not predict the exact temperature fluctuation at any specific pixel. Instead, primordial perturbations are quantum fluctuations stretched by inflation into a statistically isotropic, Gaussian random field.

Thus, the true sky signal $\boldsymbol{s}$ is itself a zero-mean random vector, $\langle \boldsymbol{s} \rangle = 0$. Cosmological parameters $\boldsymbol{\theta}$ do not determine individual pixel values; rather, they determine the signal covariance matrix:

$$\boldsymbol{S}(\boldsymbol{\theta}) \equiv \langle \boldsymbol{s}\boldsymbol{s}^\top \rangle, \tag{7.4.5}$$

which is computed from the theoretical angular power spectrum $C_\ell(\boldsymbol{\theta})$ and the telescope beam profile (which accounts for the detector's angular sensitivity and beam smoothing).

Since the observed map $\boldsymbol{d} = \boldsymbol{s} + \boldsymbol{n}$ is the sum of two independent, zero-mean Gaussian random vectors (the cosmological signal $\boldsymbol{s}$ and the detector noise $\boldsymbol{n}$), the resulting data vector $\boldsymbol{d}$ is also zero-mean Gaussian distributed:

$$\langle \boldsymbol{d} \rangle = \langle \boldsymbol{s} \rangle + \langle \boldsymbol{n} \rangle = 0, \tag{7.4.6}$$

$$\boldsymbol{C}(\boldsymbol{\theta}) \equiv \langle \boldsymbol{d}\boldsymbol{d}^\top \rangle = \boldsymbol{S}(\boldsymbol{\theta}) + \boldsymbol{N}. \tag{7.4.7}$$

The probability distribution of the observed data given parameters $\boldsymbol{\theta}$ is therefore:

$$P(\boldsymbol{d} \mid \boldsymbol{\theta}) = \frac{1}{|2\pi \boldsymbol{C}(\boldsymbol{\theta})|^{1/2}} \exp\left(-\frac{1}{2} \boldsymbol{d}^\top \boldsymbol{C}^{-1}(\boldsymbol{\theta}) \boldsymbol{d} \right). \tag{7.4.8}$$

In this case, cosmological parameters do not alter the mean of the data (which remains identically zero); instead, they enter exclusively through the total covariance matrix $\boldsymbol{C}(\boldsymbol{\theta})$.

The likelihood function. For a completed experiment, the measured data vector $\boldsymbol{d}$ is fixed and known. What varies across our analysis is the candidate cosmological model, specified by parameters $\boldsymbol{\theta}$, which generates a corresponding candidate probability distribution $P(\boldsymbol{d} \mid \boldsymbol{\theta})$.

Regarding $P(\boldsymbol{d} \mid \boldsymbol{\theta})$ as a function of the parameters $\boldsymbol{\theta}$ for the fixed observed data defines the *likelihood function*:

$$\mathcal{L}(\boldsymbol{\theta}) \equiv P(\boldsymbol{d} \mid \boldsymbol{\theta}). \tag{7.4.9}$$

Conceptually, the likelihood function does not tell us the probability of the parameters (which is the role of the Bayesian posterior); rather, *the likelihood function tells us how likely a candidate probability distribution is to have generated the observed data.* Models that assign a high probability density to the data we actually measured possess a high likelihood, whereas models under which the observed dataset would be an extraordinarily rare fluke have near-zero likelihood. Skimming the mathematical pipeline reveals the fundamental contrast between the two paradigms:

$$\text{Deterministic (Expansion):} \quad \boldsymbol{d} = \boldsymbol{\mu}(\boldsymbol{\theta}) + \boldsymbol{n} \implies \underbrace{\boldsymbol{d} - \boldsymbol{\mu}(\boldsymbol{\theta})}_{\text{residual}} \sim \mathcal{N}(0, \boldsymbol{N}) \implies \mathcal{L}(\boldsymbol{\theta}) \propto \mathrm{e}^{-\frac{1}{2}\left(\boldsymbol{d}-\boldsymbol{\mu}(\boldsymbol{\theta})\right)^\top \boldsymbol{N}^{-1} \left(\boldsymbol{d}-\boldsymbol{\mu}(\boldsymbol{\theta})\right)}, \tag{7.4.10}$$

$$\text{Stochastic (CMB Sky):} \quad \boldsymbol{d} = \boldsymbol{s} + \boldsymbol{n} \implies \begin{cases} \langle \boldsymbol{d} \rangle = 0 \\ \boldsymbol{C}(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{\theta}) + \boldsymbol{N} \end{cases} \implies \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{|2\pi \boldsymbol{C}(\boldsymbol{\theta})|^{1/2}} \mathrm{e}^{-\frac{1}{2}\boldsymbol{d}^\top \boldsymbol{C}^{-1}(\boldsymbol{\theta}) \boldsymbol{d}}. \tag{7.4.11}$$

Distinguishing between mean-shifting deterministic problems and variance-shifting stochastic problems provides the unifying key to understanding cosmological parameter estimation.

7.4.2 Temperature Fluctuation Maps and Correlation Functions

Let $\Delta T_{\rm s}(\boldsymbol{n}) = T_{\rm s}(\boldsymbol{n}) - \bar{T}_{\rm s}$ be the observed temperature fluctuation map across directions $\boldsymbol{n}$, where $\bar{T}_{\rm s}$ is the mean temperature and the subscript 's' denotes signal. The two-point correlation function $\langle \Delta T_{\rm s}(\boldsymbol{n}_1)\Delta T_{\rm s}(\boldsymbol{n}_2)\rangle$ is related to the mean squared temperature difference by:

$$\left\langle \big(T_{\rm s}(\boldsymbol{n}_1) - T_{\rm s}(\boldsymbol{n}_2)\big)^2 \right\rangle = 2\left[\langle \Delta T_{\rm s}^2\rangle - \langle \Delta T_{\rm s}(\boldsymbol{n}_1)\Delta T_{\rm s}(\boldsymbol{n}_2)\rangle\right]. \tag{7.4.12}$$

The brackets $\langle \dots \rangle$ denote an average over all pairs of directions $(\boldsymbol{n}_1, \boldsymbol{n}_2)$ with a fixed opening angle $\cos\theta = \boldsymbol{n}_1 \cdot \boldsymbol{n}_2$. Under statistical isotropy, this spatial average equates to the theoretical ensemble average.

In practice, an astronomical telescope cannot isolate an infinitesimal line of sight. Due to diffraction and the optics of reflectors and feed horns, every detector exhibits a finite *angular sensitivity profile*, described by the instrumental beam profile (or point spread function) $B(\boldsymbol{n}, \boldsymbol{n}')$.

When the telescope points toward a nominal sky direction $\boldsymbol{n}$, it does not measure radiation coming exclusively from $\boldsymbol{n}$. Instead, it collects radiation from a neighborhood of directions $\boldsymbol{n}'$ with an angular sensitivity that is highest along the central optical axis ($\boldsymbol{n}' = \boldsymbol{n}$) and diminishes smoothly away from it. The observed signal $T_{\rm s}(\boldsymbol{n})$ is therefore a sensitivity-weighted average of the true sky temperature $T(\boldsymbol{n}')$ across the beam footprint, given by the spherical convolution:

$$T_{\rm s}(\boldsymbol{n}) = \int_{S^2} B(\boldsymbol{n}, \boldsymbol{n}')\, T(\boldsymbol{n}')\, \mathrm{d}\Omega_{\boldsymbol{n}'}, \tag{7.4.13}$$

where the beam profile is normalized to unit integral over the sphere ($\int_{S^2} B(\boldsymbol{n}, \boldsymbol{n}')\, \mathrm{d}\Omega_{\boldsymbol{n}'} = 1$) to preserve the total observed flux and monopole calibration.

This finite angular sensitivity has an essential physical effect: *beam smoothing*. When sky fluctuations occur on angular scales smaller than the beam width ($\theta \lesssim \sigma_{\rm beam}$), the detector integrates across both hot and cold patches simultaneously. This spatial averaging blurs fine-scale details and dampens the measured fluctuation amplitude. As we derive below, in multipole space this angular averaging acts as a low-pass filter, exponentially attenuating the measured power spectrum at high multipoles $\ell \gtrsim 1/\sigma_{\rm beam}$.

Circularly Symmetric Beams: From Spherical Harmonics to Legendre Polynomials

To understand how beam convolution operates in harmonic space, consider a telescope pointing along direction $\boldsymbol{n}$. We establish a spherical coordinate system (θ, ϕ) centered on this optical axis, where:

- $\boldsymbol{n}$ is the optical axis (polar axis, $\theta = 0$).
- $\theta = \arccos(\boldsymbol{n} \cdot \boldsymbol{n}')$ is the polar angle measuring angular separation away from the line of sight.
- ϕ is the azimuthal angle winding around the line of sight $\boldsymbol{n}$.

In the most general case, any beam response $B(\theta, \phi)$ on the celestial sphere can be expanded in spherical harmonics:

$$B(\theta, \phi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} B_{\ell m} Y_{\ell m}(\theta, \phi), \tag{7.4.14}$$

where the spherical harmonic multipole coefficients are given by the projection:

$$B_{\ell m} = \int_{S^2} B(\theta, \phi) Y^*_{\ell m}(\theta, \phi)\, \mathrm{d}\Omega. \tag{7.4.15}$$

Now consider the physical symmetry of the optical instrument. Real telescope reflectors and circular feed horns are engineered to be azimuthally symmetric about the optical axis: the

sensitivity depends solely on the angular distance θ from the center of the beam, and does not depend on the azimuthal angle ϕ:

$$B(\theta, \phi) = B(\theta). \tag{7.4.16}$$

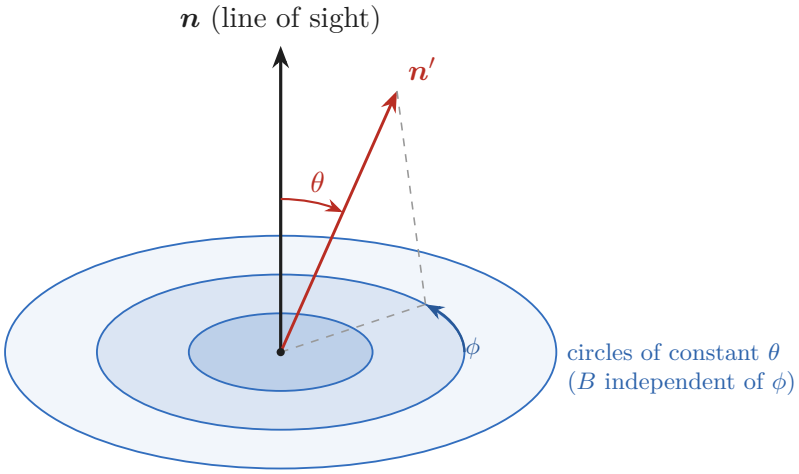


Figure 7.4: Geometry of an azimuthally symmetric beam profile. The optical axis defines the telescope line of sight $\boldsymbol{n}$. Concentric circles on the transverse footprint represent directions of constant angular separation $\theta = \arccos(\boldsymbol{n} \cdot \boldsymbol{n}')$ away from $\boldsymbol{n}$. Because the beam sensitivity $B(\theta)$ is independent of the azimuthal angle ϕ around $\boldsymbol{n}$, all $m \neq 0$ spherical harmonic modes vanish identically, reducing the expansion directly to Legendre polynomials $P_\ell(\cos\theta)$.

As depicted in Fig. 7.4, along any circle of constant θ, the sensitivity is uniform. Recalling that the spherical harmonics factorize into an associated Legendre polynomial and an azimuthal phase,

$$Y_{\ell m}(\theta, \phi) = \sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-m)!}{(\ell+m)!}} P_\ell^m(\cos\theta)\, \mathrm{e}^{im\phi}, \tag{7.4.17}$$

substituting $B(\theta, \phi) = B(\theta)$ into Eq. (7.4.15) separates the azimuthal integral completely:

$$\begin{aligned} B_{\ell m} &= \int_0^\pi \sin\theta \, \mathrm{d}\theta \int_0^{2\pi} \mathrm{d}\phi \, B(\theta) Y_{\ell m}^*(\theta, \phi) \\ &= \sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-m)!}{(\ell+m)!}} \left[\int_0^\pi B(\theta) P_\ell^m(\cos\theta) \sin\theta \, \mathrm{d}\theta\right] \underbrace{\int_0^{2\pi} \mathrm{e}^{-im\phi} \, \mathrm{d}\phi}_{2\pi\delta_{m0}} \\ &= B_{\ell 0}\, \delta_{m0}. \end{aligned} \tag{7.4.18}$$

Because all modes with $m \neq 0$ vanish identically, only the $m = 0$ mode survives. Recalling that for $m = 0$, $Y_{\ell 0}(\theta, \phi) = \sqrt{\frac{2\ell+1}{4\pi}} P_\ell(\cos\theta)$, the full spherical harmonic expansion collapses into the chained reduction:

$$\begin{aligned} B(\theta) &= \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} B_{\ell m} Y_{\ell m}(\theta, \phi) \\ &= \sum_{\ell=0}^{\infty} B_{\ell 0} \underbrace{Y_{\ell 0}(\theta)}_{\sqrt{\frac{2\ell+1}{4\pi}} P_\ell(\cos\theta)} \\ &= \sum_{\ell=0}^{\infty} \frac{2\ell+1}{4\pi} b_\ell P_\ell(\cos\theta), \end{aligned} \tag{7.4.19}$$

where the beam transfer function b_ℓ is defined by:

$$b_\ell \equiv \sqrt{\frac{4\pi}{2\ell+1}} B_{\ell 0} = 2\pi \int_0^\pi B(\theta) P_\ell(\cos\theta) \sin\theta \, \mathrm{d}\theta. \tag{7.4.20}$$

For a normalized beam ($\int B \, \mathrm{d}\Omega = 1$), the monopole coefficient is $b_0 = 1$, and higher b_ℓ quantify the transmission efficiency of the instrument at multipole ℓ.

Beam Smoothing in Harmonic Space and the Window Function

Since $\cos\theta = \boldsymbol{n} \cdot \boldsymbol{n}'$, the beam profile across arbitrary sky directions can be written in coordinate-free form as:

$$B(\boldsymbol{n}, \boldsymbol{n}') = \sum_{\ell=0}^{\infty} \frac{2\ell+1}{4\pi} b_\ell P_\ell(\boldsymbol{n} \cdot \boldsymbol{n}'). \tag{7.4.21}$$

Applying the spherical harmonic addition theorem,

$$P_\ell(\boldsymbol{n} \cdot \boldsymbol{n}') = \frac{4\pi}{2\ell+1} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\boldsymbol{n}) Y^*_{\ell m}(\boldsymbol{n}'), \tag{7.4.22}$$

the beam profile factors cleanly into:

$$B(\boldsymbol{n}, \boldsymbol{n}') = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} b_\ell Y_{\ell m}(\boldsymbol{n}) Y^*_{\ell m}(\boldsymbol{n}'). \tag{7.4.23}$$

Now substitute this factorization into the beam convolution integral for the observed sky temperature $T_{\mathrm{s}}(\boldsymbol{n})$:

$$\begin{aligned} T_{\mathrm{s}}(\boldsymbol{n}) &= \int_{S^2} B(\boldsymbol{n}, \boldsymbol{n}') \, T(\boldsymbol{n}') \, \mathrm{d}\Omega_{\boldsymbol{n}'} \\ &= \sum_{\ell m} b_\ell Y_{\ell m}(\boldsymbol{n}) \int_{S^2} Y^*_{\ell m}(\boldsymbol{n}') \left[\sum_{\ell' m'} a_{\ell' m'} Y_{\ell' m'}(\boldsymbol{n}') \right] \mathrm{d}\Omega_{\boldsymbol{n}'} \\ &= \sum_{\ell m} b_\ell Y_{\ell m}(\boldsymbol{n}) \sum_{\ell' m'} a_{\ell' m'} \underbrace{\int_{S^2} Y^*_{\ell m}(\boldsymbol{n}') Y_{\ell' m'}(\boldsymbol{n}') \, \mathrm{d}\Omega_{\boldsymbol{n}'}}_{\delta_{\ell\ell'}\delta_{mm'}} \\ &= \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} (b_\ell a_{\ell m}) Y_{\ell m}(\boldsymbol{n}). \end{aligned} \tag{7.4.24}$$

This reveals the fundamental simplicity of beam convolution: in harmonic space, spatial convolution by an azimuthally symmetric beam is simply a mode-by-mode multiplication. Skimming this chained sequence tells the entire physical story:

$$\boxed{a^{(\mathrm{s})}_{\ell m} = b_\ell a_{\ell m}} \implies \boxed{C^{(\mathrm{s})}_\ell = b^2_\ell C_\ell} \implies \boxed{W_\ell(1) = b^2_\ell.} \tag{7.4.25}$$

Evaluating the two-point temperature correlation function of the smoothed map:

$$\begin{aligned} \left\langle \frac{\Delta T_{\mathrm{s}}(\boldsymbol{n}_1) \Delta T_{\mathrm{s}}(\boldsymbol{n}_2)}{\bar{T}^2} \right\rangle &= \sum_{\ell m} \sum_{\ell' m'} \langle a^{(\mathrm{s})}_{\ell m} a^{(\mathrm{s})*}_{\ell' m'} \rangle Y_{\ell m}(\boldsymbol{n}_1) Y^*_{\ell' m'}(\boldsymbol{n}_2) \\ &= \sum_{\ell m} b^2_\ell C_\ell Y_{\ell m}(\boldsymbol{n}_1) Y^*_{\ell m}(\boldsymbol{n}_2) \\ &= \sum_{\ell} \frac{2\ell+1}{4\pi} C_\ell \, W_\ell(\boldsymbol{n}_1, \boldsymbol{n}_2), \end{aligned} \tag{7.4.26}$$

where the *window function* $W_\ell(\boldsymbol{n}_1, \boldsymbol{n}_2)$ is defined by:

$$W_\ell(\boldsymbol{n}_1, \boldsymbol{n}_2) = \int_{S^2\times S^2} B(\boldsymbol{n}_1, \boldsymbol{n}_1') B(\boldsymbol{n}_2, \boldsymbol{n}_2')\, P_\ell(\boldsymbol{n}_1' \cdot \boldsymbol{n}_2')\, \mathrm{d}\Omega_{\boldsymbol{n}_1'} \mathrm{d}\Omega_{\boldsymbol{n}_2'}. \tag{7.4.27}$$

For an azimuthally symmetric beam, the window function takes the elegant closed form:

$$W_\ell(\boldsymbol{n}_1, \boldsymbol{n}_2) = b_\ell^2 P_\ell(\boldsymbol{n}_1 \cdot \boldsymbol{n}_2), \tag{7.4.28}$$

and for the single-point autocorrelation ($\boldsymbol{n}_1 = \boldsymbol{n}_2$):

$$W_\ell(1) = b_\ell^2. \tag{7.4.29}$$

Thus, the window function $W_\ell(1)$ is simply the squared beam transfer function b_ℓ^2, establishing a direct, unbroken bridge between the physical geometry of the telescope and the observed power spectrum.

7.4.3 Example window functions

Because experimental beam widths are typically small (a few arcminutes to a degree), we can evaluate W_ℓ by approximating the sphere locally by a tangent plane orthogonal to $\boldsymbol{n}$.

Let $\boldsymbol{x}_i$ be the 2D planar displacement vector corresponding to the angle between $\boldsymbol{n}_i$ and $\boldsymbol{n}_i'$. Setting $\boldsymbol{n}_i' = (\boldsymbol{n}_i + \boldsymbol{x}_i)/\sqrt{1+x_i^2}$, the scalar product expands for $|\boldsymbol{x}_i| \ll 1$ as:

$$\begin{aligned}\boldsymbol{n}_1' \cdot \boldsymbol{n}_2' &= \frac{\boldsymbol{n}_1\cdot\boldsymbol{n}_2 + \boldsymbol{n}_1\cdot\boldsymbol{x}_2 + \boldsymbol{n}_2\cdot\boldsymbol{x}_1 + \boldsymbol{x}_1\cdot\boldsymbol{x}_2}{\sqrt{1+x_1^2}\sqrt{1+x_2^2}} \\ &\simeq \boldsymbol{n}_1\cdot\boldsymbol{n}_2\left(1 - \frac{1}{2}(x_1^2 + x_2^2)\right) + \boldsymbol{n}_1\cdot\boldsymbol{x}_2 + \boldsymbol{n}_2\cdot\boldsymbol{x}_1 + \boldsymbol{x}_1\cdot\boldsymbol{x}_2.\end{aligned} \tag{7.4.30}$$

The window function in Eq. (7.4.27) then becomes:

$$W_\ell(\boldsymbol{n}_1\cdot\boldsymbol{n}_2) = \int B(\boldsymbol{n}_1, \boldsymbol{x}_1) B(\boldsymbol{n}_2, \boldsymbol{x}_2)\, P_\ell(\boldsymbol{n}_1'\cdot\boldsymbol{n}_2')\, \mathrm{d}^2\boldsymbol{x}_1 \mathrm{d}^2\boldsymbol{x}_2. \tag{7.4.31}$$

Autocorrelation Window Function ($\boldsymbol{n}_1 = \boldsymbol{n}_2 = \boldsymbol{n}$)

When evaluating at the same point ($\boldsymbol{n}_1 = \boldsymbol{n}_2 = \boldsymbol{n}$), the displacement vectors $\boldsymbol{x}_1, \boldsymbol{x}_2$ are orthogonal to $\boldsymbol{n}$, giving $\boldsymbol{n}_1'\cdot\boldsymbol{n}_2' \simeq 1 - \frac{1}{2}|\boldsymbol{x}_1 - \boldsymbol{x}_2|^2 \simeq \cos(|\boldsymbol{x}_1 - \boldsymbol{x}_2|)$. For large ℓ and small angular separations, the Legendre polynomial is approximated by the zeroth-order Bessel function J_0:

$$P_\ell(\cos(|\boldsymbol{x}_1 - \boldsymbol{x}_2|)) \longrightarrow J_0(\ell|\boldsymbol{x}_1 - \boldsymbol{x}_2|) = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\phi\, \exp\left(-i\ell|\boldsymbol{x}_1 - \boldsymbol{x}_2|\cos\phi\right). \tag{7.4.32}$$

Let $\boldsymbol{\ell}$ be a 2D vector of magnitude ℓ oriented at angle ϕ relative to $\boldsymbol{x}_1 - \boldsymbol{x}_2$, and let $\tilde{B}(\boldsymbol{\ell})$ denote the 2D Fourier transform of the beam profile:

$$\tilde{B}(\boldsymbol{\ell}) = \int B(\boldsymbol{x})\, \mathrm{e}^{-i\boldsymbol{\ell}\cdot\boldsymbol{x}}\, \mathrm{d}^2\boldsymbol{x}. \tag{7.4.33}$$

Substituting Eq. (7.4.32) into Eq. (7.4.31) yields:

$$W_\ell(1) = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\phi\, |\tilde{B}(\boldsymbol{\ell})|^2. \tag{7.4.34}$$

The autocorrelation window function $W_\ell(1)$ is simply the angular average of the squared Fourier-transformed beam pattern.

Window Function for Separated Directions ($n_1 \neq n_2$)

For $\boldsymbol{n}_1 \neq \boldsymbol{n}_2$, let $z \equiv \boldsymbol{n}_1 \cdot \boldsymbol{n}_2 = \cos\theta$ and write $\boldsymbol{n}_1' \cdot \boldsymbol{n}_2' = z + \epsilon$, where up to second order in $\boldsymbol{x}_i$:

$$\epsilon = -\frac{1}{2} z(x_1^2 + x_2^2) + \boldsymbol{n}_1 \cdot \boldsymbol{x}_2 + \boldsymbol{n}_2 \cdot \boldsymbol{x}_1 + \boldsymbol{x}_1 \cdot \boldsymbol{x}_2. \tag{7.4.35}$$

Expanding $P_\ell(z+\epsilon)$ in a Taylor series:

$$P_\ell(z+\epsilon) \simeq P_\ell(z) + \epsilon P_\ell'(z) + \frac{1}{2}\epsilon^2 P_\ell''(z). \tag{7.4.36}$$

Inserting into Eq. (7.4.31) and retaining terms up to second order in $\boldsymbol{x}_i$, the terms linear in $\boldsymbol{x}_i$ integrate to zero by circular beam symmetry:

$$\begin{aligned} W_\ell(z) = P_\ell(z) + \frac{1}{2}\int \mathrm{d}^2\boldsymbol{x}_1 \mathrm{d}^2\boldsymbol{x}_2\, B(\boldsymbol{n}_1, \boldsymbol{x}_1) B(\boldsymbol{n}_2, \boldsymbol{x}_2) \\ \times \left[-2z(x_1^2 + x_2^2) P_\ell'(z) + \left((\boldsymbol{n}_1 \cdot \boldsymbol{x}_2)^2 + (\boldsymbol{n}_2 \cdot \boldsymbol{x}_1)^2 \right) P_\ell''(z) \right]. \end{aligned} \tag{7.4.37}$$

Decomposing $\boldsymbol{x}_i$ into components parallel to the $(\boldsymbol{n}_1, \boldsymbol{n}_2)$ plane ($\boldsymbol{e}_i$) and orthogonal to it ($\boldsymbol{m}$), $\boldsymbol{x}_i = \boldsymbol{e}_i \cos\phi_i + \boldsymbol{m} \sin\phi_i$, we find $\boldsymbol{n}_1 \cdot \boldsymbol{x}_2 = \sin\theta\cos\phi_2$ and $\boldsymbol{n}_2 \cdot \boldsymbol{x}_1 = \sin\theta\cos\phi_1$. Integrating over azimuthal angles gives:

$$\begin{aligned} W_\ell(z) = P_\ell(z) + \frac{(2\pi)^2}{4}\int_0^\infty \mathrm{d}x_1 \mathrm{d}x_2\, x_1 x_2 B(x_1) B(x_2) \\ \times \left[-2z(x_1^2 + x_2^2) P_\ell'(z) + (x_1^2 + x_2^2) \sin^2\theta\, P_\ell''(z) \right]. \end{aligned} \tag{7.4.38}$$

Using the Legendre differential equation:

$$(1 - z^2) P_\ell''(z) - 2z P_\ell'(z) = -\ell(\ell+1) P_\ell(z), \tag{7.4.39}$$

noting $\sin^2\theta = 1 - z^2$, utilizing the beam normalization $2\pi \int_0^\infty x B(x) \mathrm{d}x = 1$, and defining the beam width:

$$\sigma^2 \equiv \pi \int_0^\infty x^3 B(x)\, \mathrm{d}x, \tag{7.4.40}$$

the window function simplifies to:

$$W_\ell(z) = P_\ell(z)(1 - \sigma^2 \ell(\ell+1)) \qquad (\text{for } \sigma\ell \ll 1). \tag{7.4.41}$$

For all ℓ, this is well approximated by $W_\ell(z) = P_\ell(z) \exp(-\sigma^2 \ell(\ell+1))$.

Gaussian Beam

For a symmetric Gaussian beam with profile:

$$B(x) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2} \right), \tag{7.4.42}$$

the 2D Fourier transform is:

$$\tilde{B}(\ell) = \exp\left(-\frac{\ell^2 \sigma^2}{2} \right). \tag{7.4.43}$$

Thus, from Eq. (7.4.34):

$$W_\ell(1) = \mathrm{e}^{-\ell^2\sigma^2}, \tag{7.4.44}$$

and for separated directions:

$$W_\ell(z) = P_\ell(z) \exp\left(-\sigma^2 \ell(\ell+1) \right). \tag{7.4.45}$$

The exponential suppression factor $\mathrm{e}^{-\ell^2\sigma^2}$ illustrates the physical consequence of finite angular sensitivity: fluctuations on scales finer than the beam size ($\ell \gg 1/\sigma$, or $\theta \ll \sigma$) are smoothed away by the detector's spatial integration, exponentially dampening the measured signal.

Derivation: 2D Fourier Transform of a Gaussian Beam and Window Function
The 2D Fourier transform of $B(\boldsymbol{x}) = \frac{1}{2\pi\sigma^2}\mathrm{e}^{-x^2/(2\sigma^2)}$ in polar coordinates (x, ψ) with $\boldsymbol{\ell}\cdot\boldsymbol{x} = \ell x\cos\psi$ is:

$$\begin{aligned}\tilde{B}(\boldsymbol{\ell}) &= \int_0^\infty x\,\mathrm{d}x\int_0^{2\pi}\mathrm{d}\psi\,\frac{1}{2\pi\sigma^2}\mathrm{e}^{-x^2/(2\sigma^2)}\mathrm{e}^{-i\ell x\cos\psi}\\ &= \frac{1}{\sigma^2}\int_0^\infty x\,\mathrm{d}x\,\mathrm{e}^{-x^2/(2\sigma^2)}J_0(\ell x).\end{aligned} \tag{1}$$

Using the Weber identity $\int_0^\infty x\mathrm{e}^{-ax^2}J_0(bx)\,\mathrm{d}x = \frac{1}{2a}\mathrm{e}^{-b^2/(4a)}$ with $a = 1/(2\sigma^2)$ and $b = \ell$:

$$\tilde{B}(\boldsymbol{\ell}) = \frac{1}{\sigma^2}\left(\sigma^2\mathrm{e}^{-\ell^2\sigma^2/2}\right) = \mathrm{e}^{-\ell^2\sigma^2/2}. \tag{2}$$

Squaring and taking the angular average gives the autocorrelation window:

$$W_\ell(1) = |\tilde{B}(\ell)|^2 = \mathrm{e}^{-\ell^2\sigma^2}. \tag{3}$$

Differencing Beam

A practical challenge of using a standard Gaussian beam is that one must subtract the large mean monopole temperature ($\bar{T} = 2.7255\,\mathrm{K}$) from the raw detector data to isolate $\sim 10\,\mu\mathrm{K}$ fluctuations. Subtracting two large numbers to obtain a minuscule residual is notoriously susceptible to instrumental $1/f$ drift and calibration offsets.

To eliminate this vulnerability, experimenters often employ a **differencing beam** pattern whose spatial integral vanishes identically:

$$\int \mathrm{d}^2\boldsymbol{x}\,B(\boldsymbol{x}) = 0. \tag{7.4.46}$$

In hardware, this is achieved by differencing the signals from two horns pointing in slightly separated directions. For a single chop with throw separation $2x_0$ along the x-axis:

$$B(x,y) = \frac{1}{2\pi\sigma^2}\left[\exp\left(-\frac{(x-x_0)^2+y^2}{2\sigma^2}\right) - \exp\left(-\frac{(x+x_0)^2+y^2}{2\sigma^2}\right)\right]. \tag{7.4.47}$$

Taking the 2D Fourier transform:

$$\begin{aligned}\tilde{B}(\boldsymbol{\ell}) &= \int \mathrm{d}^2\boldsymbol{x}\,B(\boldsymbol{x})\,\mathrm{e}^{-i\boldsymbol{\ell}\cdot\boldsymbol{x}}\\ &= \mathrm{e}^{-\ell^2\sigma^2/2}\left(\mathrm{e}^{-i\ell_x x_0} - \mathrm{e}^{i\ell_x x_0}\right) = -2i\,\mathrm{e}^{-\ell^2\sigma^2/2}\sin(\ell_x x_0).\end{aligned} \tag{7.4.48}$$

In polar coordinates (ℓ, ϕ) where $\ell_x = \ell\cos\phi$, the squared magnitude is:

$$|\tilde{B}(\boldsymbol{\ell})|^2 = 4\mathrm{e}^{-\ell^2\sigma^2}\sin^2(\ell x_0\cos\phi) = 2\mathrm{e}^{-\ell^2\sigma^2}\left[1-\cos(2\ell x_0\cos\phi)\right]. \tag{7.4.49}$$

Integrating over the azimuthal angle ϕ and using the Bessel identity

$$J_0(u) = \frac{1}{2\pi}\int_0^{2\pi}\cos(u\cos\phi)\,\mathrm{d}\phi \approx P_\ell(\cos(2x_0)), \tag{7.4.50}$$

the autocorrelation window function evaluates to:

$$W_\ell(1) = \mathrm{e}^{-\sigma^2\ell^2}\left[1 - P_\ell(\cos(2x_0))\right]. \tag{7.4.51}$$

As illustrated in Fig. 7.5, the differencing beam filters out both very high multipoles ($\ell\sigma \gg 1$, where beam smoothing suppresses signal) and very low multipoles ($\ell \ll \pi/(2x_0)$, where the beam throw averages over uniform sky regions and cancels).

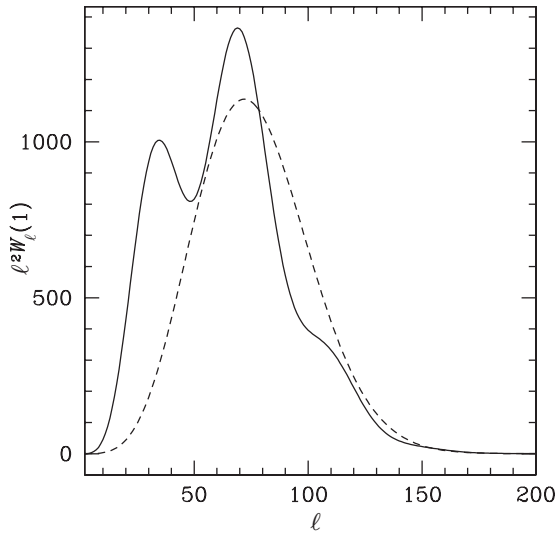


Figure 7.5: The window function $\ell^2 W_\ell(1)$ for a differencing beam of width $\sigma = 1^\circ$ for two different throw values: $x_0 = 4^\circ$ (solid line) and $x_0 = 1^\circ$ (dashed line). The high-ℓ cutoff at $\ell \sim 70$ is dictated by the Gaussian beam width σ. When the throw x_0 differs from σ, a second low-ℓ modulation appears on the left corresponding to the angular throw scale.

7.4.4 Cosmic variance

In terrestrial experiments, measurements can be repeated under identical conditions over thousands of trials to reduce statistical uncertainty. In cosmology, however, we have access to only **one** observable Universe and a single celestial sphere.

For a statistically isotropic Gaussian random field, the spherical harmonic coefficients $a_{\ell m}$ are independent complex Gaussian random variables with zero mean and variance:

$$\langle a_{\ell m} a^*_{\ell' m'} \rangle = C_\ell \, \delta_{\ell \ell'} \delta_{m m'}. \tag{7.4.52}$$

For a given multipole ℓ, full-sky coverage provides exactly $2\ell + 1$ statistically independent m-modes. The standard unbiased estimator for the power spectrum is:

$$\hat{C}_\ell \equiv \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} |a_{\ell m}|^2. \tag{7.4.53}$$

The statistical scatter of this estimator about the true ensemble mean C_ℓ is known as **cosmic variance**:

$$\sigma_\ell \equiv \frac{\sqrt{\langle (\hat{C}_\ell - C_\ell)^2 \rangle}}{C_\ell} = \sqrt{\frac{2}{2\ell + 1}}. \tag{7.4.54}$$

Derivation: Cosmic Variance from Wick's Theorem
The variance of the estimator $\hat{C}_\ell$ is:

$$\begin{aligned} \mathrm{Var}(\hat{C}_\ell) &= \langle \hat{C}_\ell^2 \rangle - \langle \hat{C}_\ell \rangle^2 \\ &= \frac{1}{(2\ell+1)^2} \sum_{m=-\ell}^{\ell} \sum_{m'=-\ell}^{\ell} \left[\langle a_{\ell m} a^*_{\ell m} a_{\ell m'} a^*_{\ell m'} \rangle - \langle |a_{\ell m}|^2 \rangle \langle |a_{\ell m'}|^2 \rangle \right]. \end{aligned} \tag{1}$$

For a complex zero-mean Gaussian field with $\langle a_{\ell m} a^*_{\ell m'} \rangle = C_\ell \delta_{m m'}$ and $\langle a_{\ell m} a_{\ell m'} \rangle =$

$(-1)^m C_\ell \delta_{m,-m'}$, Wick's theorem states:

$$\langle a_{\ell m} a^*_{\ell m} a_{\ell m'} a^*_{\ell m'} \rangle = \langle a_{\ell m} a^*_{\ell m} \rangle \langle a_{\ell m'} a^*_{\ell m'} \rangle + \langle a_{\ell m} a^*_{\ell m} \rangle \langle a^*_{\ell m} a_{\ell m'} \rangle + \langle a_{\ell m} a_{\ell m'} \rangle \langle a^*_{\ell m} a^*_{\ell m'} \rangle. \quad (2)$$

Subtracting the disconnected piece $\langle |a_{\ell m}|^2 \rangle \langle |a_{\ell m'}|^2 \rangle = C_\ell^2$, the two remaining Wick contractions yield:

$$\langle a_{\ell m} a^*_{\ell m'} \rangle \langle a^*_{\ell m} a_{\ell m'} \rangle + \langle a_{\ell m} a_{\ell m'} \rangle \langle a^*_{\ell m} a^*_{\ell m'} \rangle = C_\ell^2 \delta_{mm'} + C_\ell^2 \delta_{m,-m'}. \quad (3)$$

Summing over m and m':

$$\sum_{m,m'} \left(C_\ell^2 \delta_{mm'} + C_\ell^2 \delta_{m,-m'} \right) = (2\ell+1) C_\ell^2 + (2\ell+1) C_\ell^2 = 2(2\ell+1) C_\ell^2. \quad (4)$$

Dividing by $(2\ell+1)^2$:

$$\mathrm{Var}(\hat{C}_\ell) = \frac{2}{2\ell+1} C_\ell^2 \implies \frac{\sigma(\hat{C}_\ell)}{C_\ell} = \sqrt{\frac{2}{2\ell+1}}. \quad (5)$$

When an experiment covers only a sky fraction $f_{\mathrm{sky}} < 1$, the effective number of independent modes is reduced to $(2\ell+1) f_{\mathrm{sky}}$, increasing the variance:

$$\mathrm{Var}(\hat{C}_\ell) = \frac{2}{(2\ell+1) f_{\mathrm{sky}}} \left(C_\ell + \frac{N_\ell}{W_\ell} \right)^2. \quad (6)$$

Cosmic variance sets an irreducible lower bound on parameter uncertainties. At low multipoles (such as the quadrupole $\ell = 2$), $\sigma_2 / C_2 = \sqrt{2/5} \approx 63\%$, explaining why large-scale anomalies carry large cosmic variance uncertainties. At high multipoles ($\ell \sim 1000$), $\sigma_{1000} / C_{1000} \approx \sqrt{2/2001} \approx 3.1\%$, permitting sub-percent cosmological constraints.

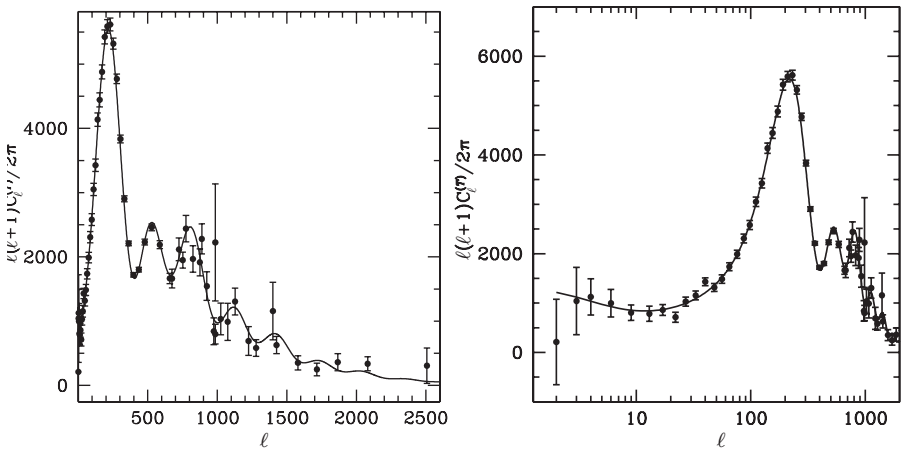


Figure 7.6: The WMAP full-sky temperature anisotropy map and observed angular power spectrum with cosmic variance error envelope.

7.5 Statistical methods

To extract optimal, unbiased constraints on cosmological parameters from observational data, robust statistical methods are required. We present the Bayesian inference framework and Fisher

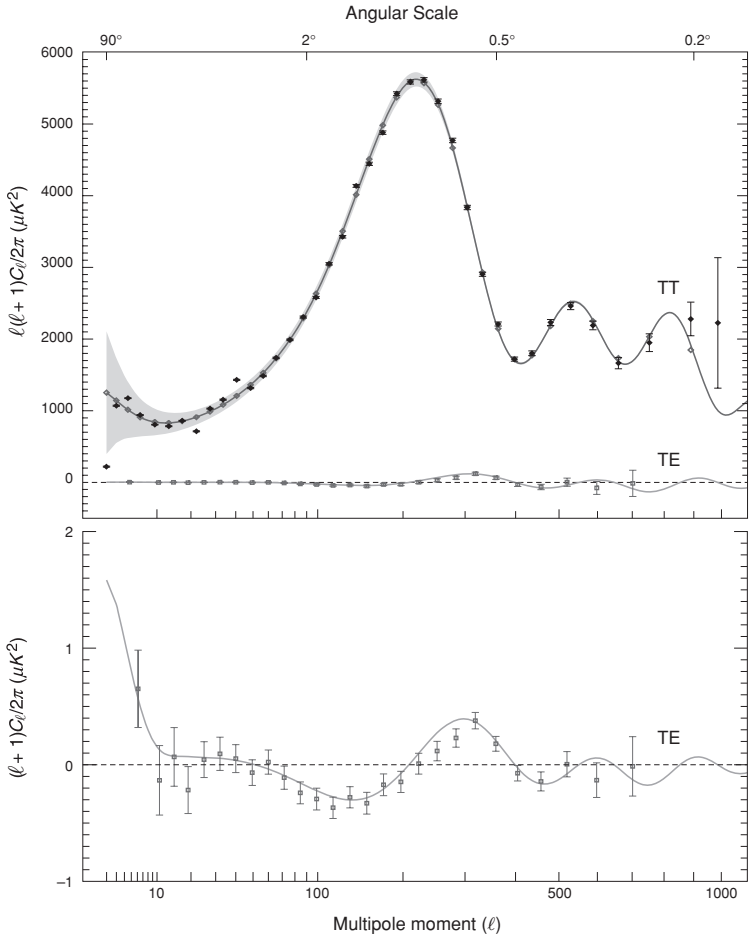


Figure 7.7: The CMB temperature anisotropy power spectrum (TT) and the temperature–polarization cross-correlation spectrum (TE) obtained from the WMAP 3-year data (from Hinshaw et al., 2007).

information formalism that underpin modern cosmological parameter analysis.

7.5.1 Bayes' Theorem and the Likelihood Function

Probability serves two fundamentally different purposes in physical science:

1. *Deduction (prediction):* Given a known physical model and its parameters, what is the probability distribution of potential experimental outcomes?

2. *Induction (inference):* Given a fixed, observed dataset that has already been recorded in the laboratory or by a telescope, which theoretical model parameters are most plausible?

In cosmological parameter estimation, our data vector $\boldsymbol{d}$ has already been gathered by telescopes such as *Planck*, WMAP, or ground-based surveys. Our objective is to infer the underlying cosmological parameters $\boldsymbol{\lambda} \equiv (\Omega_{b0}, \Omega_{c0}, H_0, n_{\rm s}, A_{\rm s}, \tau, \dots)$ that best explain this reality.

Bayes' theorem provides the rigorous mathematical apparatus for inverting conditional probabilities. In elementary probability theory, for any two propositions A and B, the joint probability satisfies:

$$P(A \cap B) = P(B \mid A)P(A) = P(A \mid B)P(B). \tag{7.5.1}$$

Identifying A with the cosmological parameters $\boldsymbol{\lambda}$ and B with the observed data $\boldsymbol{d}$ yields *Bayes'*

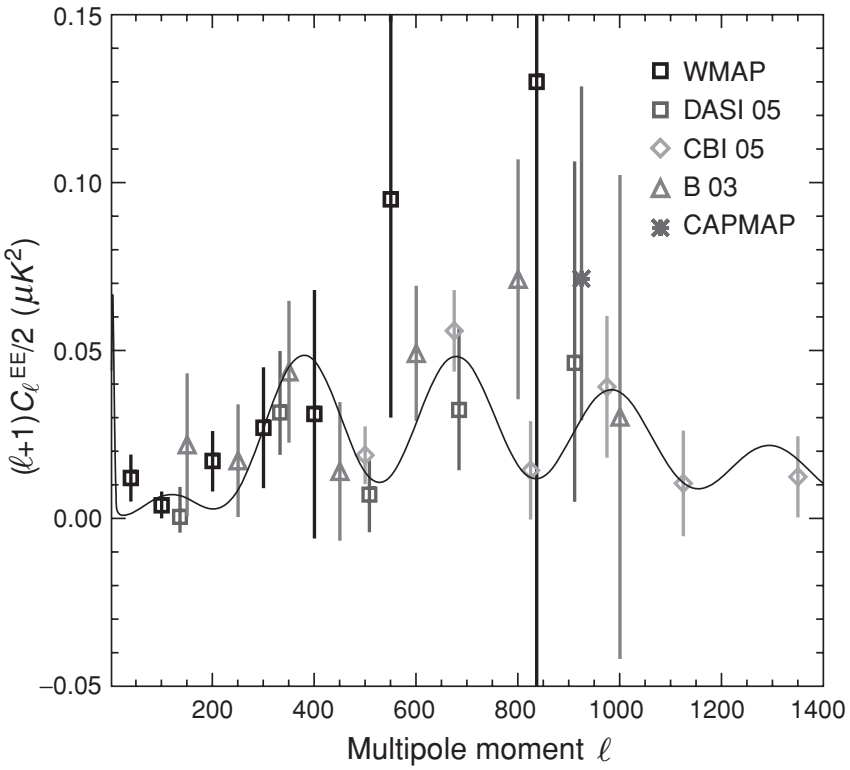


Figure 7.8: The measured EE polarization power spectrum from the WMAP 3-year data, BOOMERanG, DASI, CBI, and CAPMAP experiments (from Page et al., 2007).

theorem:

$$P(\boldsymbol{\lambda} \mid \boldsymbol{d}) = \frac{P(\boldsymbol{d} \mid \boldsymbol{\lambda})P(\boldsymbol{\lambda})}{P(\boldsymbol{d})} = \frac{\mathcal{L}(\boldsymbol{\lambda})P(\boldsymbol{\lambda})}{\mathcal{E}}, \tag{7.5.2}$$

where each factor carries a clear, intuitive physical interpretation:

- $P(\boldsymbol{\lambda} \mid \boldsymbol{d})$ is the *posterior probability distribution*: what we have learned about the cosmological parameters after assimilating the experimental data.
- $\mathcal{L}(\boldsymbol{\lambda}) \equiv P(\boldsymbol{d} \mid \boldsymbol{\lambda})$ is the *likelihood function*: evaluated for the fixed, observed data $\boldsymbol{d}$, it quantifies how likely each candidate probability distribution (specified by parameters $\boldsymbol{\lambda}$) is to have generated that observed data.
- $P(\boldsymbol{\lambda})$ is the *prior probability distribution*: our state of knowledge or physical constraints on the parameters before this experiment was performed (e.g., enforcing $\Omega_{m0} > 0$ or positivity of cosmic age).
- $\mathcal{E} \equiv P(\boldsymbol{d}) = \int \mathrm{d}\boldsymbol{\lambda}\, P(\boldsymbol{\lambda})\mathcal{L}(\boldsymbol{\lambda})$ is the *Bayesian evidence* (or marginal likelihood). For parameter estimation within a fixed cosmological model, $\mathcal{E}$ is simply a parameter-independent normalization constant ensuring $\int \mathrm{d}\boldsymbol{\lambda} P(\boldsymbol{\lambda} \mid \boldsymbol{d}) = 1$. In model comparison, however, $\mathcal{E}$ quantifies the global performance of competing cosmological scenarios (such as ΛCDM versus dynamical dark energy).

Deriving the Gaussian Likelihood Function

As introduced in §7.4.1, cosmological datasets fall into two broad statistical paradigms: stochastic fluctuations (such as CMB temperature and polarization maps) and deterministic background signals (such as Type Ia supernovae).

1. The stochastic CMB likelihood (variance-shifting). For CMB temperature anisotropy maps or harmonic multipoles, primordial perturbations are a realization of a zero-mean Gaussian random field, $\langle \boldsymbol{s} \rangle = 0$, with theoretical covariance $\boldsymbol{S}(\boldsymbol{\lambda}) = \langle \boldsymbol{s}\boldsymbol{s}^\top \rangle$. Adding independent zero-mean instrumental noise $\boldsymbol{n}$ with covariance $\boldsymbol{N}$, the observed data vector $\boldsymbol{d} = \boldsymbol{s} + \boldsymbol{n}$ has mean zero ($\langle \boldsymbol{d} \rangle = 0$) and total covariance:

$$\boldsymbol{C}(\boldsymbol{\lambda}) \equiv \langle \boldsymbol{d}\boldsymbol{d}^\top \rangle = \boldsymbol{S}(\boldsymbol{\lambda}) + \boldsymbol{N}. \tag{7.5.3}$$

Because the sum of zero-mean Gaussians is zero-mean Gaussian, the probability density of the observed data vector $\boldsymbol{d}$ given parameters $\boldsymbol{\lambda}$ is:

$$\mathcal{L}(\boldsymbol{\lambda}) \equiv P(\boldsymbol{d} \mid \boldsymbol{\lambda}) = \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{C}(\boldsymbol{\lambda})}} \exp\left(-\frac{1}{2}\boldsymbol{d}^\top \boldsymbol{C}^{-1}(\boldsymbol{\lambda})\boldsymbol{d}\right). \tag{7.5.4}$$

In the idealized case of full-sky coverage where spherical harmonic multipoles $a_{\ell m}$ (or uncorrelated spatial modes $i = 1, \ldots, N$) decouple, the covariance matrix is diagonal. Let $\sigma_i^2(\boldsymbol{\lambda}) \equiv C_{\ell_i}(\boldsymbol{\lambda}) b_{\ell_i}^2 + \sigma_{n,i}^2$ denote the total variance in mode i, combining cosmic variance and instrumental noise. Then Eq. (7.5.4) factorizes into a product of independent zero-mean Gaussians:

$$\mathcal{L}(\boldsymbol{\lambda}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2(\boldsymbol{\lambda})}} \exp\left(-\frac{d_i^2}{2\sigma_i^2(\boldsymbol{\lambda})}\right). \tag{7.5.5}$$

Notice that the observed measurements d_i appear directly as d_i^2 in the numerator: the expected temperature fluctuation across the ensemble is zero, and cosmological parameters $\boldsymbol{\lambda}$ constrain the data by controlling the width $\sigma_i^2(\boldsymbol{\lambda})$ of the distribution.

2. The deterministic signal likelihood (mean-shifting). In contrast, when analyzing measurements of background cosmic expansion $\boldsymbol{y} \equiv (y_1, \ldots, y_N)$ at redshifts z_i (such as supernova distance moduli or cosmic chronometer expansion rates), cosmological theory predicts a deterministic physical value $\boldsymbol{y}_{\rm th}(\boldsymbol{\lambda})$. The stochastic variation arises purely from experimental measurement errors $\boldsymbol{n} \sim \mathcal{N}(0, \boldsymbol{N})$.

Here, the zero-mean Gaussian quantity is the residual vector between the data and the model, $\boldsymbol{y} - \boldsymbol{y}_{\rm th}(\boldsymbol{\lambda}) = \boldsymbol{n}$, yielding the likelihood:

$$\mathcal{L}(\boldsymbol{\lambda}) = \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{N}}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{y}_{\rm th}(\boldsymbol{\lambda}))^\top \boldsymbol{N}^{-1}(\boldsymbol{y} - \boldsymbol{y}_{\rm th}(\boldsymbol{\lambda}))\right). \tag{7.5.6}$$

For uncorrelated observational errors with known standard deviations σ_i, this simplifies to:

$$\mathcal{L}(\boldsymbol{\lambda}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(y_i - y_{\rm th}(z_i; \boldsymbol{\lambda}))^2}{2\sigma_i^2}\right). \tag{7.5.7}$$

Maximizing this likelihood is mathematically equivalent to minimizing the classical sum of squared residuals $\chi^2(\boldsymbol{\lambda}) = \sum_i (y_i - y_{\rm th}(z_i; \boldsymbol{\lambda}))^2/\sigma_i^2$, as explored in detail in §7.7.2.

The Role of Priors

The choice of prior $P(\boldsymbol{\lambda})$ reflects physical plausibility:

- **Additive Parameters:** Parameters that add linearly to observables (e.g., spectral tilt $n_{\rm s}$, optical depth τ) typically adopt uniform (flat) priors: $P(\lambda) = \text{const}$.
- **Scale / Multiplicative Parameters:** Parameters that set the overall multiplicative scale (e.g., amplitude $A_{\rm s}$, Hubble constant H_0) naturally adopt logarithmic (Jeffreys) priors: $P(\lambda) \propto 1/\lambda \iff P(\ln \lambda) = \text{const}$.

When high-quality CMB data are available, the likelihood function $\mathcal{L}(\boldsymbol{\lambda})$ is sharply peaked compared to the prior, making the posterior $P(\boldsymbol{\lambda} \mid \boldsymbol{d})$ largely independent of the prior. If the posterior depends strongly on the prior, the data lack the statistical power to constrain the model.

7.5.2 Fisher matrix and parameter estimation

Maximum Likelihood and Newton–Raphson Iteration

The best-fitting cosmological model corresponds to the parameter vector $\bar{\boldsymbol{\lambda}}$ that maximizes the likelihood function $\mathcal{L}(\boldsymbol{\lambda})$—identifying the candidate probability distribution that was most likely to have generated our observed dataset—or equivalently maximizes the log-likelihood $\ln \mathcal{L}(\boldsymbol{\lambda})$:

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \lambda_i}\right|_{\bar{\boldsymbol{\lambda}}} = 0. \tag{7.5.8}$$

Expanding $\frac{\partial \ln \mathcal{L}}{\partial \lambda_i}$ to first order in a Taylor series about an initial parameter estimate $\boldsymbol{\lambda}^{(0)}$:

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \lambda_i}\right|_{\bar{\boldsymbol{\lambda}}} \simeq \left.\frac{\partial \ln \mathcal{L}}{\partial \lambda_i}\right|_{\boldsymbol{\lambda}^{(0)}} + \sum_j \left(\bar{\lambda}_j - \lambda_j^{(0)}\right) \left.\frac{\partial^2 \ln \mathcal{L}}{\partial \lambda_i \partial \lambda_j}\right|_{\boldsymbol{\lambda}^{(0)}} = 0, \tag{7.5.9}$$

which yields the parameter step $\delta\lambda_j \equiv \bar{\lambda}_j - \lambda_j^{(0)}$:

$$\delta\lambda_j \simeq -\sum_i \left(\frac{\partial^2 \ln \mathcal{L}}{\partial \lambda_i \partial \lambda_j}\right)^{-1} \frac{\partial \ln \mathcal{L}}{\partial \lambda_i}. \tag{7.5.10}$$

Derivatives of the Gaussian Log-Likelihood

For the Gaussian likelihood in Eq. (7.5.4), the log-likelihood is:

$$\ln \mathcal{L} = -\frac{1}{2} \ln(\det \boldsymbol{C}) - \frac{1}{2} \boldsymbol{d}^\top \boldsymbol{C}^{-1} \boldsymbol{d} - \frac{N}{2} \ln(2\pi). \tag{7.5.11}$$

Using the matrix identities $\partial_i \ln(\det \boldsymbol{C}) = \mathrm{Tr}(\boldsymbol{C}^{-1} \partial_i \boldsymbol{C})$ and $\partial_i \boldsymbol{C}^{-1} = -\boldsymbol{C}^{-1} (\partial_i \boldsymbol{C}) \boldsymbol{C}^{-1}$, the derivative with respect to λ_i is:

$$\frac{\partial \ln \mathcal{L}}{\partial \lambda_i} = \frac{1}{2} \left[\boldsymbol{d}^\top \boldsymbol{C}^{-1} (\partial_i \boldsymbol{C}) \boldsymbol{C}^{-1} \boldsymbol{d} - \mathrm{Tr}(\boldsymbol{C}^{-1} \partial_i \boldsymbol{C}) \right]. \tag{7.5.12}$$

Differentiating a second time gives:

$$\begin{aligned} \frac{\partial^2 \ln \mathcal{L}}{\partial \lambda_i \partial \lambda_j} = -\frac{1}{2} \Big[& \boldsymbol{d}^\top \boldsymbol{C}^{-1} (\partial_i \boldsymbol{C}) \boldsymbol{C}^{-1} (\partial_j \boldsymbol{C}) \boldsymbol{C}^{-1} \boldsymbol{d} + \boldsymbol{d}^\top \boldsymbol{C}^{-1} (\partial_j \boldsymbol{C}) \boldsymbol{C}^{-1} (\partial_i \boldsymbol{C}) \boldsymbol{C}^{-1} \boldsymbol{d} \\ & - \mathrm{Tr}(\boldsymbol{C}^{-1} (\partial_j \boldsymbol{C}) \boldsymbol{C}^{-1} (\partial_i \boldsymbol{C})) - \boldsymbol{d}^\top \boldsymbol{C}^{-1} (\partial^2_{ij} \boldsymbol{C}) \boldsymbol{C}^{-1} \boldsymbol{d} + \mathrm{Tr}(\boldsymbol{C}^{-1} \partial^2_{ij} \boldsymbol{C}) \Big]. \end{aligned} \tag{7.5.13}$$

The Fisher Information Matrix

The **Fisher Information Matrix** F_{ij} is defined as the ensemble expectation value of the second derivative (the curvature) of the log-likelihood:

$$F_{ij} \equiv \left\langle -\frac{\partial^2 \ln \mathcal{L}}{\partial \lambda_i \partial \lambda_j} \right\rangle = \left\langle \frac{\partial \ln \mathcal{L}}{\partial \lambda_i} \frac{\partial \ln \mathcal{L}}{\partial \lambda_j} \right\rangle. \tag{7.5.14}$$

Taking the expectation value of Eq. (7.5.13) using $\langle d_m d_s \rangle = C_{ms}$, we evaluate:

$$\begin{aligned}\left\langle \boldsymbol{d}^\top \boldsymbol{C}^{-1}(\partial_i \boldsymbol{C})\boldsymbol{C}^{-1}(\partial_j \boldsymbol{C})\boldsymbol{C}^{-1}\boldsymbol{d} \right\rangle &= \sum_{m,n,p,q,r,s} \langle d_m d_s \rangle C^{-1}_{mn}(\partial_i C)_{np} C^{-1}_{pq}(\partial_j C)_{qr} C^{-1}_{rs} \\ &= \sum_{m,n,p,q,r,s} C_{sm} C^{-1}_{mn}(\partial_i C)_{np} C^{-1}_{pq}(\partial_j C)_{qr} C^{-1}_{rs} \\ &= \mathrm{Tr}(\boldsymbol{C}^{-1}(\partial_i \boldsymbol{C})\boldsymbol{C}^{-1}(\partial_j \boldsymbol{C})). \end{aligned} \tag{7.5.15}$$

Similarly, $\langle \boldsymbol{d}^\top \boldsymbol{C}^{-1}(\partial^2_{ij}\boldsymbol{C})\boldsymbol{C}^{-1}\boldsymbol{d}\rangle = \mathrm{Tr}(\boldsymbol{C}^{-1}\partial^2_{ij}\boldsymbol{C})$. The second-derivative terms in Eq. (7.5.13) cancel identically against the trace term, yielding the Fisher information matrix:

$$F_{ij} = \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{C}^{-1}(\partial_i \boldsymbol{C})\boldsymbol{C}^{-1}(\partial_j \boldsymbol{C})\right]. \tag{7.5.16}$$

Properties of the Fisher Information Matrix

1. **Symmetry and Positive Definiteness:** For any non-zero vector $\boldsymbol{v}$,

$$\sum_{i,j} v_i F_{ij} v_j = \frac{1}{2}\mathrm{Tr}\left[\left(\boldsymbol{C}^{-1}\sum_i v_i \partial_i \boldsymbol{C}\right)^2\right] = \frac{1}{2}\mathrm{Tr}\left[(\boldsymbol{C}^{-1/2}(\boldsymbol{v}\cdot\nabla\boldsymbol{C})\boldsymbol{C}^{-1/2})^2\right] > 0. \quad (1)$$

   Thus, $\boldsymbol{F}$ is symmetric and strictly positive-definite.

2. **Singularity and Degeneracy:** If $\boldsymbol{F}$ has a zero eigenvalue ($\det \boldsymbol{F} = 0$) with eigenvector $\boldsymbol{\mu}$, then $\sum_i \mu_i \partial_i \boldsymbol{C} = 0$. The data are completely blind to changes along the parameter direction $\sum_i \mu_i \lambda_i$, signaling an exact parameter degeneracy.

3. **Forecasting Power:** Because F_{ij} depends solely on the covariance matrix $\boldsymbol{C}(\boldsymbol{\lambda})$ and its derivatives—not on the observed data $\boldsymbol{d}$—it can be computed prior to building an experiment to forecast expected parameter errors.

Optimal Quadratic Estimators

Replacing the sample Hessian $\partial^2 \ln \mathcal{L}/\partial\lambda_i\partial\lambda_j$ in Eq. (7.5.10) by its ensemble average $-F_{ij}$ yields the **Optimal Quadratic Estimator** iteration:

$$\lambda_i^{(1)} = \lambda_i^{(0)} + \frac{1}{2}\sum_j (F^{-1})_{ij}\left[\boldsymbol{d}^\top \boldsymbol{C}^{-1}(\partial_j \boldsymbol{C})\boldsymbol{C}^{-1}\boldsymbol{d} - \mathrm{Tr}(\boldsymbol{C}^{-1}\partial_j \boldsymbol{C})\right]_{\boldsymbol{\lambda}^{(0)}}. \tag{7.5.17}$$

Because the correction term depends quadratically on the data vector $\boldsymbol{d}$, this formulation provides a minimum-variance unbiased quadratic estimator for the parameters.

Why Grid Searches Fail

In cosmological analyses, parameter spaces typically span $M \geq 10$ dimensions (for example, physical baryon density $\Omega_\mathrm{b}h^2$, cold dark matter density $\Omega_\mathrm{cdm}h^2$, sound horizon angle θ_MC, reionization optical depth τ, scalar spectral tilt n_s, primordial amplitude A_s, neutrino mass sum $\sum m_\nu$, tensor ratio r, effective neutrino count N_eff, and dark energy equation of state w).

A coarse computational grid of only 10 points per dimension requires 10^{10} likelihood evaluations. If a fast Boltzmann solver takes 1 s per evaluation, exploring this modest grid would require:

$$t_\mathrm{grid} \approx 10^{10}\,\mathrm{s} \approx 317\,\mathrm{years!} \tag{7.5.18}$$

Grid-based searches are completely intractable in multi-dimensional cosmology. Iterative optimization, Fisher forecasting, and Markov Chain Monte Carlo (MCMC) sampling are indispensable.

Parameter Uncertainties and the Cramér–Rao Bound

To determine the error bars on estimated parameters, consider the covariance of the step $\delta\boldsymbol{\lambda}$:

$$\langle \delta\lambda_i \delta\lambda_j \rangle = \frac{1}{4}\sum_{m,n}(F^{-1})_{im}(F^{-1})_{jn}\Big\langle \Big[\boldsymbol{d}^\top \boldsymbol{C}^{-1}(\partial_m \boldsymbol{C})\boldsymbol{C}^{-1}\boldsymbol{d} - \mathrm{Tr}(\boldsymbol{C}^{-1}\partial_m \boldsymbol{C})\Big] \times \Big[\boldsymbol{d}^\top \boldsymbol{C}^{-1}(\partial_n \boldsymbol{C})\boldsymbol{C}^{-1}\boldsymbol{d} - \mathrm{Tr}(\boldsymbol{C}^{-1}\partial_n \boldsymbol{C})\Big]\Big\rangle . \tag{7.5.19}$$

Evaluating the 4-point Gaussian expectation value via Wick's theorem yields:

$$\langle \delta\lambda_i \delta\lambda_j \rangle = \sum_{m,n}(F^{-1})_{im}(F^{-1})_{jn}F_{mn} = (F^{-1})_{ij}. \tag{7.5.20}$$

Derivation: The Cramér–Rao Inequality and Marginalized Error Bars
The **Cramér–Rao inequality** states that for any unbiased estimator $\hat{\lambda}_i$ of parameter λ_i, the variance is bounded from below by the inverse Fisher matrix:

$$\mathrm{Var}(\hat{\lambda}_i) \geq (F^{-1})_{ii}. \tag{1}$$

- **Unmarginalized (Conditional) Error:** If all other parameters $\lambda_{j\neq i}$ are held fixed at their true values, the 1σ uncertainty on λ_i is:

$$\sigma_{\mathrm{unmarg}}(\lambda_i) = \frac{1}{\sqrt{F_{ii}}}. \tag{2}$$

- **Marginalized Error:** When all other parameters are integrated (marginalized) over their uncertainties, parameter degeneracies broaden the error bar to:

$$\sigma_{\mathrm{marg}}(\lambda_i) = \sqrt{(F^{-1})_{ii}} \geq \frac{1}{\sqrt{F_{ii}}}. \tag{3}$$

The ratio $\sqrt{(F^{-1})_{ii}F_{ii}} \geq 1$ measures the degree of parameter correlation; when off-diagonal elements of $\boldsymbol{F}$ are significant, marginalization substantially enlarges parameter uncertainties.

Fisher Matrix on the C_ℓ Spectrum

When the observable data are the spherical harmonic expansion coefficients $a_{\ell m}$ contaminated by instrumental pixel noise $a^{(\mathrm{n})}_{\ell m}$, the total measured coefficients are $a^{(\mathrm{tot})}_{\ell m} = a_{\ell m} + a^{(\mathrm{n})}_{\ell m}$.

Assuming independent pixel noise with variance σ^2_{n} per pixel area Ω_{pix}, the noise correlation in harmonic space evaluates to:

$$\langle a^{(\mathrm{n})}_{\ell m} a^{(\mathrm{n})*}_{\ell' m'} \rangle = \delta_{\ell\ell'}\delta_{mm'}\, w^{-1}, \tag{7.5.21}$$

where $w \equiv (\Omega_{\mathrm{pix}}\sigma^2_{\mathrm{n}})^{-1} = N_{\mathrm{pix}}/(4\pi\sigma^2_{\mathrm{n}})$ is the experimental weight per solid angle.

Because cosmological signal and detector noise are uncorrelated, the total data covariance matrix in the harmonic basis is diagonal:

$$\langle a^{(\mathrm{tot})}_{\ell m} a^{(\mathrm{tot})*}_{\ell' m'} \rangle = \delta_{\ell\ell'}\delta_{mm'}\left(C_\ell W_\ell + w^{-1}\right). \tag{7.5.22}$$

Treating the multipole bandpowers C_ℓ directly as the parameters to be estimated, Eq. (7.5.16) gives the Fisher matrix for the C_ℓ spectrum:

$$F_{\ell\ell'} = \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{C}^{-1}\left(\frac{\partial \boldsymbol{C}}{\partial C_\ell}\right)\boldsymbol{C}^{-1}\left(\frac{\partial \boldsymbol{C}}{\partial C_{\ell'}}\right)\right]$$

$$= \frac{1}{2} \sum_{m=-\ell}^{\ell} \frac{W_\ell}{C_\ell W_\ell + w^{-1}} \frac{W_\ell}{C_\ell W_\ell + w^{-1}} \delta_{\ell\ell'}$$
$$= \delta_{\ell\ell'} \frac{2\ell + 1}{2\left[C_\ell + (wW_\ell)^{-1}\right]^2}. \tag{7.5.23}$$

Because the Fisher matrix $F_{\ell\ell'}$ is diagonal in the multipole basis, the 1σ marginalized uncertainty on each measured C_ℓ is given by the celebrated **Knox formula**:

$$\delta C_\ell = \frac{1}{\sqrt{F_{\ell\ell}}} = \sqrt{\frac{2}{2\ell + 1}} \left(C_\ell + \frac{1}{wW_\ell} \right). \tag{7.5.24}$$

When the experiment covers only a fractional sky area $f_{\rm sky} < 1$, the effective number of independent harmonic modes is reduced to $(2\ell + 1) f_{\rm sky}$, modifying the error to:

$$\delta C_\ell = \sqrt{\frac{2}{(2\ell + 1) f_{\rm sky}}} \left(C_\ell + \frac{1}{wW_\ell} \right). \tag{7.5.25}$$

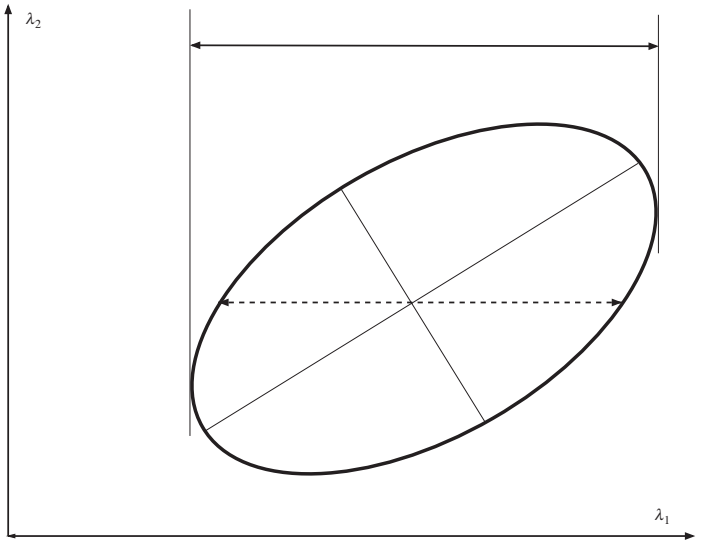


Figure 7.9: Comparison between the elliptical Gaussian Fisher matrix approximation (dashed ellipses) and the true, non-linear MCMC likelihood contours (solid curves), showing how the Fisher approximation understates errors when parameter degeneracies are curved.

Once the power spectrum uncertainties δC_ℓ are established, the Fisher matrix for the underlying cosmological parameters $\boldsymbol{\lambda}$ is obtained by applying the chain rule:

$$\tilde{F}_{ij}(\boldsymbol{\lambda}) = \sum_\ell \frac{\partial C_\ell}{\partial \lambda_i} \frac{1}{(\delta C_\ell)^2} \frac{\partial C_\ell}{\partial \lambda_j}. \tag{7.5.26}$$

While the Fisher matrix $\tilde{\boldsymbol{F}}$ provides an excellent initial forecast of experimental sensitivity, primary cosmological parameter distributions in real data often exhibit non-linear curved degeneracies (as shown in Fig. 7.9).

Worked example: an amplitude estimator from the C_ℓ spectrum

It is worth writing this Fisher machinery out as the same four-step, hand-computable recipe used for H_0 in §7.7.2: write the model, write the likelihood, differentiate, solve. Suppose we hold the *shape* of the power spectrum fixed at some fiducial template $C_\ell^{\rm fid}$ (computed once by

a Boltzmann code for a fixed $\{\Omega_{\rm b}h^2, \Omega_{\rm cdm}h^2, n_{\rm s}, \dots\}$) and ask only for the best-fitting overall amplitude A — physically, a rescaling of $A_{\rm s}$ at fixed shape:

$$C_\ell^{\rm th}(A) = A\, C_\ell^{\rm fid}. \tag{7.5.27}$$

For ℓ not too small, the bandpower estimator $\hat{C}_\ell$ of §7.4.4 is approximately Gaussian-distributed about $C_\ell^{\rm th}$ with variance $(\delta C_\ell)^2$ given by the Knox formula, Eq. (7.5.25) (evaluated at the fiducial model, since δC_ℓ itself depends only weakly on A near the best fit). The likelihood is then a product of Gaussians in $\hat{C}_\ell$, giving the χ^2

$$\chi^2(A) = \sum_\ell \frac{\left(\hat{C}_\ell - A\, C_\ell^{\rm fid}\right)^2}{(\delta C_\ell)^2}, \tag{7.5.28}$$

which has exactly the form of Eq. (7.7.3) with $\ell \leftrightarrow i$, $C_\ell^{\rm fid} \leftrightarrow E_i$, and $\delta C_\ell \leftrightarrow \sigma_i$. Differentiating and setting $\mathrm{d}\chi^2/\mathrm{d}A = 0$ gives, by the same algebra as Eq. (7.7.5),

$$\boxed{\hat{A} = \frac{\sum_\ell \hat{C}_\ell\, C_\ell^{\rm fid}/(\delta C_\ell)^2}{\sum_\ell (C_\ell^{\rm fid})^2/(\delta C_\ell)^2},} \qquad \sigma_{\hat{A}}^{-2} = \sum_\ell \frac{(C_\ell^{\rm fid})^2}{(\delta C_\ell)^2}. \tag{7.5.29}$$

The variance is not an independent result: since $\partial C_\ell^{\rm th}/\partial A = C_\ell^{\rm fid}$, it is precisely the single-parameter ($i = j = A$) case of the chain-rule Fisher matrix $\tilde{F}_{ij}$ above, so this toy amplitude estimator sits inside the same general formalism as the full multi-parameter fits of Fig. 7.10–7.11, which simply solve the analogous M-dimensional problem (for $\boldsymbol{\lambda} = \{\Omega_{\rm b}h^2, \Omega_{\rm cdm}h^2, n_{\rm s}, A_{\rm s}, \tau, \dots\}$) numerically rather than in closed form, since the individual $C_\ell(\boldsymbol{\lambda})$ depend on each parameter non-linearly. Two caveats specific to the CMB, both already flagged above, are worth restating here: (i) the Gaussian approximation for $\hat{C}_\ell$ breaks down at low ℓ, where $\hat{C}_\ell$ is really χ^2-distributed with $2\ell+1$ degrees of freedom (§7.4.4), so Planck and WMAP replace this χ^2 likelihood with the exact one at low multipoles; and (ii) a real analysis marginalizes over instrumental and foreground nuisance parameters exactly as $\mathcal{M}$ was profiled out of the supernova fit in §7.7.2.

In Fig. 7.10, we display the marginalized 2D posterior distributions obtained from the WMAP 3-year data and joint cosmological probes. In Fig. 7.11, we show the pre-launch forecasted parameter distributions for the Planck mission obtained via Markov Chain Monte Carlo sampling.

7.5.3 Model comparison and Bayesian evidence

In addition to estimating parameters within a specific cosmological model, we frequently want to compare two competing physical models, m_1 and m_2, with different parameter sets $\boldsymbol{\lambda}^{(1)} \in \mathbb{R}^{M_1}$ and $\boldsymbol{\lambda}^{(2)} \in \mathbb{R}^{M_2}$, to determine which model is favored by data.

A model with more free parameters will almost always achieve a higher maximum likelihood $\mathcal{L}_{\rm max}$ simply because it has additional degrees of freedom to fit statistical noise. However, scientific parsimony—embodied by **Occam's razor**—demands that a more complex model should be accepted only if its improvement in fit significantly outweighs the penalty of its added complexity.

In the Bayesian framework, the posterior odds ratio between two models given data $\boldsymbol{d}$ is:

$$\frac{P(m_1 \mid \boldsymbol{d})}{P(m_2 \mid \boldsymbol{d})} = \frac{P(\boldsymbol{d} \mid m_1)}{P(\boldsymbol{d} \mid m_2)} \frac{P(m_1)}{P(m_2)} = \frac{\mathcal{E}_1}{\mathcal{E}_2} \frac{P(m_1)}{P(m_2)}, \tag{7.5.30}$$

where $P(m_1)/P(m_2)$ is the prior odds ratio (usually set to unity in the absence of theoretical preference), and $\mathcal{E}_i \equiv P(\boldsymbol{d} \mid m_i)$ is the **Bayesian Evidence** (or marginal likelihood):

$$\mathcal{E}_i = \int \mathrm{d}^{M_i}\boldsymbol{\lambda}^{(i)}\, \mathcal{L}(\boldsymbol{d} \mid \boldsymbol{\lambda}^{(i)}, m_i)\, P(\boldsymbol{\lambda}^{(i)} \mid m_i). \tag{7.5.31}$$

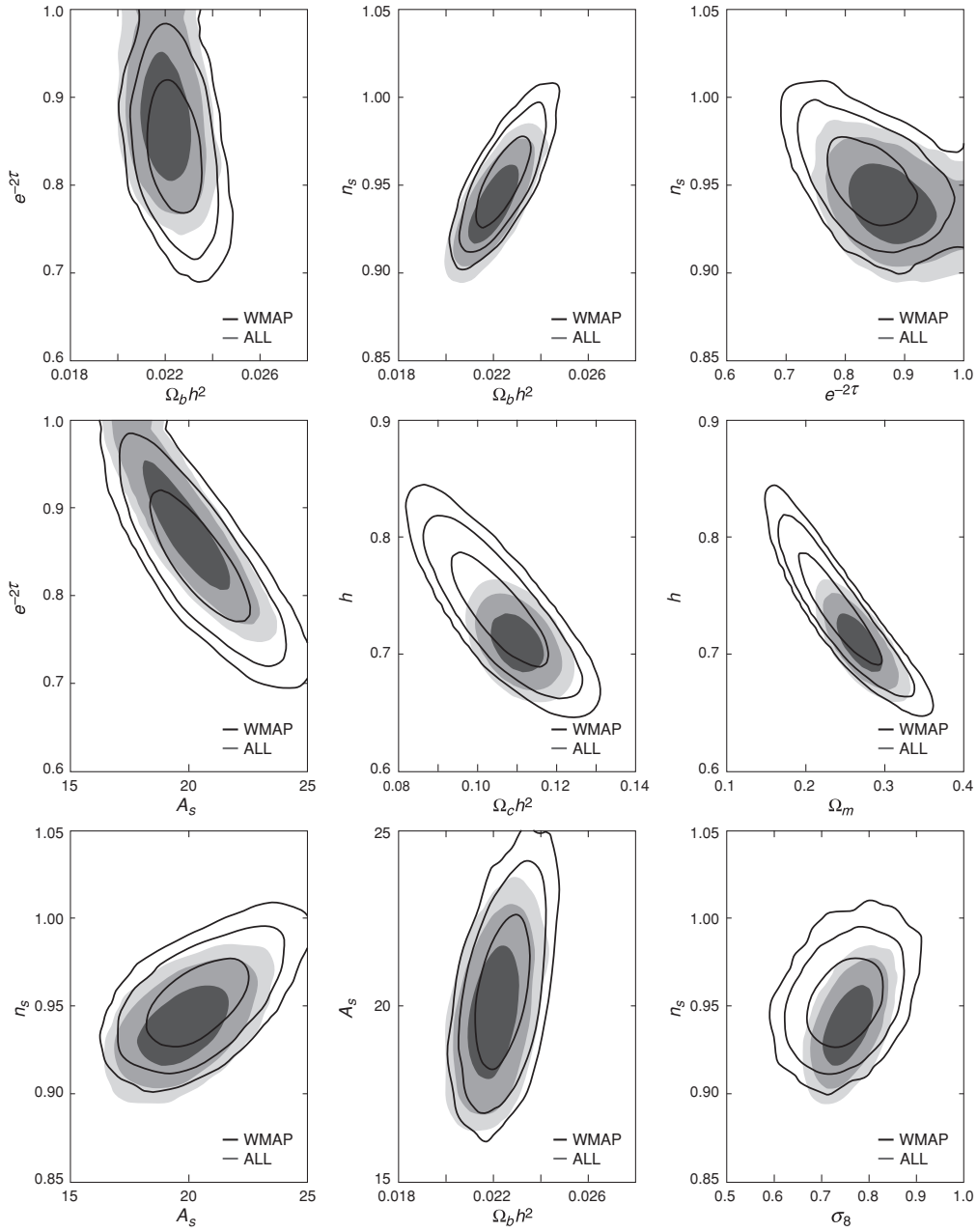


Figure 7.10: Marginalized 2D cosmological parameter posterior distributions from WMAP 3-year data alone (black contours) and combined with small-scale CMB, supernovae, and galaxy surveys (colored contours) for a flat 5-parameter ΛCDM model ($\Omega_{\rm b}h^2, \Omega_{\rm cdm}h^2, n_{\rm s}, A_{\rm s}, \tau$). Figure from Spergel et al. (2007).

The Occam Factor

If the posterior distribution is well approximated by a multivariate Gaussian centered at the best-fit point $\bar{\boldsymbol{\lambda}}^{(i)}$ with Fisher matrix $\boldsymbol{F}^{(i)}$, and the prior $P(\boldsymbol{\lambda}^{(i)} \mid m_i) = 1/\Delta\Omega_i(\text{prior})$ is uniform over a prior volume $\Delta\Omega_i(\text{prior})$, the evidence integral (7.5.31) evaluates under the Laplace approximation to:

$$\mathcal{E}_i \simeq \mathcal{L}^{(i)}_{\rm max} \cdot \frac{(2\pi)^{M_i/2}}{\sqrt{\det \boldsymbol{F}^{(i)}}} \cdot \frac{1}{\Delta\Omega_i(\text{prior})} = \mathcal{L}^{(i)}_{\rm max} \cdot \frac{\Delta\Omega_i(\text{data})}{\Delta\Omega_i(\text{prior})}. \tag{7.5.32}$$

The ratio $\Delta\Omega_i(\text{data})/\Delta\Omega_i(\text{prior}) \leq 1$ is known as the **Occam factor**. It measures the

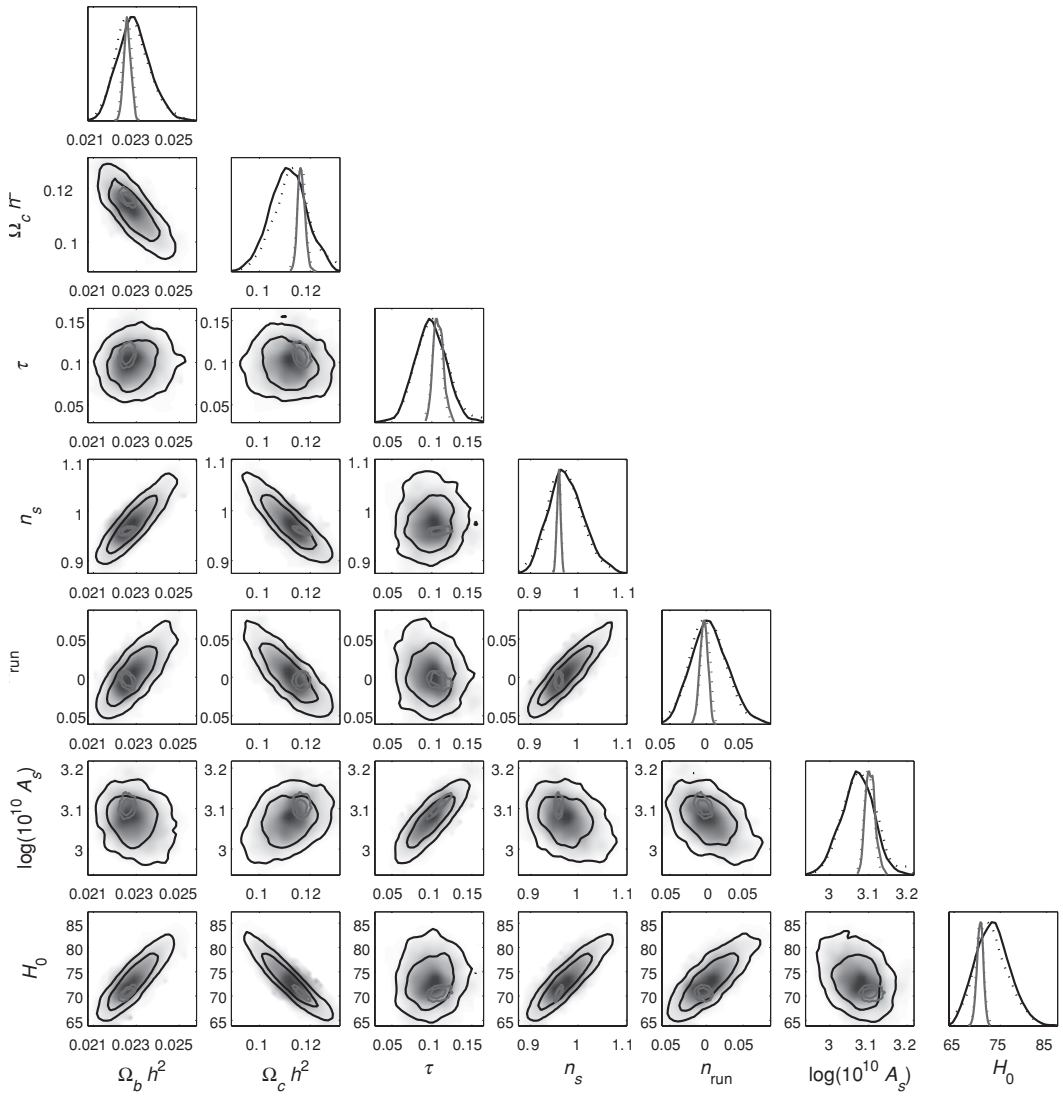


Figure 7.11: Marginalized 1D and 2D parameter probability distributions forecasted for the Planck satellite experiment for a 7-parameter flat model $(\Omega_{\rm b}h^2, \Omega_{\rm cdm}h^2, \tau, n_{\rm s}, n_{\rm run}, \ln(10^{10}A_{\rm s}), H_0)$. Figure from ESA (2005).

fraction of the prior parameter space that is consistent with the data:

- If a model introduces an unconstrained parameter that must be tuned to a narrow region to fit the data, $\Delta\Omega(\text{data})/\Delta\Omega(\text{prior}) \ll 1$, heavily penalizing the model's evidence.
- If the data fail to constrain the new parameter, the evidence drops, favoring the simpler nested model.

Fig. 7.12 illustrates the three characteristic scenarios:

- **Model (a):** Wide uninformative prior compared to posterior $\Longrightarrow$ strong Occam penalty.
- **Model (b):** Prior well matched to data posterior $\Longrightarrow$ maximal evidence.
- **Model (c):** Prior misaligned with data $\Longrightarrow$ exponentially suppressed likelihood.

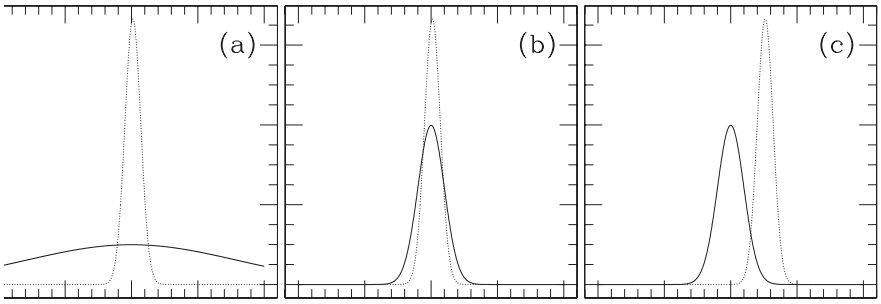


Figure 7.12: Three 1-parameter models with their prior (solid lines) and posterior (dotted curves). All three models have comparable maximum likelihoods. In (a), the prior volume is much wider than the posterior, leading to a small Occam factor. In (b) and (c), the widths are comparable, but model (c) lacks overlap between prior and data. Model (b) has the largest Bayesian evidence.

Nested Models in Cosmology

A common problem in cosmological model selection involves **nested models**, where a simpler model m_1 is obtained by fixing one or more parameters of a broader model m_2 (e.g., flat ΛCDM with $\Omega_K = 0$ nested inside non-flat models with free Ω_K, or scalar perturbations with tensor ratio $r = 0$ nested inside models with free r).

For example, current CMB data constrain spatial curvature to $\Omega_K = -0.01 \pm 0.03$. Whether flat space ($\Omega_K = 0$) is favored depends on the prior adopted for Ω_K:

- If the prior is chosen broadly as $\Omega_K \in [-1, 1]$, the open model receives an Occam penalty of $\sigma_K/\Delta\Omega_K \approx 0.03/2 = 0.015$, strongly favoring the flat model.
- If an inflationary prior $\Omega_K \in [-0.05, 0.05]$ is adopted, the Occam factor is $\approx 0.03/0.10 = 0.3$, leaving both models equally plausible.

7.5.4 Markov chain Monte Carlo methods (MCMC)

When the posterior probability distribution $P(\boldsymbol{\lambda} \mid \boldsymbol{d})$ is non-Gaussian, or when multi-dimensional degeneracies create curved, non-linear manifolds, the Fisher matrix approximation fails. In such regimes, **Markov Chain Monte Carlo (MCMC)** algorithms provide the gold standard for mapping the posterior.

Phase-Space Ensembles, Trajectory Sampling, and the Monte Carlo Principle

To build a rigorous physical intuition for why stochastic Monte Carlo sampling works so effectively in multi-dimensional parameter estimation, it is instructive to connect the procedure to the concept of an **ensemble** in classical statistical mechanics.

In statistical mechanics, an ensemble represents the formal sample space Ω of all possible microscopic configurations accessible to a physical system. The individual elements (microstates) of this sample space are points in the system's phase space $\Gamma = (\boldsymbol{q}, \boldsymbol{p})$. Under Hamiltonian dynamics, the microstate evolves along a continuous one-dimensional trajectory $\gamma(t) = (\boldsymbol{q}(t), \boldsymbol{p}(t))$ in phase space, naturally parameterized by time t (or by a dynamical phase angle $\theta(t) = \omega t$).

By the ergodic hypothesis, the long-term time average of any macroscopic observable along a single trajectory equals its statistical ensemble average over phase space:

$$\langle A \rangle_{\text{time}} \equiv \lim_{T\to\infty} \frac{1}{T} \int_0^T A(\boldsymbol{q}(t), \boldsymbol{p}(t))\, \mathrm{d}t = \int_\Gamma A(\boldsymbol{q}, \boldsymbol{p})\, \rho(\boldsymbol{q}, \boldsymbol{p})\, \mathrm{d}\boldsymbol{q}\, \mathrm{d}\boldsymbol{p} \equiv \langle A \rangle_{\text{ensemble}}. \tag{7.5.33}$$

Now suppose we wish to determine the spatial probability distribution $P(\boldsymbol{q})$ of the system, but we lack *a priori* knowledge of the analytical density $\rho(\boldsymbol{q}, \boldsymbol{p})$. A powerful Monte Carlo technique is to exploit the deterministic mapping $\boldsymbol{q}(t)$ by treating the trajectory parameter itself as a uniform random variable. Because the system sweeps through its orbit at a uniform rate in phase θ, we draw N independent, uniformly distributed random phase samples:

$$\theta_i \sim \text{Uniform}(0, 2\pi), \qquad i = 1, \ldots, N. \tag{7.5.34}$$

Pushing these random phases through the kinematic equation of motion $\boldsymbol{q}_i = \boldsymbol{q}(\theta_i)$ generates a synthetic ensemble of spatial microstates. Constructing a normalized histogram of $\{\boldsymbol{q}_i\}$ directly reconstructs the true spatial probability distribution $P(\boldsymbol{q})$ without needing to invert or integrate the distribution analytically.

Canonical Example: The Simple Harmonic Oscillator. Consider a one-dimensional classical harmonic oscillator of mass m, amplitude A, and natural frequency ω. Its phase-space trajectory is given by:

$$x(t) = A\cos(\omega t) \equiv A\cos\theta, \qquad p(t) = -m\omega A\sin(\omega t) \equiv -m\omega A\sin\theta. \tag{7.5.35}$$

Because the phase $\theta \in [0, 2\pi)$ is uniformly distributed with density $P(\theta) = \frac{1}{2\pi}$, the conservation of probability requires $P(x)|\,\mathrm{d}x\,| = 2P(\theta)|\,\mathrm{d}\theta\,| = \frac{\mathrm{d}\theta}{\pi}$, where the factor of 2 accounts for the two branches $\theta \in [0, \pi]$ and $\theta \in [\pi, 2\pi)$ that map onto each position $x \in (-A, A)$. Differentiating $x(\theta) = A\cos\theta$ gives $|\,\mathrm{d}x\,/\,\mathrm{d}\theta\,| = A\sin\theta = \sqrt{A^2 - x^2}$. The exact classical spatial probability density is therefore:

$$P(x) = \frac{1}{\pi\left|\frac{\mathrm{d}x}{\mathrm{d}\theta}\right|} = \frac{1}{\pi\sqrt{A^2 - x^2}} \qquad (|x| < A). \tag{7.5.36}$$

The probability density diverges at the classical turning points $x = \pm A$. Physically, this divergence reflects the fact that the particle's velocity $v(x) = \dot{x} = -\omega\sqrt{A^2 - x^2}$ vanishes at the turning points ($v \to 0$), so the particle spends the maximum fraction of each oscillation period near its extreme displacements.

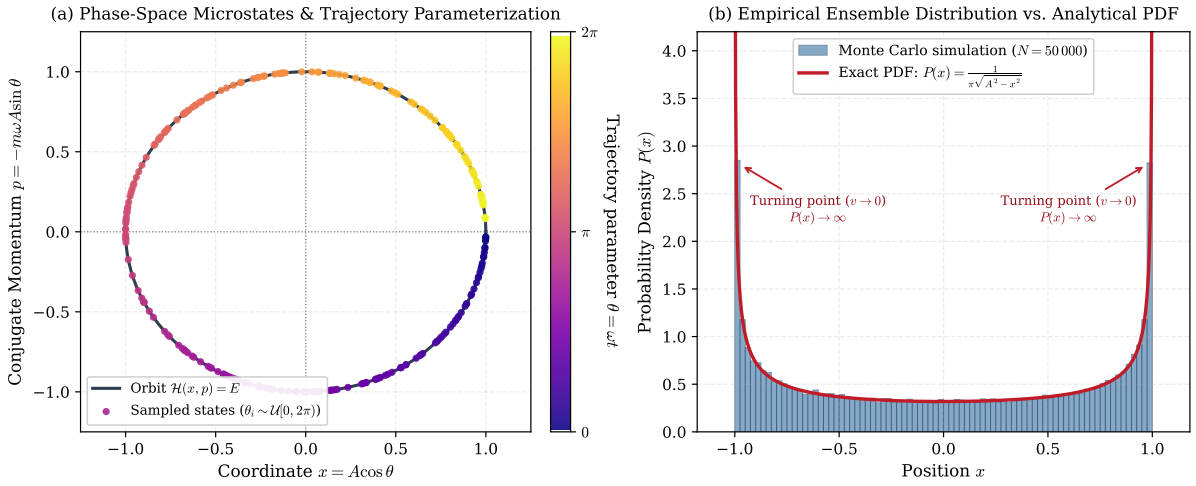


Figure 7.13: Monte Carlo simulation of the phase-space ensemble for a classical harmonic oscillator ($A = 1, \omega = 1$). (a) Phase-space orbit $\mathcal{H}(x, p) = E$ with $N = 250$ microstates sampled uniformly along the trajectory parameter $\theta = \omega t \sim \mathcal{U}[0, 2\pi)$. (b) Empirical histogram of $N = 50\,000$ simulated position samples $x_i = A\cos\theta_i$ (blue bars) compared against the exact analytical probability density $P(x) = \frac{1}{\pi\sqrt{A^2 - x^2}}$ (red curve), illustrating the characteristic U-shaped profile and turning-point kinematic slowdown.

In Listing 7.1, we provide a Python implementation demonstrating this trajectory-sampling Monte Carlo technique.

```python
import numpy as np
import matplotlib.pyplot as plt

# 1. Exact classical spatial probability density: P(x) = 1 / (pi * sqrt(A^2 - x^2))
def analytical_pdf(x, A=1.0):
    pdf = np.full_like(x, np.nan, dtype=np.float64)
    inside = np.abs(x) < A
    pdf[inside] = 1.0 / (np.pi * np.sqrt(A**2 - x[inside]**2))
    return pdf

# 2. Sample microstate ensemble by uniformly parameterizing the phase-space trajectory
def simulate_ensemble(A=1.0, N=50000):
    theta = np.random.uniform(0.0, 2.0 * np.pi, size=N)  # Random phase theta = omega * t
    x = A * np.cos(theta)                                # Position on orbit
    p = -A * np.sin(theta)                               # Conjugate momentum
    return x, p

# 3. Execution and empirical density reconstruction
if __name__ == "__main__":
    A, N = 1.0, 50000
    x_samples, p_samples = simulate_ensemble(A, N)
    x_grid = np.linspace(-0.998 * A, 0.998 * A, 1000)
    pdf_exact = analytical_pdf(x_grid, A)

    # Plot empirical histogram against exact analytical PDF
    plt.figure(figsize=(7, 4.2), dpi=300)
    plt.hist(x_samples, bins=80, density=True, range=(-A, A),
             color="steelblue", alpha=0.6, edgecolor="white", lw=0.4,
             label=f"Monte Carlo Samples ($N = {N:,}$)")
    plt.plot(x_grid, pdf_exact, "r-", lw=2.2,
             label=r"Exact PDF: $P(x) = \frac{1}{\pi\sqrt{A^2-x^2}}$")
    plt.xlabel(r"Position $x$")
    plt.ylabel(r"Probability Density $P(x)$")
    plt.title(r"Harmonic Oscillator Phase-Space Ensemble Sampling")
    plt.legend(loc="upper center")
    plt.tight_layout()
    plt.show()
```

Listing 7.1: Vectorized Monte Carlo simulation and analytical comparison for the classical harmonic oscillator phase-space ensemble.

This elementary one-dimensional example encapsulates the grand philosophical insight of Monte Carlo methods in modern cosmological data analysis: when the analytical target distribution $P(\boldsymbol{\lambda} \mid \boldsymbol{d})$ is high-dimensional and non-trivial, we can explore and reconstruct the entire ensemble by generating samples along parameter paths through stochastic exploration.

The Target Distribution and Monte Carlo Integration

Given data $\boldsymbol{d}$ and prior $P(\boldsymbol{\lambda})$, the unnormalized posterior probability density is:

$$P^*(\boldsymbol{\lambda}) \equiv \mathcal{L}(\boldsymbol{d} \mid \boldsymbol{\lambda})\, P(\boldsymbol{\lambda}). \tag{7.5.37}$$

The true normalized posterior is $p(\boldsymbol{\lambda}) = P^*(\boldsymbol{\lambda})/\mathcal{E}$.

Our goal is to construct a representative sample $\mathcal{S} = \{\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}, \ldots, \boldsymbol{\lambda}^{(R)}\}$ of points in parameter space such that the number of sample points in any volume element $\mathrm{d}^M\boldsymbol{\lambda}$ is proportional to $p(\boldsymbol{\lambda})\mathrm{d}^M\boldsymbol{\lambda}$.

With such a representative chain in hand:

1. **Expectation Values and Integrals:** Any integral $\langle f \rangle = \int f(\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) \mathrm{d}^M \boldsymbol{\lambda}$ is estimated simply by the sample mean:
$$\langle f \rangle \simeq \frac{1}{R} \sum_{k=1}^{R} f(\boldsymbol{\lambda}^{(k)}). \tag{7.5.38}$$
2. **Marginalization:** To obtain the 1D or 2D marginalized posterior distribution of a parameter subset $\boldsymbol{\mu} = (\lambda_1, \lambda_2)$, we simply project (histogram) the sampled points onto those coordinates. The integral over all remaining parameters $\int \mathrm{d}\lambda_3 \dots \mathrm{d}\lambda_M p(\boldsymbol{\lambda})$ is computed automatically by ignoring those coordinates in the sample!

The Metropolis–Hastings Algorithm

The standard algorithm for generating a representative Markov chain is the **Metropolis–Hastings algorithm**:

1. Start at an arbitrary parameter position $\boldsymbol{\lambda}^{(1)}$ within the prior support.
2. At iteration step t, with current state $\boldsymbol{\lambda}^{(t)}$, draw a candidate proposal $\boldsymbol{\lambda}^*$ from a proposal probability density $Q(\boldsymbol{\lambda}^* \mid \boldsymbol{\lambda}^{(t)})$.
3. Compute the Hastings acceptance ratio r:
$$r = \frac{P^*(\boldsymbol{\lambda}^*)\, Q(\boldsymbol{\lambda}^{(t)} \mid \boldsymbol{\lambda}^*)}{P^*(\boldsymbol{\lambda}^{(t)})\, Q(\boldsymbol{\lambda}^* \mid \boldsymbol{\lambda}^{(t)})}. \tag{7.5.39}$$
For a symmetric proposal distribution, $Q(\boldsymbol{\lambda}^* \mid \boldsymbol{\lambda}^{(t)}) = Q(\boldsymbol{\lambda}^{(t)} \mid \boldsymbol{\lambda}^*)$, this simplifies to the standard Metropolis ratio:
$$r = \frac{P^*(\boldsymbol{\lambda}^*)}{P^*(\boldsymbol{\lambda}^{(t)})}. \tag{7.5.40}$$
4. Accept the proposed state with probability $\alpha = \min(1, r)$:
   - Draw a random number $u \sim \mathrm{Uniform}(0, 1)$.
   - If $u \leq \alpha$, accept the proposal and set $\boldsymbol{\lambda}^{(t+1)} = \boldsymbol{\lambda}^*$.
   - If $u > \alpha$, reject the proposal and retain the current state: $\boldsymbol{\lambda}^{(t+1)} = \boldsymbol{\lambda}^{(t)}$.
5. Repeat for R iterations.

Derivation: Detailed Balance and Biasing the Random Walker
Why does the Metropolis–Hastings acceptance rule guarantee that the random walker samples the target posterior $p(\boldsymbol{\lambda})$?
Consider a Markov chain with transition probability $T(\boldsymbol{\lambda} \to \boldsymbol{\lambda}')$ defined as the probability of proposing $\boldsymbol{\lambda}'$ times the probability of accepting it:
$$T(\boldsymbol{\lambda} \to \boldsymbol{\lambda}') = Q(\boldsymbol{\lambda}' \mid \boldsymbol{\lambda})\, \alpha(\boldsymbol{\lambda} \to \boldsymbol{\lambda}') \qquad (\text{for } \boldsymbol{\lambda}' \neq \boldsymbol{\lambda}). \tag{1}$$
A sufficient condition for $p(\boldsymbol{\lambda})$ to be the unique stationary distribution of the chain is the condition of **Detailed Balance**:
$$p(\boldsymbol{\lambda})\, T(\boldsymbol{\lambda} \to \boldsymbol{\lambda}') = p(\boldsymbol{\lambda}')\, T(\boldsymbol{\lambda}' \to \boldsymbol{\lambda}). \tag{2}$$
Substituting $T(\boldsymbol{\lambda} \to \boldsymbol{\lambda}')$:
$$p(\boldsymbol{\lambda})\, Q(\boldsymbol{\lambda}' \mid \boldsymbol{\lambda})\, \alpha(\boldsymbol{\lambda} \to \boldsymbol{\lambda}') = p(\boldsymbol{\lambda}')\, Q(\boldsymbol{\lambda} \mid \boldsymbol{\lambda}')\, \alpha(\boldsymbol{\lambda}' \to \boldsymbol{\lambda}). \tag{3}$$

Without loss of generality, assume $p(\boldsymbol{\lambda}')Q(\boldsymbol{\lambda} \mid \boldsymbol{\lambda}') \leq p(\boldsymbol{\lambda})Q(\boldsymbol{\lambda}' \mid \boldsymbol{\lambda})$. Then:

$$\alpha(\boldsymbol{\lambda} \to \boldsymbol{\lambda}') = \frac{p(\boldsymbol{\lambda}')Q(\boldsymbol{\lambda} \mid \boldsymbol{\lambda}')}{p(\boldsymbol{\lambda})Q(\boldsymbol{\lambda}' \mid \boldsymbol{\lambda})} \leq 1, \qquad \alpha(\boldsymbol{\lambda}' \to \boldsymbol{\lambda}) = 1. \tag{4}$$

Multiplying:

$$\begin{aligned} p(\boldsymbol{\lambda})Q(\boldsymbol{\lambda}' \mid \boldsymbol{\lambda})\,\alpha(\boldsymbol{\lambda} \to \boldsymbol{\lambda}') &= p(\boldsymbol{\lambda})Q(\boldsymbol{\lambda}' \mid \boldsymbol{\lambda}) \left(\frac{p(\boldsymbol{\lambda}')Q(\boldsymbol{\lambda} \mid \boldsymbol{\lambda}')}{p(\boldsymbol{\lambda})Q(\boldsymbol{\lambda}' \mid \boldsymbol{\lambda})}\right) \\ &= p(\boldsymbol{\lambda}')Q(\boldsymbol{\lambda} \mid \boldsymbol{\lambda}') = p(\boldsymbol{\lambda}')Q(\boldsymbol{\lambda} \mid \boldsymbol{\lambda}')\,\alpha(\boldsymbol{\lambda}' \to \boldsymbol{\lambda}). \end{aligned} \tag{5}$$

Detailed balance holds identically!
Physical Intuition: An unguided random walker moving according to $Q(\boldsymbol{\lambda}' \mid \boldsymbol{\lambda})$ would explore all regions uniformly. The Metropolis–Hastings acceptance probability $\alpha = \min(1, P^*(\boldsymbol{\lambda}^*)/P^*(\boldsymbol{\lambda}^{(t)}))$ acts as a statistical bias:

- Steps toward *higher* posterior probability ($P^*(\boldsymbol{\lambda}^*) \geq P^*(\boldsymbol{\lambda}^{(t)})$) are **always accepted** ($\alpha = 1$).
- Steps toward *lower* probability ($P^*(\boldsymbol{\lambda}^*) < P^*(\boldsymbol{\lambda}^{(t)})$) are accepted only with probability $r < 1$. When rejected, the walker remains at the current state, adding another copy of $\boldsymbol{\lambda}^{(t)}$ to the chain.

This precisely biases the walker into spending time in each volume element in exact proportion to $p(\boldsymbol{\lambda})\mathrm{d}^M\boldsymbol{\lambda}$, converting a blind random walk into an ergodic posterior sampler.

Practical MCMC Diagnostics and Convergence

To ensure that an MCMC chain has converged to a faithful posterior sample:

- **Burn-in Period:** The initial segment of the chain (typically 1000–5000 steps) reflects the arbitrary starting point rather than the stationary distribution and must be discarded.
- **Proposal Tuning:** The proposal distribution Q is typically chosen as a multivariate Gaussian whose covariance is tuned to the inverse Fisher matrix $\boldsymbol{F}^{-1}$. Rotating parameter coordinates to the eigenvectors of the covariance matrix decouples correlations and reduces the number of samples required from $R_{\min} \propto \delta_{\mathrm{m}}^2$ to $R_{\min} \propto D$. The step size is scaled to achieve an optimal acceptance rate of $\approx$ 20–30% in high dimensions (Roberts et al., 1997).
- **Gelman–Rubin Diagnostic ($\hat{R}$):** To verify convergence and avoid getting trapped in local maxima or narrow degenerate canyons, multiple independent chains are launched with dispersed initial points. The variance between chains is compared to the variance within chains via the potential scale reduction factor $\hat{R}$; convergence is declared when $\hat{R} - 1 < 0.01$ for all parameters.

7.6 Degeneracies

A fundamental limitation in cosmological parameter estimation from CMB data alone is the existence of **parameter degeneracies**—continuous trajectories in multi-dimensional parameter space along which the CMB power spectrum remains nearly invariant.

7.6.1 Curvature (The Geometrical Degeneracy)

The most prominent and physically important degeneracy is the **geometrical degeneracy** involving spatial curvature Ω_K, dark energy Ω_Λ, and the Hubble constant h.

Recall that the physical sound horizon at decoupling $r_\mathrm{s}(z_\mathrm{dec})$ and the detailed evolution of perturbations prior to recombination are entirely determined by local microphysics and the early physical matter densities:

$$\omega_\mathrm{b} \equiv \Omega_\mathrm{b} h^2, \qquad \omega_\mathrm{cdm} \equiv \Omega_\mathrm{cdm} h^2, \qquad \omega_\gamma \equiv \Omega_\gamma h^2. \tag{7.6.1}$$

If we consider two different cosmological models that share identical values of $(\omega_\mathrm{b}, \omega_\mathrm{cdm}, \omega_\gamma)$, their early-universe acoustic oscillations, baryon drag, and Silk damping scales will be completely indistinguishable.

The angular locations of the acoustic peaks on the sky are given by:

$$\theta_n = \frac{\lambda_n}{D_\mathrm{A}(z_\mathrm{dec})} \implies \ell_n \simeq \frac{n\pi D_\mathrm{A}(z_\mathrm{dec})}{r_\mathrm{s}(z_\mathrm{dec})}. \tag{7.6.2}$$

Consequently, if we alter the spatial curvature Ω_K, dark energy density Ω_Λ, and Hubble constant h in such a manner that the comoving angular diameter distance $D_\mathrm{A}(z_\mathrm{dec})$ remains strictly constant, the entire high-ℓ acoustic peak sequence ℓ_n is held invariant!

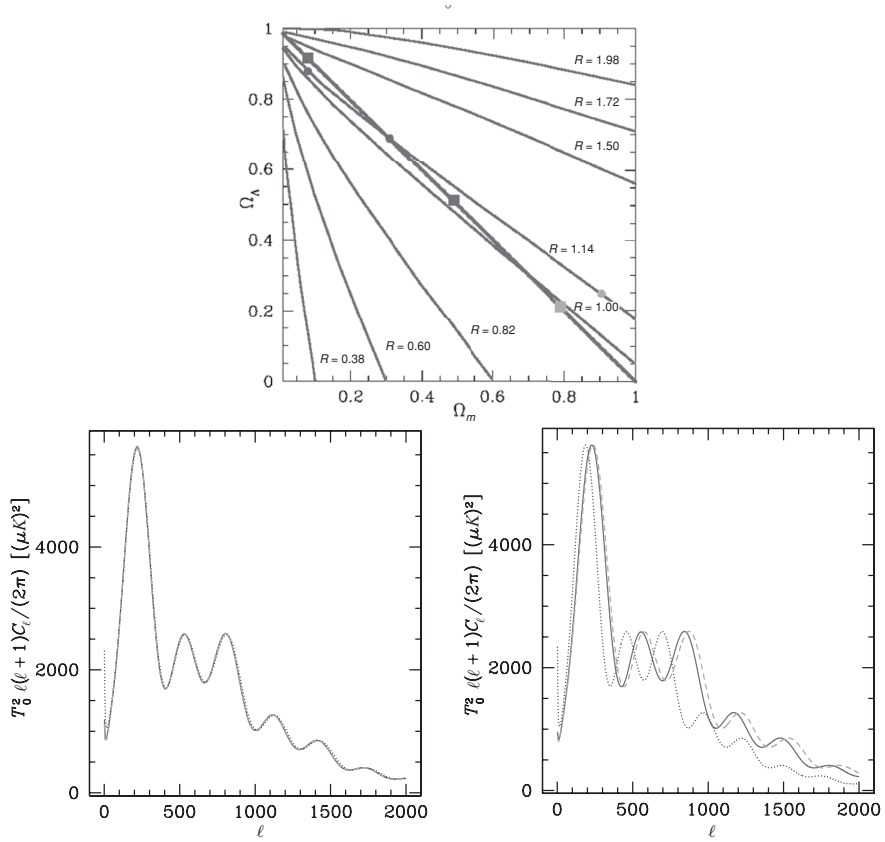


Figure 7.14: The geometrical degeneracy. Upper panel: The degeneracy lines of constant angular diameter distance $D_\mathrm{A}(z_\mathrm{dec})$ in the $(\Omega_\mathrm{m}, \Omega_\Lambda)$ plane for fixed $\Omega_\mathrm{b} h^2 = 0.022$ and $\Omega_\mathrm{m} h^2 = 0.12$. Lower panel: The corresponding relation between Ω_m and the Hubble parameter h. Along each curve, the CMB power spectrum at $\ell \geq 50$ is virtually identical.

The comoving angular diameter distance to decoupling is:

$$D_\mathrm{A}(z_\mathrm{dec}) = \frac{c}{(1+z_\mathrm{dec}) H_0 \sqrt{|\Omega_K|}}\, \chi_K\left(\sqrt{|\Omega_K|}\int_0^{z_\mathrm{dec}} \frac{\mathrm{d}z}{\sqrt{\Omega_\mathrm{m}(1+z)^3 + \Omega_\Lambda + \Omega_K(1+z)^2}}\right). \tag{7.6.3}$$

As illustrated in Fig. 7.14, one can smoothly trade off spatial curvature Ω_K against dark energy Ω_Λ along lines of constant D_A. The only differences between models along this degeneracy line appear at very low multipoles ($\ell \lesssim 20$) via the late Integrated Sachs–Wolfe (ISW) effect sourced by the time variation of gravitational potentials when dark energy or curvature becomes important. However, cosmic variance at $\ell \leq 20$ is so large that primary CMB temperature data alone cannot break this degeneracy.

To break the geometrical degeneracy, one must combine the CMB with low-redshift geometric probes, such as Baryon Acoustic Oscillations (BAO), Type Ia Supernovae (SNe Ia), or direct local measurements of the Hubble constant H_0.

7.6.2 Scalar Spectral Index, Tensor Component, and Optical Depth

A second important degeneracy involves the primordial scalar spectral index n_s, the tensor-to-scalar ratio r, and the reionization optical depth τ.

Primordial tensor perturbations (gravitational waves) contribute to CMB temperature anisotropies only on angular scales that were super-Hubble prior to decoupling ($\ell \lesssim 80$); on smaller scales, tensor modes enter the horizon during the radiation era and decay rapidly. Thus, adding a tensor component boosts the low-ℓ Sachs–Wolfe plateau without altering the acoustic peaks.

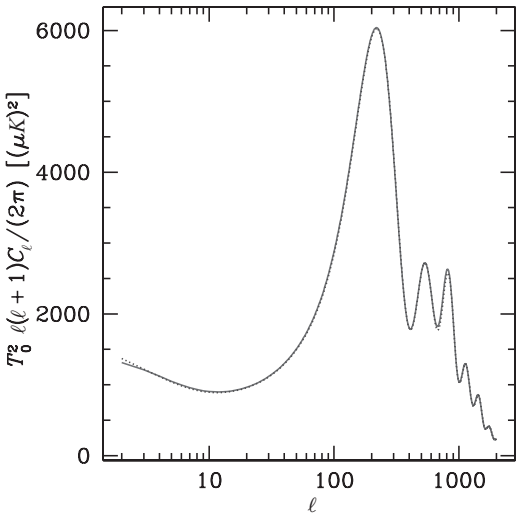


Figure 7.15: Comparison between a purely scalar model with red tilt $n_\mathrm{s} = 0.96$ (solid line) and a model with a significant tensor contribution of $r = 0.3$ and scale-invariant tilt $n_\mathrm{s} = 1.0$ (dotted line). Both models lie within 1σ–2σ of concordance parameters and produce nearly identical temperature power spectra.

A red scalar tilt ($n_\mathrm{s} < 1$) tilts the entire primordial spectrum, enhancing power at low multipoles relative to high multipoles. As shown in Fig. 7.15, a model with $r = 0.3$ and $n_\mathrm{s} = 1.0$ produces a temperature power spectrum virtually indistinguishable from a purely scalar model with $n_\mathrm{s} = 0.96$.

Furthermore, Thomson rescattering during reionization damps all acoustic peaks by a factor $\mathrm{e}^{-2\tau}$ ($\ell \gg 10$), which can be compensated by scaling up the primordial amplitude A_s (the $A_\mathrm{s}\mathrm{e}^{-2\tau}$ degeneracy).

Breaking the Degeneracy with Polarization

These degeneracies are broken definitively by polarization observations:

1. **Large-Scale E-Mode Polarization:** Rescattering off free electrons during reionization generates a distinctive bump in the EE power spectrum at multipoles $\ell \sim 2$–10 with amplitude $C_\ell^{EE} \propto \tau^2$, measuring τ directly and breaking the $A_\mathrm{s}\mathrm{e}^{-2\tau}$ degeneracy.

2. **Primordial B-Mode Polarization:** Because scalar perturbations cannot produce curl-type polarization at linear order, detecting B-modes at $\ell \sim 80$ uniquely measures the tensor amplitude r, isolating gravitational waves and testing the inflationary consistency relation:
$$r = -8\, n_\mathrm{T}. \tag{7.6.4}$$

7.6.3 Initial Conditions: Adiabatic vs. Isocurvature Modes

Standard single-field slow-roll inflation generates purely adiabatic initial fluctuations. However, multi-field inflation or early-universe phase transitions can excite non-adiabatic **isocurvature modes**.

For a cosmological fluid containing photons, baryons, cold dark matter, and neutrinos, there exist four regular perturbation modes:

1. The **Adiabatic mode** (all species share the same fractional perturbation $\delta\rho_i/\dot{\rho}_i$).

2. The **Cold Dark Matter Isocurvature mode** (entropy perturbation $\mathcal{S}_{\mathrm{c}\gamma} = \delta_\mathrm{c} - \frac{3}{4}\delta_\gamma$).

3. The **Neutrino Density Isocurvature mode** ($\mathcal{S}_{\nu\gamma} = \delta_\nu - \delta_\gamma$).

4. The **Neutrino Velocity Isocurvature mode** (relative velocity perturbation between neutrinos and the photon–baryon fluid).

Allowing for arbitrary cross-correlations between these four modes introduces a 4×4 symmetric positive semi-definite matrix of primordial amplitudes (10 free parameters) plus four individual spectral tilts.

Isocurvature perturbations shift the phase of the acoustic oscillations: for isocurvature modes, the acoustic peaks appear at $c_\mathrm{s} k_n \eta_\mathrm{dec} = (2n+1)\pi/2$, shifting peak locations and changing odd/even peak height ratios. When isocurvature fractions are left free, the acoustic peaks no longer uniquely determine the angular diameter distance, severely degrading parameter constraints from CMB temperature data alone.

However, while isocurvature modes contribute substantially to large-scale CMB anisotropies, they generate negligible matter density fluctuations on large scales. Comparing CMB data with the large-scale matter power spectrum from galaxy surveys (SDSS and 2dFGRS) places stringent bounds on non-adiabatic modes, demonstrating that the observed Universe is consistent with purely adiabatic initial perturbations and limiting any isocurvature fraction to $\leq 20\%$ (Trotta, 2006).

Running of the Spectral Index

When testing inflation, one often parameterizes a scale-dependent tilt by allowing for a **running spectral index** $\alpha \equiv \mathrm{d}n_\mathrm{s}/\mathrm{d}\ln k$:

$$k^3 P_\Phi(k) = A_* \left(\frac{k}{k_*}\right)^{n_\mathrm{s} - 1 + \frac{1}{2}\alpha \ln(k/k_*)}, \tag{7.6.5}$$

where k_* is a chosen pivot scale. Simple slow-roll inflation predicts running to be second-order in slow-roll parameters ($\alpha \sim \mathcal{O}(10^{-3})$); detecting a large running ($\alpha \sim -0.05$) would rule out standard single-field models.

Table 7.1: Joint parameter constraints for a flat ΛCDM model with purely scalar perturbations from WMAP 3-year data alone and combined with small-scale CMB experiments (CBI, VSA, ACBAR, BOOMERanG) and the 2dFGRS galaxy survey (from Spergel et al., 2007).

Parameter	WMAP only	WMAP+CBI+VSA	WMAP+ACBAR+B03	WMAP+2dFGRS
$100\,\Omega_{\rm b}h^2$	$2.233^{+0.072}_{-0.091}$	$2.212^{+0.066}_{-0.084}$	$2.231^{+0.070}_{-0.088}$	$2.223^{+0.066}_{-0.083}$
$\Omega_{\rm m}h^2$	$0.1268^{+0.0072}_{-0.0095}$	$0.1233^{+0.0070}_{-0.0086}$	$0.1259^{+0.0077}_{-0.0095}$	$0.1262^{+0.0045}_{-0.0062}$
h	$0.734^{+0.028}_{-0.038}$	$0.743^{+0.027}_{-0.037}$	$0.739^{+0.028}_{-0.038}$	$0.732^{+0.018}_{-0.025}$
$A_{\rm s}$	$0.801^{+0.043}_{-0.054}$	$0.796^{+0.042}_{-0.052}$	$0.798^{+0.046}_{-0.054}$	$0.799^{+0.042}_{-0.051}$
τ	$0.088^{+0.028}_{-0.034}$	$0.088^{+0.027}_{-0.033}$	$0.088^{+0.030}_{-0.033}$	$0.083^{+0.027}_{-0.031}$
$n_{\rm s}$	$0.951^{+0.015}_{-0.019}$	$0.947^{+0.014}_{-0.017}$	$0.951^{+0.015}_{-0.020}$	$0.948^{+0.014}_{-0.018}$
σ_8	$0.744^{+0.050}_{-0.060}$	$0.722^{+0.043}_{-0.053}$	$0.739^{+0.047}_{-0.059}$	$0.737^{+0.033}_{-0.045}$
$\Omega_{\rm m}$	$0.238^{+0.030}_{-0.041}$	$0.226^{+0.026}_{-0.036}$	$0.233^{+0.029}_{-0.041}$	$0.236^{+0.016}_{-0.024}$

Table 7.2: As in Table 7.1, but combining WMAP 3-year data with SDSS galaxy clustering, Type Ia Supernovae (SNLS and SN Gold), and weak gravitational lensing (CFHTLS) (from Spergel et al., 2007).

Parameter	WMAP+SDSS	WMAP+SNLS	WMAP+SN Gold	WMAP+CFHTLS
$100\,\Omega_{\rm b}h^2$	$2.233^{+0.062}_{-0.086}$	$2.233^{+0.069}_{-0.088}$	$2.227^{+0.065}_{-0.082}$	$2.255^{+0.062}_{-0.083}$
$\Omega_{\rm m}h^2$	$0.1329^{+0.0056}_{-0.0075}$	$0.1295^{+0.0056}_{-0.0072}$	$0.1349^{+0.0056}_{-0.0071}$	$0.1408^{+0.0034}_{-0.0050}$
h	$0.709^{+0.024}_{-0.032}$	$0.723^{+0.021}_{-0.030}$	$0.701^{+0.020}_{-0.026}$	$0.687^{+0.016}_{-0.024}$
$A_{\rm s}$	$0.813^{+0.042}_{-0.052}$	$0.808^{+0.044}_{-0.051}$	$0.827^{+0.045}_{-0.053}$	$0.846^{+0.037}_{-0.047}$
τ	$0.079^{+0.029}_{-0.032}$	$0.085^{+0.028}_{-0.032}$	$0.079^{+0.028}_{-0.034}$	$0.088^{+0.026}_{-0.032}$
$n_{\rm s}$	$0.948^{+0.015}_{-0.018}$	$0.950^{+0.015}_{-0.019}$	$0.946^{+0.015}_{-0.019}$	$0.953^{+0.015}_{-0.019}$
σ_8	$0.772^{+0.036}_{-0.048}$	$0.758^{+0.038}_{-0.052}$	$0.784^{+0.035}_{-0.049}$	$0.826^{+0.022}_{-0.035}$
$\Omega_{\rm m}$	$0.266^{+0.026}_{-0.036}$	$0.249^{+0.024}_{-0.031}$	$0.276^{+0.023}_{-0.031}$	$0.299^{+0.019}_{-0.025}$

7.7 Complementary observations

Although CMB anisotropies provide our most pristine cosmological dataset, external astrophysical measurements are essential to break degeneracies and independently cross-check the cosmological model.

7.7.1 The Hubble Parameter $H(z)$ and Type Ia Supernovae

Measuring cosmological distances is notoriously challenging; historically, Edwin Hubble originally overestimated the expansion rate H_0 by a factor of 7 due to Cepheid distance calibration errors.

In an expanding universe, distance measurements rely on **standard candles** (objects of known intrinsic luminosity L) or **standard rulers** (objects of known physical length R). For small redshifts ($z \lesssim 0.1$), the luminosity distance $D_{\rm L}(z)$ and angular diameter distance $D_{\rm A}(z)$ reduce to the linear Hubble relation cz/H_0. At higher redshifts, the luminosity distance depends on the cosmological geometry:

$$D_{\rm L}(z) = \frac{c\,(1+z)}{H_0\sqrt{|\Omega_K|}}\,\chi_K\left(H_0\int_0^z \frac{{\rm d}z'}{H(z')}\right). \tag{7.7.1}$$

Type Ia Supernovae (SNe Ia) serve as standardizable candles. Arising from carbon-oxygen white dwarfs in binary systems that accrete mass until reaching the Chandrasekhar stability limit ($M_{\rm Ch} \approx 1.4\,M_\odot$), their thermonuclear explosion releases a remarkably consistent

peak luminosity. The small intrinsic dispersion is corrected using the empirical Phillips relation between peak luminosity and light-curve decline rate, yielding distance precisions of $\sim 7\%$.

High-redshift supernova surveys (SNLS, ESSENCE, HST) measuring $D_{\rm L}(z)$ up to $z \approx 1.7$ provided the historical breakthrough discovery of the accelerated cosmic expansion, requiring dark energy ($\Omega_\Lambda \approx 0.7$) with negative pressure $w \approx -1$.

7.7.2 Worked example: constructing a maximum-likelihood estimator

Sections 7.5.1–7.5.2 derived the likelihood and Fisher-matrix machinery in full generality, for an arbitrary data vector $\boldsymbol{d}$ and an arbitrary $\boldsymbol{\lambda}$-dependent covariance $\boldsymbol{C}(\boldsymbol{\lambda})$. It is worth seeing that machinery collapse to the completely explicit, hand-computable form it takes in an actual distance-ladder analysis, using $H(z)$ and Type Ia supernova data as in §7.7.1. The recipe is always the same four steps: (1) write down the observational model, (2) write down the likelihood, (3) set its derivative to zero, and (4) solve — analytically if the model is linear in the parameter, numerically otherwise.

The observational model

Consider N independent measurements y_i at redshifts z_i, each with Gaussian uncertainty σ_i. Two standard examples:

- **Cosmic chronometers** (differential ages of passively evolving galaxies) measure $H(z_i)$ directly, with model $H_{\rm th}(z_i; H_0, \Omega_{m0}) = H_0\, E(z_i; \Omega_{m0})$ and $E(z)$ the normalized expansion rate of Eq. (2.3.19).
- **Type Ia supernovae** measure distance moduli $\mu_i = m_i - M$, with theoretical model $\mu_{\rm th}(z; H_0, \boldsymbol{\Omega}) = 5\log_{10}(D_{\rm L}(z; \boldsymbol{\Omega})/{\rm Mpc}) + 25$, $D_{\rm L}$ given by Eq. (7.7.1).

From the likelihood to χ^2

For uncorrelated Gaussian errors, this corresponds directly to the deterministic mean-shifting likelihood introduced in Eq. (7.5.7) of §7.5.1 (where parameters shift the expected value $y_{\rm th}(z_i; \boldsymbol{\theta})$ against fixed instrumental uncertainties σ_i, rather than shifting the variance as in the CMB covariance matrix $\boldsymbol{C}(\boldsymbol{\lambda})$):

$$\ln L(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{N}\left[\frac{y_i - y_{\rm th}(z_i;\boldsymbol{\theta})}{\sigma_i}\right]^2 - \sum_{i=1}^{N}\ln\left(\sqrt{2\pi}\sigma_i\right). \tag{7.7.2}$$

Maximizing $\ln L$ is equivalent to minimizing the weighted sum of squared residuals, the χ^2 objective function

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N}\frac{\left(y_i - y_{\rm th}(z_i;\boldsymbol{\theta})\right)^2}{\sigma_i^2}, \tag{7.7.3}$$

or, if the data points possess correlated measurement noise described by an instrumental covariance matrix $\boldsymbol{N}$, its matrix generalization $\chi^2(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{y}_{\rm th})^{\top}\boldsymbol{N}^{-1}(\boldsymbol{y} - \boldsymbol{y}_{\rm th})$, as given in Eq. (7.5.6).

A closed-form estimator: H_0 from cosmic chronometers

If a parameter enters the model *linearly*, differentiating Eq. (7.7.3) gives a closed-form estimator directly. Holding Ω_{m0} (and hence $E_i \equiv E(z_i; \Omega_{m0})$) fixed, $\chi^2(H_0) = \sum_i (H_i - H_0 E_i)^2/\sigma_i^2$, so that

$$\frac{{\rm d}\chi^2}{{\rm d}H_0} = -2\sum_{i=1}^{N}\frac{(H_i - H_0 E_i)E_i}{\sigma_i^2} = 0 \implies \sum_{i=1}^{N}\frac{H_i E_i}{\sigma_i^2} = H_0\sum_{i=1}^{N}\frac{E_i^2}{\sigma_i^2}, \tag{7.7.4}$$

giving the maximum-likelihood estimator

$$\boxed{\hat{H}_0 = \frac{\sum_{i=1}^{N} H_i E_i/\sigma_i^2}{\sum_{i=1}^{N} E_i^2/\sigma_i^2}.} \tag{7.7.5}$$

This is an explicit recipe: evaluate $E(z_i;\Omega_{m0})$ at the measured redshifts, form the two weighted sums, and their ratio is $\hat{H}_0$ — no numerical minimizer is needed.

Profiling out a nuisance parameter: supernova absolute magnitude

In supernova cosmology, H_0 and the absolute magnitude M never appear separately: they enter $\mu_{\rm th}$ only through the combination $\mathcal{M} \equiv M - 5\log_{10} H_0 + 25$, since $\mu_{\rm th}(z_i) = 5\log_{10} D_L(z_i;\boldsymbol{\Omega}) + \mathcal{M}$ with $D_L(z) \equiv (H_0/c)\, d_L(z)$ independent of H_0. Writing the residual $\Delta_i \equiv \mu_i - 5\log_{10} D_L(z_i) - \mathcal{M}$, the χ^2 is quadratic in $\mathcal{M}$,

$$\chi^2(\mathcal{M},\boldsymbol{\Omega}) = \sum_{i=1}^{N} \frac{\Delta_i^2}{\sigma_i^2} = A - 2B\mathcal{M} + C\mathcal{M}^2, \tag{7.7.6}$$

with $A = \sum_i (\mu_i - 5\log_{10} D_L(z_i))^2/\sigma_i^2$, $B = \sum_i (\mu_i - 5\log_{10} D_L(z_i))/\sigma_i^2$, $C = \sum_i 1/\sigma_i^2$. Setting $\partial\chi^2/\partial\mathcal{M} = -2B + 2C\mathcal{M} = 0$ gives $\hat{\mathcal{M}} = B/C$, and substituting back yields the **profile likelihood**

$$\chi^2_{\rm profile}(\boldsymbol{\Omega}) = A - \frac{B^2}{C}, \tag{7.7.7}$$

independent of $\mathcal{M}$ (and hence of H_0 and M separately). This is exactly how supernova data constrain Ω_{m0} and $\Omega_{\Lambda 0}$ without ever needing an independent prior on H_0.

Variance of the estimator, and when to go numerical

The uncertainty on $\hat{H}_0$ follows from the same curvature argument as the Cramér–Rao bound of §7.5.2: for a model linear in the mean at fixed noise, $-\partial^2 \ln L/\partial H_0^2 = \sum_i E_i^2/\sigma_i^2$ plays the role of the Fisher information F (the mean-shift counterpart of Eq. (7.5.16), which was instead built for a $\boldsymbol{\lambda}$-dependent *covariance*), so that

$$\sigma_{\hat{H}_0} = \left(\sum_{i=1}^{N} \frac{E_i^2}{\sigma_i^2} \right)^{-1/2}, \tag{7.7.8}$$

saturating the Cramér–Rao bound exactly, as it must for a linear Gaussian problem. When the parameters enter *nonlinearly* — e.g. Ω_{m0} varied simultaneously with H_0 inside the integral defining $E(z)$ — $\nabla\chi^2 = 0$ no longer has a closed-form solution, and one falls back on the Newton–Raphson iteration of Eq. (7.5.10) or on full MCMC sampling (§7.5.4) of the kind used throughout the rest of this chapter.

7.7.3 Big Bang Nucleosynthesis (BBN)

During primordial nucleosynthesis ($T \sim 0.1\,\text{MeV}$, $t \sim 3\,\text{min}$, $z \sim 2.3 \times 10^9$), nuclear reactions fuse protons and neutrons into light elements: ^{4}He, D, ^{3}He, and ^{7}Li.

The primordial Deuterium abundance (D/H) is sensitive to the physical baryon density $\Omega_{\rm b}h^2$: because Deuterium burning into ^{4}He proceeds via two-body nuclear reactions, a higher baryon density accelerates Deuterium destruction, resulting in a lower relic Deuterium fraction. High-redshift quasar absorption systems measure:

$$0.017 \leq \Omega_{\rm b}h^2 \leq 0.024 \qquad (95\%\ \text{confidence level}). \tag{7.7.9}$$

The remarkable concordance between this BBN constraint (set when the Universe was 3 minutes old) and the CMB acoustic peak constraint (set at 380,000 years) provides a foundational triumph of modern cosmology.

Furthermore, the primordial ^{4}He mass fraction $Y_{\rm p} \approx 0.25$ depends on the neutron-to-proton ratio at freeze-out ($T_{\rm fo} \sim 0.7\,\mathrm{MeV}$), which is governed by the competition between weak interaction rates and the Hubble expansion rate $H \propto \sqrt{g_* \rho_{\rm rel}}$. This constrains the effective number of light relativistic neutrino species during BBN to:

$$N_\nu = 3.24 \pm 1.2 \qquad (95\%\ \text{confidence level}), \tag{7.7.10}$$

in full agreement with the precision electroweak measurement from the invisible Z-boson decay width at CERN/LEP:

$$N_\nu = 2.99 \pm 0.016. \tag{7.7.11}$$

7.7.4 The Galaxy Distribution and Large-Scale Structure

Observations of the spatial distribution of galaxies provide a direct probe of late-time ($z \lesssim 1$) large-scale structure (LSS), acting as an essential counterpart to the early-universe ($z \approx 1100$) snapshot delivered by the CMB. However, inferring the fundamental matter power spectrum $P_{\rm m}(k)$ from galaxy surveys introduces two key physical challenges:

1. **Non-Linear Gravitational Clustering:** On small spatial scales corresponding to comoving wavenumbers $k \gtrsim 0.2\,h\,\mathrm{Mpc}^{-1}$, the matter density contrast reaches order unity ($\delta \sim 1$) and higher. In this regime, linear perturbation theory breaks down completely. While linear theory achieves sub-percent theoretical precision ($< 1\%$), non-linear perturbation theory and numerical N-body simulations are hard-pressed to reach an accuracy better than 10%. Consequently, precision cosmological parameter estimation from galaxy surveys is strictly restricted to linear and quasi-linear scales ($k \lesssim 0.15\,h\,\mathrm{Mpc}^{-1}$).

2. **Galaxy Bias b:** Galaxies are discrete, luminous tracers residing within dense dark matter halos; they do not trace the continuous matter distribution in an unbiased fashion. In the linear regime on large scales, this mismatch is parameterized by the **galaxy bias factor** b:

$$\delta_{\rm g}(\boldsymbol{x}) = b\,\delta_{\rm m}(\boldsymbol{x}) \implies P_{\rm g}(k) = b^2 P_{\rm m}(k). \tag{7.7.12}$$

   Although b is approximately constant on very large scales, it generically exhibits scale and redshift dependence on intermediate and small scales.

The Shape Parameter Γ

Despite the uncertainty in galaxy bias, galaxy redshift surveys provide a robust determination of the **shape parameter** Γ, which sets the turnover scale $k_{\rm eq}$ where the linear matter power spectrum transitions from its super-horizon primordial growth ($P_{\rm m} \propto k^{n_{\rm s}}$ for $k \ll k_{\rm eq}$) to its radiation-era sub-horizon stagnation ($P_{\rm m} \propto k^{n_{\rm s}-4}$ for $k \gg k_{\rm eq}$).

The turnover wavenumber corresponds to the comoving Hubble horizon at matter-radiation equality $\eta_{\rm eq}$. The shape parameter Γ is defined relative to the turnover scale $k_{\rm m1}$ in a fiducial Einstein–de Sitter universe with $\Omega_{\rm m} = 1$ and $h = 1$:

$$k_{\rm eq} = \frac{k_{\rm m1}}{\Gamma}, \qquad \Gamma \simeq \Omega_{\rm m} h. \tag{7.7.13}$$

Observational determinations of the kink position yield $\Gamma \simeq 0.2$, confirming that $\Omega_{\rm m} h \approx 0.2$ and decisively ruling out a matter-dominated flat universe ($\Omega_{\rm m} = 1, h \approx 0.7 \implies \Gamma \approx 0.7$).

Neither asymptotic regime ($k^{n_{\rm s}}$ or $k^{n_{\rm s}-4}$) is directly observed in pristine form in galaxy power spectra:

- On small scales ($k \gg k_{\rm eq}$), non-linear gravitational collapse injects additional power, causing the spectrum to flatten significantly before reaching the asymptotic $k^{n_{\rm s}-4}$ slope.
- On large scales ($k \ll k_{\rm eq}$), surveys are limited by finite cosmic volume and discrete Poisson sampling (**shot noise** $P_{\rm shot} = 1/\bar{n}_{\rm g}$). Because shot noise is scale-independent (white noise), it inevitably overwhelms the scale-invariant $\propto k^{n_{\rm s}}$ power spectrum on sufficiently large scales.

Power Spectrum Normalization and σ_8

The amplitude of matter fluctuations on a characteristic scale R is quantified by the root-mean-square density fluctuation inside a spherical volume of radius R:

$$\left\langle \left(\frac{\Delta M}{M}\right)^2 \right\rangle_R = \left\langle \left(\int W_R(\boldsymbol{x}-\boldsymbol{y})\, \delta_{\rm m}(\boldsymbol{y})\, \mathrm{d}^3\boldsymbol{y} \right)^2 \right\rangle, \tag{7.7.14}$$

where $W_R(\boldsymbol{x})$ is a spatial window function. Applying the convolution theorem in Fourier space:

$$\int W_R(\boldsymbol{x}-\boldsymbol{y})\, \delta_{\rm m}(\boldsymbol{y})\, \mathrm{d}^3\boldsymbol{y} = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3}\, \mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{x}}\, W_R(k)\, \delta_{\rm m}(\boldsymbol{k}), \tag{7.7.15}$$

the variance reduces to a 1D integral over the matter power spectrum:

$$\left\langle \left(\frac{\Delta M}{M}\right)^2 \right\rangle_R = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3}\, |W_R(k)|^2\, P_{\rm m}(k) = \frac{1}{2\pi^2}\int_0^\infty k^2\, \mathrm{d}k\, |W_R(k)|^2\, P_{\rm m}(k). \tag{7.7.16}$$

The standard cosmological amplitude parameter is σ_8, defined as the rms fluctuation in a top-hat sphere of radius $R = 8\,h^{-1}\,\mathrm{Mpc}$:

$$\sigma_8^2 \equiv \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3}\, |W_{8\,h^{-1}\mathrm{Mpc}}(k)|^2\, P_{\rm m}(k). \tag{7.7.17}$$

While determining σ_8 directly from galaxy clustering requires breaking the degeneracy with the galaxy bias b ($\sigma_{8,\rm g} = b\,\sigma_8$), linear perturbation theory accurately predicts σ_8 from the CMB primordial amplitude $A_{\rm s}$ and spectral index $n_{\rm s}$, yielding $\sigma_8 = 0.74 \pm 0.05$ (WMAP 3-year).

Derivation: Spherical Top-Hat Window Function and σ_8 Normalization
Consider a spherical top-hat window function of radius R in real space:

$$W_R(\boldsymbol{x}) = \frac{1}{V_R}\Theta(R-|\boldsymbol{x}|) = \frac{3}{4\pi R^3}\Theta(R-|\boldsymbol{x}|). \tag{1}$$

Its Fourier transform is obtained by integrating in spherical polar coordinates (r,θ,ϕ) with $\boldsymbol{k}\cdot\boldsymbol{x} = kr\cos\theta \equiv kr\mu$:

$$\begin{aligned} W_R(k) &= \int \mathrm{d}^3\boldsymbol{x}\, W_R(\boldsymbol{x})\, \mathrm{e}^{-i\boldsymbol{k}\cdot\boldsymbol{x}} \\ &= \frac{3}{4\pi R^3}\int_0^R r^2\,\mathrm{d}r \int_{-1}^1 \mathrm{d}\mu\, \mathrm{e}^{-ikr\mu}\int_0^{2\pi}\mathrm{d}\phi \\ &= \frac{3}{2R^3}\int_0^R r^2\,\mathrm{d}r \left[\frac{\mathrm{e}^{-ikr\mu}}{-ikr}\right]_{-1}^{1} \\ &= \frac{3}{kR^3}\int_0^R r\sin(kr)\,\mathrm{d}r. \end{aligned} \tag{2}$$

Integrating by parts with $u = r$ and $\mathrm{d}v = \sin(kr)\,\mathrm{d}r$:

$$\int_0^R r\sin(kr)\,\mathrm{d}r = \left[-\frac{r\cos(kr)}{k}\right]_0^R + \frac{1}{k}\int_0^R \cos(kr)\,\mathrm{d}r$$
$$= -\frac{R\cos(kR)}{k} + \frac{\sin(kR)}{k^2}. \tag{3}$$

Substituting this back into the Fourier transform yields:

$$W_R(k) = \frac{3\big(\sin(kR) - kR\cos(kR)\big)}{(kR)^3} = \frac{3j_1(kR)}{kR}, \tag{4}$$

where $j_1(x) = (\sin x - x\cos x)/x^2$ is the spherical Bessel function of order 1.
In the large-scale limit ($kR \to 0$):

$$W_R(k) = \frac{3}{(kR)^3}\left[\left(kR - \frac{(kR)^3}{6} + \ldots\right) - kR\left(1 - \frac{(kR)^2}{2} + \ldots\right)\right]$$
$$= \frac{3}{(kR)^3}\left[\frac{(kR)^3}{3} + \mathcal{O}\big((kR)^5\big)\right] \to 1. \tag{5}$$

In the small-scale limit ($kR \gg 1$), $W_R(k) \propto (kR)^{-2}\cos(kR)$, ensuring rapid convergence of the σ_8^2 integral:

$$\sigma_8^2 = \frac{1}{2\pi^2}\int_0^\infty k^2\,\mathrm{d}k\left(\frac{3j_1(8\,h^{-1}\mathrm{Mpc}\cdot k)}{8\,h^{-1}\mathrm{Mpc}\cdot k}\right)^2 P_\mathrm{m}(k). \tag{6}$$

Neutrino Mass Bounds and Baryon Acoustic Oscillations

The matter power spectrum is also exquisitely sensitive to light massive neutrinos. Relativistic neutrinos free stream out of gravitational wells below their free-streaming scale k_fs, suppressing sub-horizon matter growth by $\Delta P_\mathrm{m}/P_\mathrm{m} \approx -8\Omega_\nu/\Omega_\mathrm{m}$. Galaxy redshift surveys (such as the 2dFGRS and the Sloan Digital Sky Survey, SDSS; Fig. 7.16) establish rigorous upper bounds $\sum m_\nu < 1\,\mathrm{eV}$.

Finally, sound waves propagating through the coupled photon-baryon plasma prior to decoupling freeze out at the baryon drag epoch ($z_\mathrm{drag} \approx 1060$), imprinting a characteristic spherical acoustic shell of comoving radius $r_\mathrm{s}(z_\mathrm{drag}) \approx 150\,\mathrm{Mpc}$ around overdensities. These **Baryon Acoustic Oscillations (BAO)** are observed in large galaxy surveys (Eisenstein et al., 2005; Cole et al., 2005). Because the sound horizon r_s is calibrated to sub-percent precision by the CMB, BAO serves as a cosmological **standard ruler**, enabling direct geometric measurements of the angular diameter distance $D_\mathrm{A}(z)$ and Hubble parameter $H(z)$ across cosmological epochs.

7.7.5 Measuring the Late ISW Effect via CMB–LSS Cross-Correlations

The integrated Sachs–Wolfe (ISW) effect describes the net energy shift experienced by CMB photons traversing time-dependent gravitational potentials:

$$\left(\frac{\Delta T}{T}\right)_\mathrm{ISW}(\boldsymbol{x}_0, \hat{\boldsymbol{n}}) = 2\int_{\eta_\mathrm{dec}}^{\eta_0} \Phi'(\boldsymbol{x}_0 - \hat{\boldsymbol{n}}(\eta_0 - \eta), \eta)\,\mathrm{d}\eta, \tag{7.7.18}$$

where η is conformal time, $\Phi' \equiv \partial\Phi/\partial\eta$, and we have set $\Phi = \Psi$ (vanishing anisotropic stress).

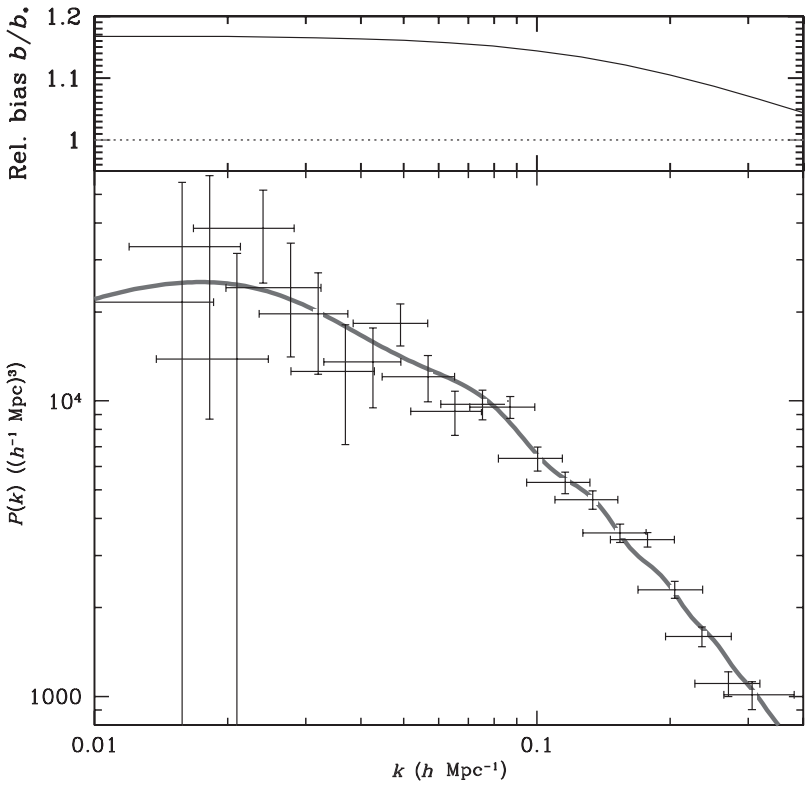


Figure 7.16: The galaxy power spectrum from the Sloan Digital Sky Survey (SDSS) corrected by the modeled galaxy bias (top). The solid curve represents the best-fitting linear ΛCDM cosmological model (adapted from Tegmark et al., 2004).

In a flat universe, the Bardeen gravitational potential is related to the comoving matter overdensity $D(\boldsymbol{k}, t)$ via the relativistic Poisson equation:

$$\Phi(\boldsymbol{k}, \eta) = -\frac{4\pi G_{\mathrm{N}} a^2 \bar{\rho}_{\mathrm{m}}}{k^2} \delta_{\mathrm{m}}(\boldsymbol{k}, \eta) = -\frac{3H_0^2 \Omega_{\mathrm{m}}(1+z)}{2k^2} \delta_{\mathrm{m}}(\boldsymbol{k}, \eta). \tag{7.7.19}$$

In a purely matter-dominated Einstein–de Sitter universe ($\Omega_{\mathrm{m}} = 1$), linear perturbations grow in direct proportion to the scale factor ($\delta_{\mathrm{m}} \propto a \propto (1+z)^{-1}$). Hence, the product $a^2 \bar{\rho}_{\mathrm{m}} \delta_{\mathrm{m}} \propto a^{-1} \delta_{\mathrm{m}}$ is strictly constant in time, meaning $\Phi' = 0$: *there is no late ISW effect in a matter-dominated universe.*

However, at late times ($z \lesssim 1$), dark energy ($\Omega_\Lambda \approx 0.7$) accelerates cosmic expansion and suppresses the growth of linear matter perturbations ($\delta_{\mathrm{m}}(a) < a$). Consequently, gravitational potential wells decay ($\Phi' > 0$), imparting a net blue-shift to CMB photons traversing deep potential wells.

Because the decaying potentials correspond precisely to the evolving large-scale matter distribution at $z \lesssim 1$, the late ISW effect induces a non-zero cross-correlation between CMB temperature anisotropies and foreground galaxy overdensities.

Derivation of the CMB–LSS Cross-Power Spectrum

The density fluctuation at redshift z in direction $\hat{\boldsymbol{n}}'$ is:

$$\begin{aligned}\delta_{\mathrm{m}}(\boldsymbol{x}_0, \hat{\boldsymbol{n}}', z) &= \delta_{\mathrm{m}}\big(\boldsymbol{x}_0 - \hat{\boldsymbol{n}}'(\eta(z) - \eta_0), \eta(z)\big) \\ &= \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} \, \mathrm{e}^{-i\boldsymbol{k}\cdot(\boldsymbol{x}_0 - \hat{\boldsymbol{n}}'(\eta(z)-\eta_0))} \, \delta_{\mathrm{m}}(\boldsymbol{k}, \eta(z))\end{aligned}$$

$$= -\frac{2}{3H_0^2\Omega_{\rm m}(1+z)} \int \frac{{\rm d}^3\boldsymbol{k}}{(2\pi)^3}\, k^2\, {\rm e}^{-i\boldsymbol{k}\cdot(\boldsymbol{x}_0 - \hat{\boldsymbol{n}}'(\eta(z)-\eta_0))}\, \Phi(\boldsymbol{k}, \eta(z)). \tag{7.7.20}$$

Expanding the cross-correlation between the matter overdensity and the ISW temperature fluctuation in Legendre polynomials:

$$\left\langle \delta_{\rm m}(\boldsymbol{x}_0, \hat{\boldsymbol{n}}', z) \left(\frac{\Delta T}{T}\right)_{\rm ISW}(\boldsymbol{x}_0, \hat{\boldsymbol{n}}) \right\rangle = \frac{1}{4\pi} \sum_\ell (2\ell+1)\, C_\ell^{Tg}(z)\, P_\ell(\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{n}}'). \tag{7.7.21}$$

We parameterize the temporal evolution of the potential by:

$$\Phi(\boldsymbol{k}, \eta) = g(\eta, k)\, \Phi_{\rm in}(\boldsymbol{k}), \tag{7.7.22}$$

where $g(\eta) = D_1(\eta)/a(\eta)$ is the potential growth suppression factor (Fig. 7.17).

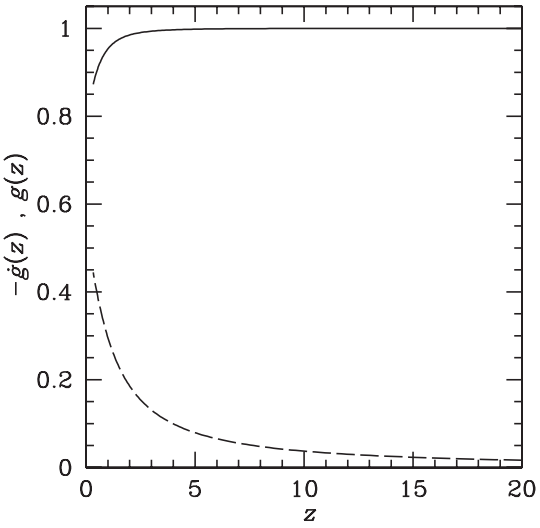


Figure 7.17: The gravitational potential growth function $g(z)$ (solid line) and its time derivative $\dot{g}(z)$ (dashed line) as functions of redshift in a flat ΛCDM universe with $\Omega_\Lambda = 0.7$ and $\Omega_{\rm m} = 0.3$. Notice that $g(z)$ deviates significantly from unity only at late times ($z \lesssim 2$) when dark energy dominates.

Using the Rayleigh plane-wave expansion into spherical Bessel functions j_ℓ and spherical harmonics, the exact angular cross-power spectrum is:

$$C_\ell^{Tg}(z) = -\frac{8g(\eta(z))}{3\pi\Omega_{\rm m}H_0^2(1+z)} \int_0^\infty {\rm d}k\, k^4\, P_\Phi(k)\, j_\ell(k(\eta_0 - \eta(z))) \int_{\eta_{\rm dec}}^{\eta_0} {\rm d}\eta\, g'(\eta)\, j_\ell(k(\eta_0 - \eta)). \tag{7.7.23}$$

Here $P_\Phi(k) = 2\pi^2 A_{\rm s}^2 (k/H_0)^{n_{\rm s}-1} k^{-3}$ is the primordial potential power spectrum.

Derivation: CMB–LSS Angular Cross-Correlation Power Spectrum $C_\ell^{Tg}(z)$
Substituting the Fourier representations of $D(\boldsymbol{x}_0, \hat{\boldsymbol{n}}', z)$ and $(\Delta T/T)_{\rm ISW}(\hat{\boldsymbol{n}})$ into the two-point expectation value:

$$\left\langle \delta_{\rm m}(\hat{\boldsymbol{n}}', z) \left(\frac{\Delta T}{T}\right)_{\rm ISW}(\hat{\boldsymbol{n}}) \right\rangle = -\frac{4g(\eta(z))}{3H_0^2\Omega_{\rm m}(1+z)} \int_{\eta_{\rm dec}}^{\eta_0} {\rm d}\eta\, g'(\eta) \int \frac{{\rm d}^3\boldsymbol{k}}{(2\pi)^3} \int \frac{{\rm d}^3\boldsymbol{k}'}{(2\pi)^3}\, k'^2$$
$$\times\, {\rm e}^{-i\boldsymbol{k}'\cdot(\boldsymbol{x}_0 - \hat{\boldsymbol{n}}'(\eta(z)-\eta_0))}\, {\rm e}^{-i\boldsymbol{k}\cdot(\boldsymbol{x}_0 - \hat{\boldsymbol{n}}(\eta_0-\eta))}\, \langle \Phi_{\rm in}(\boldsymbol{k}')\Phi_{\rm in}(\boldsymbol{k})\rangle. \tag{1}$$

Using statistical homogeneity $\langle \Phi_{\rm in}(\boldsymbol{k}')\Phi_{\rm in}(\boldsymbol{k})\rangle = (2\pi)^3\delta(\boldsymbol{k}+\boldsymbol{k}')P_\Phi(k)$, setting $\boldsymbol{k}' = -\boldsymbol{k}$:

$$\left\langle \delta_{\rm m}(\hat{\boldsymbol{n}}', z)\left(\frac{\Delta T}{T}\right)_{\rm ISW}(\hat{\boldsymbol{n}})\right\rangle = -\frac{4g(\eta(z))}{3H_0^2\Omega_{\rm m}(1+z)}\int_{\eta_{\rm dec}}^{\eta_0}\mathrm{d}\eta\, g'(\eta)\int\frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3}\,k^2\,P_\Phi(k) \times \mathrm{e}^{i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}'(\eta_0-\eta(z))}\,\mathrm{e}^{-i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}(\eta_0-\eta)}. \tag{2}$$

Expanding each plane wave into spherical harmonics and spherical Bessel functions:

$$\mathrm{e}^{i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}'(\eta_0-\eta(z))} = 4\pi\sum_{\ell',m'} i^{\ell'}\, j_{\ell'}(k(\eta_0-\eta(z)))\, Y^*_{\ell'm'}(\hat{\boldsymbol{k}})\, Y_{\ell'm'}(\hat{\boldsymbol{n}}'), \tag{3}$$

$$\mathrm{e}^{-i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}(\eta_0-\eta)} = 4\pi\sum_{\ell,m}(-i)^{\ell}\, j_{\ell}(k(\eta_0-\eta))\, Y_{\ell m}(\hat{\boldsymbol{k}})\, Y^*_{\ell m}(\hat{\boldsymbol{n}}). \tag{4}$$

Integrating over directions in $\boldsymbol{k}$-space using the orthogonality of spherical harmonics $\int \mathrm{d}\Omega_{\hat{\boldsymbol{k}}}\, Y^*_{\ell'm'}(\hat{\boldsymbol{k}})Y_{\ell m}(\hat{\boldsymbol{k}}) = \delta_{\ell\ell'}\delta_{mm'}$:

$$\begin{aligned}\int \mathrm{d}\Omega_{\hat{\boldsymbol{k}}}\,\mathrm{e}^{i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}'(\eta_0-\eta(z))}\,\mathrm{e}^{-i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}(\eta_0-\eta)} &= (4\pi)^2\sum_{\ell,m} j_\ell(k(\eta_0-\eta(z)))\, j_\ell(k(\eta_0-\eta))\, Y^*_{\ell m}(\hat{\boldsymbol{n}})Y_{\ell m}(\hat{\boldsymbol{n}}') \\ &= 4\pi\sum_{\ell}(2\ell+1)\, j_\ell(k(\eta_0-\eta(z)))\, j_\ell(k(\eta_0-\eta))\, P_\ell(\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{n}}'),\end{aligned} \tag{5}$$

where we applied the spherical harmonic addition theorem $\sum_m Y^*_{\ell m}(\hat{\boldsymbol{n}})Y_{\ell m}(\hat{\boldsymbol{n}}') = \frac{2\ell+1}{4\pi}P_\ell(\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{n}}')$.
Substituting this back and equating term-by-term with Eq. (7.7.21) immediately yields:

$$C_\ell^{Tg}(z) = -\frac{8g(\eta(z))}{3\pi\Omega_{\rm m}H_0^2(1+z)}\int_0^\infty \mathrm{d}k\, k^4\, P_\Phi(k)\, j_\ell(k(\eta_0-\eta(z)))\int_{\eta_{\rm dec}}^{\eta_0}\mathrm{d}\eta\, g'(\eta)\, j_\ell(k(\eta_0-\eta)). \tag{6}$$

Limiting Behavior and Observational Strategy

The integration kernel $K(k,\ell,z) \equiv j_\ell(k(\eta_0-\eta(z)))\int_{\eta_{\rm dec}}^{\eta_0}\mathrm{d}\eta\, g'(\eta)j_\ell(k(\eta_0-\eta))$ peaks strongly when the arguments of the spherical Bessel functions match:

$$k_{\rm max} \sim \frac{\ell}{\eta_0-\eta(z)}, \qquad \Delta k_{\rm max} \sim \frac{\pi}{\eta_0-\eta(z)}. \tag{7.7.24}$$

Approximating $j_\ell(x_{\rm max}) \sim 1/\ell$, the cross-spectrum scales as:

$$C_\ell^{Tg}(z) \simeq -\frac{8(\eta_0-\eta(z))^{-2}A_{\rm s}^2}{3\Omega_{\rm m}H_0^2(1+z)\ell^2}\left(\frac{\ell}{H_0(\eta_0-\eta(z))}\right)^{n_{\rm s}-1}. \tag{7.7.25}$$

In the low-redshift limit ($z \ll 1$) where $H_0(\eta_0-\eta(z)) \sim z$:

$$\ell^2 C_\ell^{Tg}(z) \simeq -\frac{8A_{\rm s}^2(\ell/z)^{n_{\rm s}-1}}{3\Omega_{\rm m}(1+z)z^2}. \tag{7.7.26}$$

To remain strictly within the linear regime ($k_{\rm max} \lesssim 0.1\,h\,\mathrm{Mpc}^{-1}$), the minimum redshift at a given multipole must satisfy $z_{\rm min} \gtrsim \ell/300$. An optimal observational compromise occurs around $z \sim 0.2$, allowing measurement of the late ISW cross-correlation up to $\ell_{\rm max} \sim 60$. Multiple detections at the 2–3σ level have been established by cross-correlating WMAP/Planck CMB maps with SDSS, NVSS, and 2MASS galaxy catalogues (Padmanabhan et al., 2005; Pietrobon et al., 2006), providing direct geometric confirmation of dark energy.

7.7.6 The Lyman-α Forest

As light from high-redshift quasars ($z \sim 2$–4) travels toward Earth, it passes through intergalactic clouds of neutral hydrogen (HI). These cool, moderately overdense gas structures absorb photons at the Lyman-α rest-frame wavelength $\lambda_\alpha = 1215.67\,\text{Å}$, producing a dense sequence of absorption lines in the observed quasar spectrum at $\lambda_{\rm obs} = (1+z)\lambda_\alpha$ known as the **Lyman-α forest** (Fig. 7.18).

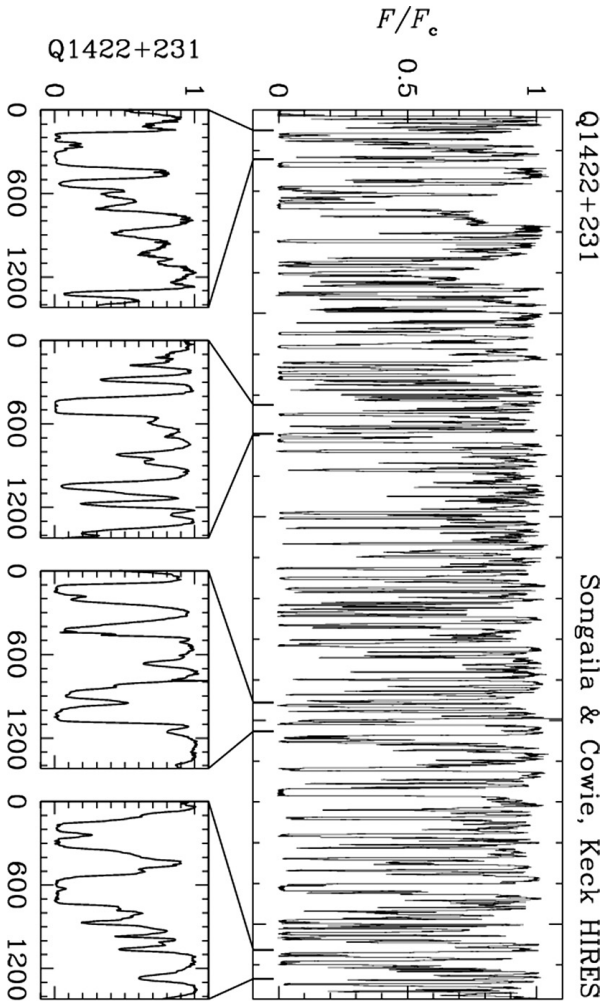


Figure 7.18: The Lyman-α forest in the high-resolution Keck HIRES spectrum of the distant quasar QSO 1422+231 at emission redshift $z = 3.61$ (from Songaila & Cowie, 1996).

The observational quantity is the transmitted flux fraction $F(\lambda) = \mathrm{e}^{-\tau(\lambda)}$, where $\tau(\lambda)$ is the Ly-α optical depth. In photoionization equilibrium with the ionizing metagalactic UV background ($\Gamma_{\rm UV}$), the neutral hydrogen density scales as:

$$n_{\rm HI} \propto \frac{\rho_{\rm b}^2\, T^{-0.7}}{\Gamma_{\rm UV}} \propto \rho_{\rm m}^{2-0.7(\gamma-1)}, \tag{7.7.27}$$

where $T \propto \rho_{\rm b}^{\gamma-1}$ is the gas temperature-density relation ($\gamma \approx 1.6$).

If density perturbations along the line of sight were purely linear, the two-point correlation of overdensities at redshifts z_1 and z_2 would factorize cleanly:

$$\begin{aligned}\langle \delta_{\rm m}(z_1)\delta_{\rm m}(z_2)\rangle &= \langle \delta_{\rm m}(\boldsymbol{x}_0 - \hat{\boldsymbol{n}}(\eta_0-\eta(z_1)),\eta(z_1))\,\delta_{\rm m}(\boldsymbol{x}_0-\hat{\boldsymbol{n}}(\eta_0-\eta(z_2)),\eta(z_2))\rangle \\ &= D_1(z_1)\,D_1(z_2)\,C_{\rm in}(|\eta(z_1)-\eta(z_2)|),\end{aligned} \tag{7.7.28}$$

where $T(z)$ is the linear growth factor and $C_{\rm in}$ is the primordial correlation function. For typical separations $\Delta z = |z_1 - z_2| \approx 0.01$ at mean redshift $z = 2.5$, the comoving spatial separation is $|\eta(z_1) - \eta(z_2)| \approx 1.5\,h^{-1}\,\mathrm{Mpc}$.

Cosmological Significance: The High-k Lever Arm

Because the Lyman-α forest probes comoving scales $k \sim 1$–$10\,h\,\mathrm{Mpc}^{-1}$ at high redshifts $z \sim 2$–4 (where perturbations are still relatively close to the linear regime), it extends the observational baseline of the matter power spectrum by nearly three decades beyond CMB acoustic peaks. This extended "lever arm" provides decisive constraining power for:

- The scalar spectral index $n_{\rm s}$ and its scale dependence (**running**) $\alpha \equiv \mathrm{d}n_{\rm s}/\mathrm{d}\ln k$.
- The overall fluctuation amplitude σ_8.
- Relic neutrino masses: Hydrodynamical simulations matched to SDSS Ly-α forest flux power spectra constrain $\sum m_\nu < 0.17\,\mathrm{eV}$ and find no evidence for running spectral tilt (McDonald et al., 2005; Seljak et al., 2006).

7.7.7 Weak Gravitational Lensing

Gravitational lensing refers to the deflection of light from background sources by foreground mass concentrations. When the gravitational deflection is mild, background galaxies are not multiplied into separate images but are coherently distorted, transforming circular sources into slightly tangential ellipses (**cosmic shear**).

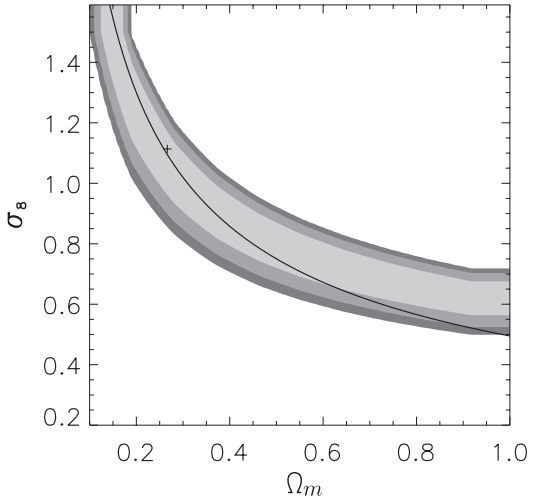


Figure 7.19: Cosmological constraints in the $(\Omega_{\rm m}, \sigma_8)$ plane from cosmic shear measurements (Maoli et al., 2001). The light, medium, and dark shaded regions enclose the 1σ, 2σ, and 3σ confidence contours. The cross marks the best-fit model, and the solid curve displays the constraint inferred from local galaxy cluster abundances.

The intrinsic shape dispersion of galaxies ($\sigma_\epsilon \sim 0.3$) far exceeds the lensing-induced shear ($\gamma \sim 0.01$). Nevertheless, because intrinsic galaxy orientations are uncorrelated across large physical separations, cosmic shear is measured with high statistical significance by cross-correlating the ellipticities of millions of background galaxies.

Cosmic Shear and the $\Omega_{\rm m}$–σ_8 Degeneracy

The fundamental advantage of weak lensing over galaxy clustering is that *gravitational lensing responds directly to the total matter distribution (dark matter + baryons)*, making it completely independent of galaxy bias b.

Lensing shear is determined by the transverse tidal tensor of the projected gravitational potential $\Phi \propto \Omega_{\rm m}\delta_{\rm m} \propto \Omega_{\rm m}\sigma_8$. Integrating Φ along the line of sight reduces the parameter dependence to approximately $\sigma_8\Omega_{\rm m}^{0.5}$. Cosmic shear surveys therefore produce characteristic banana-shaped degeneracy contours in the $(\Omega_{\rm m}, \sigma_8)$ plane (Fig. 7.19):

$$\sigma_8\,\Omega_{\rm m}^{0.47} \simeq 0.59 \pm 0.03. \tag{7.7.29}$$

Derivation: Weak Lensing Convergence Power Spectrum via Limber's Approximation

In the weak lensing regime, the deflection of a light ray from a source at comoving distance $\chi_{\rm s}$ is governed by the projected lensing potential $\psi(\boldsymbol{\theta})$:

$$\psi(\boldsymbol{\theta}) = \frac{2}{c^2}\int_0^{\chi_{\rm s}} {\rm d}\chi\, \frac{\chi_{\rm s}-\chi}{\chi\,\chi_{\rm s}}\,\Phi(\chi\boldsymbol{\theta},\chi). \tag{1}$$

The **convergence field** $\kappa(\boldsymbol{\theta}) \equiv \frac{1}{2}\nabla^2_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})$ measures the local isotropic magnification. Using the Poisson equation $\nabla^2\Phi = \frac{3}{2}H_0^2\Omega_{\rm m}a^{-1}\delta_{\rm m}$:

$$\kappa(\boldsymbol{\theta}) = \frac{3H_0^2\Omega_{\rm m}}{2c^2}\int_0^{\chi_{\rm s}} {\rm d}\chi\, \frac{\chi(\chi_{\rm s}-\chi)}{\chi_{\rm s}\,a(\chi)}\,\delta_{\rm m}(\chi\boldsymbol{\theta},\chi) \equiv \int_0^{\chi_{\rm s}} {\rm d}\chi\, W_\kappa(\chi)\,\delta_{\rm m}(\chi\boldsymbol{\theta},\chi), \tag{2}$$

where $W_\kappa(\chi) = \frac{3H_0^2\Omega_{\rm m}}{2c^2}\frac{\chi(\chi_{\rm s}-\chi)}{\chi_{\rm s}a(\chi)}$ is the lensing efficiency kernel.

Under Limber's approximation (valid for $\ell \gg 1$ where line-of-sight fluctuations with $k_\parallel \neq 0$ cancel by destructive interference), the 2D angular power spectrum $P_\kappa(\ell)$ of the convergence field is the line-of-sight projection of the 3D matter power spectrum $P_{\rm m}(k,\chi)$:

$$P_\kappa(\ell) = \int_0^{\chi_{\rm s}} \frac{{\rm d}\chi}{\chi^2}\, W_\kappa^2(\chi)\, P_{\rm m}\left(k=\frac{\ell}{\chi},\chi\right). \tag{3}$$

Since $W_\kappa \propto \Omega_{\rm m}$ and $P_{\rm m} \propto \sigma_8^2$, the convergence power spectrum scales strictly as $P_\kappa(\ell) \propto \Omega_{\rm m}^2\sigma_8^2$, explaining why weak lensing provides an exceptionally clean, unbiased measurement of the matter fluctuation amplitude.

7.7.8 Galaxy Clusters

Rich galaxy clusters ($M \sim 10^{14}$–$10^{15}\,h^{-1}\,M_\odot$) represent the largest virialized, gravitationally bound structures in the Universe. Because clusters form at the extreme exponential tail of the cosmic density field, their spatial number density as a function of mass and redshift provides an extraordinarily sensitive test of cosmological parameters.

For a Gaussian matter density field, the probability density of observing an overdensity $\delta_{\rm m}$ smoothed on mass scale $M = \frac{4\pi}{3}\bar{\rho}_{\rm m}\lambda^3$ at redshift z is:

$$P(\delta_{\rm m}, z) = \frac{1}{\sqrt{2\pi}\sigma(\lambda,z)}\exp\left(-\frac{\delta_{\rm m}^2}{2\sigma^2(\lambda,z)}\right), \tag{7.7.30}$$

where $\sigma^2(\lambda,z) = D_1^2(z)\sigma^2(\lambda,0)$ is the linear variance on scale λ.

According to the **Press–Schechter formalism** (Press & Schechter, 1974), a region collapses to form a bound cluster when its linearly extrapolated overdensity exceeds the critical collapse

threshold $\delta_{\rm c} \simeq 1.686$ (derived from the spherical top-hat collapse model). The probability that an overdensity has collapsed is:

$$P(\delta_{\rm m} > \delta_{\rm c} \mid \lambda, z) = \frac{1}{\sqrt{2\pi}\sigma(\lambda, z)} \int_{\delta_{\rm c}}^{\infty} {\rm e}^{-\delta_{\rm m}^2/(2\sigma^2(\lambda,z))}\, {\rm d}\delta_{\rm m}$$
$$= 1 - {\rm erf}\left(\frac{\delta_{\rm c}}{\sqrt{2}\sigma(\lambda, z)}\right) = {\rm erfc}\left(\frac{\delta_{\rm c}}{\sqrt{2}\sigma(\lambda, z)}\right). \tag{7.7.31}$$

Derivation: Spherical Top-Hat Collapse and the Press–Schechter Mass Function

Consider a spherical overdense perturbation in a flat matter-dominated universe. By Birkhoff's theorem, the interior evolves like a closed FLRW universe with parametric cycloid equations:

$$r(\theta) = A(1 - \cos\theta), \qquad t(\theta) = B(\theta - \sin\theta), \qquad A^3 = GMB^2. \tag{1}$$

1. Turnaround and Collapse Times:

- Maximum expansion (turnaround) occurs at $\theta = \pi$: $r_{\rm ta} = 2A$, $t_{\rm ta} = \pi B$.
- Full collapse occurs at $\theta = 2\pi$: $r_{\rm coll} = 0$, $t_{\rm coll} = 2\pi B = 2t_{\rm ta}$.

2. Linear Density Contrast at Collapse: Taylor expanding $r(\theta)$ and $t(\theta)$ for $\theta \ll 1$:

$$r(t) = \frac{A}{2}\left(\frac{6t}{B}\right)^{2/3}\left[1 - \frac{1}{20}\left(\frac{6t}{B}\right)^{2/3} + \dots\right]. \tag{2}$$

The physical density inside the sphere is $\rho(t) = \frac{3M}{4\pi r^3(t)}$. The background density scales as $\bar{\rho}(t) = \frac{1}{6\pi G t^2}$. The exact linear density contrast $\delta_{\rm lin}(t)$ is:

$$\delta_{\rm lin}(t) = \frac{3}{5}\left(\frac{3}{4}\right)^{2/3}(\theta - \sin\theta)^{2/3}. \tag{3}$$

At collapse ($\theta = 2\pi$):

$$\delta_{\rm c} \equiv \delta_{\rm lin}(t_{\rm coll}) = \frac{3}{5}\left(\frac{3}{4}\right)^{2/3}(2\pi)^{2/3} = \frac{3(12\pi)^{2/3}}{20} \approx 1.686. \tag{4}$$

3. Differential Halo Mass Function: Assigning the collapsed fraction $F(>M) = 2P(\delta_{\rm m} > \delta_{\rm c})$ (where the factor of 2 accounts for the cloud-in-cloud resolution / excursion set normalization), the comoving number density of halos of mass M per unit mass interval is:

$$\frac{{\rm d}n(M,z)}{{\rm d}M} = \frac{\bar{\rho}_{\rm m}}{M}\left|\frac{{\rm d}F}{{\rm d}M}\right| = \sqrt{\frac{2}{\pi}}\,\frac{\bar{\rho}_{\rm m}}{M}\,\frac{\delta_{\rm c}}{\sigma^2(M,z)}\left|\frac{{\rm d}\sigma(M,z)}{{\rm d}M}\right| \exp\left(-\frac{\delta_{\rm c}^2}{2\sigma^2(M,z)}\right). \tag{5}$$

Cosmological Constraints from Cluster Abundance

Because the halo mass function depends exponentially on $\delta_{\rm c}/\sigma(M,z)$, the abundance of massive clusters as a function of redshift provides tight constraints:

1. **Cluster Abundance and $\sigma_8\Omega_{\rm m}^{0.6}$:** In an Einstein–de Sitter universe ($\Omega_{\rm m} = 1$), structure growth continues unabated to $z = 0$, producing rapid evolution in cluster abundance. In contrast, in a low-density ΛCDM universe ($\Omega_{\rm m} \approx 0.3$), growth ceases at $z \sim (\Omega_\Lambda/\Omega_{\rm m})^{1/3} -$

$1 \approx 0.3$, leading to mild cluster evolution. Calibrating local cluster counts with numerical simulations yields (Pierpaoli et al., 2001):

$$\sigma_8\,\Omega_{\rm m}^{0.6} = 0.495^{+0.034}_{-0.037}, \tag{7.7.32}$$

which intersects the cosmic shear degeneracy line in Fig. 7.19 to break parameter degeneracies.

2. **Cluster Velocity Dispersions:** For virialized clusters, the velocity dispersion $\sigma_v^2 \sim GM/R_{\rm vir}$ directly constrains the velocity power spectrum $P_V \propto \Omega_{\rm m}^{1.2}\sigma_8^2$.

3. **Baryon Fraction $f_{\rm b} \equiv \Omega_{\rm b}/\Omega_{\rm m}$:** X-ray bremsstrahlung emission ($L_X \propto \int n_{\rm e}^2 T^{1/2}{\rm d}V$) measures the intra-cluster gas mass $M_{\rm gas}$, while weak lensing and galaxy kinematics measure the total cluster mass $M_{\rm tot}$. Because clusters are enormous potential wells that retain all accreted primordial gas, the cluster baryon fraction mirrors the cosmic mean:

$$f_{\rm b} \equiv \frac{M_{\rm b}}{M_{\rm tot}} \simeq \frac{\Omega_{\rm b}}{\Omega_{\rm m}} \approx 0.10\text{–}0.15. \tag{7.7.33}$$

Combining this with the BBN/CMB baryon density $\Omega_{\rm b}h^2 \approx 0.022$ historically provided one of the earliest direct proofs that $\Omega_{\rm m} \approx 0.25\text{–}0.30 \ll 1$ (Bahcall & Cen, 1992).

7.8 Sources: Active Seeds and Topological Defects

Throughout preceding chapters, we operated under the standard inflationary paradigm: small, Gaussian initial metric perturbations were generated during an early epoch of accelerated expansion ($t \sim 10^{-35}$ s) on super-Hubble scales. After inflation, these perturbations evolve purely passively according to the linear homogeneous Einstein–Boltzmann system, with power spectra determined by deterministic transfer functions:

$$P_\Phi(k,\eta) = g^2(k,\eta)\,P_\Phi(k). \tag{7.8.1}$$

There exists an alternative, physically motivated mechanism for cosmic structure formation: **active sources** (or **seeds**). An active seed is an inhomogeneous and anisotropic matter component that makes up only a tiny fraction of the cosmic energy budget ($\rho_{\rm seed} \ll \rho_{\rm background}$) and interacts with ordinary matter and radiation purely gravitationally.

Unlike passive primordial perturbations, the energy-momentum tensor $\Theta_{\mu\nu}$ of an active seed acts as an inhomogeneous source term in the Einstein equations, continuously generating metric and fluid fluctuations throughout the history of the Universe.

7.8.1 Topological Defects

Topological defects represent the primary physical realization of active seeds. They arise naturally during cosmological phase transitions in Grand Unified Theories (GUTs) via the **Kibble mechanism** (Kibble, 1976).

Consider a complex scalar Higgs field ϕ with Lagrangian:

$$\mathcal{L} = -\frac{1}{2}g^{\mu\nu}(\nabla_\mu\phi)^\dagger(\nabla_\nu\phi) - V(\phi,T). \tag{7.8.2}$$

At temperatures above the critical temperature $T_{\rm c}$, the effective potential has a unique, symmetric minimum at $\phi = 0$. When the temperature drops below $T_{\rm c}$, the potential transitions to a "Mexican hat" shape (Fig. 7.20):

$$V(\phi) = \frac{\lambda}{4}\left(|\phi|^2 - v^2\right)^2, \tag{7.8.3}$$

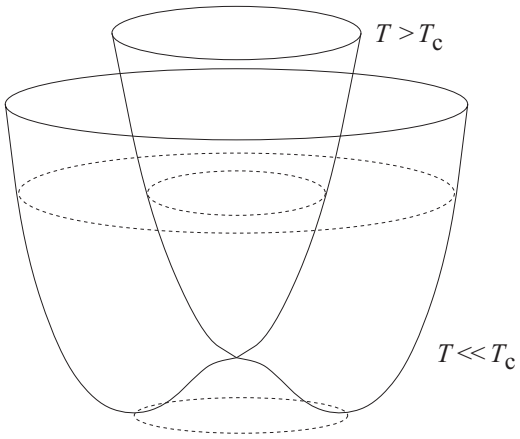


Figure 7.20: The finite-temperature effective potential $V(\phi, T)$ of a complex Higgs field for $T > T_c$ (single symmetric vacuum) and $T \ll T_c$ (spontaneously broken $U(1)$ symmetry with vacuum manifold $\mathcal{S}^1$).

with an entire continuous manifold of degenerate vacuum states $\mathcal{M} = \mathcal{S}^1 = \{\phi : |\phi| = v\}$.

Because regions separated by distances greater than the Hubble horizon $H^{-1}(t)$ are causally disconnected, the Higgs field chooses random, uncorrelated vacuum phases $\theta(\boldsymbol{x})$ in different horizon volumes. When tracing a closed spatial loop spanning several horizon patches, the phase mapping $\phi : \mathcal{S}^1_{\text{space}} \to \mathcal{M}$ may possess a non-zero winding number $n \in \mathbb{Z}$. By topological continuity, the field is forced to leave the vacuum manifold and vanish ($\phi = 0$) at some point inside the loop, trapping potential energy along a 1D curve. This filament of trapped energy is a **cosmic string** (Fig. 7.21).

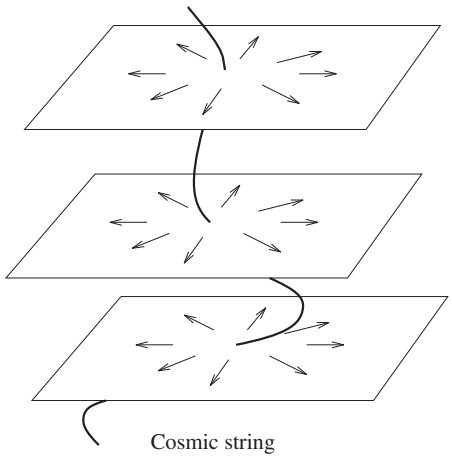


Figure 7.21: A cosmic string in 3D space with the corresponding spatial configuration of the complex Higgs field indicated by arrows.

The type of defect formed at a phase transition is dictated by the homotopy groups $\pi_n(\mathcal{M})$ of the vacuum manifold:

1. **Domain Walls ($\pi_0(\mathcal{M}) \neq \mathbb{I}$):** Arise from the breaking of a discrete symmetry (e.g., $\mathbb{Z}_2$). Domain walls have surface energy density $\sigma \sim \eta^3$, and their average energy density dilutes as $\rho_{\text{wall}} \propto a^{-1}$. They rapidly dominate the cosmic energy density and disrupt cosmic expansion, decisively ruling out stable domain walls (the Zel'dovich problem).

2. **Cosmic Strings ($\pi_1(\mathcal{M}) \neq \mathbb{I}$):** Arise from broken $U(1)$ abelian symmetries. As the universe expands, long strings intercommute upon collision, chopping off closed string loops that oscillate and decay by emitting gravitational waves. This energy loss allows the string network to attain a self-similar **scaling regime**, in which the string energy density maintains a constant ratio to the background cosmic energy density:

$$\frac{\rho_{\rm string}}{\rho_{\rm tot}} \simeq 4\pi G v^2 = 4\pi \left(\frac{v}{M_{\rm Pl}}\right)^2 \equiv \epsilon. \tag{7.8.4}$$

   To generate gravitational potentials of order $\Phi \sim 10^{-5}$, the symmetry-breaking scale must be $v \sim 10^{-3} M_{\rm Pl} \sim 10^{16}\,\text{GeV}$—the Grand Unification (GUT) scale.

3. **Monopoles ($\pi_2(\mathcal{M}) \neq \mathbb{I}$):** Local (gauged) monopoles are point-like defects whose density dilutes as non-relativistic matter $\rho_{\rm m} \propto a^{-3}$, overclosing the early universe (the monopole problem resolved by cosmic inflation). Global monopoles, however, annihilate efficiently via long-range gradient forces and attain a scaling density $\rho_{\rm M} \propto t^{-2}$.

4. **Textures ($\pi_3(\mathcal{M}) \neq \mathbb{I}$):** Arise in 4-component global scalar fields ($O(4) \to O(3)$ symmetry breaking). Textures are unstable, dynamic configurations that continuously collapse and unwind at the speed of light, achieving scaling without forming permanent topological singular lines.

7.8.2 Causal Scaling Seeds

In any cosmological scenario without super-luminal inflation, physical correlations cannot extend beyond the causal horizon. The two-point correlation function of any component of the seed energy-momentum tensor $\Theta_{\mu\nu}$ must vanish identically for spatial separations $r > 2\eta$ (compact support in real space).

By the Paley–Wiener theorem (Reed & Simon, 1980), the Fourier transform of a function with compact support is **analytic** in the complex wavevector $\boldsymbol{k}$. Furthermore, because the conformal Hubble time η is the only characteristic physical scale in the scaling regime, the dimensionless correlators depend on scale k and time t exclusively through the dimensionless scaling variable $z \equiv k\eta$.

We parameterize the seed energy-momentum tensor and its equal-time correlator as:

$$\Theta_{\mu\nu}(\boldsymbol{k}, \eta) = M^2\, \theta_{\mu\nu}(\boldsymbol{k}, \eta), \tag{7.8.5}$$

$$\langle \theta_{\mu\nu}(\boldsymbol{k}, \eta)\, \theta^*_{\rho\lambda}(\boldsymbol{k}', \eta) \rangle = (2\pi)^3\, C_{\mu\nu\rho\lambda}(k, \eta)\, \delta(\boldsymbol{k} - \boldsymbol{k}'). \tag{7.8.6}$$

Dimensional analysis requires $C_{\mu\nu\rho\lambda}(k, \eta)$ to have dimensions of inverse length (t^{-1}), taking the general form:

$$C_{0000} = \frac{1}{t} F_1(kt), \tag{7.8.7}$$

$$C_{0i0j} = \frac{1}{t}\left[t^2 k_i k_j F_2(kt) + (kt)^2 \delta_{ij} F_3(kt)\right], \tag{7.8.8}$$

where $F_n(z)$ are analytic functions of $z^2 = (k\eta)^2$ that decay to zero for $z \to \infty$. Energy-momentum conservation $\nabla^\mu \Theta_{\mu\nu} = 0$ reduces the full tensor correlator to five independent scalar functions of z.

Derivation: Scale-Invariant Sachs–Wolfe Plateau from Causal Scaling Seeds
The Bardeen metric potentials Φ and Ψ are sourced by the seed energy-momentum tensor through the relativistic Poisson equations $k^2\Phi, k^2\Psi \sim \epsilon\theta$. Their power spectrum takes

the scaling form:

$$\langle (\Phi+\Psi)(\boldsymbol{k},\eta)\,(\Phi+\Psi)^*(\boldsymbol{k}',\eta)\rangle = \epsilon^2(2\pi)^3\delta(\boldsymbol{k}-\boldsymbol{k}')\frac{F(z^2)}{k^4\eta^4} \equiv (2\pi)^3\delta(\boldsymbol{k}-\boldsymbol{k}')P(k,\eta). \quad (1)$$

The large-scale Sachs–Wolfe temperature anisotropy is given by:

$$\frac{\Delta T}{T}(\boldsymbol{k},\hat{\boldsymbol{n}}) = \mathrm{e}^{i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}(\eta_0-\eta_{\rm dec})}(\Phi+\Psi)(\boldsymbol{k},\eta_{\rm dec}) + \int_{\eta_{\rm dec}}^{\eta_0} \mathrm{e}^{i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}(\eta_0-t)}\partial_t(\Phi+\Psi)(\boldsymbol{k},\eta)\,\mathrm{d}t. \quad (2)$$

Differentiating the plane-wave prefactor:

$$\partial_t\left[\mathrm{e}^{i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}(\eta_0-t)}(\Phi+\Psi)\right] = \mathrm{e}^{i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}(\eta_0-t)}\left[-i(\boldsymbol{k}\cdot\hat{\boldsymbol{n}})(\Phi+\Psi)+\partial_t(\Phi+\Psi)\right]. \quad (3)$$

For super-horizon modes ($k\eta<1$), the time derivative $\partial_t(\Phi+\Psi)$ dominates over the gradient term $\boldsymbol{k}\cdot\hat{\boldsymbol{n}}(\Phi+\Psi)$. We can therefore approximate the integrand by the total derivative. Because the source potentials decay rapidly inside the horizon ($k\eta\gg1$), we integrate from $\eta_{\rm dec}$ up to horizon entry $\eta\approx 1/k$:

$$\begin{aligned}\frac{\Delta T}{T}(\boldsymbol{k},\hat{\boldsymbol{n}}) &\simeq \mathrm{e}^{i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}(\eta_0-\eta_{\rm dec})}(\Phi+\Psi)(\boldsymbol{k},\eta_{\rm dec}) + \left[\mathrm{e}^{i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}(\eta_0-t)}(\Phi+\Psi)(\boldsymbol{k},\eta)\right]_{\eta_{\rm dec}}^{1/k}\\ &= \mathrm{e}^{i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}(\eta_0-1/k)}(\Phi+\Psi)(\boldsymbol{k},1/k). \end{aligned} \quad (4)$$

The ordinary Sachs–Wolfe surface term at $\eta_{\rm dec}$ cancels identically against the lower boundary! The resulting two-point correlation is:

$$\left\langle \frac{\Delta T}{T}(\boldsymbol{k},\hat{\boldsymbol{n}})\left(\frac{\Delta T}{T}\right)^*(\boldsymbol{k}',\hat{\boldsymbol{n}}')\right\rangle \simeq \mathrm{e}^{i\boldsymbol{k}\cdot(\hat{\boldsymbol{n}}-\hat{\boldsymbol{n}}')(\eta_0-1/k)}\frac{\epsilon^2}{k^3}F(1)\,(2\pi)^3\delta(\boldsymbol{k}-\boldsymbol{k}'). \quad (5)$$

Expanding in spherical harmonics and Legendre polynomials using the Rayleigh plane-wave identity:

$$C_\ell \simeq \frac{2}{\pi}\epsilon^2F(1)\int_0^\infty\frac{\mathrm{d}k}{k}\,j_\ell^2(k\eta_0) = \frac{\epsilon^2F(1)}{\pi\ell(\ell+1)}. \quad (6)$$

Thus, causal scaling seeds generic predict a **strictly scale-invariant Sachs–Wolfe plateau**:

$$\ell(\ell+1)C_\ell = \frac{\epsilon^2F(1)}{\pi} = \text{constant}. \quad (7)$$

7.8.3 Calculating CMB Anisotropies from Sources

To perform an exact computation of CMB anisotropies and polarization from active sources, the full linear Einstein–Boltzmann system is written in Fourier space as an inhomogeneous matrix differential equation:

$$\mathcal{D}\boldsymbol{X}(\boldsymbol{k},\eta) = \epsilon\boldsymbol{S}(\boldsymbol{k},\eta), \quad (7.8.9)$$

where $\mathcal{D}$ is the first-order differential operator in conformal time,

$$\boldsymbol{X}(\boldsymbol{k},\eta) = \left(\delta_{\rm m},\delta_\gamma,v_\gamma,\Theta_\ell^{(m)},E_\ell^{(m)},B_\ell^{(m)},\dots\right)^\top \quad (7.8.10)$$

is the generalized cosmological perturbation vector, and $\boldsymbol{S}(\boldsymbol{k},\eta)$ is the active source vector constructed from linear combinations of the seed energy-momentum tensor components $\theta_{\mu\nu}(\boldsymbol{k},\eta)$.

The general solution is obtained via the Green's function $\boldsymbol{G}(\boldsymbol{k},\eta_0,\eta)$ of the operator $\mathcal{D}$:

$$\boldsymbol{X}(\boldsymbol{k},\eta_0) = \epsilon\int_{\eta_{\rm in}}^{\eta_0}\mathrm{d}\eta\,\boldsymbol{G}(\boldsymbol{k},\eta_0,\eta)\,\boldsymbol{S}(\boldsymbol{k},\eta) + \boldsymbol{X}_0(\boldsymbol{k},\eta_0), \quad (7.8.11)$$

where $\boldsymbol{X}_0$ is the homogeneous inflationary solution. Since inflation and post-inflationary defect formation are statistically uncorrelated, their cross-correlation vanishes, and the resulting power spectra simply add: $C_\ell^{\rm tot} = C_\ell^{\rm inf} + C_\ell^{\rm strings}$.

Focusing on the source-induced component, the ensemble-averaged perturbation covariance matrix is:

$$\langle X_i(\boldsymbol{k},\eta_0) X_j^*(\boldsymbol{k}',\eta_0)\rangle = \epsilon^2 \int_{\eta_{\rm in}}^{\eta_0} \mathrm{d}\eta \int_{\eta_{\rm in}}^{\eta_0} \mathrm{d}\eta'\, G_{im}(\boldsymbol{k},\eta_0,\eta) G_{jn}^*(\boldsymbol{k}',\eta_0,\eta')\, \langle S_m(\boldsymbol{k},\eta) S_n^*(\boldsymbol{k}',\eta')\rangle. \tag{7.8.12}$$

Evaluating Eq. (7.8.12) requires the **Unequal-Time Correlators (UETCs)** of the source:

$$\langle S_i(\boldsymbol{k},\eta) S_j^*(\boldsymbol{k}',\eta')\rangle = (2\pi)^3\, (\eta\eta')^{-p/2}\, F_{ij}\left(k\sqrt{\eta\eta'}, \frac{\eta}{\eta'}\right) \delta(\boldsymbol{k}-\boldsymbol{k}'). \tag{7.8.13}$$

Here $r \equiv \eta/\eta'$ is the time ratio and $z \equiv k\sqrt{\eta\eta'}$ is the geometric mean scaling wavenumber. The UETC functions $F_{ij}(z,r)$ are symmetric under $r \to 1/r$, analytic in z^2, and decay rapidly when $z \gg 1$ or $r \gg 1$ (Fig. 7.22).

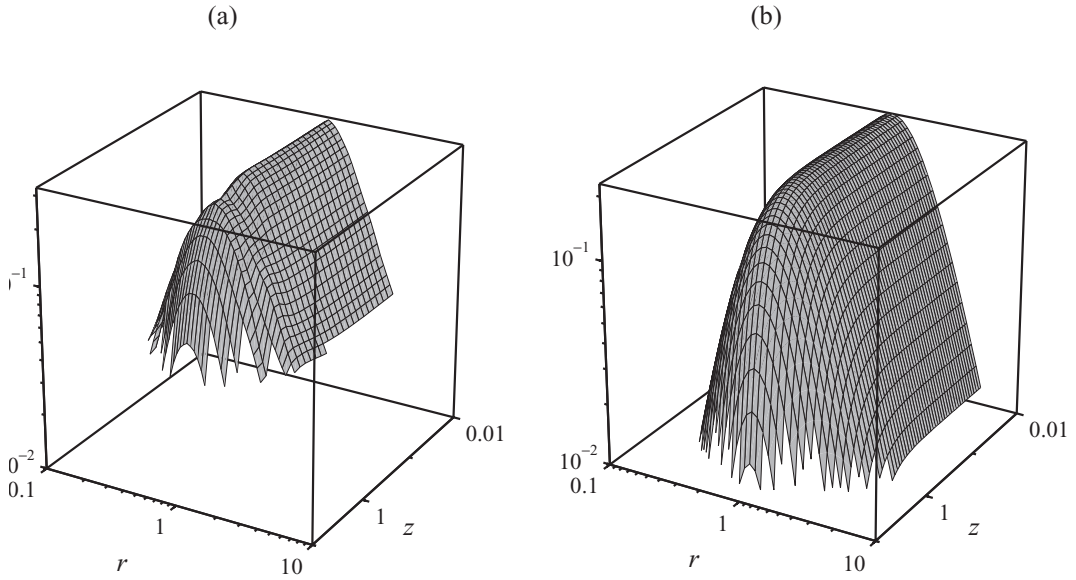


Figure 7.22: The source unequal-time correlation function $F(r,z)$ for vector perturbations as a function of $z = k\sqrt{\eta\eta'}$ and $r = t/t'$: (a) measured from numerical field simulations of a 4-component global texture, and (b) semi-analytical large-N limit (adapted from Durrer et al., 2002).

7.8.4 Decoherence of Acoustic Oscillations

The unequal-time correlation function $F(\eta,\eta',k)$ forms a positive semi-definite, symmetric integral kernel. By Mercer's theorem, it admits an exact spectral decomposition in terms of orthonormal temporal eigenmodes $v_n(k,\eta)$ and non-negative eigenvalues $\lambda_n^2(k)$:

$$F(\eta,\eta',k) = \sum_{n=1}^{\infty} \lambda_n^2(k)\, v_n(k,\eta)\, v_n^*(k,\eta'). \tag{7.8.14}$$

Substituting this spectral decomposition into the power spectrum integral Eq. (7.8.12), the 2D time double-integral factorizes into an incoherent sum over independent 1D integrals:

$$P_X(k) = \sum_{n=1}^{\infty} \lambda_n^2(k) \left| \int_{\eta_{\rm in}}^{\eta_0} \mathrm{d}\eta\, G(k,\eta_0,\eta)\, v_n(k,\eta) \right|^2. \tag{7.8.15}$$

Derivation: Spectral Decomposition and the Mechanism of Acoustic Decoherence

Why do active seeds wash out the acoustic peaks observed in inflationary models?

1. Inflationary Coherence: In standard inflationary cosmology, all Fourier modes of a given wavelength $\lambda = 2\pi/k$ cross the horizon with identical, synchronized initial phases ($\delta_k = 0$). In the acoustic plasma, the photon density contrast oscillates as a coherent cosine standing wave:

$$\delta_\gamma(k,\eta) \simeq A(k)\cos(c_s k\eta). \tag{1}$$

At the decoupling epoch $\eta_{\rm dec}$, modes with $k_m = m\pi/(c_s\eta_{\rm dec})$ simultaneously reach maximum compression or rarefaction ($\cos(c_s k_m \eta_{\rm dec}) = \pm 1$), producing sharp acoustic peaks in the angular power spectrum C_ℓ.

2. Active Seed Incoherence (Decoherence): In contrast, active sources continuously inject energy and generate perturbations throughout cosmic history. The resulting photon perturbation is an incoherent superposition of responses to sources acting at different epochs t_n:

$$\delta_\gamma(k,\eta) = \sum_{n=1}^{\infty} \lambda_n(k)\, A_n(k)\cos\left(c_s k\eta - \delta_n(k)\right), \tag{2}$$

where each eigenmode $v_n(k,\eta)$ imparts a distinct, time-dependent phase shift $\delta_n(k)$. When computing the total power spectrum:

$$P_\gamma(k,\eta_{\rm dec}) = \sum_{n=1}^{\infty} \lambda_n^2(k)\, A_n^2(k)\cos^2\left(c_s k\eta_{\rm dec} - \delta_n(k)\right). \tag{3}$$

Because different eigenmodes possess different phase shifts $\delta_n(k)$, the peaks and troughs of individual terms in the sum cancel out. Summing over a large number of active eigenmodes ($N \gtrsim 10$) washes out the sharp oscillatory acoustic peaks, replacing them with a single broad, smoothed **acoustic hump** (Albrecht et al., 1996; Durrer et al., 2002).

7.8.5 Observational Results and Constraints on Defects

CMB anisotropies seeded by causal scaling defects exhibit three defining observational signatures:

1. **Prominent Vector Modes:** While vector (vorticity) perturbations generated during inflation decay rapidly as a^{-2} and leave no observable trace, active scaling seeds continuously generate vector perturbations on the horizon scale. Vector modes contribute an amplitude comparable to scalars ($C_\ell^{(V)} \sim C_\ell^{(S)}$; Fig. 7.23).

2. **Tensor Perturbations:** Defect dynamics produce a stochastic background of gravitational waves with $C_2^{(T)}/C_2^{(S)} \sim 0.25$.

3. **Acoustic Hump without Oscillatory Peaks:** Because perturbations are isocurvature-like and subject to decoherence, the spectrum features a scale-invariant Sachs–Wolfe plateau at low ℓ, followed by a single broad hump at $\ell \sim 300$–500 rather than alternating acoustic peaks.

Cosmological Implications

Precision WMAP, ACT, SPT, and Planck CMB observations have revealed multiple sharp, alternating acoustic peaks and a pronounced acoustic phase shift that are in perfect agreement

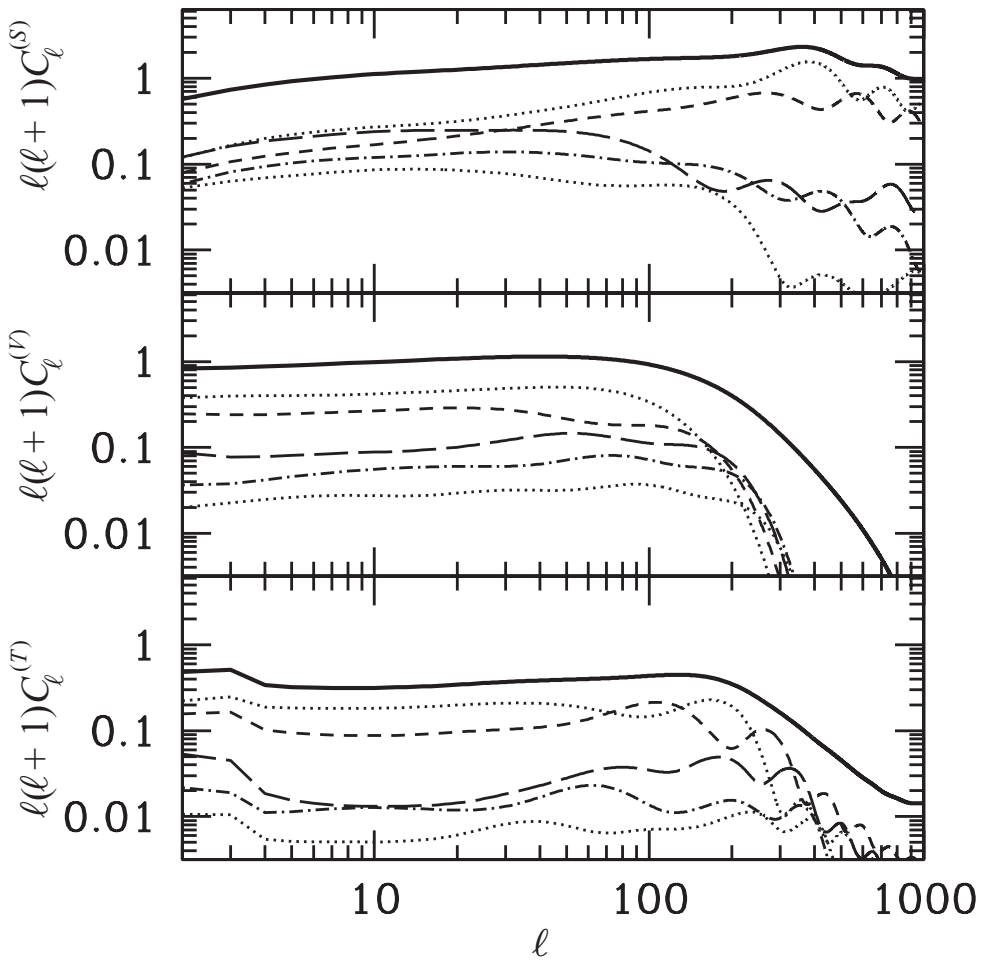


Figure 7.23: The scalar (S), vector (V), and tensor (T) contributions to the CMB temperature power spectrum $\ell(\ell+1)C_\ell$ for a 4-component global texture model. Dashed curves indicate contributions from individual single eigenfunctions, while the bold solid curve shows their total sum over 100 modes. While individual eigenmodes display oscillations, they wash out into a broad hump in the sum (data courtesy of Bevis et al., 2004).

with coherent adiabatic inflationary perturbations. Pure topological defect models are therefore **decisively ruled out** as the primary source of cosmic structure.

However, in hybrid Grand Unified models where a symmetry-breaking phase transition occurs at the end of inflation, cosmic strings can contribute a sub-dominant fraction to CMB fluctuations. Current observational data limit the cosmic string tension $G\mu$ and dimensionless defect strength ϵ to:

$$G\mu \lesssim 10^{-7}, \qquad \epsilon \equiv 4\pi G v^2 \lesssim 10^{-5}, \tag{7.8.16}$$

restricting topological defects to contribute less than $\sim$ 2–5% of the total CMB temperature anisotropy power.

Chapter 8

Gravitational Lensing and Dark Matter

8.1 Gravitational lensing and its applications

Light trajectories are deflected by any inhomogeneous gravitational field, as established in the foundational framework of general relativity (see Chapter 1). Historically, this phenomenon provided the very first empirical validation of Einsteinian gravity during the 1919 solar eclipse expedition led by Dyson, Eddington, and Davidson. In contemporary astrophysics and cosmology, gravitational deflection of light has matured into one of the most powerful and versatile observational probes available. Because photons respond directly to the total metric perturbations generated by all energy-momentum sources regardless of their physical state or electromagnetic emission, gravitational lensing provides an unbiased, geometrically calibrated map of the cosmic matter distribution, and in particular of the dark matter component that dominates the clustered mass budget of the Universe. The astronomical applications of gravitational lensing encompass

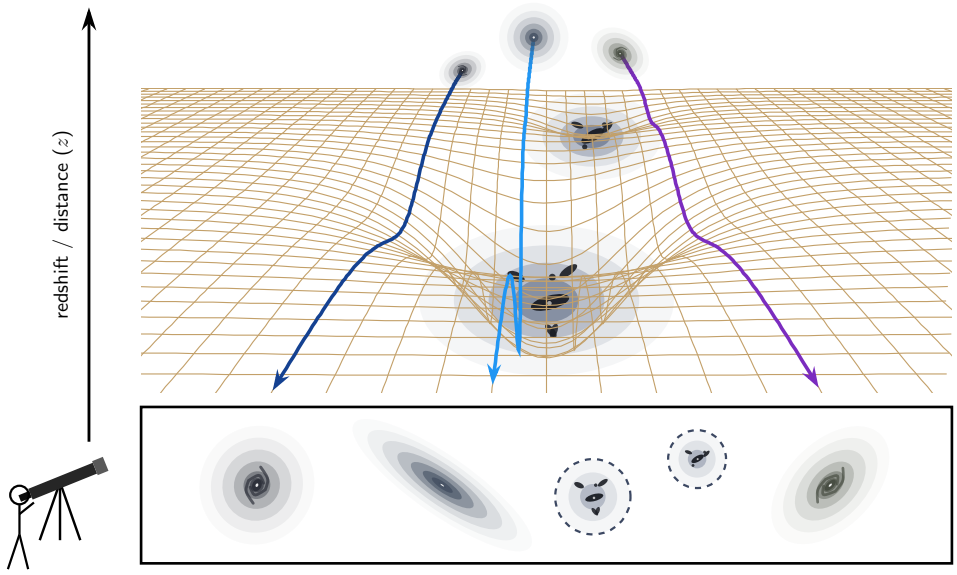


Figure 8.1: Schematic overview of cosmic shear and gravitational lensing by large-scale structure. Light rays from distant background galaxies (blue and purple trajectories) propagate across cosmological distances (redshift z) and are deflected by the gravitational potential wells of clustered foreground galaxies and dark matter halos. In the observer's field of view (bottom panel), light from the distant background galaxies is tangentially sheared into coherent ellipses around the foreground lens clusters and dark matter halos (indicated by the dashed circles causing the gravitational depressions above).

Table 8.1: Classification of the principal gravitational lensing regimes categorized by source and lens configurations, along with their characteristic observational manifestations and cosmological applications. Lensing regimes are designated as strong lensing (SL), microlensing (ML), weak lensing (WL), and statistical weak lensing (stWL).

Source	**Lens**	**Observational Effects**	**Astrophysical Applications**
Star (Galactic)	Star (Galactic)	ML: photometric light curves	Extrasolar planets, Galactic structure
Star (Extragalactic)	Compact object	ML: transient amplification	Dark matter in Milky Way halo
Galaxy	Galaxy	stWL: tangential shear, ellipticity bias	Galaxy halo profiles, galaxy bias
Galaxy	Cluster	SL: giant arcs, multiple images WL: arclets, magnification bias	Cluster mass distribution, Ω_Λ, dark matter substructure, redshifts
Galaxy	LSS	stWL: cosmic shear, two-point stats	Cosmological parameters $(\sigma_8, \Omega_\mathrm{m})$, matter power spectrum evolution
Quasar	Galaxy	SL: multiple images, Einstein rings, time delays	Absolute galactic mass, substructure, Hubble constant H_0, dark energy Ω_Λ
Quasar	LSS	stWL: cosmic magnification, quasar–galaxy cross-correlations	Cosmological parameters, bias, large-scale density fluctuations
CMB last scattering	LSS	stWL: smoothing of acoustic peaks, generation of B-mode polarization	Cosmological parameters, neutrino mass, primordial power spectrum tilt

an extraordinary dynamic range of physical length scales and mass regimes, from terrestrial planet searches via stellar microlensing to cosmic shear surveys probing large-scale structure and the cosmic microwave background (CMB). Table 9.1 provides a systematic classification of the primary gravitational lensing regimes and their astrophysical applications. Comprehensive expositions of the theoretical foundations, historical developments, and observational techniques of gravitational lensing may be found in the classic monographs of Refs. [1,2] and the extensive review articles of Refs. [3,4].

8.1.1 Gravitational lensing in the thin-lens regime

Light deflection by a point-like mass

Deflection of light in curved spacetime We begin by establishing how curved spacetime deflects light rays from first principles. In standard astrophysical configurations, the gravitational fields produced by localized deflectors (such as stars, galaxies, and galaxy clusters) are weak, $|\Phi|/c^2 \ll 1$, and internal velocities are non-relativistic, $v \ll c$. Under these conditions, the spacetime metric can be accurately described by the perturbed Minkowski metric in Newtonian gauge (conformal Newtonian coordinates):

$$\mathrm{d}s^2 = -\left(1 + \frac{2\Phi}{c^2}\right) c^2\,\mathrm{d}t^2 + \left(1 - \frac{2\Phi}{c^2}\right)(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2), \tag{8.1.1}$$

where $\Phi(\boldsymbol{r})$ is the Newtonian gravitational potential ($\Phi < 0$ for an attractive mass distribution).

Photons propagate along null geodesics satisfying $\mathrm{d}s^2 = 0$. Setting $\mathrm{d}s^2 = 0$ in Eq. (8.1.1):

$$0 = -\left(1 + \frac{2\Phi}{c^2}\right) c^2\,\mathrm{d}t^2 + \left(1 - \frac{2\Phi}{c^2}\right) |\,\mathrm{d}\boldsymbol{x}\,|^2. \tag{8.1.2}$$

Rearranging for the coordinate speed of light $v \equiv |\,\mathrm{d}\boldsymbol{x}\,|/\,\mathrm{d}t$:

$$v^2 = c^2 \frac{1 + 2\Phi/c^2}{1 - 2\Phi/c^2} \approx c^2 \left(1 + \frac{2\Phi}{c^2}\right)^2 \approx c^2 \left(1 + \frac{4\Phi}{c^2}\right), \tag{8.1.3}$$

where we have expanded to first order in the small parameter $|\Phi|/c^2 \ll 1$. Taking the square root gives:

$$v(\boldsymbol{r}) \approx c \left(1 + \frac{2\Phi(\boldsymbol{r})}{c^2}\right). \tag{8.1.4}$$

Because $\Phi < 0$, the coordinate velocity of light is strictly less than c. Spacetime around a mass distribution acts as an optically dense medium with an effective index of refraction $n(\boldsymbol{r})$:

$$n(\boldsymbol{r}) \equiv \frac{c}{v(\boldsymbol{r})} \approx 1 - \frac{2\Phi(\boldsymbol{r})}{c^2} > 1. \tag{8.1.5}$$

Physical Remark: The Factor of Two. In a purely Newtonian corpuscular calculation (as conceived by Soldner in 1801), light is treated as a stream of massive particles moving at speed c. In that case, only the temporal metric component $g_{00} = -(1+2\Phi/c^2)$ contributes, leading to $n_{\text{Newton}} = 1 - \Phi/c^2$. In General Relativity, the spatial metric components $g_{ij} = (1 - 2\Phi/c^2)\delta_{ij}$ contribute an equal amount due to the spatial curvature induced by the mass. The relativistic deflection is therefore precisely twice the Newtonian corpuscular value:

$$\alpha_{\text{GR}} = 2\alpha_{\text{Newton}}. \tag{1}$$

This factor of 2 was famously confirmed by the 1919 solar eclipse measurements, cementing General Relativity over Newtonian gravity.

By Fermat's Principle, light rays follow stationary trajectories of the arrival time t:

$$t = \frac{1}{c}\int n(\boldsymbol{x})\ \mathrm{d}l = \frac{1}{c}\int \left(1 - \frac{2\Phi}{c^2}\right) \mathrm{d}l\,. \tag{8.1.6}$$

Applying the Euler–Lagrange variational equations to the path $\boldsymbol{x}(l)$ parameterized by Euclidean arc length l, the ray trajectory satisfies:

$$\frac{\mathrm{d}}{\mathrm{d}l}\left(n\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}l}\right) = \boldsymbol{\nabla} n \approx -\frac{2}{c^2}\boldsymbol{\nabla}\Phi. \tag{8.1.7}$$

Because astronomical deflection angles are minute ($\alpha \ll 1\,\text{rad}$), the trajectory deviates only infinitesimally from a straight line. We can therefore evaluate the integral along an unperturbed straight path oriented along the optical axis z at physical impact parameter vector $\boldsymbol{b} = (x, y)$ in the transverse plane. This constitutes the *Born approximation*. Integrating the transverse gradient along the unperturbed line of sight yields the total deflection angle vector $\boldsymbol{\alpha}(\boldsymbol{b})$:

$$\boldsymbol{\alpha}(\boldsymbol{b}) = -\int_{-\infty}^{\infty} \boldsymbol{\nabla}_\perp n\ \mathrm{d}z = \frac{2}{c^2}\int_{-\infty}^{\infty} \boldsymbol{\nabla}_\perp \Phi(\boldsymbol{b}, z)\ \mathrm{d}z\,, \tag{8.1.8}$$

where $\boldsymbol{\nabla}_\perp \equiv (\partial/\partial x, \partial/\partial y)$ represents the two-dimensional spatial gradient perpendicular to the line of sight.

For an isolated point mass M located at the origin, the Newtonian potential at physical distance $r = \sqrt{b^2 + z^2}$ is:

$$\Phi(\boldsymbol{b}, z) = -\frac{G_{\text{N}} M}{\sqrt{b^2 + z^2}}. \tag{8.1.9}$$

Taking the transverse gradient:

$$\boldsymbol{\nabla}_\perp \Phi(\boldsymbol{b}, z) = \frac{G_{\text{N}} M \boldsymbol{b}}{(b^2 + z^2)^{3/2}}. \tag{8.1.10}$$

Substituting into Eq. (8.1.8) and evaluating the line-of-sight integral:

$$\boldsymbol{\alpha}(\boldsymbol{b}) = \frac{2G_{\text{N}} M \boldsymbol{b}}{c^2}\int_{-\infty}^{\infty} \frac{\mathrm{d}z}{(b^2 + z^2)^{3/2}} = \frac{4G_{\text{N}} M}{c^2 b^2}\boldsymbol{b} = \frac{2R_{\text{Sch}}}{b}\hat{\boldsymbol{b}}, \tag{8.1.11}$$

where $R_{\text{Sch}} \equiv 2G_{\text{N}} M/c^2$ is the Schwarzschild radius of the lens, and $\hat{\boldsymbol{b}} \equiv \boldsymbol{b}/b$ is the unit vector directed radially outward from the deflector.

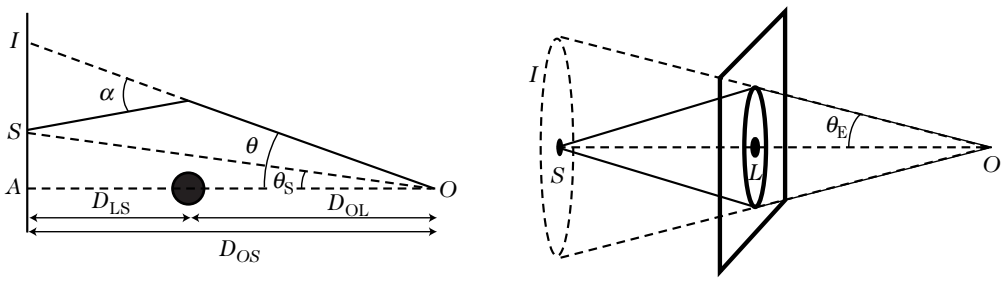


Figure 8.2: (Left) Geometrical configuration of a gravitational lensing system in the thin-lens approximation. L denotes the localized deflector, S the true source position, and I the apparent image position. The unperturbed source angle is θ_S, the observed image angle is θ, and the physical deflection angle is α. The relevant cosmological angular diameter distances are $D_{\rm OL}$, $D_{\rm OS}$, and $D_{\rm LS}$. (Right) When the source, deflector, and observer are perfectly collinear ($\theta_S = 0$), rotational symmetry around the optical axis produces an annular image known as an Einstein ring of angular radius $\theta_{\rm E}$.

Proof of the Line-of-Sight Integral. Substituting $z = b\tan\phi$ with $\mathrm{d}z = b\sec^2\phi\,\mathrm{d}\phi$:

$$\int_{-\infty}^{\infty}\frac{\mathrm{d}z}{(b^2+z^2)^{3/2}} = \int_{-\pi/2}^{\pi/2}\frac{b\sec^2\phi\,\mathrm{d}\phi}{b^3\sec^3\phi} = \frac{1}{b^2}\int_{-\pi/2}^{\pi/2}\cos\phi\,\mathrm{d}\phi = \frac{1}{b^2}\left[\sin\phi\right]_{-\pi/2}^{\pi/2} = \frac{2}{b^2}. \quad (1)$$

Multiplying by $\frac{2G_{\rm N}M\boldsymbol{b}}{c^2}$ directly confirms Eq. (8.1.11):

$$\boldsymbol{\alpha}(\boldsymbol{b}) = \frac{2G_{\rm N}M\boldsymbol{b}}{c^2}\left(\frac{2}{b^2}\right) = \frac{4G_{\rm N}M}{c^2 b}\hat{\boldsymbol{b}} = \frac{2R_{\rm Sch}}{b}\hat{\boldsymbol{b}}. \quad (2)$$

The Lens Equation The fundamental equation governing gravitational lensing establishes the mapping between the unperturbed angular position of a light source and its observed angular position on the celestial sphere (see Fig. 8.2). We define the optical axis as the fiducial line connecting the observer O to the center of the lens L. The source plane and the lens plane are defined as two transverse Euclidean planes oriented perpendicular to the optical axis, situated at angular diameter distances $D_{\rm OS}$ and $D_{\rm OL}$ from the observer, respectively. The angular diameter distance from the lens to the source plane is denoted by $D_{\rm LS}$. In general curved spacetimes, angular diameter distances are non-additive ($D_{\rm OS} \neq D_{\rm OL} + D_{\rm LS}$); their precise values depend on the background cosmological metric and the respective redshifts $z_{\rm L}$ and $z_{\rm S}$.

Referring to the geometric ray-tracing diagram of Fig. 8.2, let A denote the intersection of the optical axis with the source plane. The transverse physical position of the apparent image on the source plane is given by the vector sum $AI = AS + SI$. In the small-angle approximation characteristic of astrophysical deflections ($\theta, \theta_S, \alpha \ll 1$), the transverse separation between the optical axis and the source is $AS = \theta_S D_{\rm OS}$, while the projected image position satisfies $AI = \theta D_{\rm OS}$. The transverse displacement induced by the deflection angle α across the distance $D_{\rm LS}$ between lens and source is $SI = \alpha D_{\rm LS}$. Equating these geometric segments yields the scalar lens equation:

$$AI = AS + SI \implies \theta D_{\rm OS} = \theta_S D_{\rm OS} + \alpha D_{\rm LS} \implies \theta = \theta_S + \frac{D_{\rm LS}}{D_{\rm OS}}\alpha(\theta). \quad (8.1.12)$$

Substituting the point-mass deflection angle $\alpha(\theta) = \frac{4G_{\rm N}M}{c^2 b} = \frac{2R_{\rm Sch}}{D_{\rm OL}\theta}$ into Eq. (8.1.12) gives:

$$\theta = \theta_S + \frac{2R_{\rm Sch}}{\mathcal{D}}\frac{1}{\theta}, \quad (8.1.13)$$

where we have defined the effective geometrical distance factor:

$$\mathcal{D} \equiv \frac{D_{\rm OS} D_{\rm OL}}{D_{\rm LS}}. \tag{8.1.14}$$

Pedagogical Bridge: From the Gaussian Thin-Lens Formula to Gravitational Lensing

To see the deep mathematical and physical connection between gravitational lensing and introductory geometric optics, consider a standard thin glass lens of focal length f. In elementary optics, a ray passing through a thin lens at physical height ξ is refracted by a deflection angle proportional to its height:

$$\alpha_{\rm glass}(\xi) = \frac{\xi}{f}. \tag{1}$$

Referring to the ray geometry in Fig. 8.2, let $u \equiv D_{\rm LS}$ denote the object (source) distance from the lens, and $v \equiv D_{\rm OL}$ denote the image (observer) distance from the lens. In Euclidean space, the total observer-source distance is $D_{\rm OS} = u + v$. Substituting the physical impact parameter $\xi = D_{\rm OL}\theta = v\theta$ and the deflection angle $\alpha(\theta) = v\theta/f$ into the general ray-tracing relation $\theta_S = \theta - \frac{D_{\rm LS}}{D_{\rm OS}}\alpha(\theta)$:

$$\theta_S = \theta\left(1 - \frac{D_{\rm LS} D_{\rm OL}}{D_{\rm OS} f}\right) = \theta\left(1 - \frac{uv}{(u+v)f}\right). \tag{2}$$

For an on-axis point source ($\theta_S = 0$) to form a focused image at the observer ($\theta \neq 0$), the factor in parentheses must vanish identically:

$$1 - \frac{uv}{(u+v)f} = 0 \iff \frac{1}{f} = \frac{u+v}{uv} = \frac{1}{u} + \frac{1}{v}. \tag{3}$$

This is precisely the Gaussian thin-lens equation $\frac{1}{u} + \frac{1}{v} = \frac{1}{f}$ familiar from undergraduate physics. However, while a manufactured glass lens has a fixed focal length $f = \xi/\alpha_{\rm glass}(\xi)$, in gravitational lensing the effective focal length:

$$f(\xi) \equiv \frac{\xi}{\alpha(\xi)} = \frac{c^2\xi^2}{4G_{\rm N}M(\leq \xi)} \tag{4}$$

is set entirely by the gravitational deflector's mass distribution $M(\leq \xi)$.

Einstein radius Due to the perfect rotational symmetry of the point-mass potential about the optical axis, a source positioned exactly on the optical axis ($\theta_S = 0$) cannot produce isolated point images. Instead, rays emanating from the source and passing through the lens plane at a constant critical impact parameter are deflected symmetrically toward the observer, forming an annular luminous feature known as the *Einstein ring*. Setting $\theta_S = 0$ in Eq. (8.1.13), the angular radius of this ring evaluates to:

$$\theta_{\rm E} = \sqrt{\frac{4G_{\rm N}M}{c^2}\frac{D_{\rm LS}}{D_{\rm OS}D_{\rm OL}}} = \sqrt{\frac{2R_{\rm Sch}}{\mathcal{D}}}. \tag{8.1.15}$$

The Einstein angular radius $\theta_{\rm E}$ establishes the characteristic angular scale of any gravitational lens configuration. The corresponding physical radius in the lens plane is $R_{\rm E} = D_{\rm OL}\theta_{\rm E}$. If a source lies within the angular radius $\theta_S \lesssim \theta_{\rm E}$, the lens induces strong gravitational lensing characterized by multiple images and high magnification; conversely, for distant source alignments $\theta_S \gg \theta_{\rm E}$, the deflection is minute and the system resides in the weak lensing regime.

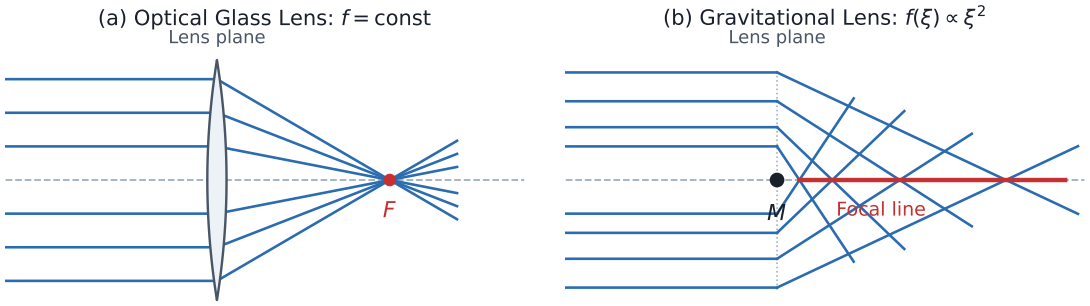


Figure 8.3: Comparison of focal properties between an optical thin glass lens and a gravitational point-mass lens. (a) A manufactured thin glass lens has linear deflection $\alpha \propto \xi$, resulting in a constant focal length f and a unique focal point F. (b) A gravitational point lens has inverse deflection $\alpha(\xi) \propto 1/\xi$, so the effective focal length $f(\xi) \propto \xi^2$ varies with impact parameter ξ. Rays at different radii focus at different points along the optical axis, forming a continuous focal line instead of a single focal point.

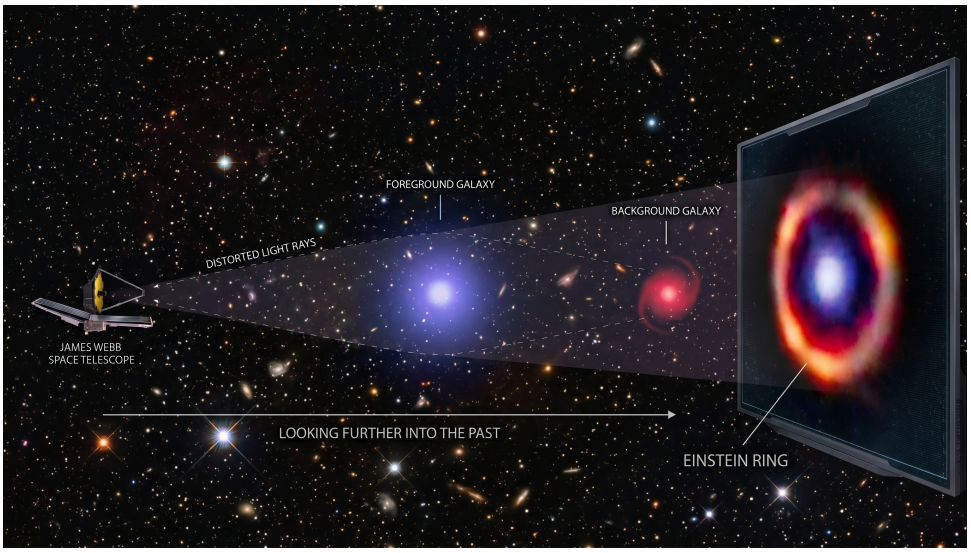


Figure 8.4: Geometry of an Einstein ring produced by gravitational lensing when the source, lens, and observer are perfectly aligned. The lensed image of the background source forms a ring at the Einstein radius θ_E around the lens.

The characteristic numerical order of magnitude of the Einstein radius for a solar-mass star acting as a Galactic microlens ($M \sim M_\odot, \mathcal{D} \sim 10\,\text{kpc}$) is:

$$\theta_{\text{E,star}} = 0.9 \left(\frac{M}{M_\odot}\right)^{1/2} \left(\frac{\mathcal{D}}{10\,\text{kpc}}\right)^{-1/2} \text{mas}, \tag{8.1.16}$$

whereas for an entire galaxy acting as a strong deflector at cosmological distances ($M \sim 10^{11} M_\odot, \mathcal{D} \sim 1\,\text{Gpc}$):

$$\theta_{\text{E,gal}} = 0.9 \left(\frac{M}{10^{11} M_\odot}\right)^{1/2} \left(\frac{\mathcal{D}}{1\,\text{Gpc}}\right)^{-1/2} \text{arcsec}, \tag{8.1.17}$$

and for a rich galaxy cluster ($M \sim 10^{15} M_\odot, \mathcal{D} \sim 1\,\text{Gpc}$):

$$\theta_{\text{E,cluster}} \approx 30 \left(\frac{M}{10^{15} M_\odot}\right)^{1/2} \left(\frac{\mathcal{D}}{1\,\text{Gpc}}\right)^{-1/2} \text{arcsec}. \tag{8.1.18}$$

Image positions To understand why a gravitational deflector generates multiple images, recall the behavior of an ordinary thin glass lens in geometric optics. A glass convex lens deflects a ray incident at height h linearly: $\alpha(h) \approx h/f$. Its local refractive power:

$$P \equiv \frac{\mathrm{d}\alpha}{\mathrm{d}h} = \frac{1}{f} = \text{const} \tag{8.1.19}$$

is constant across the entire aperture. Because the deflection angle is strictly linear in height, all parallel rays converge at a single focal point, and the ray-tracing relation $y_{\text{source}} = h(1 - D/f)$ is linear, forming a single sharp image.

In gravitational lensing, the ray equation relating the true source position θ_S and observed image angle θ on the sky is the identical geometric ray trace:

$$\theta_S = \theta - \alpha(\theta). \tag{8.1.20}$$

For a point mass M, however, Einstein's deflection angle is inversely proportional to the impact parameter:

$$\alpha(\theta) = \frac{\theta_{\mathrm{E}}^2}{\theta}. \tag{8.1.21}$$

A gravitational lens is therefore a fundamentally non-linear, *over-focusing* lens: its local focal power:

$$P(\theta) \equiv \frac{\mathrm{d}\alpha}{\mathrm{d}\theta} = -\frac{\theta_{\mathrm{E}}^2}{\theta^2} \tag{8.1.22}$$

changes at every radius, diverging near the center and decaying outward (optically, this behaves like looking through the flared foot of a wine glass). Rays passing close to the mass are deflected far more violently than rays passing far out.

Because the deflection $\alpha(\theta) \propto 1/\theta$ is non-linear, the ray condition $\theta - \theta_S = \theta_{\mathrm{E}}^2/\theta$ becomes a quadratic equation:

$$\theta = \theta_S + \frac{\theta_{\mathrm{E}}^2}{\theta} \iff \theta^2 - \theta_S\theta - \theta_{\mathrm{E}}^2 = 0. \tag{8.1.23}$$

This quadratic equation in θ possesses two real solutions for any non-zero source position θ_S:

$$\theta_\pm = \frac{1}{2}\left(\theta_S \pm \sqrt{\theta_S^2 + 4\theta_{\mathrm{E}}^2}\right). \tag{8.1.24}$$

Introducing the dimensionless source position $u \equiv \theta_S/\theta_{\mathrm{E}}$ and image positions $u_\pm \equiv \theta_\pm/\theta_{\mathrm{E}}$:

$$u_\pm = \frac{1}{2}\left(u \pm \sqrt{u^2 + 4}\right). \tag{8.1.25}$$

A point mass always produces exactly two images:

1. Image θ_+ is situated outside the Einstein ring ($\theta_+ > \theta_{\mathrm{E}}$) on the same side of the lens center as the unlensed source ($\theta_+ \cdot \theta_S > 0$).

2. Image θ_- is situated inside the Einstein ring ($|\theta_-| < \theta_{\mathrm{E}}$) on the opposite side of the lens center ($\theta_- \cdot \theta_S < 0$).

The total angular separation between the two images is:

$$\Delta\theta \equiv \theta_+ - \theta_- = \sqrt{\theta_S^2 + 4\theta_{\mathrm{E}}^2} \geq 2\theta_{\mathrm{E}}. \tag{8.1.26}$$

The angular separation between the two images can never be smaller than twice the Einstein radius ($2\theta_{\mathrm{E}}$). The three-dimensional ray-tracing geometry and the resulting face-on sky plane distortion are illustrated in Fig. 8.5.

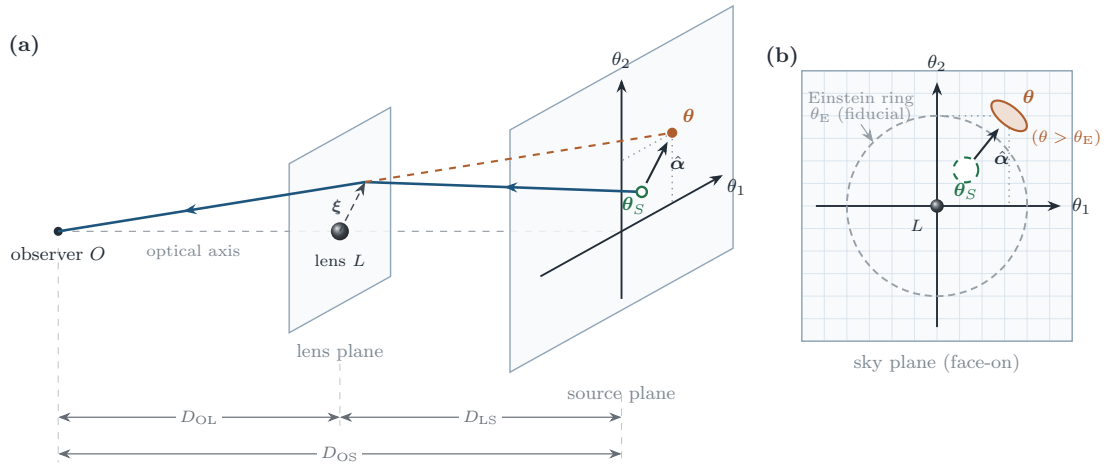


Figure 8.5: (a) Three-dimensional geometric configuration of a gravitational lensing system in the thin-lens approximation. Light rays from a distant source S at unperturbed angular position $\boldsymbol{\theta}_S$ are deflected at impact parameter $\boldsymbol{\xi}$ by a localized deflector L on the lens plane, arriving at observer O along the apparent direction $\boldsymbol{\theta} = \boldsymbol{\theta}_S + \hat{\boldsymbol{\alpha}}$. Cosmological angular diameter distances are denoted $D_{\rm OL}$, $D_{\rm LS}$, and $D_{\rm OS}$. (b) Sky plane seen face-on by the observer in angular coordinates (θ_1, θ_2). The dashed circle denotes the fiducial Einstein ring of angular radius $\theta_{\rm E}$ (the configuration obtained when the source is centered on the optical axis, $\boldsymbol{\theta}_S = 0$). For an off-axis source ($\boldsymbol{\theta}_S \neq 0$), the primary image $\boldsymbol{\theta}$ is displaced strictly outside the Einstein ring ($\theta > \theta_{\rm E}$) and distorted tangentially into an ellipse by the tidal gravitational shear of the deflector.

Magnification and Parity Because gravitational light deflection preserves photon frequency and phase-space distribution, Liouville's theorem guarantees that gravitational lensing conserves the intrinsic surface brightness (specific intensity) of the source:

$$I(\boldsymbol{\theta}) = I_S(\boldsymbol{\theta}_S(\boldsymbol{\theta})). \tag{8.1.27}$$

Consequently, any change in the total received flux is purely geometric, arising from the stretching or compression of the solid angle subtended by the image relative to the source:

$$\mu \equiv \frac{\mathrm{d}\Omega}{\mathrm{d}\Omega_S} = \frac{\theta}{\theta_S}\frac{\mathrm{d}\theta}{\mathrm{d}\theta_S}. \tag{8.1.28}$$

Differentiating the lens equation $\theta_S(\theta) = \theta - \theta_{\rm E}^2/\theta$ gives $\mathrm{d}\theta_S/\mathrm{d}\theta = 1 + \theta_{\rm E}^2/\theta^2$. Substituting into Eq. (8.1.28) yields the local magnification as a function of image position:

$$\mu(\theta) = \frac{\theta}{\theta - \theta_{\rm E}^2/\theta}\left(1 + \frac{\theta_{\rm E}^2}{\theta^2}\right)^{-1} = \left[1 - \left(\frac{\theta_{\rm E}}{\theta}\right)^4\right]^{-1}. \tag{8.1.29}$$

Evaluating this at the two image roots $\theta_\pm/\theta_{\rm E} = \frac{1}{2}(u \pm \sqrt{u^2+4})$ gives the individual image magnifications:

$$\mu_\pm = \frac{1}{2} \pm \frac{u^2+2}{2u\sqrt{u^2+4}}. \tag{8.1.30}$$

Because the interior image forms inside the Einstein ring ($|\theta_-| < \theta_{\rm E}$), Eq. (8.1.29) implies $\mu_- < 0$. The negative sign indicates an inversion of parity: image I_- is a mirror-reflected reproduction of the source profile (see Fig. 8.6). The physically observable total flux amplification is the sum of the absolute magnifications of both images:

$$\mu = |\mu_+| + |\mu_-| = \mu_+ - \mu_- = \frac{u^2+2}{u\sqrt{u^2+4}}. \tag{8.1.31}$$

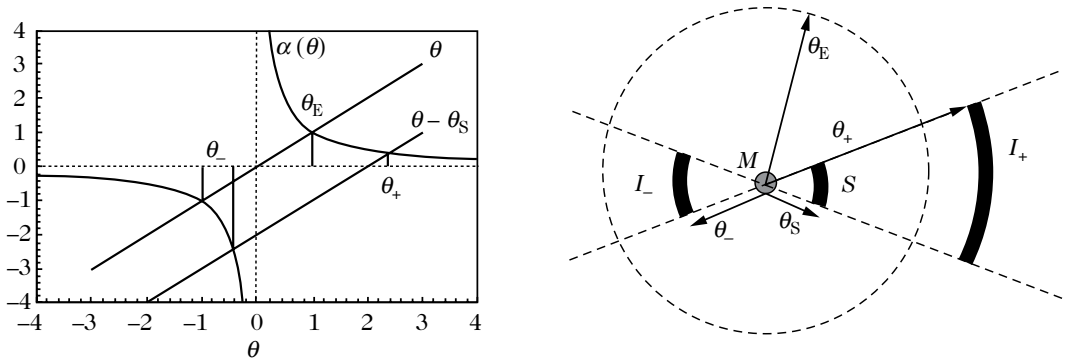


Figure 8.6: (Left) Graphical solution of the point-mass lens equation $\theta - \theta_S = \theta_{\rm E}^2/\theta$. The intersection points between the hyperbola $\alpha(\theta)$ and the straight line $\theta - \theta_S$ identify the angular positions of the two images θ_+ and θ_-. (Right) Spatial configuration and parity of the two images. Image I_+ outside the Einstein ring preserves intrinsic parity ($\mu_+ > 0$), whereas image I_- inside the Einstein ring has inverted parity ($\mu_- < 0$), corresponding to a reflection across the tangential axis.

Remarkably, the signed algebraic sum satisfies:

$$\mu_+ + \mu_- = 1. \tag{8.1.32}$$

The signed magnifications always sum to unity: the extra flux gained by the positive-parity image exceeds the unlensed source flux by exactly the amount radiated into the negative-parity counter-image.

Inspection of Eq. (8.1.31) reveals that $\mu > 1$ for all finite impact parameters u. For a source situated on the Einstein ring ($u = 1$), the magnification evaluates to $\mu = 3/\sqrt{5} \simeq 1.34$, corresponding to a photometric brightening of ≈ 0.32 magnitudes. In the limit where the source approaches the optical axis ($u \to 0$), the point-source magnification diverges as $\mu \propto u^{-1}$. This formal divergence is an artifact of the point-source and geometric-optics approximations; for any physical source with finite angular radius θ_*, the peak magnification remains finite and scales as $\mu_{\rm max} \sim 2\theta_{\rm E}/\theta_*$.

Geometric optics perspective: magnification and parity.
In one-dimensional optics, lateral magnification is $m = {\rm d}y_{\rm image}/{\rm d}y_{\rm source}$. For a circularly symmetric two-dimensional lens, solid angle magnification factorizes into two orthogonal one-dimensional stretches:

$$\mu_t \equiv \frac{\theta}{\theta_S} \quad \text{(tangential stretch)}, \qquad \mu_r \equiv \frac{{\rm d}\theta}{{\rm d}\theta_S} = \left(1 + \frac{\theta_{\rm E}^2}{\theta^2}\right)^{-1} \quad \text{(radial stretch)}. \tag{1}$$

Their product directly yields the total point-mass magnification:

$$\mu = \mu_r \cdot \mu_t = \left(1 + \frac{\theta_{\rm E}^2}{\theta^2}\right)^{-1} \left(\frac{\theta}{\theta_S}\right) = \left[1 - \left(\frac{\theta_{\rm E}}{\theta}\right)^4\right]^{-1}. \tag{2}$$

Parity is simply the sign of the magnification, matching the upright ($m > 0$) versus inverted ($m < 0$) image condition in introductory optics:

- Outer image θ_+: Both θ_+ and θ_S lie on the same side ($\theta_+ > 0, \theta_S > 0$), so $\mu_t > 0$ and $\mu_r > 0 \implies \mu_+ > 0$ (upright image, positive parity).

- Inner image θ_-: The ray crosses to the opposite side of the optical axis ($\theta_- < 0$ while $\theta_S > 0$). Thus $\mu_t = \theta_-/\theta_S < 0$, giving:

$$\mu_- = \mu_r \cdot \mu_t < 0. \tag{3}$$

  The negative magnification indicates an inverted image, flipped along the tangential direction.

Deflection by an extended mass distribution

Linearized gravity and the validity of superposition Before generalizing the point-mass deflection law to extended, continuous astrophysical bodies, an essential theoretical principle must be emphasized. In full General Relativity, Einstein's field equations:

$$G_{\mu\nu} = \frac{8\pi G_{\mathrm{N}}}{c^4} T_{\mu\nu} \tag{8.1.33}$$

are fundamentally non-linear partial differential equations: the gravitational field itself carries energy and momentum, so the spacetime metric sourced by two masses is not simply the sum of their individual metrics. In general spacetimes, the principle of linear superposition strictly fails.

The reason we are mathematically and physically justified in treating an extended astrophysical deflector (such as a galaxy or galaxy cluster) as a linear superposition of infinitesimal point masses $\mathrm{d}M = \rho(\boldsymbol{r}')\mathrm{d}^3r'$ lies in the *weak-field, linearized limit* of general relativity. In astrophysical lensing configurations, two key physical conditions are satisfied to extraordinary precision:

1. *Weak gravitational potentials:* The Newtonian potential is everywhere minute compared to the square of the speed of light:

$$\frac{|\Phi|}{c^2} \sim 10^{-6} \quad \text{(for individual galaxies)}, \qquad \frac{|\Phi|}{c^2} \sim 10^{-4} \quad \text{(for rich galaxy clusters)}. \tag{8.1.34}$$

   Non-linear gravitational self-interaction corrections scale as $\mathcal{O}(|\Phi|^2/c^4) \lesssim 10^{-8}$ and are completely negligible for all observational purposes outside the immediate event horizons of black holes.

2. *Non-relativistic internal velocities:* The characteristic velocity dispersion σ_v of stars and galaxies within dark matter halos satisfies $v^2/c^2 \sim (\sigma_v/c)^2 \sim (10^2\text{–}10^3\,\mathrm{km/s})^2/c^2 \sim 10^{-7}\text{–}10^{-5} \ll 1$, ensuring that retardation effects and gravitomagnetic frame-dragging are highly suppressed.

Under these conditions, the spacetime metric is accurately described by the linearly perturbed conformal Newtonian metric (8.1.1). The field equations decouple and reduce to the linear 3D Newtonian Poisson equation:

$$\nabla^2\Phi(\boldsymbol{r}) = 4\pi G_{\mathrm{N}}\rho(\boldsymbol{r}). \tag{8.1.35}$$

Because the Poisson differential operator ∇^2 is strictly linear, the potential of an arbitrary continuous mass distribution $\rho(\boldsymbol{r})$ is the exact Green's function linear superposition of the potentials of its constituent point-mass elements:

$$\Phi(\boldsymbol{r}) = -G_{\mathrm{N}} \int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|}\, \mathrm{d}^3r' \,. \tag{8.1.36}$$

Furthermore, because astronomical deflection angles are minute ($\alpha \lesssim 1''$ for galaxies and $\lesssim 30''$ for clusters), photon trajectories deviate only infinitesimally from straight lines. Under this Born

approximation, the line-of-sight deflection integral (8.1.8):

$$\boldsymbol{\alpha}(\boldsymbol{b}) = \frac{2}{c^2}\int_{-\infty}^{\infty} \boldsymbol{\nabla}_{\perp}\Phi(\boldsymbol{b}, z)\, \mathrm{d}z \tag{8.1.37}$$

is a linear functional of the potential Φ. Substituting the superposition integral (8.1.36) into $\boldsymbol{\alpha}(\boldsymbol{b})$ directly validates the principle of linear superposition for gravitational light deflection: the total deflection vector is the vector sum of the independent deflections produced by every mass element in the deflector.

Continuous matter distribution and surface mass density While the point-mass idealization is mathematically elegant and appropriate for compact stellar objects or primordial black holes, real astrophysical deflectors—such as galaxies, groups, and massive galaxy clusters—are extended, continuous three-dimensional mass distributions. Consequently, the scalar deflection formalism must be generalized to accommodate extended mass distributions $\rho(\boldsymbol{r}) = \rho(\boldsymbol{b}, z)$, where $\boldsymbol{b} = (x, y)$ denotes the two-dimensional transverse coordinate vector and z is the Cartesian coordinate aligned along the unperturbed optical axis.

Because astronomical deflection angles are exceedingly small ($\alpha \lesssim 1''$ for individual galaxies and $\alpha \lesssim 30''$ for massive clusters), the path of the deflected photon deviates only marginally from a Euclidean straight line throughout the lensing region. Consequently, the deflection integral (8.1.8) can be accurately evaluated along the unperturbed ray path. This procedure constitutes the *Born approximation*, which is rigorously justified for all astrophysical lensing configurations outside the immediate event horizons of black holes.

Furthermore, the effective deflection region is confined to a longitudinal interval $\Delta z \sim \pm b$ around the lens, which is negligibly small compared to the gigaparsec cosmological distances $D_{\mathrm{OL}}, D_{\mathrm{OS}}, D_{\mathrm{LS}}$. We may therefore invoke the *thin-lens approximation*, projecting the entire three-dimensional mass density $\rho_{\mathrm{L}}(\boldsymbol{b}, z)$ into a two-dimensional *surface mass density* $\Sigma(\boldsymbol{b})$ localized entirely on the lens plane:

$$\Sigma(\boldsymbol{b}) \equiv \int_{-\infty}^{\infty} \rho_{\mathrm{L}}(\boldsymbol{b}, z)\, \mathrm{d}z\,. \tag{8.1.38}$$

For a point-mass deflector, this surface density reduces to $\Sigma(\boldsymbol{b}) = M\delta^{(2)}(\boldsymbol{b})$, or in angular coordinates, $\Sigma(\boldsymbol{\theta}) = (M/D_{\mathrm{OL}}^2)\delta^{(2)}(\boldsymbol{\theta})$.

Superposition and the 2D Green's function Under the thin-lens approximation, each infinitesimal surface element $\mathrm{d}M = \Sigma(\boldsymbol{b}')\mathrm{d}^2b'$ acts as an independent point-like deflector. By the principle of linear superposition in the weak-field regime, the total deflection vector $\boldsymbol{\alpha}(\boldsymbol{b})$ is obtained by integrating the differential deflections over the entire deflector plane:

$$\boldsymbol{\alpha}(\boldsymbol{b}) = \frac{4G_{\mathrm{N}}}{c^2}\int \Sigma(\boldsymbol{b}')\frac{\boldsymbol{b}-\boldsymbol{b}'}{|\boldsymbol{b}-\boldsymbol{b}'|^2}\, \mathrm{d}^2b'\,. \tag{8.1.39}$$

To relate this result directly to field theory, we decompose the three-dimensional Laplacian into its transverse and longitudinal components: $\nabla^2\Phi = \nabla_{\perp}^2\Phi + \partial^2\Phi/\partial z^2$. The Newtonian Poisson equation $\nabla^2\Phi = 4\pi G_{\mathrm{N}}\rho_{\mathrm{L}}$ can then be rewritten as:

$$\nabla_{\perp}^2\Phi = 4\pi G_{\mathrm{N}}\rho_{\mathrm{L}}(\boldsymbol{b}, z) - \frac{\partial^2\Phi}{\partial z^2}. \tag{8.1.40}$$

Taking the two-dimensional divergence of the deflection angle (8.1.8) and substituting Eq. (8.1.40):

$$\boldsymbol{\nabla}_{\perp}\cdot\boldsymbol{\alpha} = \frac{2}{c^2}\int_{-\infty}^{\infty}\left[4\pi G_{\mathrm{N}}\rho_{\mathrm{L}}(\boldsymbol{b}, z) - \frac{\partial^2\Phi}{\partial z^2}\right]\mathrm{d}z\,. \tag{8.1.41}$$

For an isolated gravitational deflector, the gravitational potential and its gradient vanish at longitudinal infinity ($\lim_{z\to\pm\infty}\partial\Phi/\partial z = 0$), so the boundary term $\int_{-\infty}^{\infty}(\partial^2\Phi/\partial z^2)\,\mathrm{d}z =$

$[\partial\Phi/\partial z]_{-\infty}^{+\infty} = 0$ vanishes identically. Invoking the definition of the surface mass density $\Sigma(\boldsymbol{b}) = \int_{-\infty}^{\infty} \rho_{\rm L}(\boldsymbol{b}, z)\, \mathrm{d}z$ yields the two-dimensional Poisson-type divergence equation:

$$\boldsymbol{\nabla}_{\perp} \cdot \boldsymbol{\alpha}(\boldsymbol{b}) = \frac{8\pi G_{\rm N}}{c^2} \Sigma(\boldsymbol{b}). \tag{8.1.42}$$

In two dimensions, the Green's function for the divergence equation $\boldsymbol{\nabla} \cdot \boldsymbol{G}(\boldsymbol{x}) = 2\pi\delta^{(2)}(\boldsymbol{x})$ is $\boldsymbol{G}(\boldsymbol{x}) = \boldsymbol{x}/|\boldsymbol{x}|^2$. Convolving the source term $\frac{8\pi G_{\rm N}}{c^2}\Sigma(\boldsymbol{b})$ with $\boldsymbol{G}/(2\pi)$ immediately confirms the superposition formula (8.1.39).

Angular coordinates and the critical surface mass density To transition to observable angular coordinates on the celestial sphere, we substitute $\boldsymbol{b} = D_{\rm OL}\boldsymbol{\theta}$ and $\boldsymbol{b}' = D_{\rm OL}\boldsymbol{\theta}'$. The vector lens equation becomes:

$$\boldsymbol{\theta} = \boldsymbol{\theta}_S + \frac{D_{\rm LS}}{D_{\rm OS}} \boldsymbol{\alpha}(D_{\rm OL}\boldsymbol{\theta}) \equiv \boldsymbol{\theta}_S + \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}), \tag{8.1.43}$$

where $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) \equiv (D_{\rm LS}/D_{\rm OS})\boldsymbol{\alpha}(D_{\rm OL}\boldsymbol{\theta})$ is the *reduced deflection angle*. Substituting Eq. (8.1.39) into $\hat{\boldsymbol{\alpha}}$:

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \frac{4G_{\rm N}}{c^2} \frac{D_{\rm OL} D_{\rm LS}}{D_{\rm OS}} \int \mathrm{d}^2\theta'\, \Sigma(D_{\rm OL}\boldsymbol{\theta}') \frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{|\boldsymbol{\theta} - \boldsymbol{\theta}'|^2}. \tag{8.1.44}$$

We now define the fundamental *critical surface mass density* $\Sigma_{\rm crit}$:

$$\Sigma_{\rm crit} \equiv \frac{c^2}{4\pi G_{\rm N}} \frac{D_{\rm OS}}{D_{\rm OL} D_{\rm LS}} \simeq 0.35 \left(\frac{\mathcal{D}}{1\,\mathrm{Gpc}}\right)^{-1} \mathrm{g \cdot cm^{-2}}, \tag{8.1.45}$$

where $\mathcal{D} \equiv D_{\rm OL} D_{\rm LS}/D_{\rm OS}$ represents the effective lensing distance. The reduced deflection angle then takes the canonical convolution form:

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \frac{1}{\pi} \int \mathrm{d}^2\theta' \, \frac{\Sigma(\boldsymbol{\theta}')}{\Sigma_{\rm crit}} \frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{|\boldsymbol{\theta} - \boldsymbol{\theta}'|^2} = \frac{1}{\pi} \int \mathrm{d}^2\theta' \, \kappa(\boldsymbol{\theta}') \frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{|\boldsymbol{\theta} - \boldsymbol{\theta}'|^2}, \tag{8.1.46}$$

where $\kappa(\boldsymbol{\theta}) \equiv \Sigma(\boldsymbol{\theta})/\Sigma_{\rm crit}$ is the dimensionless *convergence*.

A hypothetical deflector possessing a spatially uniform surface density equal to $\Sigma_{\rm crit}$ produces a scaled deflection angle $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, which leads to $\boldsymbol{\theta}_S = 0$ for all observed angles $\boldsymbol{\theta}$. Such a system acts as a perfect, stigmatic optical lens with a well-defined focal length. Conversely, realistic astrophysical mass distributions possess non-uniform surface densities that induce anisotropic deformations and image astigmatism. It is instructive to compare $\Sigma_{\rm crit}$ with typical surface mass densities encountered in galaxies ($\Sigma \sim M/R^2$):

$$\Sigma_{\rm gal} \simeq 0.3 \left(\frac{M}{10^{11} M_\odot}\right) \left(\frac{R_{\rm gal}}{10\,\mathrm{kpc}}\right)^{-2} \mathrm{g \cdot cm^{-2}}, \tag{8.1.47}$$

and in rich galaxy clusters:

$$\Sigma_{\rm cluster} \simeq 0.03 \left(\frac{M}{10^{15} M_\odot}\right) \left(\frac{R_{\rm cluster}}{3\,\mathrm{Mpc}}\right)^{-2} \mathrm{g \cdot cm^{-2}}. \tag{8.1.48}$$

Circularly symmetric extended mass distributions The simplest, most fundamental class of extended deflectors consists of circularly symmetric (axisymmetric) systems, where the projected surface mass density depends only on the angular radius from the lens center: $\Sigma(\boldsymbol{\theta}) = \Sigma(\theta)$, with $\theta \equiv |\boldsymbol{\theta}|$.

By rotational invariance around the optical axis, the deflection vector $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ must be strictly radial, pointing toward the center of the lens:

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \hat{\alpha}(\theta)\, \frac{\boldsymbol{\theta}}{\theta}. \tag{8.1.49}$$

To determine the scalar deflection angle $\hat{\alpha}(\theta)$, we express the two-dimensional Green's function integral (8.1.46) in polar coordinates (θ', ϕ') on the lens plane:

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \frac{1}{\pi} \int_0^\infty \theta' \, \mathrm{d}\theta' \, \frac{\Sigma(\theta')}{\Sigma_{\mathrm{crit}}} \int_0^{2\pi} \mathrm{d}\phi' \, \frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{|\boldsymbol{\theta} - \boldsymbol{\theta}'|^2}. \tag{8.1.50}$$

For any circularly symmetric surface mass density $\Sigma(\theta')$, the angular integral satisfies:

$$\boldsymbol{I}(\boldsymbol{\theta}, \theta') \equiv \int_0^{2\pi} \frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{|\boldsymbol{\theta} - \boldsymbol{\theta}'|^2} \, \mathrm{d}\phi' = \begin{cases} \dfrac{2\pi}{\theta^2} \, \boldsymbol{\theta}, & \theta' < \theta, \\ 0, & \theta' > \theta. \end{cases} \tag{1}$$

Proof. Align the angular coordinate system such that $\boldsymbol{\theta} = (\theta, 0)$ lies along the horizontal axis, so that $\boldsymbol{\theta}' = (\theta' \cos\phi', \theta' \sin\phi')$. The transverse component of the kernel is an odd function of ϕ', $\sin\phi'/(\theta^2 + \theta'^2 - 2\theta\theta' \cos\phi')$, and its integral over $[0, 2\pi]$ vanishes identically by symmetry. The radial component along $\hat{\boldsymbol{\theta}}$ is:

$$I_r(\theta, \theta') = \int_0^{2\pi} \frac{\theta - \theta' \cos\phi'}{\theta^2 + \theta'^2 - 2\theta\theta' \cos\phi'} \, \mathrm{d}\phi' \, . \tag{2}$$

We evaluate this integral algebraically by decomposing the numerator:

$$\frac{\theta - \theta' \cos\phi'}{\theta^2 + \theta'^2 - 2\theta\theta' \cos\phi'} = \frac{1}{2\theta} \left[1 + \frac{\theta^2 - \theta'^2}{\theta^2 + \theta'^2 - 2\theta\theta' \cos\phi'} \right]. \tag{3}$$

The second term in brackets is the standard Poisson kernel on the circle. Using Poisson's integral formula:

$$\int_0^{2\pi} \frac{\theta^2 - \theta'^2}{\theta^2 + \theta'^2 - 2\theta\theta' \cos\phi'} \, \mathrm{d}\phi' = \begin{cases} +2\pi, & \theta' < \theta, \\ -2\pi, & \theta' > \theta. \end{cases} \tag{4}$$

Substituting this result into the radial integral:

- For $\theta' < \theta$ (mass ring interior to the light ray):

$$I_r = \frac{1}{2\theta}[2\pi + 2\pi] = \frac{2\pi}{\theta}. \tag{5}$$

- For $\theta' > \theta$ (mass ring exterior to the light ray):

$$I_r = \frac{1}{2\theta}[2\pi - 2\pi] = 0. \tag{6}$$

Multiplying by the unit vector $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}/\theta$ completes the proof of Eq. (1).

The physical consequences of the 2D shell theorem are profound:

1. All matter residing outside the angular impact radius of the light path ($\theta' > \theta$) exerts *identically zero net deflection* on the photon.

2. All matter situated inside the circle ($\theta' < \theta$) produces a net gravitational deflection *identical to that of a point mass* equal to the total enclosed projected mass placed at the center of the lens!

Substituting Eq. (1) into Eq. (8.1.50), the scalar deflection angle simplifies to:

$$\hat{\alpha}(\theta) = \frac{1}{\pi}\int_0^\theta \theta' \, \mathrm{d}\theta' \, \frac{\Sigma(\theta')}{\Sigma_{\rm crit}} \left(\frac{2\pi}{\theta}\right) = \frac{2}{\theta}\int_0^\theta \theta' \, \kappa(\theta') \, \mathrm{d}\theta' \,. \tag{8.1.51}$$

We now define the total two-dimensional physical mass enclosed within the cylinder of physical radius $\xi = D_{\rm OL}\theta$:

$$M_{\rm 2D}(<\theta) \equiv 2\pi D_{\rm OL}^2 \int_0^\theta \theta' \, \Sigma(\theta') \, \mathrm{d}\theta' \,, \tag{8.1.52}$$

and the corresponding dimensionless enclosed mass $m(\theta)$:

$$m(\theta) \equiv \frac{M_{\rm 2D}(<\theta)}{\pi D_{\rm OL}^2 \Sigma_{\rm crit}} = 2\int_0^\theta \theta' \, \kappa(\theta') \, \mathrm{d}\theta' \,. \tag{8.1.53}$$

We also define the mean surface mass density $\bar{\Sigma}(<\theta)$ and mean convergence $\bar{\kappa}(\theta)$ enclosed within angular radius θ:

$$\bar{\kappa}(\theta) \equiv \frac{\bar{\Sigma}(<\theta)}{\Sigma_{\rm crit}} = \frac{M_{\rm 2D}(<\theta)}{\pi (D_{\rm OL}\theta)^2 \Sigma_{\rm crit}} = \frac{m(\theta)}{\theta^2} = \frac{2}{\theta^2}\int_0^\theta \theta' \, \kappa(\theta') \, \mathrm{d}\theta' \,. \tag{8.1.54}$$

Using these fundamental definitions, the deflection angle of any circularly symmetric extended lens takes the remarkably compact canonical forms:

$$\hat{\alpha}(\theta) = \frac{4G_{\rm N} M_{\rm 2D}(<\theta)}{c^2 D_{\rm OL}\theta} \frac{D_{\rm LS}}{D_{\rm OS}} = \frac{m(\theta)}{\theta} = \theta \, \bar{\kappa}(\theta). \tag{8.1.55}$$

Extended lens equation and Einstein radius Substituting the scalar deflection angle (8.1.55) into the scalar lens equation $\theta_S = \theta - \hat{\alpha}(\theta)$, we obtain the *lens equation for an extended circularly symmetric mass distribution*:

$$\theta - \theta_S = \frac{m(\theta)}{\theta} = \theta \, \bar{\kappa}(\theta). \tag{8.1.56}$$

Comparing Eq. (8.1.56) with the point-mass lens equation $\theta - \theta_S = \theta_{\rm E}^2/\theta$, we see that the constant point-mass parameter $\theta_{\rm E}^2$ is replaced by the scale-dependent enclosed mass function $m(\theta) = \theta^2 \bar{\kappa}(\theta)$.

When the source lies directly on the optical axis ($\theta_S = 0$), rotational symmetry dictates that the image forms a complete circular ring: the *Einstein ring*. Setting $\theta_S = 0$ in Eq. (8.1.56) yields the condition for the Einstein radius $\theta_{\rm E}$:

$$\theta_{\rm E} = \frac{m(\theta_{\rm E})}{\theta_{\rm E}} \implies \theta_{\rm E}^2 = m(\theta_{\rm E}) \iff \bar{\kappa}(\theta_{\rm E}) = 1. \tag{8.1.57}$$

Expressing this in terms of physical mass densities:

$$\bar{\Sigma}(<\theta_{\rm E}) = \Sigma_{\rm crit} \iff M_{\rm 2D}(<\theta_{\rm E}) = \pi (D_{\rm OL}\theta_{\rm E})^2 \Sigma_{\rm crit}. \tag{8.1.58}$$

Fundamental Theorem of Gravitational Lensing Rings. For any arbitrary circularly symmetric mass distribution, the angular radius of the Einstein ring $\theta_{\rm E}$ is precisely the radius inside which the **mean projected surface mass density equals the critical surface mass density**:

$$\bar{\Sigma}(<\theta_{\rm E}) = \Sigma_{\rm crit}. \tag{1}$$

Measuring the Einstein radius $\theta_{\rm E}$ of an astrophysical lens immediately determines the total enclosed projected mass $M_{\rm 2D}(<\theta_{\rm E})$, completely independent of the internal radial density profile $\rho(r)$!

Magnification and critical curves of extended lenses Because a circularly symmetric lens maps concentric circles centered on the optical axis into concentric circles, the two-dimensional transformation is diagonal in polar coordinates (θ, ϕ), with solid angle element $\mathrm{d}\Omega = \theta\,\mathrm{d}\theta\,\mathrm{d}\phi$. The observed flux magnification factor μ factorizes into radial and tangential components:

$$\mu = \frac{\mathrm{d}\Omega}{\mathrm{d}\Omega_S} = \frac{\theta}{\theta_S}\frac{\mathrm{d}\theta}{\mathrm{d}\theta_S} \equiv \mu_t\,\mu_r. \tag{8.1.59}$$

The two principal magnification factors are:

1. **Radial magnification μ_r:** Governs the stretching of light bundles along the radial direction connecting the image to the lens center:

$$\mu_r \equiv \left(\frac{\mathrm{d}\theta_S}{\mathrm{d}\theta}\right)^{-1} = \left(1 - \frac{\mathrm{d}\hat{\alpha}}{\mathrm{d}\theta}\right)^{-1}. \tag{8.1.60}$$

   Differentiating the deflection angle $\hat{\alpha}(\theta) = m(\theta)/\theta$:

$$\frac{\mathrm{d}\hat{\alpha}}{\mathrm{d}\theta} = \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{m(\theta)}{\theta}\right) = \frac{1}{\theta}\frac{\mathrm{d}m}{\mathrm{d}\theta} - \frac{m(\theta)}{\theta^2}. \tag{8.1.61}$$

   From the definition of $m(\theta)$ in Eq. (8.1.53), by Leibniz's rule $\mathrm{d}m\,/\,\mathrm{d}\theta = 2\theta\kappa(\theta)$. Therefore:

$$\frac{\mathrm{d}\hat{\alpha}}{\mathrm{d}\theta} = 2\kappa(\theta) - \bar{\kappa}(\theta). \tag{8.1.62}$$

   Substituting this derivative into the radial magnification:

$$\mu_r = \frac{1}{1 - [2\kappa(\theta) - \bar{\kappa}(\theta)]}. \tag{8.1.63}$$

2. **Tangential magnification μ_t:** Governs the stretching of light bundles perpendicular to the radial direction (along the azimuth):

$$\mu_t \equiv \frac{\theta}{\theta_S} = \left(1 - \frac{\hat{\alpha}(\theta)}{\theta}\right)^{-1} = \frac{1}{1 - \bar{\kappa}(\theta)}. \tag{8.1.64}$$

The asymmetry between radial and tangential stretching arises from the tidal gravitational shearing of light rays across the finite beam. For any circularly symmetric lens, defining the scalar tidal shear as the difference between the mean enclosed convergence and the local convergence:

$$\gamma(\theta) \equiv \bar{\kappa}(\theta) - \kappa(\theta), \tag{8.1.65}$$

The radial and tangential magnifications then take the symmetric canonical forms:

$$\mu_r = \frac{1}{1 - \kappa(\theta) - \gamma(\theta)}, \qquad \mu_t = \frac{1}{1 - \kappa(\theta) + \gamma(\theta)}. \tag{8.1.66}$$

The total flux magnification μ is the product of the radial and tangential eigenvalues:

$$\mu(\theta) = \mu_r\,\mu_t = \frac{1}{(1 - \bar{\kappa}(\theta))(1 - [2\kappa(\theta) - \bar{\kappa}(\theta)])} = \frac{1}{(1 - \kappa(\theta))^2 - \gamma^2(\theta)}. \tag{8.1.67}$$

The locus of points in the lens plane where the magnification formally diverges defines the **critical curves**. For circularly symmetric lenses, these decompose into two distinct geometric circles:

- **Tangential critical curve:** Defined by $\mu_t^{-1} = 0 \iff \bar{\kappa}(\theta_t) = 1$. As established in Eq. (8.1.57), this corresponds precisely to the **Einstein radius**: $\theta_t = \theta_{\mathrm{E}}$. Mapping this curve back to the source plane via the lens equation yields the *tangential caustic*:

$$\theta_{S,t} = \theta_{\mathrm{E}}(1 - \bar{\kappa}(\theta_{\mathrm{E}})) = 0. \tag{8.1.68}$$

  The tangential caustic collapses to a **single point** at the origin ($\theta_S = 0$). When an unlensed source crosses this caustic point, its image stretches into a complete, brilliant Einstein ring.

- **Radial critical curve:** Defined by $\mu_r^{-1} = 0 \iff 2\kappa(\theta_r) - \bar{\kappa}(\theta_r) = 1 \iff \kappa(\theta_r) + \gamma(\theta_r) = 1$. The corresponding caustic in the source plane is a circle of radius:

$$\theta_{S,r} = |\theta_r - \hat{\alpha}(\theta_r)| = \theta_r |1 - \bar{\kappa}(\theta_r)|. \tag{8.1.69}$$

  When a source crosses the radial caustic $\theta_{S,r}$, a pair of images merges and dissolves in the radial direction, giving rise to radially elongated arcs.

Image multiplicity and the odd-number theorem A fundamental difference between an idealized point mass and a realistic extended lens lies in the number and distribution of multiple images. For a point mass, the central surface density diverges ($\Sigma \propto \delta^{(2)}(\boldsymbol{\theta})$), yielding $m(\theta) = \theta_{\mathrm{E}}^2 = \text{const}$. As established in Eq. (8.1.24), this guarantees exactly two images for every source position.

Now consider a realistic, smooth extended lens with a finite, non-singular central core radius θ_c, such that the central convergence remains finite: $\kappa(0) \equiv \Sigma(0)/\Sigma_{\mathrm{crit}} < \infty$. As $\theta \to 0$, the enclosed mass scales as $m(\theta) \approx \kappa(0)\theta^2$, which yields a linear deflection angle:

$$\lim_{\theta \to 0} \hat{\alpha}(\theta) = \kappa(0)\, \theta. \tag{8.1.70}$$

Substituting this asymptotic behavior into the scalar lens equation $\theta_S(\theta) = \theta - \hat{\alpha}(\theta)$:

$$\left. \frac{\mathrm{d}\theta_S}{\mathrm{d}\theta} \right|_{\theta=0} = 1 - \kappa(0). \tag{8.1.71}$$

We now distinguish two fundamentally different physical regimes:

1. **Subcritical lens ($\kappa(0) < 1 \iff \Sigma(0) < \Sigma_{\mathrm{crit}}$):** The slope at the origin is positive: $\mathrm{d}\theta_S \,/\, \mathrm{d}\theta > 0$. If $\kappa(\theta)$ decreases monotonically, $\mathrm{d}\theta_S \,/\, \mathrm{d}\theta > 0$ everywhere. The function $\theta_S(\theta)$ is strictly monotonic, meaning that for each source position θ_S, there exists **exactly one image**. A subcritical lens cannot produce multiple images or strong lensing caustics.

2. **Supercritical lens ($\kappa(0) > 1 \iff \Sigma(0) > \Sigma_{\mathrm{crit}}$):** The slope at the origin is negative: $1 - \kappa(0) < 0$. However, at large angular radii ($\theta \gg \theta_c$), the enclosed mass $m(\theta)$ approaches a constant (or grows slower than θ^2), so $\hat{\alpha}(\theta) \to 0$ or approaches a constant, ensuring that $\theta_S(\theta) \to \theta$ with positive slope $+1$. Because $\theta_S(\theta)$ starts at $\theta_S(0) = 0$ with negative slope and eventually rises with positive slope, the function $\theta_S(\theta)$ must possess a local minimum at radius θ_r (the radial critical curve, where $\mathrm{d}\theta_S \,/\, \mathrm{d}\theta = 0$).

Consequently, for any source situated inside the radial caustic ($|\theta_S| < \theta_{S,r}$), the horizontal line $\theta_S = \text{const}$ intersects the curve $\theta_S(\theta)$ at **three distinct points**:

- Image $\theta_1 > \theta_{\mathrm{E}}$: Outer image situated outside the Einstein ring, preserving intrinsic parity ($\mu_1 > 0$).

- Image θ_2 ($\theta_r < |\theta_2| < \theta_{\mathrm{E}}$): Intermediate image situated inside the Einstein ring, with inverted parity ($\mu_2 < 0$).

- Image θ_3 ($|\theta_3| < \theta_r$): Interior image situated in the deep central core of the deflector, preserving intrinsic parity ($\mu_3 > 0$).

This illustrates a general topological theorem of gravitational optics:

> *The Burke–Blandford Odd-Number Theorem (1981).* Every non-singular, transparent, and bounded gravitational deflector produces an **odd number** of images ($N = 1, 3, 5, \dots$) for all source positions not lying precisely on a caustic.

The puzzle of the missing third image If non-singular lenses universally produce three images, why are observed strong lensing systems almost exclusively detected as **doubles** or **quadruples**? The answer lies in the magnification of the central image θ_3. In the deep core ($\theta_3 \ll \theta_c$), we have $\bar{\kappa}(\theta_3) \approx \kappa(0)$ and $2\kappa(\theta_3) - \bar{\kappa}(\theta_3) \approx \kappa(0)$. Substituting these into Eq. (8.1.67):

$$\mu_3 \approx \frac{1}{(1-\kappa(0))^2}. \tag{8.1.72}$$

For massive elliptical galaxies and galaxy clusters that act as strong lenses, the central surface mass density is exceedingly supercritical: $\kappa(0) \sim 10 - 100$. The magnification of the central image evaluates to:

$$\mu_3 \sim \frac{1}{\kappa(0)^2} \sim 10^{-2} - 10^{-4} \ll 1. \tag{8.1.73}$$

The central third image is *heavily demagnified* and suppressed by factors of hundreds to thousands relative to the outer images! Furthermore, this faint image is situated directly behind the luminous stellar bulge and supermassive black hole of the lens galaxy, rendering its observational detection virtually impossible in optical bands (though high-dynamic-range VLBI radio observations have succeeded in detecting faint core images in a handful of systems, placing tight limits on central core densities).

Core models: The non-singular isothermal sphere (NIS) A realistic analytic model of an extended galaxy halo with a smooth core is the *non-singular isothermal sphere* (NIS, or pseudo-isothermal sphere), described by the three-dimensional density:

$$\rho(r) = \frac{\rho_0}{1 + (r/r_c)^2}, \tag{8.1.74}$$

where r_c denotes the physical core radius. Integrating along the line of sight z:

$$\Sigma(\xi) = \int_{-\infty}^{\infty} \frac{\rho_0 \, \mathrm{d}z}{1 + (\xi^2 + z^2)/r_c^2} = \frac{\pi \rho_0 r_c}{\sqrt{1 + (\xi/r_c)^2}} \equiv \frac{\Sigma_0}{\sqrt{1 + (\theta/\theta_c)^2}}, \tag{8.1.75}$$

where $\theta_c \equiv r_c / D_{\mathrm{OL}}$ is the angular core radius and $\Sigma_0 \equiv \pi \rho_0 r_c$ is the central surface mass density. The convergence profile is:

$$\kappa(\theta) = \frac{\kappa_0}{\sqrt{1 + (\theta/\theta_c)^2}}, \qquad \kappa_0 \equiv \frac{\Sigma_0}{\Sigma_{\mathrm{crit}}}. \tag{8.1.76}$$

The enclosed two-dimensional mass evaluates analytically:

$$M_{\mathrm{2D}}(<\theta) = 2\pi D_{\mathrm{OL}}^2 \Sigma_0 \int_0^{\theta} \frac{\theta' \, \mathrm{d}\theta'}{\sqrt{1 + (\theta'/\theta_c)^2}} = 2\pi D_{\mathrm{OL}}^2 \Sigma_0 \theta_c^2 \left(\sqrt{1 + \frac{\theta^2}{\theta_c^2}} - 1 \right). \tag{8.1.77}$$

The dimensionless enclosed mass is $m(\theta) = 2\kappa_0 \theta_c^2 (\sqrt{1 + \theta^2/\theta_c^2} - 1)$, yielding the scalar deflection angle:

$$\hat{\alpha}(\theta) = \frac{m(\theta)}{\theta} = \frac{2\kappa_0 \theta_c^2}{\theta} \left(\sqrt{1 + \frac{\theta^2}{\theta_c^2}} - 1 \right). \tag{8.1.78}$$

Examining the asymptotic behavior of the NIS deflection angle (8.1.78):

- **Outer regime ($\theta \gg \theta_c$):** Far outside the core, $\sqrt{1+\theta^2/\theta_c^2} \approx \theta/\theta_c \gg 1$. The deflection angle asymptotes to a constant:

$$\lim_{\theta \gg \theta_c} \hat{\alpha}(\theta) = 2\kappa_0 \theta_c \equiv \theta_{\mathrm{E}}. \tag{8.1.79}$$

  This recovers the constant deflection $\hat{\alpha} = \theta_{\mathrm{E}}$ characteristic of the singular isothermal sphere (SIS), whose dynamical derivation from self-gravitating hydrostatic equilibrium and velocity dispersion σ_v is established in Section 8.1.1.

- **Inner core regime ($\theta \ll \theta_c$):** Expanding the square root in a Taylor series: $\sqrt{1+\theta^2/\theta_c^2} - 1 \approx \frac{\theta^2}{2\theta_c^2} - \frac{\theta^4}{8\theta_c^4}$. The deflection angle scales linearly:

$$\lim_{\theta \ll \theta_c} \hat{\alpha}(\theta) = \kappa_0 \theta. \tag{8.1.80}$$

  This linear deflection produces the demagnified central third image characteristic of non-singular extended mass distributions.

General case: Projected potential, shear, and magnification tensor

Projected potential Because the deflection field $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ is curl-free ($\boldsymbol{\nabla}_2 \times \hat{\boldsymbol{\alpha}} = 0$), it can be represented as the angular gradient of a scalar potential. We define the effective two-dimensional *projected lensing potential* $\psi(\boldsymbol{\theta})$ by integrating the three-dimensional Newtonian potential along the line of sight:

$$\psi(\boldsymbol{\theta}) \equiv \frac{D_{\mathrm{LS}}}{D_{\mathrm{OL}} D_{\mathrm{OS}}} \frac{2}{c^2} \int_{-\infty}^{\infty} \Phi(D_{\mathrm{OL}}\boldsymbol{\theta}, z)\, \mathrm{d}z\,. \tag{8.1.81}$$

Recalling the vector identity $\boldsymbol{\nabla}_2 \ln\theta = \boldsymbol{\theta}/\theta^2$, the projected potential is directly related to the deflection angle through:

$$\boldsymbol{\nabla}_2 \psi(\boldsymbol{\theta}) = \frac{D_{\mathrm{LS}}}{D_{\mathrm{OS}}} \boldsymbol{\alpha}(\boldsymbol{\theta}) = \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}). \tag{8.1.82}$$

Taking the two-dimensional divergence of Eq. (8.1.82) and invoking Eq. (8.1.42) together with the identity $\nabla^2 \ln\theta = 2\pi\delta^{(2)}(\boldsymbol{\theta})$, we obtain the two-dimensional Poisson equation governing the lensing potential:

$$\frac{1}{2}\nabla^2 \psi(\boldsymbol{\theta}) = \frac{\Sigma(\boldsymbol{\theta})}{\Sigma_{\mathrm{crit}}} \equiv \kappa(\boldsymbol{\theta}). \tag{8.1.83}$$

Proof. Taking the angular Laplacian $\nabla^2 \equiv \boldsymbol{\nabla}_2 \cdot \boldsymbol{\nabla}_2$ of Eq. (8.1.82) with $\boldsymbol{\nabla}_2 = D_{\mathrm{OL}} \boldsymbol{\nabla}_\perp$:

$$\nabla^2 \psi(\boldsymbol{\theta}) = \boldsymbol{\nabla}_2 \cdot \left[\frac{D_{\mathrm{LS}}}{D_{\mathrm{OS}}} \boldsymbol{\alpha}(\boldsymbol{\theta})\right] = D_{\mathrm{OL}} \frac{D_{\mathrm{LS}}}{D_{\mathrm{OS}}} \boldsymbol{\nabla}_\perp \cdot \boldsymbol{\alpha}. \tag{1}$$

Substituting the divergence from Eq. (8.1.42), $\boldsymbol{\nabla}_\perp \cdot \boldsymbol{\alpha} = \frac{8\pi G_{\mathrm{N}}}{c^2}\Sigma(\boldsymbol{\theta})$:

$$\nabla^2 \psi(\boldsymbol{\theta}) = D_{\mathrm{OL}} \frac{D_{\mathrm{LS}}}{D_{\mathrm{OS}}} \frac{8\pi G_{\mathrm{N}}}{c^2} \Sigma(\boldsymbol{\theta}) = 2\left[\frac{4\pi G_{\mathrm{N}}}{c^2} \frac{D_{\mathrm{OL}} D_{\mathrm{LS}}}{D_{\mathrm{OS}}}\right] \Sigma(\boldsymbol{\theta}). \tag{2}$$

Comparing the bracketed factor with the definition of Σ_{crit} in Eq. (8.1.45):

$$\frac{4\pi G_{\mathrm{N}}}{c^2} \frac{D_{\mathrm{OL}} D_{\mathrm{LS}}}{D_{\mathrm{OS}}} = \frac{1}{\Sigma_{\mathrm{crit}}} \quad \Longrightarrow \quad \frac{1}{2}\nabla^2 \psi(\boldsymbol{\theta}) = \frac{\Sigma(\boldsymbol{\theta})}{\Sigma_{\mathrm{crit}}} = \kappa(\boldsymbol{\theta}). \tag{3}$$

This completes the derivation of Eq. (8.1.83).

Integrating the two-dimensional Poisson equation (8.1.83) using the logarithmic Green's function in the plane gives the explicit potential:

$$\psi(\boldsymbol{\theta}) = \frac{1}{\pi} \int \mathrm{d}^2\theta'\, \kappa(\boldsymbol{\theta}') \ln|\boldsymbol{\theta} - \boldsymbol{\theta}'|. \tag{8.1.84}$$

Magnification and distortion Consider an extended light source observed through the deflector. In accordance with surface brightness conservation [Eq. (8.1.27)], the image intensity distribution $I(\boldsymbol{\theta})$ satisfies:

$$I(\boldsymbol{\theta}) = I_S[\boldsymbol{\theta}_S(\boldsymbol{\theta})] = I_S[\boldsymbol{\theta}_S^0 + \mathcal{A}(\boldsymbol{\theta}_0) \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0)], \tag{8.1.85}$$

where we have Taylor-expanded the coordinate mapping $\boldsymbol{\theta}_S(\boldsymbol{\theta})$ to first order around a reference image center $\boldsymbol{\theta}_0$, with $\boldsymbol{\theta}_S^0 = \boldsymbol{\theta}_0 - \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}_0)$.

The distortion matrix $\mathcal{A}(\boldsymbol{\theta})$ (also termed the inverse magnification matrix) is the Jacobian matrix of the lensing coordinate transformation $\boldsymbol{\theta} \mapsto \boldsymbol{\theta}_S(\boldsymbol{\theta})$; physically, it describes the local deformation of an infinitesimal patch of the sky under differential gravitational deflection:

$$\mathrm{d}\boldsymbol{\theta}_S = \mathcal{A}(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\,, \qquad \mathcal{A}(\boldsymbol{\theta}) \equiv \frac{\partial \boldsymbol{\theta}_S}{\partial \boldsymbol{\theta}}. \tag{8.1.86}$$

Differentiating the vector lens equation $\boldsymbol{\theta}_S = \boldsymbol{\theta} - \boldsymbol{\nabla}\psi(\boldsymbol{\theta})$ yields the matrix components:

$$\mathcal{A}_{ab}(\boldsymbol{\theta}) = \frac{\partial \theta_S^a}{\partial \theta^b} = \delta_{ab} - \frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \theta^a \partial \theta^b} \equiv \delta_{ab} - \psi_{,ab}. \tag{8.1.87}$$

Equivalently, the lens equation is the stationarity condition $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \tau_{\mathrm{Fermat}} = \mathbf{0}$ of the dimensionless Fermat potential:

$$\tau_{\mathrm{Fermat}}(\boldsymbol{\theta}, \boldsymbol{\theta}_S) \equiv \frac{1}{2}|\boldsymbol{\theta} - \boldsymbol{\theta}_S|^2 - \psi(\boldsymbol{\theta}), \tag{8.1.88}$$

which represents the arrival time surface of light rays [Eq. (8.1.108)]. In this variational formulation, the distortion matrix $\mathcal{A}_{ab}$ is recognized as the Hessian matrix of the arrival time surface:

$$\mathcal{A}_{ab}(\boldsymbol{\theta}) = \frac{\partial^2 \tau_{\mathrm{Fermat}}}{\partial \theta^a \partial \theta^b} = \delta_{ab} - \psi_{,ab}. \tag{8.1.89}$$

Why are the physical lensing parameters defined as convergence κ and shear components (γ_1, γ_2)? Any real 2×2 matrix has four real degrees of freedom, spanned by the Pauli spin matrices $\{\sigma_0, \sigma_1, \sigma_2, \sigma_3\}$. Because partial derivatives commute ($\psi_{,ab} = \psi_{,ba}$), the distortion matrix $\mathcal{A}$ is strictly symmetric. The antisymmetric basis element $\varepsilon = i\sigma_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ corresponds to net image rotation (vorticity), which vanishes identically because gravitational deflection is the gradient of a single scalar potential ($\boldsymbol{\nabla} \times \hat{\boldsymbol{\alpha}} = \mathbf{0}$).

The three remaining symmetric basis matrices $(\sigma_0, \sigma_3, \sigma_1)$ define the irreducible decomposition of any gravitational distortion:

$$\mathcal{A} = \begin{pmatrix} 1 - \kappa - \gamma_1 & -\gamma_2 \\ -\gamma_2 & 1 - \kappa + \gamma_1 \end{pmatrix} = (1 - \kappa)\sigma_0 - \gamma_1 \sigma_3 - \gamma_2 \sigma_1, \tag{8.1.90}$$

where

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{8.1.91}$$

Expanding $\mathcal{A}$ in this basis immediately reveals the physical meaning of each coefficient:

1. $(1 - \kappa)$ is the projection onto the identity σ_0: the *convergence* $\kappa \equiv \frac{1}{2}\nabla^2 \psi = \Sigma / \Sigma_{\mathrm{crit}}$ provides isotropic focusing that rescales the image area by $(1 - \kappa)^{-2}$ while preserving circular shapes.

2. $-\gamma_1$ is the projection onto σ_3: the quadrupole tidal stretch along the Cartesian coordinate axes, $\gamma_1 = \frac{1}{2}(\psi_{,11} - \psi_{,22}) = \gamma \cos(2\phi)$.

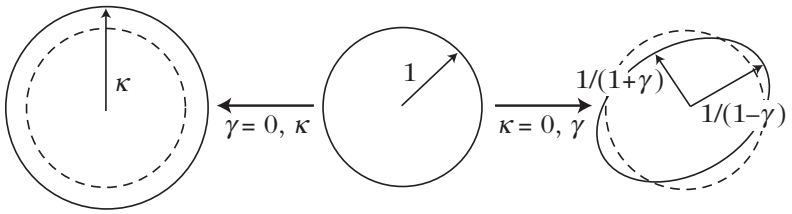


Figure 8.7: Distortion of an intrinsically circular source of unit radius under the separate actions of convergence κ and shear γ. Pure convergence ($\gamma = 0$) rescales the source isotropically into a magnified circle of radius $(1-\kappa)^{-1}$. Pure shear ($\kappa = 0$) distorts the circular source into an ellipse with semi-minor axis $(1+\gamma)^{-1}$ and semi-major axis $(1-\gamma)^{-1}$, conserving the orientation angle ϕ.

3. $-\gamma_2$ is the projection onto σ_1: the quadrupole tidal stretch along the 45° diagonal axes, $\gamma_2 = \psi_{,12} = \gamma \sin(2\phi)$.

Here $\gamma = \sqrt{\gamma_1^2 + \gamma_2^2}$ is the shear magnitude and ϕ is the orientation angle of the tidal ellipse.

As illustrated in Fig. 8.7, an intrinsically circular source of unit radius is mapped by $\mathcal{A}$ into an ellipse whose semi-major and semi-minor axes are oriented along the eigenvectors of $\mathcal{A}$, with lengths $(1-\kappa-\gamma)^{-1}$ and $(1-\kappa+\gamma)^{-1}$, respectively. The physical surface area of the image S is related to that of the source S_0 by the magnification factor μ:

$$S = \mu S_0, \qquad \mu = \frac{1}{\det \mathcal{A}} = \frac{1}{(1-\kappa)^2 - \gamma^2}. \tag{8.1.92}$$

Because differential deflection alters the solid angle subtended by the image while preserving surface brightness, the total received flux is magnified by μ. The distortion matrix $\mathcal{A}$ is the inverse of the amplification matrix $\mathcal{M} \equiv \mathcal{A}^{-1}$, giving $\mu = \det \mathcal{M}$.

Factoring out the isotropic magnification $(1-\kappa)$, the distortion matrix can be written in terms of the *reduced shear* $\boldsymbol{g} \equiv \boldsymbol{\gamma}/(1-\kappa)$:

$$\mathcal{A} = (1-\kappa) \begin{pmatrix} 1-g_1 & -g_2 \\ -g_2 & 1+g_1 \end{pmatrix}. \tag{8.1.93}$$

Because $(1-\kappa)$ rescales image size without altering image shape, observational measurements of galaxy ellipticities probe the reduced shear $\boldsymbol{g}$ directly, rather than $\boldsymbol{\gamma}$ or κ in isolation.

The Mass-Sheet Degeneracy To understand the profound observational crisis posed by the *Mass-Sheet Degeneracy* (Falco, Gorenstein, & Shapiro 1985; Schneider & Seitz 1995), consider what happens in elementary geometric optics when we observe a moving vehicle through a lens. In the thin-lens formula:

$$\frac{1}{f} = \frac{1}{v} + \frac{1}{u}, \qquad m \equiv \frac{y_{\rm image}}{y_{\rm source}} = \frac{v}{u}, \tag{8.1.94}$$

the apparent transverse velocity of an image is obtained directly by differentiating $y_{\rm image} = m y_{\rm source}$ with respect to time:

$$v_{\rm image,\perp} = m\, v_{\rm source,\perp}, \tag{8.1.95}$$

while differentiating the lens formula $\frac{1}{f} = \frac{1}{v} + \frac{1}{u}$ along the optical axis yields the longitudinal image velocity:

$$-\frac{1}{v^2}\frac{\mathrm{d}v}{\mathrm{d}t} - \frac{1}{u^2}\frac{\mathrm{d}u}{\mathrm{d}t} = 0 \implies \frac{\mathrm{d}v}{\mathrm{d}t} = -\left(\frac{v}{u}\right)^2 \frac{\mathrm{d}u}{\mathrm{d}t} = -m^2\, v_{\rm source,\parallel}. \tag{8.1.96}$$

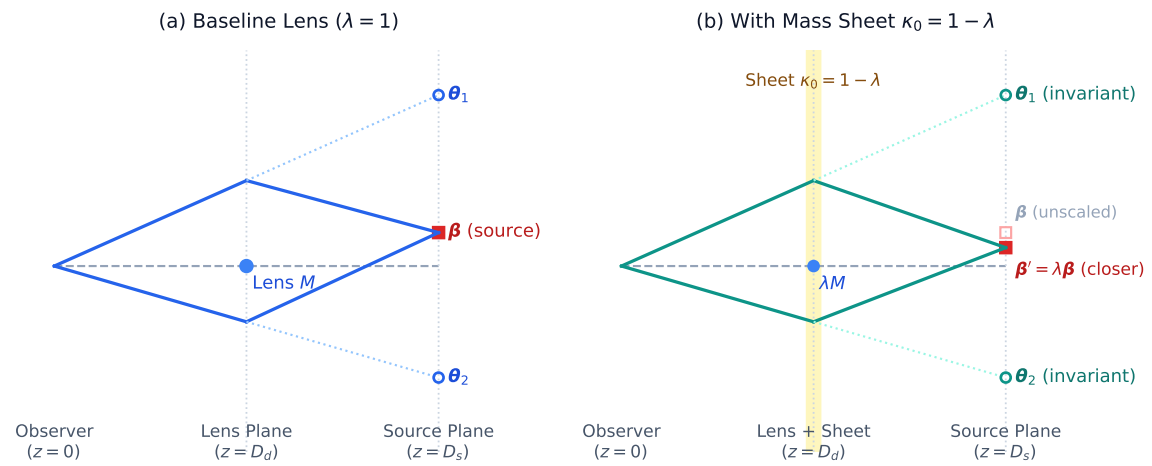


Figure 8.8: The geometric mechanism of the Mass-Sheet Degeneracy. (a) Baseline gravitational lens ($\lambda = 1$). A distant source at true celestial coordinate $\boldsymbol{\beta}$ emits light rays deflected by lens mass M, arriving at the observer at image angles $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. (b) Lens with an added uniform mass sheet $\kappa_0 = 1 - \lambda$ (shaded yellow band) and rescaled lens mass λM. If the true unlensed source is scaled inward to $\boldsymbol{\beta}' = \lambda\boldsymbol{\beta}$, the linear outward deflection $(1 - \lambda)\boldsymbol{\theta}$ of the mass sheet perfectly cancels the shift. The rays arrive at the observer at the exact same image angles $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, rendering the mass sheet completely invisible to image astrometry.

In both directions, the apparent velocity seen on a screen is not the true physical speed of the vehicle: it is directly scaled by the optical magnification m (or m^2). If an observer measures the image velocity on a detector without knowing the true magnification of the lens, their inferred velocity of the vehicle is directly biased by m.

In cosmology, the expansion rate of the universe (the Hubble constant H_0) is the cosmic speed we seek to measure, and the time delay Δt between multiply-imaged quasars is our clock. Because cosmological distances scale inversely with the expansion rate ($D_{\Delta t} \propto c/H_0$), the time delay directly measures physical distance:

$$\Delta t = \frac{1 + z_d}{c} D_{\Delta t} \Delta\tau(\boldsymbol{\theta}) \propto \frac{1}{H_0} \Delta\tau. \tag{8.1.97}$$

Now ask the question in words: *what happens if an unseen, perfectly uniform sheet of dark matter lies across the lens plane?*

A uniform mass sheet has constant surface mass density $\Sigma_0 = (1 - \lambda)\Sigma_{\rm crit}$, contributing a constant convergence $\kappa_0 = 1 - \lambda$. By 2D Gauss's law, a uniform sheet of mass produces a gravitational deflection that pushes light radially outward in direct proportion to radius:

$$\Delta\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = (1 - \lambda)\boldsymbol{\theta}. \tag{8.1.98}$$

It acts as a giant optical zoom lens that expands every ray outward by $(1 - \lambda)\boldsymbol{\theta}$.

As illustrated in Fig. 8.8, suppose nature scales the deflector's mass by λ while adding the uniform mass sheet:

$$\kappa_\lambda(\boldsymbol{\theta}) = \lambda\kappa(\boldsymbol{\theta}) + (1 - \lambda), \qquad \psi_\lambda(\boldsymbol{\theta}) = \lambda\psi(\boldsymbol{\theta}) + \frac{1 - \lambda}{2}|\boldsymbol{\theta}|^2. \tag{8.1.99}$$

The total deflection angle becomes:

$$\hat{\boldsymbol{\alpha}}_\lambda(\boldsymbol{\theta}) = \boldsymbol{\nabla}\psi_\lambda(\boldsymbol{\theta}) = \lambda\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) + (1 - \lambda)\boldsymbol{\theta}. \tag{8.1.100}$$

Now consider what happens to the observed image positions $\boldsymbol{\theta}$. If the true, unobservable source position is scaled inward by λ ($\boldsymbol{\theta}'_S = \lambda\boldsymbol{\theta}_S$), the lens equation becomes:

$$\boldsymbol{\theta} - \hat{\boldsymbol{\alpha}}_\lambda(\boldsymbol{\theta}) = \boldsymbol{\theta} - [\lambda\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) + (1 - \lambda)\boldsymbol{\theta}] = \lambda[\boldsymbol{\theta} - \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})] = \lambda\boldsymbol{\theta}_S = \boldsymbol{\theta}'_S. \tag{8.1.101}$$

The observed image positions $\boldsymbol{\theta}$ do not shift by a single arcsecond! The linear outward deflection produced by the mass sheet perfectly balances the inward displacement of the source.

Nor is this conspiracy confined to image positions: every other classical optical observable on the sky is equally blind to the mass sheet. Consider the distorted shapes of lensed background galaxies. Because the lens mass scales as λ, the shear tensor scales in direct proportion: $\boldsymbol{\gamma}_\lambda = \lambda\boldsymbol{\gamma}$. Yet the isotropic denominator of the distortion matrix scales identically: $1 - \kappa_\lambda = 1 - [\lambda\kappa + (1 - \lambda)] = \lambda(1 - \kappa)$. The observable reduced shear is therefore strictly invariant:

$$\boldsymbol{g}_\lambda \equiv \frac{\boldsymbol{\gamma}_\lambda}{1 - \kappa_\lambda} = \frac{\lambda\boldsymbol{\gamma}}{\lambda(1 - \kappa)} = \boldsymbol{g}. \tag{8.1.102}$$

Every background galaxy retains the exact same observed ellipticity. Similarly, the local magnification $\mu = 1/\det\mathcal{A}$ scales across the entire lens plane as $\mu_\lambda = \lambda^{-2}\mu$. Because all images are amplified by the identical global factor λ^{-2}, every relative flux ratio μ_A/μ_B is preserved without distortion.

An observer equipped with the sharpest telescope measuring image positions, image shapes, and relative flux ratios is entirely powerless to detect the mass sheet. A single lens configuration admits a continuous, one-parameter family of physically distinct models parametrized by $\lambda \in (0, 1]$, each matching every piece of imaging data with equal perfection.

Here the optical analogy of the moving vehicle reveals its full cosmological consequence. Recall that differentiating the lens formula showed that the apparent speed of a vehicle viewed through a lens is scaled by the optical magnification. In cosmological time-delay cosmography, the cosmic expansion rate H_0 is the velocity we wish to infer from the arrival-time differences between multiple images. Substituting the transformed deflection potential into the Fermat potential [Eq. (8.1.88)] gives:

$$\tau_\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}'_S) = \frac{1}{2}|\boldsymbol{\theta} - \lambda\boldsymbol{\theta}_S|^2 - \left[\lambda\psi(\boldsymbol{\theta}) + \frac{1 - \lambda}{2}|\boldsymbol{\theta}|^2\right] = \lambda\tau(\boldsymbol{\theta}, \boldsymbol{\theta}_S) + \frac{1}{2}(\lambda^2 - \lambda)|\boldsymbol{\theta}_S|^2. \tag{8.1.103}$$

Because the source-position term is independent of $\boldsymbol{\theta}$, it drops out when taking differences between distinct image positions $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$. The differential time delay between multiple images therefore scales in direct proportion to λ:

$$\Delta t_\lambda = \lambda\Delta t. \tag{8.1.104}$$

Because the physical time delay Δt is fixed by real clocks, an incorrect choice of λ directly biases the inferred Hubble constant:

$$H_0^{\text{inferred}} = \lambda H_0^{\text{true}}. \tag{8.1.105}$$

Just as measuring the speed of a car through an uncalibrated zoom lens yields an image velocity biased by the magnification, an unrecognized 20% mass sheet ($\lambda = 0.8$) suppresses the inferred expansion rate from $H_0 = 73\,\text{km/s/Mpc}$ down to $58\,\text{km/s/Mpc}$ without leaving a single trace in the image astrometry.

How can cosmology escape this blind spot? Because pure ray deflection is blind to the mass sheet, breaking the degeneracy requires external physical anchors that probe the gravitational field independently of optics. Two anchors are paramount:

1. *Stellar kinematics of the lens galaxy.* While ray optics measures only the dimensionless combination $\kappa \propto M/D$, the stellar velocity dispersion σ_v of the deflector galaxy probes the true depth of its 3D gravitational potential via the virial theorem ($\sigma_v^2 \sim GM/R_{\text{eff}}$). Because the true lens galaxy mass scales as $M_\lambda = \lambda M$, the velocity dispersion scales as $\sigma_{v,\lambda}^2 = \lambda\sigma_v^2$. Measuring σ_v spectroscopically provides an absolute mass scale, pinning down λ and restoring the unbiased measurement of H_0.

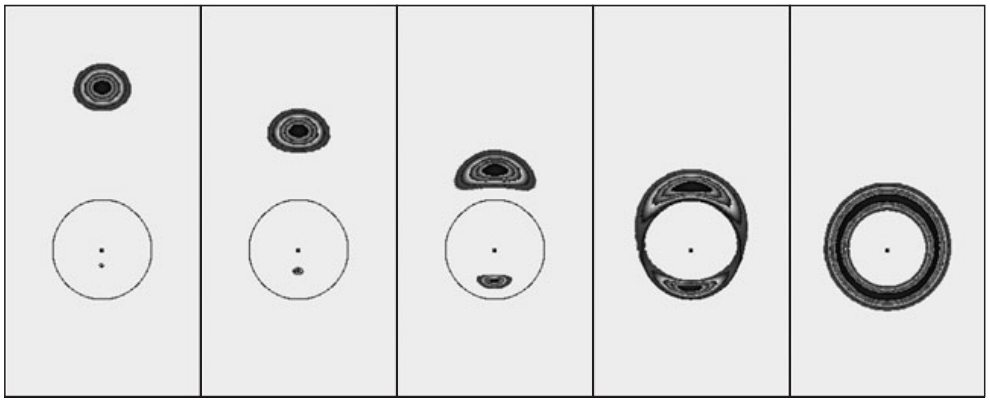


Figure 8.9: Progressive distortion of an extended, circularly symmetric source passing behind a point-like deflector, demonstrating the transition between weak and strong lensing. The dashed circle depicts the Einstein radius $\theta_{\rm E}$. At large impact parameters ($u \gg 1$), the source suffers slight elliptical elongation. As the source approaches the Einstein ring ($u \sim 1$), it stretches into pronounced curved arclets and forms a faint interior counter-image. Upon crossing inside the Einstein circle, the two images merge into a luminous, closed Einstein ring when collinearity is attained ($u = 0$). From Ref. [3].

2. *Standard candles.* If the multiply-imaged background source is a Type Ia supernova whose unlensed peak luminosity is known from standard candle calibration, the observed apparent brightness of any image determines the absolute magnification $\mu_\lambda = \lambda^{-2}\mu$, fixing λ directly without relying on mass models.

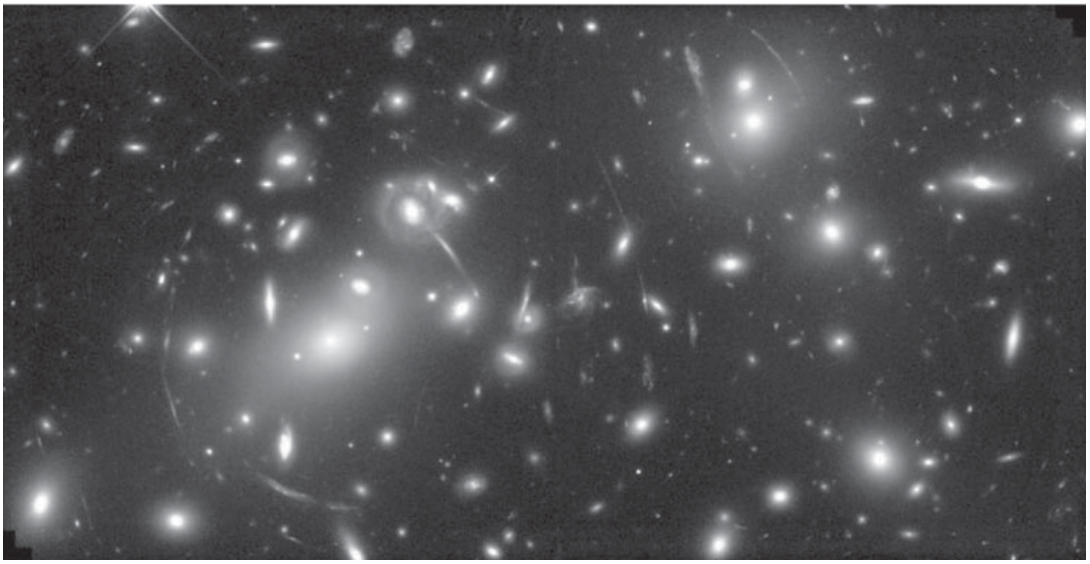
Different lensing regimes The qualitative morphological behavior of a gravitational lens is primarily dictated by the dimensionless ratio $\Sigma/\Sigma_{\rm crit} = \kappa$:

- **Supercritical / Strong lensing regime** ($\Sigma/\Sigma_{\rm crit} > 1$): The lens central density exceeds the critical threshold. Convergence and shear are of order unity ($\kappa \sim 1, \gamma \sim 1$). The mapping $\boldsymbol{\theta}_S \mapsto \boldsymbol{\theta}$ becomes non-bijective, producing multiple discrete images, giant arcs, and Einstein rings.
- **Subcritical / Weak lensing regime** ($\Sigma/\Sigma_{\rm crit} < 1$): Across the entire deflector, $\kappa \ll 1$ and $\gamma \ll 1$. Multiple images cannot form; background sources experience subtle, coherent ellipticities of order 1% that cannot be recognized individually by eye and must be measured statistically across vast ensembles of background galaxies.
- **Arclet / Intermediate regime**: Occurs in peripheral regions of supercritical clusters where $\kappa, \gamma \lesssim 1$. Background sources are visibly stretched into tangential arclets without forming closed Einstein rings or multiple counter-images.

The locus of points in the lens plane where the Jacobian determinant vanishes:

$$\det \mathcal{A}(\boldsymbol{\theta}) = (1-\kappa)^2 - \gamma^2 = 0, \tag{8.1.106}$$

defines smooth closed curves known as *critical curves*. Points on critical curves exhibit formally divergent magnification $\mu \to \infty$. The mapping of these critical curves onto the source plane via the lens equation $\boldsymbol{\beta}(\boldsymbol{\theta}) = \boldsymbol{\theta} - \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ forms closed curves containing singular cusps, termed *caustics*. Caustics demarcate boundaries across which the image multiplicity jumps by ± 2: when a source crosses from outside to inside a caustic, two new highly magnified, opposite-parity



Galaxy Cluster Abell 2218 **HST** · WFPC2
NASA, A, Fruchter and the ERO Team (STScI, ST-ECF) · STScI-PRC00-08

Figure 8.10: Hubble Space Telescope (WFPC2) image of the rich galaxy cluster Abell 2218, providing a spectacular textbook demonstration of gravitational lensing phenomena across all three regimes: prominent giant arcs and multiple images near the dense cluster core, weakly distorted arclets in the intermediate virial region, and coherent cosmic shear in the distant cluster outskirts. (Credit: A. Fruchter, ERO Team, STScI/ST-ECF, NASA).

images are created near the corresponding critical curve. A source located near a smooth fold caustic produces a pair of bright merging images; a source situated near a caustic cusp produces three prominent merging images forming a giant arc.

As established in Section 8.1.1 via vector field index topology, any smooth, non-singular, transparent mass distribution produces an odd total number of images $(2N+1)$. We now demonstrate how Fermat's principle of stationary arrival time provides a complementary variational framework explaining both the spatial image locations and their parity classifications.

Time delay and Fermat's principle Using the potential gradient relation $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \boldsymbol{\nabla}_2\psi(\boldsymbol{\theta})$ [Eq. (8.1.82)], the vector lens equation (8.1.43) can be recast in the variational form:

$$\boldsymbol{\nabla}_2\left[\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_S)^2 - \psi(\boldsymbol{\theta})\right] = 0. \tag{8.1.107}$$

In accordance with Fermat's principle, actual light rays correspond to stationary paths of the arrival time surface. The bracketed term is directly proportional to the total travel time delay $\tau(\boldsymbol{\theta}, \boldsymbol{\theta}_S)$ between the deflected ray and the unperturbed geometric trajectory:

$$\tau(\boldsymbol{\theta}, \boldsymbol{\theta}_S) = \frac{1+z_\mathrm{L}}{c}\frac{D_\mathrm{OL}D_\mathrm{OS}}{D_\mathrm{LS}}\left[\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_S)^2 - \psi(\boldsymbol{\theta})\right]. \tag{8.1.108}$$

The cosmological distance prefactor defines the *time-delay distance*:

$$D_{\Delta t} \equiv (1+z_\mathrm{L})\frac{D_\mathrm{OL}D_\mathrm{OS}}{D_\mathrm{LS}} \propto H_0^{-1}. \tag{8.1.109}$$

Proof. The total travel time of a photon traveling from source S at redshift $z_{\rm S}$ through lens L at redshift $z_{\rm L}$ to observer O consists of two contributions: a geometric path length difference $\Delta L_{\rm geom}$ and a gravitational Shapiro time delay $\Delta t_{\rm grav}$ caused by the gravitational potential well:

$$t = \int \frac{\mathrm{d}l}{c/n}, \qquad n = 1 - \frac{2\Phi}{c^2}. \tag{1}$$

The geometric path length difference between the deflected trajectory (passing through $\boldsymbol{\xi} = D_{\rm OL}\boldsymbol{\theta}$) and the straight unperturbed path connecting O to $\boldsymbol{\eta} = D_{\rm OS}\boldsymbol{\theta}_S$ evaluates in the paraxial approximation to:

$$\Delta L_{\rm geom} = \frac{1}{2} D_{\rm OL}\theta^2 + \frac{1}{2} D_{\rm LS}(\theta - \theta_S)^2 \frac{D_{\rm OS}}{D_{\rm LS}} - \frac{1}{2} D_{\rm OS}\theta_S^2 = \frac{1}{2}\frac{D_{\rm OL} D_{\rm OS}}{D_{\rm LS}}(\boldsymbol{\theta} - \boldsymbol{\theta}_S)^2. \tag{2}$$

The gravitational Shapiro delay acquired along the line of sight across the lens potential is:

$$\Delta t_{\rm grav} = -\frac{2}{c^3}\int_{-\infty}^{\infty} \Phi(D_{\rm OL}\boldsymbol{\theta}, z)\, \mathrm{d}z = -\frac{1}{c}\frac{D_{\rm OL} D_{\rm OS}}{D_{\rm LS}}\psi(\boldsymbol{\theta}). \tag{3}$$

Combining these two contributions and accounting for the cosmological time dilation factor $(1 + z_{\rm L})$ between the lens epoch and the observer frame:

$$\tau(\boldsymbol{\theta}, \boldsymbol{\theta}_S) = \frac{1 + z_{\rm L}}{c}\left[\Delta L_{\rm geom} + c\Delta t_{\rm grav}\right] = \frac{D_{\Delta t}}{c}\left[\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_S)^2 - \psi(\boldsymbol{\theta})\right]. \tag{4}$$

This completes the derivation of Eq. (8.1.108).

The arrival time surface $\tau(\boldsymbol{\theta}, \boldsymbol{\theta}_S)$ defined by Eq. (8.1.108) possesses several profound physical and geometric properties:

- **Observable time delays**: The differential time delay between two observed images $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ of the same variable source is given by:

$$\Delta\tau_{AB} = \tau(\boldsymbol{\theta}_A, \boldsymbol{\theta}_S) - \tau(\boldsymbol{\theta}_B, \boldsymbol{\theta}_S) = \frac{D_{\Delta t}}{c}\left[\tau_{\rm Fermat}(\boldsymbol{\theta}_A) - \tau_{\rm Fermat}(\boldsymbol{\theta}_B)\right]. \tag{8.1.110}$$

  Any intrinsic luminosity variation of the source (e.g., a supernova outburst or quasar flare) is registered first at the image corresponding to the global minimum of the travel-time surface. Because $D_{\Delta t} \propto H_0^{-1}$, measuring $\Delta\tau_{AB}$ together with a well-constrained lens mass model provides a direct, geometric measurement of the Hubble constant H_0 that is completely independent of the local cosmic distance ladder (Refsdal 1964).

- **Classification of stationary points**: The stationary points of τ fall into three topological classes determined by the eigenvalues of the Hessian matrix $\mathcal{T}_{ab} \equiv \partial^2\tau/\partial\theta^a\partial\theta^b$, which is directly proportional to the distortion matrix $\mathcal{A}_{ab}$:

  - *Type I (Minima)*: Both eigenvalues of $\mathcal{A}$ are strictly positive ($\det\mathcal{A} > 0$ and $\mathrm{Tr}\,\mathcal{A} > 0$). These images have positive intrinsic parity.
  - *Type II (Saddle points)*: The two eigenvalues have opposite signs ($\det\mathcal{A} < 0$). These images suffer a parity flip.
  - *Type III (Maxima)*: Both eigenvalues are negative ($\det\mathcal{A} > 0$ and $\mathrm{Tr}\,\mathcal{A} < 0$). These images preserve parity but are typically strongly de-magnified in realistic smooth lens potentials.

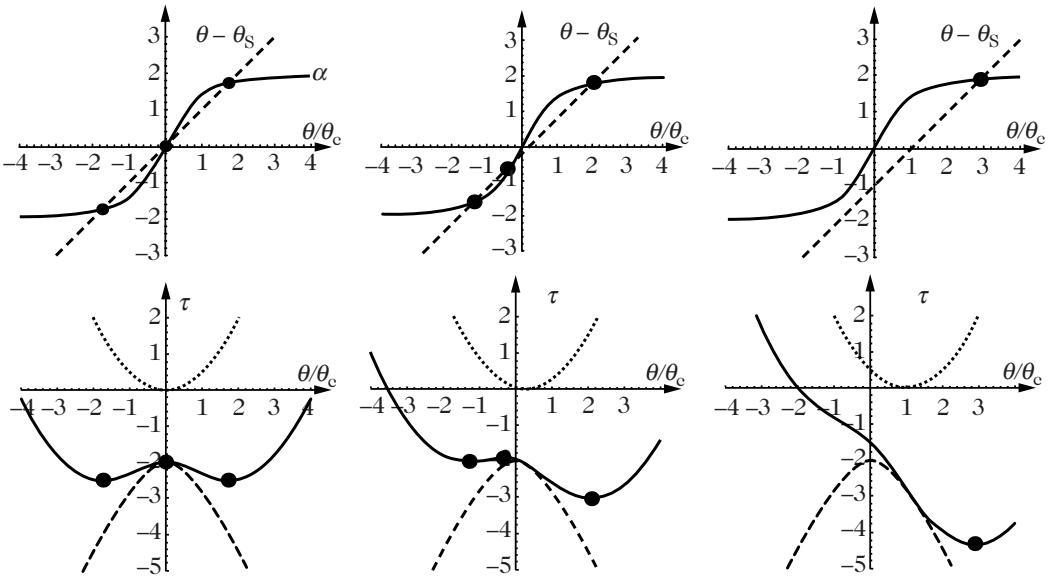


Figure 8.11: Determination of image multiplicities for a smooth isothermal lens model. (Top) Graphical solutions of the vector lens equation $\boldsymbol{\theta} - \boldsymbol{\theta}_S = \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ for three source alignments: $\theta_S/\theta_c = 0$, 0.3, and 1.0 (left to right). (Bottom) Corresponding arrival time profiles $\tau(\theta, \theta_S)$. The dashed parabola represents the geometric path length difference, while the lower dashed curve depicts the negative gravitational Shapiro potential. The solid curve represents the total arrival time, whose local extrema identify the image positions and their respective magnification types.

Lens models

Real astrophysical deflectors—whether individual early-type galaxies, spiral disks, or massive galaxy clusters—are not point masses. They are extended, complex systems composed of baryonic stars and gas embedded within vast, collisionless dark matter halos. Because dark matter cannot be photographed directly, astronomers must infer its two-dimensional projected mass distribution $\Sigma(\boldsymbol{\theta})$ by fitting parameterized physical profiles to observed image positions, shapes, and relative fluxes.

A successful lens model must balance two competing virtues: physical realism (faithfully reflecting the equilibrium states of collisionless dark matter) and mathematical tractability (yielding transparent, calculable deflection fields and magnifications). We examine the foundational model families that form the bedrock of modern gravitational lensing.

The Singular Isothermal Sphere (SIS) The simplest realistic model for a galaxy-scale deflector is the *Singular Isothermal Sphere* (SIS). Picture a spherical ensemble of collisionless particles (stars and dark matter) whose velocity distribution is isotropic and characterized by a spatially uniform one-dimensional velocity dispersion σ_v. Treating the collisionless particles as a self-gravitating gas with effective pressure $P(r) = \rho(r)\sigma_v^2$, hydrostatic equilibrium requires:

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{G_{\mathrm{N}} M(r)\rho(r)}{r^2} \implies \frac{r^2}{\rho}\frac{\mathrm{d}\rho}{\mathrm{d}r} = -\frac{G_{\mathrm{N}}}{\sigma_v^2} M(r). \tag{8.1.111}$$

Differentiating with respect to r using $\mathrm{d}M\,/\,\mathrm{d}r = 4\pi r^2 \rho(r)$ yields the hydrostatic balance equation:

$$\frac{\mathrm{d}}{\mathrm{d}r}\left[\frac{r^2}{\rho}\frac{\mathrm{d}\rho}{\mathrm{d}r}\right] = -\frac{4\pi G_{\mathrm{N}}}{\sigma_v^2} r^2 \rho(r). \tag{8.1.112}$$

Seeking a scale-free power-law solution $\rho(r) = Ar^{-\gamma}$, direct substitution confirms balance if and only if $\gamma = 2$ and $A = \sigma_v^2/(2\pi G_{\rm N})$, yielding the singular isothermal density profile:

$$\rho(r) = \frac{\sigma_v^2}{2\pi G_{\rm N}} \frac{1}{r^2}. \tag{8.1.113}$$

Integrating this density outward, the enclosed three-dimensional mass scales linearly with radius: $M(r) = 2\sigma_v^2 r/G_{\rm N}$. The circular orbital velocity of a test particle is therefore strictly independent of radius:

$$v_c^2(r) = \frac{G_{\rm N} M(r)}{r} = 2\sigma_v^2 = \text{const}, \tag{8.1.114}$$

providing the foundational dynamical explanation for the asymptotically flat rotation curves observed in spiral galaxies.

Projecting $\rho(r)$ along the line-of-sight coordinate z across a projected physical impact parameter ξ, where $r = \sqrt{\xi^2 + z^2}$, the surface mass density evaluates to:

$$\Sigma(\xi) = \int_{-\infty}^{\infty} \rho\left(\sqrt{\xi^2 + z^2}\right) \mathrm{d}z = \frac{\sigma_v^2}{2\pi G_{\rm N}} \int_{-\infty}^{\infty} \frac{\mathrm{d}z}{\xi^2 + z^2} = \frac{\sigma_v^2}{2G_{\rm N}} \frac{1}{\xi}. \tag{8.1.115}$$

The enclosed two-dimensional projected mass within radius ξ is $M_{\rm 2D}(<\xi) = 2\pi \int_0^{\xi} \Sigma(\xi')\xi' \,\mathrm{d}\xi' = \pi\sigma_v^2 \xi/G_{\rm N}$. Applying Gauss's law for circularly symmetric deflectors, the physical deflection angle is:

$$\alpha(\xi) = \frac{4G_{\rm N}}{c^2 \xi} M_{\rm 2D}(<\xi) = 4\pi \frac{\sigma_v^2}{c^2} = 1.4'' \left(\frac{\sigma_v}{220\,\mathrm{km/s}}\right)^2. \tag{8.1.116}$$

This is a remarkable result: because the enclosed projected mass grows in exact linear proportion to the impact parameter ($M_{\rm 2D} \propto \xi$), the radial coordinate ξ cancels out completely. The deflection angle of an SIS lens is strictly constant everywhere on the sky!

In optical terms, an SIS acts as an *axicon* (a conical glass prism). Unlike a manufactured thin lens (which bends rays linearly, $\alpha \propto \theta$) or a point mass (which bends rays inversely, $\alpha \propto 1/\theta$), an SIS deflects all rays toward the center by the identical angle:

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \theta_{\rm E} \frac{\boldsymbol{\theta}}{|\boldsymbol{\theta}|}, \qquad \theta_{\rm E} \equiv 4\pi \frac{\sigma_v^2}{c^2} \frac{D_{\rm LS}}{D_{\rm OS}}. \tag{8.1.117}$$

The vector lens equation $\boldsymbol{\theta}_S = \boldsymbol{\theta} - \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ simplifies along the source-lens axis to the one-dimensional piecewise relation:

$$\theta_S = \theta - \theta_{\rm E} \,\mathrm{sgn}(\theta). \tag{8.1.118}$$

Consider what happens as a source is positioned at angular offset $\theta_S \geq 0$:

1. *Multiple images ($\theta_S < \theta_{\rm E}$):* If the source lies inside the Einstein ring, Eq. (8.1.118) admits two solutions:

   $$\theta_+ = \theta_S + \theta_{\rm E} \quad \text{(outside the ring, same side)}, \qquad \theta_- = \theta_S - \theta_{\rm E} \quad \text{(inside the ring, opposite side)}. \tag{8.1.119}$$

   The angular separation between the two images is strictly invariant:

   $$\Delta\theta \equiv \theta_+ - \theta_- = 2\theta_{\rm E}, \tag{8.1.120}$$

   independent of the source position θ_S!

2. *Single image ($\theta_S > \theta_{\rm E}$):* If the source lies outside the Einstein ring, the negative branch $\theta_- = \theta_S - \theta_{\rm E} > 0$ contradicts the hypothesis $\theta < 0$, leaving only a single un-inverted image at $\theta_+ = \theta_S + \theta_{\rm E}$.

Differentiating the coordinate mapping, the individual image magnifications are:

$$\mu_{\pm} = \frac{\theta_{\pm}}{\theta_S}\frac{\mathrm{d}\theta_{\pm}}{\mathrm{d}\theta_S} = \frac{\theta_S \pm \theta_{\mathrm{E}}}{\theta_S}(1) = 1 \pm \frac{\theta_{\mathrm{E}}}{\theta_S}. \tag{8.1.121}$$

Image θ_+ is upright ($\mu_+ > 1$), while image θ_- is inverted ($\mu_- < 0$). The net signed magnification is $\mu_+ + \mu_- = 2$, meaning the deflector consistently amplifies the received flux by a net factor of 2 compared to the unlensed source, with the interior counter-image vanishing as $\theta_S \to \theta_{\mathrm{E}}$.

Core Softening and the Odd-Number Image Theorem The central density singularity $\rho(r) \to \infty$ of the SIS is unphysical. Real galaxies possess finite central stellar densities and smooth potential cores. A more realistic parametrization is the *Non-Singular Isothermal Sphere* (NIS), which introduces a finite core radius $r_c \equiv D_{\mathrm{OL}}\theta_c$:

$$\rho(r) = \frac{\sigma_v^2}{2\pi G_{\mathrm{N}}}\frac{1}{r^2 + r_c^2}, \qquad \Sigma(\theta) = \frac{\sigma_v^2}{2G_{\mathrm{N}}D_{\mathrm{OL}}}\frac{1}{\sqrt{\theta^2 + \theta_c^2}}. \tag{8.1.122}$$

For an NIS, the deflection angle smoothly interpolates between linear deflection in the core and constant deflection in the isothermal envelope:

$$\hat{\alpha}(\theta) = \theta_{\mathrm{E}}\frac{\sqrt{\theta^2 + \theta_c^2} - \theta_c}{\theta}. \tag{8.1.123}$$

Inside the flat core ($\theta \ll \theta_c$), the deflection is linear: $\hat{\alpha}(\theta) \approx \frac{\theta_{\mathrm{E}}}{2\theta_c}\theta$. As illustrated in Fig. 8.11, because the deflection vanishes smoothly at the origin rather than jumping discontinuously by $2\theta_{\mathrm{E}}$, a source aligned near the optical axis intercepts the deflection curve three times, producing three images:

- Two outer images ($\theta_1 > \theta_{\mathrm{E}}$ and $\theta_2 < 0$) corresponding to the familiar positive-parity minimum and negative-parity saddle point of the arrival-time surface.
- A third central image ($\theta_3 \approx 0$) located in the deep core of the lens.

This demonstrates the general *Odd-Number Image Theorem* of gravitational lensing: any non-singular, transparent gravitational deflector with finite total mass produces an odd number of images ($N = 2k + 1$). Why, then, do strong galaxy lenses routinely appear as image pairs or quadruplets rather than triplets or quintuplets? The answer lies in the extreme demagnification of the central image. In the core of a massive galaxy, the convergence $\kappa(0) \sim 10$–100 is strongly supercritical. Differentiating the lens equation, the magnification of the central image scales as $\mu_3 \sim [1 - \kappa(0)]^{-2} \lesssim 10^{-2}$–$10^{-4}$. The odd third image is buried beneath the bright stellar core of the lens galaxy, rendering it virtually undetectable in ordinary optical surveys.

The Navarro–Frenk–White (NFW) Profile While isothermal profiles provide an excellent match to individual galaxy dynamics, cosmological N-body simulations of collisionless Cold Dark Matter (CDM) reveal that dark matter halos do not follow a single power law across all scales. Instead, halos spanning from dwarf galaxies to massive clusters collapse into a universal, two-scale equilibrium profile (Navarro, Frenk, & White 1996, 1997):

$$\rho_{\mathrm{NFW}}(r) = \frac{\rho_0}{\left(\frac{r}{r_s}\right)\left(1 + \frac{r}{r_s}\right)^2}, \tag{8.1.124}$$

where r_s is the characteristic scale radius where the logarithmic slope is isothermal ($\mathrm{d}\ln\rho/\mathrm{d}\ln r = -2$), and ρ_0 is the central density normalization.

The NFW profile exhibits three distinct physical regimes:

- *Inner cusp ($r \ll r_s$):* The density scales as $\rho \propto r^{-1}$. This cusp is shallower than the isothermal $\rho \propto r^{-2}$. When projected along the line of sight, the central surface mass density diverges only logarithmically: $\Sigma(\theta) \propto -\ln\theta$. Consequently, the deflection angle $\hat{\alpha}(\theta)$ vanishes at the origin ($\hat{\alpha} \to 0$ as $\theta \to 0$).
- *Isothermal transition ($r \sim r_s$):* The profile matches the isothermal slope $\rho \propto r^{-2}$, where strong lensing arcs and Einstein rings typically form.
- *Outer virial envelope ($r \gg r_s$):* The density steepens to $\rho \propto r^{-3}$. The projected surface mass density falls as $\Sigma(\theta) \propto \theta^{-2}$, and the cosmic shear decays rapidly as $\gamma(\theta) \propto \theta^{-2}$.

Projecting $\rho_{\rm NFW}$ along the line of sight at projected impact parameter $\xi = D_{\rm OL}\theta$ and scaling by the critical surface density $\Sigma_{\rm crit}$ yields the closed-form convergence in terms of the dimensionless radius $x \equiv \xi/r_s = \theta/\theta_s$ (Bartelmann 1996):

$$\kappa_{\rm NFW}(x) = \frac{2\rho_0 r_s}{\Sigma_{\rm crit}} \frac{f(x)}{x^2-1}, \qquad \bar{\kappa}_{\rm NFW}(x) = \frac{4\rho_0 r_s}{\Sigma_{\rm crit}} \frac{g(x)}{x^2}, \tag{8.1.125}$$

where $\bar{\kappa}(x) \equiv \frac{2}{x^2}\int_0^x x'\kappa(x')\,{\rm d}x'$ is the mean enclosed convergence within radius x. The auxiliary radial functions $f(x)$ and $g(x)$ evaluate analytically across the scale radius $x = 1$:

$$f(x) = \begin{cases} 1 - \dfrac{2}{\sqrt{1-x^2}} \operatorname{artanh}\sqrt{\dfrac{1-x}{1+x}}, & (x<1), \\ \dfrac{1}{3}, & (x=1), \\ 1 - \dfrac{2}{\sqrt{x^2-1}} \arctan\sqrt{\dfrac{x-1}{x+1}}, & (x>1), \end{cases} \tag{8.1.126}$$

and $g(x) = \ln(x/2) + [1 - f(x)]$. For any circularly symmetric lens, the observable tangential shear is given by the Gauss flux difference:

$$\gamma_t(x) = \bar{\kappa}(x) - \kappa(x). \tag{8.1.127}$$

As illustrated in Fig. 8.12, while the SIS model predicts identical, scale-free profiles $\kappa(\theta) = \gamma(\theta) = \theta_{\rm E}/(2\theta)$ throughout all space, the NFW profile captures the physical turnover from the cuspy core to the steep $\gamma \propto \theta^{-2}$ decay in the halo outskirts. This makes NFW the indispensable workhorse for modeling weak lensing shear profiles in galaxy clusters.

Elliptical Deflectors, External Shear, and the Einstein Cross Real galaxies and clusters are not circularly symmetric: tidal torques from neighboring large-scale structures, triaxial dark matter halos, and anisotropic stellar velocity dispersions flatten deflectors into elliptical systems. Non-axisymmetry enters through two distinct physical mechanisms:

1. *Intrinsic lens ellipticity:* The surface mass density or deflection potential of the galaxy itself is flattened, parametrized with axis ratio $q \in (0, 1]$ and eccentricity $\epsilon \equiv (1-q)/(1+q)$:

$$\psi_{\rm ell}(\boldsymbol{\theta}) = \theta_{\rm E}\left[\theta_c^2 + (1-\epsilon)\theta_1^2 + (1+\epsilon)\theta_2^2\right]^{1/2}. \tag{8.1.128}$$

2. *External tidal shear:* The lens galaxy is embedded within an external gravitational tidal field sourced by a parent cluster, group, or line-of-sight cosmic structure. To quadrupole order, the environmental contribution is:

$$\psi_{\rm ext}(\boldsymbol{\theta}) = \frac{1}{2}\gamma_{\rm ext}\left(\theta_1^2 - \theta_2^2\right)\cos(2\phi_\gamma) + \gamma_{\rm ext}\theta_1\theta_2\sin(2\phi_\gamma), \tag{8.1.129}$$

where $\gamma_{\rm ext}$ is the tidal shear amplitude and ϕ_γ is its orientation.

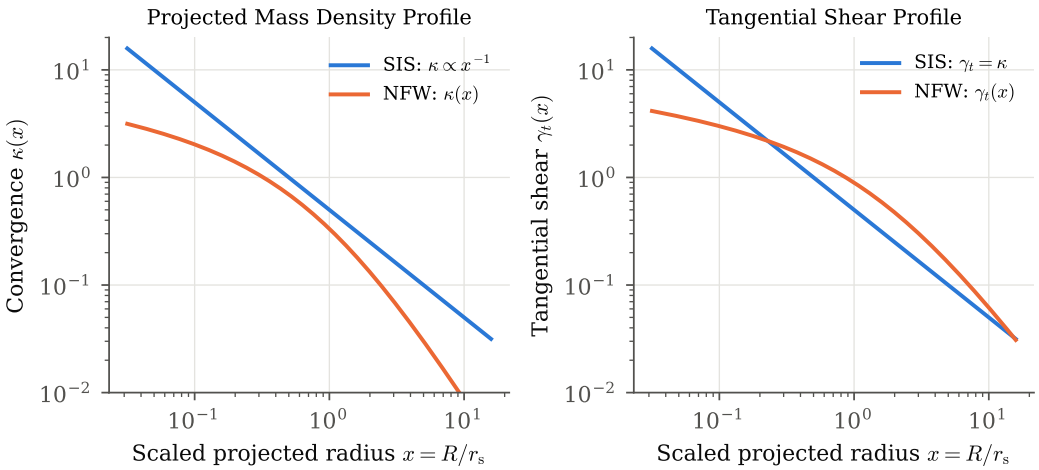


Figure 8.12: Comparison of convergence $\kappa(\theta)$ (solid curves) and shear $\gamma(\theta)$ (dashed curves) for a Singular Isothermal Sphere (SIS) versus a Navarro–Frenk–White (NFW) dark matter halo profile. For the SIS model, $\kappa(\theta) = \gamma(\theta) \propto \theta^{-1}$ everywhere. For the NFW model, the central cuspy core produces $\kappa \propto \theta^{-1}$ as $\theta \to 0$, while transitioning to a steeper fall-off $\gamma \propto \theta^{-2}$ in the outer virial regime.

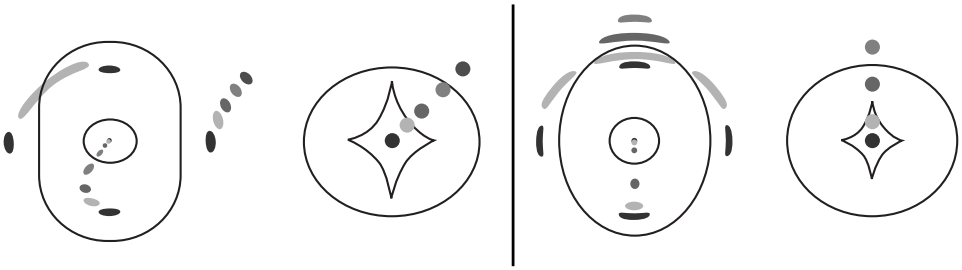


Figure 8.13: Critical curves (lens plane, left panels) and caustic lines (source plane, right panels) for representative elliptical gravitational lens models. When a point source enters the inner astroid (diamond) caustic, image multiplicity transitions from two to four prominent images forming an Einstein cross. From Ref. [4].

Both intrinsic ellipticity and external shear introduce a quadrupole perturbation into the Fermat travel-time surface. This produces a profound topological catastrophe on the source plane (see Fig. 8.13). In a circularly symmetric lens, the tangential caustic is degenerate: it is a single mathematical point centered on the optical axis ($\boldsymbol{\theta}_S = \mathbf{0}$). Breaking circular symmetry unfolds this point caustic into a four-cusped diamond astroid (the tangential caustic) surrounded by an outer oval curve (the radial caustic).

The number of observable images depends strictly on the source's position relative to these caustic boundaries:

- Outside both caustics: The source is weakly deflected, forming a single un-inverted image.
- Crossing the outer radial caustic: A pair of images (one positive-parity minimum and one negative-parity saddle point) is born along the minor axis, producing 3 images in total.
- Entering the diamond astroid caustic: A second pair of images is born along the orthogonal axis. The resulting configuration consists of 4 bright, highly magnified outer images arranged symmetrically around the deflector—forming the celebrated *Einstein Cross*—plus 1 highly demagnified image at the center, yielding 5 images in total.

Table 8.2: Properties of canonical spherically symmetric gravitational lens models.

Lens Model	**Potential** $\psi(\theta)$	**Deflection** $\alpha(\theta)$	**Convergence** $\kappa(\theta)$
Point-like mass	$\theta_{\rm E}^2 \ln\theta$	$\theta_{\rm E}^2/\theta$	0 (0 for $\theta > 0$)
Singular isothermal sphere (SIS)	$\theta_{\rm E}\theta$	$\theta_{\rm E}$	$\frac{1}{2}\theta_{\rm E}/\theta$
Non-singular isothermal sphere (NIS)	$\theta_{\rm E}\sqrt{\theta_c^2+\theta^2}$	$\theta_{\rm E}\frac{\sqrt{\theta_c^2+\theta^2}-\theta_c}{\theta}$	$\frac{1}{2}\frac{\theta_{\rm E}}{\sqrt{\theta_c^2+\theta^2}}$
Navarro–Frenk–White (NFW)	Eq. (8.1.125)	$\bar{\kappa}(\theta)\theta$	$\kappa_{\rm NFW}(\theta)$
Uniform mass sheet	$\frac{1}{2}\kappa_0\theta^2$	$\kappa_0\theta$	κ_0

Famous astrophysical examples, such as the quadruply imaged quasar Q2237+0305 (Huchra's Lens) and the cloverleaf quasar H1413+117, are natural realizations of sources residing inside this central diamond caustic.

Summary of Canonical Lens Models Table 9.2 compiles the projected potential $\psi(\theta)$, deflection angle $\alpha(\theta)$, and convergence $\kappa(\theta)$ for the principal analytical lens models used in modern astrophysics.

8.1.2 Lensing by galaxies and galaxy clusters

The mathematical machinery developed in the preceding sections finds its realization across a spectacular hierarchy of cosmic scales. Depending on the mass and physical extent of the deflector relative to the impact parameter of passing light rays, gravitational lensing manifests in three distinct regimes:

1. *Microlensing:* Deflectors are compact stellar-mass or sub-stellar objects ($M \sim 10^{-6}$–$10\,M_\odot$). While the angular separation of the resulting images ($\sim 1\,$mas) is too minute to resolve with optical telescopes, the transient, achromatic flux amplification produced as the lens sweeps across a background star is readily detectable.

2. *Galaxy-scale strong lensing:* Deflectors are massive early-type galaxies ($M \sim 10^{11}$–$10^{13}\,M_\odot$). Passing light rays are split into distinct, resolvable multiple images (Einstein rings, quads, and double quasars) with image separations on the scale of arcseconds, providing direct probes of dark matter halo profiles and the expansion rate H_0.

3. *Cluster-scale strong lensing:* Deflectors are massive galaxy clusters ($M \sim 10^{14}$–$10^{15}\,M_\odot$). Their deep, extended gravitational potentials bend light from entire background fields into giant, warped tangential arcs and multiple image systems across tens of arcseconds, acting as natural cosmic telescopes.

Gravitational microlensing in the Milky Way

In 1986, Bohdan Paczyński recognized that if the dark matter halo of the Milky Way were composed of Massive Astrophysical Compact Halo Objects (MACHOs)—such as black holes, neutron stars, or brown dwarfs—these compact bodies would occasionally pass in front of stars in neighboring satellite galaxies (such as the Large and Small Magellanic Clouds), generating transient brightening events.

The angular Einstein radius for a solar-mass compact lens within the Galactic halo ($D_{\rm OL} \sim 10\,$kpc) deflecting light from a source in the LMC ($D_{\rm OS} \sim 50\,$kpc) is:

$$\theta_{\rm E} = \sqrt{\frac{4G_{\rm N}M}{c^2}\frac{D_{\rm LS}}{D_{\rm OL}D_{\rm OS}}} \approx 0.64\,\text{mas}\left(\frac{M}{M_\odot}\right)^{1/2}. \tag{8.1.130}$$

Because this milliarcsecond image separation is far below the angular resolution of optical telescopes, the multiple images cannot be resolved individually. Instead, an observer measures the time-varying combined flux of the unresolved images.

The characteristic event duration τ_*, or Einstein crossing time, is the time required for the source to traverse the transverse physical radius of the Einstein ring $R_{\rm E} = D_{\rm OL}\theta_{\rm E}$ at relative transverse velocity $v_\perp$:

$$\tau_* = \frac{D_{\rm OL}\theta_{\rm E}}{v_\perp} \approx 0.214 \left(\frac{M}{M_\odot}\right)^{1/2} \left(\frac{D_{\rm OL}}{10\,\mathrm{kpc}}\right)^{1/2} \left(\frac{D_{\rm LS}}{D_{\rm OS}}\right)^{1/2} \left(\frac{v_\perp}{200\,\mathrm{km\,s^{-1}}}\right)^{-1} \text{years}. \tag{8.1.131}$$

For typical halo kinematics ($v_\perp \sim 200\,\mathrm{km\,s^{-1}}$), the event timescale scales directly as the square root of the lens mass:

- Terrestrial-mass free-floating planets ($M \sim 10^{-6}\,M_\odot$): $\tau_* \sim$ hours.
- Brown dwarfs and low-mass stars ($M \sim 10^{-2}$–$0.1\,M_\odot$): $\tau_* \sim$ days to weeks.
- Solar-mass stars and stellar black holes ($M \sim 1$–$10\,M_\odot$): $\tau_* \sim$ months to a year.

The Paczyński light curve Assuming uniform rectilinear relative motion between observer, lens, and background source, the dimensionless angular separation $u(t) \equiv \theta_S(t)/\theta_{\rm E}$ in units of the Einstein radius follows the Pythagorean trajectory:

$$u(t) = \sqrt{u_{\rm min}^2 + \left(\frac{t - t_0}{\tau_*}\right)^2}, \tag{8.1.132}$$

where t_0 is the moment of closest impact and $u_{\rm min}$ is the minimum impact parameter.

Derivation of the Rectilinear Trajectory $u(t)$. Choose a 2D celestial coordinate system centered on the deflecting lens $(0, 0)$. Under uniform rectilinear relative motion at transverse velocity $v_\perp$, let the closest approach occur at time t_0 with impact angle $\theta_{\rm min}$ aligned along the y-axis: $\boldsymbol{\theta}_S(t_0) = (0, \theta_{\rm min})$.
At any arbitrary time t, the relative angular velocity along the transverse x-axis is $\dot\theta_\perp = v_\perp/D_{\rm OL}$. The time-dependent 2D source position is:

$$\boldsymbol{\theta}_S(t) = \left(\dot\theta_\perp(t - t_0),\, \theta_{\rm min}\right). \tag{1}$$

The Euclidean angular separation $\theta_S(t) = |\boldsymbol{\theta}_S(t)|$ between source and lens follows the Pythagorean relation:

$$\theta_S(t) = \sqrt{\theta_{\rm min}^2 + \left[\dot\theta_\perp(t - t_0)\right]^2}. \tag{2}$$

Dividing by the angular Einstein radius $\theta_{\rm E}$ and substituting the Einstein crossing time $\tau_* \equiv \theta_{\rm E}/\dot\theta_\perp = D_{\rm OL}\theta_{\rm E}/v_\perp$:

$$u(t) \equiv \frac{\theta_S(t)}{\theta_{\rm E}} = \sqrt{\left(\frac{\theta_{\rm min}}{\theta_{\rm E}}\right)^2 + \left[\frac{\dot\theta_\perp(t - t_0)}{\theta_{\rm E}}\right]^2} = \sqrt{u_{\rm min}^2 + \left(\frac{t - t_0}{\tau_*}\right)^2}, \tag{3}$$

where $u_{\rm min} \equiv \theta_{\rm min}/\theta_{\rm E}$.

Recalling the total point-mass magnification $\mu = |\mu_+| + |\mu_-|$ from Eq. (8.1.31), substituting $u(t)$ yields the classical *Paczyński light curve*:

$$A(t) = \frac{u^2(t) + 2}{u(t)\sqrt{u^2(t) + 4}}. \tag{8.1.133}$$

Derivation of the Paczyński Magnification $A(t)$. For a point mass, the lens equation is $\theta_S = \theta - \theta_{\rm E}^2/\theta$. In dimensionless coordinates $u \equiv \theta_S/\theta_{\rm E}$ and $x \equiv \theta/\theta_{\rm E}$, this reads $u = x - 1/x$, or equivalently:

$$x^2 - ux - 1 = 0 \implies x_\pm = \frac{u \pm \sqrt{u^2+4}}{2}. \tag{1}$$

Because surface brightness is conserved, the magnification of an image at coordinate x is the solid-angle Jacobian:

$$\mu(x) = \frac{\theta}{\theta_S}\frac{\mathrm{d}\theta}{\mathrm{d}\theta_S} = \frac{x}{u}\frac{\mathrm{d}x}{\mathrm{d}u} = \frac{x}{x-1/x}\left(1+\frac{1}{x^2}\right)^{-1} = \frac{x^4}{x^4-1}. \tag{2}$$

Evaluating $\mu(x)$ at the two image roots $x_\pm$:

$$\mu_\pm = \frac{1}{2} \pm \frac{u^2+2}{2u\sqrt{u^2+4}}. \tag{3}$$

Because the interior image x_- forms inside the Einstein ring ($|x_-| < 1$), its magnification is negative ($\mu_- < 0$), reflecting parity inversion. In microlensing, the microarcsecond image separation $\Delta\theta \sim \theta_{\rm E}$ cannot be resolved by optical telescopes; the detector registers the combined unresolved flux of both images. The total amplification factor is the sum of their absolute fluxes:

$$A(u) = |\mu_+| + |\mu_-| = \mu_+ - \mu_- = \left(\frac{1}{2} + \frac{u^2+2}{2u\sqrt{u^2+4}}\right) - \left(\frac{1}{2} - \frac{u^2+2}{2u\sqrt{u^2+4}}\right) = \frac{u^2+2}{u\sqrt{u^2+4}}. \tag{4}$$

Evaluating along the rectilinear trajectory $u(t)$ gives Eq. (8.1.133).

As shown in Fig. 8.14, the amplification curve displays distinctive physical features:

1. *Time symmetry:* The light curve is strictly symmetric about the peak at $t = t_0$, reflecting the time-reversal symmetry of the straight-line trajectory across the circularly symmetric Einstein disk.

2. *Achromaticity:* Because photons of all frequencies follow identical null geodesics in general relativity, the magnification $A(t)$ is entirely independent of observing wavelength.

3. *Singular peak for central impact:* As $u \to 0$, the magnification diverges as $A \propto 1/u$, corresponding to the point-caustic singularity at the center of the lens plane. In reality, the peak magnification is capped by the finite angular size of the source star.

Optical depth and event rate To detect microlensing, an observer must know the probability that a given background star falls within the Einstein disk of a foreground deflector at any moment. This dimensionless probability defines the *microlensing optical depth* τ.

The cross-section $\sigma(> A)$ of an individual point-mass deflector is defined as the physical area on the lens plane within which a source must lie to experience a magnification exceeding threshold A:

$$\sigma(> A) = 2\frac{M}{\Sigma_{\rm crit}}\frac{1}{A^2 - 1 + A\sqrt{A^2-1}}. \tag{8.1.134}$$

Derivation of the Microlensing Cross-Section. Inverting the point-mass magnification

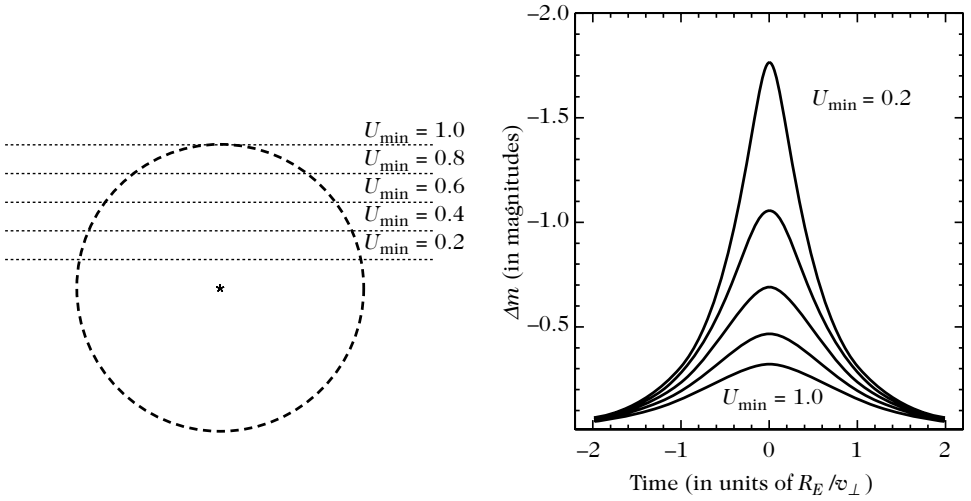


Figure 8.14: Theoretical Paczyński microlensing light curves $\Delta m(t) = -2.5\log_{10} A(t)$ plotted as a function of time normalized to the Einstein radius crossing time $\tau_* = R_{\rm E}/v_\perp$, for impact parameters $u_{\rm min} = 0.2, 0.4, 0.6, 0.8, 1.0$. (Left) Trajectories across the projected Einstein ring.

relation $A = (u^2+2)/(u\sqrt{u^2+4})$ [Eq. (8.1.31)] for u^2 yields:

$$u^2(A) = 2\left[\frac{A}{\sqrt{A^2-1}} - 1\right] = \frac{2}{A^2 - 1 + A\sqrt{A^2-1}}. \tag{1}$$

The corresponding physical cross-sectional area on the lens plane within impact radius $\xi = D_{\rm OL}\theta_{\rm E} u$ is:

$$\sigma(>A) = \pi\left[D_{\rm OL}\theta_{\rm E} u(A)\right]^2 = \pi D_{\rm OL}^2 \theta_{\rm E}^2 u^2(A). \tag{2}$$

Recalling the critical surface mass density $\Sigma_{\rm crit} = \frac{c^2}{4\pi G_{\rm N}}\frac{D_{\rm OS}}{D_{\rm OL}D_{\rm LS}}$, the physical area of the Einstein ring satisfies:

$$\pi D_{\rm OL}^2\theta_{\rm E}^2 = \pi D_{\rm OL}^2\left(\frac{4G_{\rm N}M}{c^2}\frac{D_{\rm LS}}{D_{\rm OL}D_{\rm OS}}\right) = \frac{M}{\Sigma_{\rm crit}}. \tag{3}$$

Multiplying by $u^2(A)$ completes the proof of Eq. (8.1.134):

$$\sigma(>A) = \frac{M}{\Sigma_{\rm crit}}u^2(A) = 2\frac{M}{\Sigma_{\rm crit}}\frac{1}{A^2-1+A\sqrt{A^2-1}}. \tag{4}$$

Consider a pencil beam of solid angle $\Delta\Omega$ toward a source at distance $D_{\rm OS}$. At distance $D_{\rm OL}$, this solid angle encompasses a transverse area element $\Delta\Omega D_{\rm OL}^2$. If the spatial number density of compact lenses of mass M is $n(D_{\rm OL})$, the number of deflectors in a volume element $\mathrm{d}V = \Delta\Omega D_{\rm OL}^2\,\mathrm{d}D_{\rm OL}$ is $\mathrm{d}N = n(D_{\rm OL})\,\mathrm{d}V$. Summing the projected cross-sections of all deflectors along the line of sight:

$$\tau \equiv \frac{1}{\Delta\Omega D_{\rm OS}^2}\int \mathrm{d}N\left[\pi D_{\rm OS}^2\theta_{\rm E}^2\right] = \int_0^{D_{\rm OS}} n(D_{\rm OL})\left[\pi D_{\rm OL}^2\theta_{\rm E}^2\right]\mathrm{d}D_{\rm OL}\,. \tag{8.1.135}$$

Substituting $\theta_{\rm E}^2 = \frac{4G_{\rm N}M}{c^2}\frac{D_{\rm LS}}{D_{\rm OL}D_{\rm OS}}$ and defining the mass density of compact objects $\rho_{\rm comp}(D_{\rm OL}) \equiv n(D_{\rm OL})M$:

$$\tau = \frac{4\pi G_{\rm N}}{c^2}\int_0^{D_{\rm OS}} \rho_{\rm comp}(D_{\rm OL})\frac{D_{\rm OL}D_{\rm LS}}{D_{\rm OS}}\,\mathrm{d}D_{\rm OL}\,. \tag{8.1.136}$$

Remarkably, the individual lens mass M cancels out completely! The optical depth depends solely on the integrated mass density along the line of sight.

Introducing the normalized fractional distance $x \equiv D_{\rm OL}/D_{\rm OS}$, and using Euclidean distances $D_{\rm LS} = D_{\rm OS}(1-x)$ (valid within the local Galactic halo):

$$\tau = \frac{4\pi G_{\rm N} D_{\rm OS}^2}{c^2} \int_0^1 \rho_{\rm comp}(x)\, x(1-x)\, {\rm d}x\,. \tag{8.1.137}$$

Using the differential coordinate relation ${\rm d}D_{\rm OL} = D_{\rm OS}\,{\rm d}x$ and substituting $D_{\rm OL} = xD_{\rm OS}$ and $D_{\rm LS} = (1-x)D_{\rm OS}$ into Eq. (8.1.136):

$$\begin{aligned} \tau &= \frac{4\pi G_{\rm N}}{c^2} \int_0^1 \rho_{\rm comp}(x) \left[\frac{(xD_{\rm OS})(D_{\rm OS}(1-x))}{D_{\rm OS}}\right] (D_{\rm OS}\,{\rm d}x) \\ &= \frac{4\pi G_{\rm N} D_{\rm OS}^2}{c^2} \int_0^1 \rho_{\rm comp}(x) x(1-x)\,{\rm d}x\,. \end{aligned} \tag{1}$$

The geometric kernel $x(1-x)$ vanishes at the observer ($x=0$) and at the source ($x=1$), reaching its peak at $x = 1/2$. A deflector positioned halfway between observer and source therefore provides the maximum possible lensing optical depth.

For a constant halo density ρ_0, the integral evaluates to $\int_0^1 x(1-x)\,{\rm d}x = 1/6$, yielding $\tau = \frac{2\pi}{3} G_{\rm N}\rho_0 D_{\rm OS}^2/c^2 \approx 10^{-6}$–$10^{-7}$ toward the Magellanic Clouds. The total number of events with magnification exceeding A relates directly to the optical depth:

$$N(>A) = \frac{2\tau}{A^2 - 1 + A\sqrt{A^2-1}}. \tag{8.1.138}$$

For $A = 1.34$ ($u \le 1$), $N \approx \tau \sim 10^{-6}$, confirming that millions of background stars must be surveyed nightly to detect microlensing.

Observational limits on MACHOs Beginning in the early 1990s, extensive monitoring campaigns—most prominently the MACHO, EROS, and OGLE collaborations—observed stars in the LMC, SMC, and Galactic bulge. Lensing events were separated from intrinsic stellar variability (pulsators, flares, cataclysmic variables) by requiring strict time-symmetry, achromaticity across multiple color filters, and unrepeated, solitary occurrence.

As summarized in Fig. 9.11, the combined EROS and MACHO surveys detected far fewer events than predicted for a halo comprised of compact objects. Across the mass window $10^{-7}\,M_\odot \le M \le 10$–$30\,M_\odot$, compact objects contribute less than 10% (and in the stellar range $< 2\%$) to the total dark matter halo of the Milky Way. This decisive null result provided fundamental empirical proof that the cosmological dark matter cannot be made of baryonic compact remnants or primordial stellar-mass black holes, pointing inexorably toward non-baryonic particle dark matter (such as WIMPs or axions). Today, microlensing is primarily utilized as a premier method for discovering exoplanets (particularly cold planets orbiting beyond the snow line) and detecting isolated stellar-mass black holes in the Galactic disk.

Galaxy-scale strong lensing

The discovery of the doubly-imaged quasar QSO 0957+561 ($z_S = 1.41, z_{\rm L} = 0.36$) by Walsh, Carswell, and Weymann in 1979 confirmed Einstein's prediction on cosmological scales. Strong lensing by individual galaxies serves two distinct cosmological purposes: it maps the total mass and density profiles of early-type galaxies (independent of whether matter is luminous or dark), and it provides a direct, one-step geometrical measurement of the Hubble constant through time-delay cosmography.

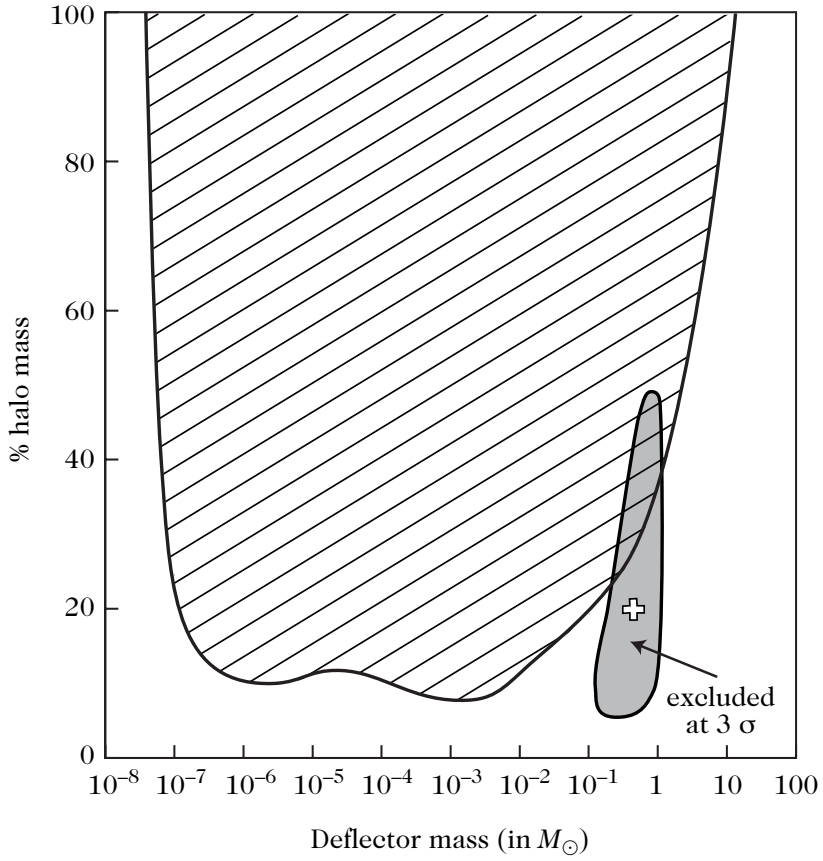


Figure 8.15: Observational constraints from the MACHO (shaded grey) and EROS (hatched) surveys on the halo mass fraction comprised of compact objects (modeled with total halo mass $4 \times 10^{11} M_\odot$ within 50 kpc). The allowed parameter space lies strictly below the upper envelope curve, ruling out planetary to solar-mass MACHOs as the primary constituent of Galactic dark matter. From Ref. [15].

Probing galaxy structure Because strong lensing depends only on gravity, it provides an un-biased mass scale for galaxies:

- *Total enclosed mass:* The angular separation between multiple images $\Delta\theta \approx 2\theta_{\rm E}$ measures the total projected mass enclosed within the Einstein radius:

$$M_{\rm 2D}(< \theta_{\rm E}) = \pi (D_{\rm OL}\theta_{\rm E})^2 \Sigma_{\rm crit}. \tag{8.1.139}$$

  For typical elliptical galaxies, $\Delta\theta \sim 1''$–$3''$, revealing enclosed masses $M(< \theta_{\rm E}) \sim 10^{11}$–$10^{12}\, M_\odot$.

- *Isothermal total profile (the "bulge-halo conspiracy"):* Combining strong lensing (which measures projected mass within $\sim R_{\rm E}$) with stellar dynamics (which measures central velocity dispersion σ_v) and weak lensing (which measures outer halo shear) demonstrates that the total density profile of early-type galaxies is exceptionally close to isothermal ($\rho_{\rm tot} \propto r^{-2}$) from sub-kiloparsec scales out to hundreds of kiloparsecs, even though the stars ($\rho_* \propto r^{-3}$ to r^{-4}) and dark matter ($\rho_{\rm DM} \propto r^{-1}$ to r^{-3}) have completely different radial profiles.

- *Cuspy galactic cores:* The non-detection of demagnified central images in quasar lenses

requires that early-type galaxies possess dense, cuspy cores with core radii $r_c \lesssim 100$–$200\,\mathrm{pc}$, ruling out large constant-density cores.

Time-delay cosmography and the Hubble constant When the background source is variable (such as an active galactic nucleus or a gravitationally lensed supernova), fluctuations in source luminosity travel along different null geodesic paths, arriving at the observer at different times.

From Fermat's principle (Section 8.1.1), the arrival time at image position $\boldsymbol{\theta}$ relative to an undeflected ray is:

$$\tau(\boldsymbol{\theta}, \boldsymbol{\theta}_S) = \frac{1+z_\mathrm{L}}{c}\frac{D_\mathrm{OL}D_\mathrm{OS}}{D_\mathrm{LS}}\left[\frac{1}{2}|\boldsymbol{\theta}-\boldsymbol{\theta}_S|^2 - \psi(\boldsymbol{\theta})\right]. \tag{8.1.140}$$

The physical time delay $\Delta t_{ij} \equiv \tau(\boldsymbol{\theta}_i) - \tau(\boldsymbol{\theta}_j)$ between two images $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ is determined by the *time-delay distance* $D_{\Delta t}$:

$$c\,\Delta t_{ij} = D_{\Delta t}\,\Delta\phi(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j), \tag{8.1.141}$$

where $\Delta\phi(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \equiv \left[\frac{1}{2}|\boldsymbol{\theta}_i-\boldsymbol{\theta}_S|^2 - \psi(\boldsymbol{\theta}_i)\right] - \left[\frac{1}{2}|\boldsymbol{\theta}_j-\boldsymbol{\theta}_S|^2 - \psi(\boldsymbol{\theta}_j)\right]$ is the dimensionless Fermat potential difference, and $D_{\Delta t}$ is the cosmological distance combination:

$$D_{\Delta t} \equiv (1+z_\mathrm{L})\frac{D_\mathrm{OL}D_\mathrm{OS}}{D_\mathrm{LS}}. \tag{8.1.142}$$

Because every angular diameter distance scales inversely with the Hubble constant ($D \propto c/H_0$), the time-delay distance is inversely proportional to H_0:

$$D_{\Delta t} \propto H_0^{-1}. \tag{8.1.143}$$

Consequently, measuring the time delay Δt_{ij} between image light curves and modeling the lens potential to determine $\Delta\phi$ provides a direct, absolute anchor for H_0, entirely independent of the local cosmic distance ladder (Cepheids and Type Ia supernovae) and the CMB sound horizon.

For the benchmark Singular Isothermal Sphere (SIS), the differential time delay between the two collinear images θ_1 and θ_2 evaluates to:

$$c\,\delta\tau = 8\pi\left(\frac{\sigma_v^2}{c^2}\right)^2(1+z_\mathrm{L})D_\mathrm{OL}\theta_\mathrm{E}. \tag{8.1.144}$$

Derivation of the SIS Time Delay. From Eq. (8.1.108), the differential travel time between images θ_1 and θ_2 is:

$$c\,\delta\tau = (1+z_\mathrm{L})\frac{D_\mathrm{OL}D_\mathrm{OS}}{D_\mathrm{LS}}\left\{\left[\frac{1}{2}(\theta_1-\theta_S)^2 - \frac{1}{2}(\theta_2-\theta_S)^2\right] - [\psi(\theta_1)-\psi(\theta_2)]\right\}. \tag{1}$$

For an SIS lens, the image positions lie at $\theta_1 = \theta_S + \theta_\mathrm{E}$ and $\theta_2 = \theta_S - \theta_\mathrm{E}$, which implies $\theta_1 - \theta_S = \theta_\mathrm{E}$ and $\theta_2 - \theta_S = -\theta_\mathrm{E}$. The geometric term vanishes identically:

$$\frac{1}{2}(\theta_1-\theta_S)^2 - \frac{1}{2}(\theta_2-\theta_S)^2 = \frac{1}{2}\theta_\mathrm{E}^2 - \frac{1}{2}(-\theta_\mathrm{E})^2 = 0. \tag{2}$$

The time delay is therefore driven entirely by the difference in the deflection potential $\psi(\theta) = \theta_\mathrm{E}\theta$:

$$\psi(\theta_1) - \psi(\theta_2) = \theta_\mathrm{E}(\theta_1-\theta_2) = \theta_\mathrm{E}\,[(\theta_S+\theta_\mathrm{E}) - (\theta_S-\theta_\mathrm{E})] = 2\theta_\mathrm{E}^2. \tag{3}$$

Substituting $\theta_\mathrm{E} = (D_\mathrm{LS}/D_\mathrm{OS})4\pi(\sigma_v^2/c^2)$ into the prefactor:

$$c\,\delta\tau = (1+z_\mathrm{L})\frac{D_\mathrm{OL}D_\mathrm{OS}}{D_\mathrm{LS}}(2\theta_\mathrm{E}^2) = 2(1+z_\mathrm{L})D_\mathrm{OL}\left(4\pi\frac{\sigma_v^2}{c^2}\right)\theta_\mathrm{E} = 8\pi\left(\frac{\sigma_v^2}{c^2}\right)^2(1+z_\mathrm{L})D_\mathrm{OL}\theta_\mathrm{E}. \tag{4}$$

This confirms Eq. (8.1.144).

For QSO 0957+561, the measured time delay of $\Delta t = (417 \pm 3)$ days yielded an early estimate of $h \approx 0.70$–0.75. Modern cosmological programs—notably the H0LiCOW and TDCOSMO collaborations—have monitored samples of quadruply lensed variable quasars with Hubble Space Telescope imaging and Keck adaptive optics, finding:

$$H_0 = 73.3^{+1.7}_{-1.8}\,\mathrm{km\,s^{-1}\,Mpc^{-1}}, \tag{8.1.145}$$

in sharp agreement with local distance ladder measurements ($73.04 \pm 1.04\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ from the SH0ES project) and in persistent $\sim 5\sigma$ tension with the early-universe sound-horizon inference from Planck CMB observations ($67.4 \pm 0.5\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$).

Cosmological optical depth of compact objects Extending the microlensing optical depth formalism to cosmological distances, the assumption $D_{\mathrm{LS}} = D_{\mathrm{OS}} - D_{\mathrm{OL}}$ must be replaced by relativistic angular diameter distances. For a population of compact objects with uniform cosmological density parameter Ω_{M0}:

$$\tau(z_{\mathrm{S}}) = \frac{3\Omega_{\mathrm{M0}}}{2} \int_0^{z_{\mathrm{S}}} \frac{D_{\mathrm{LS}}(z, z_{\mathrm{S}})}{D_{\mathrm{OS}}(z_{\mathrm{S}})} \frac{D_{\mathrm{OL}}(z)}{D_{H_0}} \frac{\mathrm{d}z}{E(z)}, \tag{8.1.146}$$

where $D_{H_0} \equiv c/H_0$ and $E(z) \equiv H(z)/H_0$. In an Einstein–de Sitter universe ($E(z) = (1+z)^{3/2}$), the integral evaluates analytically:

$$\tau(z_{\mathrm{S}}) = 3\frac{\Omega_{\mathrm{M0}}}{z_{\mathrm{S}}} \left[\left(z_{\mathrm{S}} + 2 + 2\sqrt{1+z_{\mathrm{S}}} \right) \ln(1+z_{\mathrm{S}}) - 4z_{\mathrm{S}} \right]. \tag{8.1.147}$$

Derivation in an Einstein–de Sitter Universe. In a flat, matter-dominated universe ($\Omega_{\mathrm{M0}} = 1$), $E(z) = (1+z)^{3/2}$ and the comoving distance is $r(z) = \frac{2c}{H_0}[1 - (1+z)^{-1/2}]$. The angular diameter distances are:

$$D_{\mathrm{OL}}(z) = \frac{r(z)}{1+z}, \qquad D_{\mathrm{OS}}(z_{\mathrm{S}}) = \frac{r(z_{\mathrm{S}})}{1+z_{\mathrm{S}}}, \qquad D_{\mathrm{LS}}(z, z_{\mathrm{S}}) = \frac{r(z_{\mathrm{S}}) - r(z)}{1+z_{\mathrm{S}}}. \tag{1}$$

Substituting these into Eq. (8.1.146):

$$\tau(z_{\mathrm{S}}) = \frac{3\Omega_{\mathrm{M0}}}{2} \int_0^{z_{\mathrm{S}}} \frac{r(z_{\mathrm{S}}) - r(z)}{r(z_{\mathrm{S}})} \frac{r(z)}{(1+z)D_{H_0}} \frac{\mathrm{d}z}{(1+z)^{3/2}}. \tag{2}$$

Substituting $u = \sqrt{1+z}$ with $\mathrm{d}z = 2u\,\mathrm{d}u$ and carrying out the rational function integration reproduces Eq. (8.1.147).

The observed incidence of cosmological microlensing establishes the constraint $\Omega_{\mathrm{M0}} < 0.05$ for compact masses in the range $M \in [10^6, 10^{12}]\,M_\odot$.

Cosmological volume test from lens statistics The probability that a distant quasar is strongly lensed depends on the comoving volume element $\mathrm{d}V_c\,/\,\mathrm{d}z$ and the spatial density of deflecting galaxies. In universes with a positive cosmological constant Λ, cosmic acceleration expands the volume element at high redshifts:

$$\mathrm{d}V_c\,(z) \propto \frac{r^2(z)}{H(z)}\,\mathrm{d}z\,. \tag{8.1.148}$$

Because high-z volume is significantly larger in a Λ-dominated universe than in an Einstein–de Sitter universe, the expected number of gravitationally lensed quasars increases dramatically with $\Omega_{\Lambda 0}$. The observed frequency of lensed quasars in optical and radio surveys (such as CLASS) established an early, independent upper bound $\Omega_{\Lambda 0} < 0.7$, providing critical support for the concordance ΛCDM model.

Lensing by galaxy clusters

Galaxy clusters are the most massive collapsed, virialized structures in the universe ($M \sim 10^{14}$–$10^{15}\, M_\odot$). When an aligned background galaxy lies behind a cluster core, the tremendous surface mass density—often exceeding the critical threshold $\Sigma_{\rm crit}$ over physical radii of several hundred kiloparsecs—generates striking, highly distorted arcs, arclets, and multiple images spanning tens of arcseconds across the sky.

Cluster mass reconstruction Gravitational lensing serves as the ultimate arbiter of cluster mass distributions because it relies purely on the metric distortion of spacetime, requiring no assumptions regarding the dynamical equilibrium or physical state of the gas:

- *Strong lensing core mass:* The radius of giant tangential arcs provides a direct geometrical measure of the cluster's Einstein radius $\theta_{\rm E}$. The projected mass enclosed within the arc is simply:
$$M(<\theta_{\rm E}) = \pi (D_{\rm OL}\theta_{\rm E})^2 \Sigma_{\rm crit}. \tag{8.1.149}$$
For rich clusters like Abell 1689 or the "El Gordo" cluster, $M(<\theta_{\rm E}) \sim 10^{14}\, M_\odot$ within $R \sim 150$–$200\,$kpc.
- *Direct proof of dark matter (The Bullet Cluster):* In merging clusters such as 1E 0657-56 (the Bullet Cluster), gravitational lensing mass maps (derived from weak and strong lensing) reveal that the gravitational potential peaks are centered on the collisionless galaxies and dark matter halos, while X-ray observations (Chandra) show that the collisional baryonic gas (comprising 90% of ordinary matter) was shock-decelerated and separated from the dark halos. This spatial offset provides smoking-gun empirical evidence that the vast majority of cosmic matter is collisionless dark matter rather than modified gravity (MOND).

Magnification bias Gravitational lensing alters the observed spatial density of background sources through two competing physical mechanisms:

1. *Flux amplification:* Lensing magnifies the apparent flux of every background source ($f = \mu f_0$), pushing intrinsically faint sources that would otherwise fall below the survey flux limit $f_{\rm lim}$ into the detected catalog. This tends to increase the observed surface density.
2. *Area dilation:* Lensing expands the celestial patch of solid angle ($\mathrm{d}\Omega = \mu \mathrm{d}\Omega_0$), spreading background sources over a larger sky area. This geometrical stretching dilutes the apparent number density per unit solid angle by a factor of $1/\mu$.

Let $n_0(f, z)\, \mathrm{d}f\, \mathrm{d}z$ be the intrinsic, unlensed number density of galaxies per unit solid angle with flux in the interval $[f, f + \mathrm{d}f]$ and redshift in $[z, z + \mathrm{d}z]$. Since observed flux scales as $f = \mu f_0$ and the solid angle stretches by $\mathrm{d}\Omega = \mu\, \mathrm{d}\Omega_0$, the lensed number density in direction $\boldsymbol{\theta}$ becomes:
$$n(f, z) = \frac{1}{\mu(\boldsymbol{\theta}, z)} n_0 \left(\frac{f}{\mu(\boldsymbol{\theta}, z)}, z \right). \tag{8.1.150}$$
Integrating over redshift and expressing galaxy counts in terms of apparent magnitude $m = -2.5 \log_{10} f + \text{const}$, suppose the cumulative number count follows a local power law with logarithmic slope $\alpha \equiv \mathrm{d} \log_{10} n_0 \,/\, \mathrm{d}m$. The lensed galaxy count evaluates to:
$$n(m, \boldsymbol{\theta}) = n_0(m)\, [\mu(\boldsymbol{\theta})]^{2.5\alpha - 1}. \tag{8.1.151}$$

Derivation of the Magnification Bias Power Law. By definition of the slope α, the unlensed cumulative count scales as $n_0(<m) \propto 10^{\alpha m} \propto f^{-2.5\alpha}$. Under magnification μ, the threshold flux is lowered to f/μ, corresponding to an effective fainter magnitude $m' = m + 2.5\log_{10}\mu$. Applying the cumulative magnification relation:

$$n(<m, \boldsymbol{\theta}) = \frac{1}{\mu} n_0\left(<m + 2.5\log_{10}\mu\right) = \frac{1}{\mu} n_0(<m) 10^{2.5\alpha \log_{10}\mu} = n_0(<m)\mu^{2.5\alpha - 1}. \quad (1)$$

For differential counts with constant α, the identical power-law scaling applies, establishing Eq. (8.1.151).

Equation (8.1.151) encapsulates the phenomenon of *magnification bias*:

- *Critical slope ($\alpha = 0.4$):* When the logarithmic count slope is exactly $\alpha = 0.4$, the power exponent $2.5\alpha - 1 = 0$. Flux amplification and area dilation cancel out perfectly, leaving the observed surface density unaltered ($n = n_0$).
- *Steep count slope ($\alpha > 0.4$):* For bright quasars or submillimeter galaxies whose unlensed luminosity function is very steep, flux boosting overwhelms area dilution ($2.5\alpha - 1 > 0$). Foreground clusters act as positive concentrators, producing an overdensity of background sources correlated with the foreground mass.
- *Shallow count slope ($\alpha < 0.4$):* For faint optical galaxies near typical survey depth limits, the luminosity function flattens ($\alpha \sim 0.1$–0.2). Area dilation wins, creating a physical deficit or "depletion hole" in the background galaxy count behind the cluster center.

Cosmic telescopes and the high-redshift universe By distorting spacetime, massive clusters function as natural cosmic telescopes with magnifications $\mu \sim 10$–100 along critical lines. This natural amplification enables the James Webb Space Telescope (JWST) and Hubble to observe ultra-faint, high-redshift galaxies ($z \sim 7$–13) and individual magnified supergiant stars at cosmological distances that would otherwise lie well beyond the reach of human instrumentation.

8.1.3 Gravitational distortion by the large-scale structure

Beyond localized thin-screen lenses, light rays propagating across the cosmological volume are continuously sheared and focused by the inhomogeneous large-scale matter distribution. Although individual distortions are minute ($\sim 1\%$) and do not produce multiple images, the coherent alignment of background galaxy shapes (cosmic shear) provides an exquisite, unbiased tracer of cosmological mass perturbations.

Sachs equation

We analyze the deformation of a bundle of null geodesics converging at the observer O (see Fig. 8.16). Let $\bar{x}^\mu(\lambda)$ denote the worldline of a reference null geodesic parameterized by affine parameter λ, normalized such that $\lambda = 0$ at the observer O and increases toward the past. Any adjacent ray in the bundle is parameterized as:

$$x^\mu(\lambda) = \bar{x}^\mu(\lambda) + \xi^\mu(\lambda), \quad (8.1.152)$$

where $\xi^\mu(\lambda)$ is the infinitesimal connecting vector (geodesic deviation vector).

At the observer O, we erect an orthonormal Sachs tetrad $\{k^\mu, u^\mu, n_1^\mu, n_2^\mu\}$, where $k^\mu \equiv \mathrm{d}\bar{x}^\mu / \mathrm{d}\lambda$ is the wave vector tangent to the reference ray, u^μ is the 4-velocity of the observer ($u^\mu u_\mu = -1$),

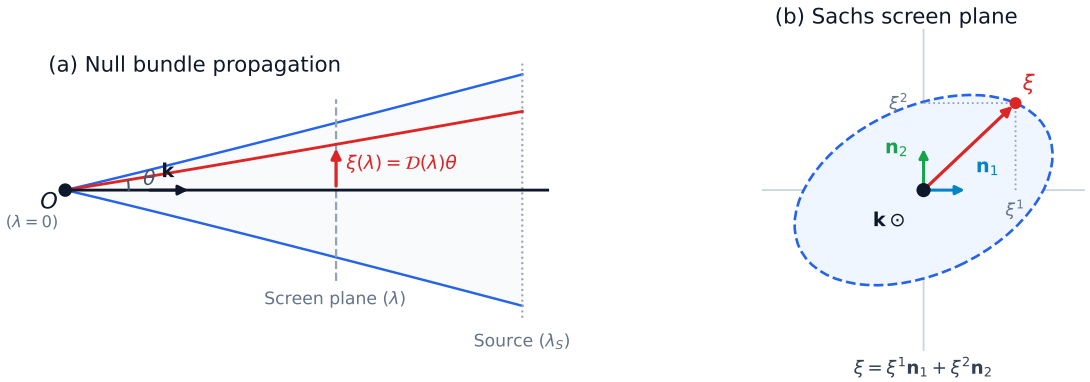


Figure 8.16: Geometric derivation and screen plane projection of the Sachs optical deviation equation. (a) A bundle of null geodesics converging at the observer vertex O ($\lambda = 0$), parameterized by affine parameter λ increasing into the past toward the source plane λ_S. The reference ray has tangent wavevector $\mathbf{k}$, with initial opening angle $\boldsymbol{\theta}$ mapped along the trajectory to the transverse separation $\boldsymbol{\xi}(\lambda) = \mathcal{D}(\lambda)\boldsymbol{\theta}$ by the Jacobi matrix $\mathcal{D}(\lambda)$. (b) Face-on view of the transverse Sachs screen plane orthogonal to $\mathbf{k}$ ($\mathbf{k}\odot$), spanned by the parallel-transported orthonormal basis vectors $\mathbf{n}_1, \mathbf{n}_2$. The transverse deviation vector decomposes as $\boldsymbol{\xi} = \xi^1\mathbf{n}_1 + \xi^2\mathbf{n}_2$, projecting the 4D geodesic deviation into a 2D optical deviation equation on the screen.

and $\{n_a^\mu\}_{a=1,2}$ are two spacelike unit vectors spanning the screen plane orthogonal to the line of sight:

$$k^\mu k_\mu = 0, \qquad u^\mu u_\mu = -1, \qquad n_a^\mu n_{\mu b} = \delta_{ab}, \qquad n_a^\mu k_\mu = n_a^\mu u_\mu = 0. \tag{8.1.153}$$

Parallel transporting the screen vectors along the reference geodesic ($k^\nu \nabla_\nu n_a^\mu = 0$), the Sachs frame remains orthonormal everywhere along the path. Decomposing the separation vector ξ^μ onto this basis:

$$\xi^\mu = \xi^0 k^\mu + \sum_{a=1}^{2} \xi^a n_a^\mu. \tag{8.1.154}$$

Because shifts along k^μ merely correspond to relabeling points along the same null geodesic, we set $\xi^0 = 0$ without loss of generality.

The evolution of ξ^μ is governed by the geodesic deviation equation:

$$\frac{\mathrm{D}^2}{\mathrm{d}\lambda^2}\xi^\mu = R^\mu{}_{\nu\alpha\beta} k^\nu k^\alpha \xi^\beta, \tag{8.1.155}$$

where $\mathrm{D}/\,\mathrm{d}\lambda \equiv k^\nu \nabla_\nu$. Projecting onto the screen vectors $\{n_a^\mu\}$, Eq. (8.1.155) reduces to the two-dimensional Sachs deviation equation:

$$\frac{\mathrm{d}^2\xi^a}{\mathrm{d}\lambda^2} = \mathcal{R}^a{}_b \xi^b, \qquad \mathcal{R}^a{}_b \equiv R^\mu{}_{\nu\alpha\beta} k^\nu k^\alpha n_{a\mu} n_b^\beta, \tag{8.1.156}$$

where $\mathcal{R}^a{}_b(\lambda)$ is the optical tidal matrix.

Decomposing the Riemann tensor into the trace-free Weyl tensor $C_{\mu\nu\alpha\beta}$, the Ricci tensor $R_{\mu\nu}$, and the scalar curvature R, the optical matrix takes the explicit form:

$$\mathcal{R}^a{}_b = -\frac{1}{2} R_{\mu\nu} k^\mu k^\nu \delta^a_b + C^\mu{}_{\nu\alpha\beta} k^\nu k^\alpha n_{a\mu} n_b^\beta. \tag{8.1.157}$$

Physical Decomposition of the Optical Matrix. As illustrated in Fig. 8.7, the two terms in Eq. (8.1.157) have profoundly different physical origins:

1. **Ricci Focusing**: The trace term $-\frac{1}{2}R_{\mu\nu}k^\mu k^\nu \delta^a_b = -4\pi G_{\rm N} T_{\mu\nu}k^\mu k^\nu \delta^a_b$ is sourced directly by the local stress-energy tensor traversed by the photon beam. By Einstein's equations, ordinary matter ($T_{\mu\nu}k^\mu k^\nu \geq 0$) generates an attractive, isotropic convergence that focuses rays, reducing beam cross-section without introducing shear. This directly sources cosmological convergence κ.
2. **Weyl Tidal Shearing**: The trace-free term $C^\mu{}_{\nu\alpha\beta}k^\nu k^\alpha n_{a\mu}n^\beta_b$ represents the gravitational tidal field generated by mass concentrations situated outside the light bundle. It induces an anisotropic, volume-preserving distortion of the beam cross-section, transforming circular wavefronts into ellipses. This directly sources cosmic shear $\boldsymbol{\gamma}$.

To solve the Sachs deviation equation (8.1.156), we trace light rays backward from the observer (O) at $\lambda = 0$ toward the source plane. Because all observed rays emanate from a single vertex point at the observer, the bundle satisfies the initial boundary condition:

$$\xi^a(0) = 0. \tag{8.1.158}$$

The initial rate of transverse ray separation per unit affine parameter defines the observed celestial angular position vector:

$$\left.\frac{\mathrm{d}\xi^a}{\mathrm{d}\lambda}\right|_0 \equiv \theta^a. \tag{8.1.159}$$

Because the Sachs equation is a linear, homogeneous second-order differential equation for $\xi^a(\lambda)$, any solution satisfying the vertex condition $\xi^a(0) = 0$ is uniquely determined by and linearly proportional to the initial angular direction θ^b:

$$\xi^a(\lambda) = \mathcal{D}^a{}_b(\lambda)\theta^b, \tag{8.1.160}$$

where $\mathcal{D}^a{}_b(\lambda)$ is the 2×2 Jacobi matrix mapping the observed sky angle to the transverse separation of the beam at affine distance λ.

Substituting Eq. (8.1.160) directly into the Sachs equation (8.1.156):

$$\left[\frac{\mathrm{d}^2\mathcal{D}^a{}_b(\lambda)}{\mathrm{d}\lambda^2} - \mathcal{R}^a{}_c(\lambda)\mathcal{D}^c{}_b(\lambda)\right]\theta^b = 0. \tag{8.1.161}$$

Because this relation must hold for any arbitrary incoming ray direction θ^b, the Jacobi matrix itself satisfies the Sachs matrix differential equation:

$$\frac{\mathrm{d}^2\mathcal{D}^a{}_b}{\mathrm{d}\lambda^2} = \mathcal{R}^a{}_c\mathcal{D}^c{}_b, \tag{8.1.162}$$

subject to the initial conditions at the observer vertex O ($\lambda = 0$):

$$\mathcal{D}^a{}_b(0) = 0, \qquad \left.\frac{\mathrm{d}\mathcal{D}^a{}_b}{\mathrm{d}\lambda}\right|_0 = \delta^a_b. \tag{8.1.163}$$

Integrating Eq. (8.1.156) directly twice from $\lambda = 0$ to λ with these initial conditions yields the equivalent Volterra integral equation:

$$\xi^a(\lambda) = \lambda\theta^a + \int_0^\lambda \mathrm{d}\lambda'(\lambda - \lambda')\mathcal{R}^a{}_b(\lambda')\xi^b(\lambda'). \tag{8.1.164}$$

In flat spacetime ($\mathcal{R}^a{}_b = 0$), the integral vanishes and rays propagate straight, $\xi^a(\lambda) = \lambda\theta^a$. The integral term thus isolates the cumulative transverse gravitational deflection accumulated along the line of sight, with $(\lambda - \lambda')$ acting as the geometric lever arm.

At the source plane $\lambda = \lambda_S$, the true unlensed angular position θ_S^a of the source is defined by dividing the physical separation $\xi^a(\lambda_S)$ by the background angular diameter distance $D_A(\lambda_S)$:

$$\theta_S^a \equiv \frac{\xi^a(\lambda_S)}{D_A(\lambda_S)}. \tag{8.1.165}$$

Substituting the Jacobi solution (8.1.160) into Eq. (8.1.165) relates the true source position linearly to the observed image position:

$$\theta_S^a = \mathcal{A}^a{}_b \theta^b, \qquad \mathcal{A}^a{}_b \equiv \frac{\mathcal{D}^a{}_b(\lambda_S)}{D_A(\lambda_S)}, \tag{8.1.166}$$

where $\mathcal{A}^a{}_b$ is the cosmological deformation (amplification) matrix decomposed into convergence κ and cosmic shear γ_{ab} according to Eq. (8.1.90).

Application to cosmology

Homogeneous and isotropic background In an unperturbed Friedmann–Lemaître spacetime with spatial curvature $K \in \{-1, 0, +1\}$, the metric is conformal to static spacetime: $\mathrm{d}s^2 = a^2(\eta)[-\,\mathrm{d}\eta^2 + \gamma_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j]$. Because conformal transformations preserve null geodesics, the unperturbed optical matrix reduces to pure spatial curvature:

$$\mathcal{R}^{(0)} = -K\mathcal{I}. \tag{8.1.167}$$

The Sachs matrix differential equation (8.1.162) reduces to the scalar radial distance equation:

$$\frac{\mathrm{d}^2}{\mathrm{d}\chi^2} f_K(\chi) + K f_K(\chi) = 0, \qquad f_K(0) = 0, \quad f_K'(0) = 1, \tag{8.1.168}$$

whose solution is the comoving angular diameter distance:

$$f_K(\chi) = \begin{cases} \dfrac{1}{\sqrt{K}} \sin\!\left(\sqrt{K}\chi\right), & (K > 0), \\ \chi, & (K = 0), \\ \dfrac{1}{\sqrt{-K}} \sinh\!\left(\sqrt{-K}\chi\right), & (K < 0). \end{cases} \tag{8.1.169}$$

Perturbed cosmological metric Introducing scalar metric perturbations in the longitudinal gauge:

$$\mathrm{d}s^2 = a^2(\eta)\left[-(1+2\Psi)\,\mathrm{d}\eta^2 + (1-2\Phi)\gamma_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j\right], \tag{8.1.170}$$

where Φ and Ψ are the gauge-invariant Bardeen potentials. In the absence of anisotropic stress (standard collisionless cold dark matter and baryonic gas), Einstein's equations require $\Phi = \Psi$.

Integrating the Sachs equation to first order in the metric perturbations yields the explicit line-of-sight expression for the magnification matrix:

$$\mathcal{A}_{ab}(\boldsymbol{\theta}) = \delta_{ab} - \frac{2}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\, \frac{f_K(\chi) f_K(\chi_s - \chi)}{f_K(\chi_s)} \frac{\partial^2 \Phi}{\partial x^a \partial x^b}(f_K(\chi)\boldsymbol{\theta}, \chi). \tag{8.1.171}$$

Equivalently, expressing the tidal distortion in terms of angular coordinates $\boldsymbol{\theta}$ on the observer's sky, the geometric weight transfers $f_K(\chi)$ from the numerator to the denominator:

$$\mathcal{A}_{ab}(\boldsymbol{\theta}) = \delta_{ab} - \frac{2}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\, \frac{f_K(\chi_s - \chi)}{f_K(\chi_s) f_K(\chi)} \frac{\partial^2 \Phi}{\partial \theta^a \partial \theta^b}(f_K(\chi)\boldsymbol{\theta}, \chi), \tag{8.1.172}$$

from which the effective projected lensing potential in curved spacetime is identified as:

$$\psi(\boldsymbol{\theta}) = \frac{2}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\, \frac{f_K(\chi_s - \chi)}{f_K(\chi_s) f_K(\chi)} \Phi(f_K(\chi)\boldsymbol{\theta}, \chi). \tag{8.1.173}$$

Derivation of the Angular Magnification Matrix and Lensing Potential

The evolution of the 2×2 Jacobi deviation matrix $\mathcal{D}(\chi)$ along a null geodesic is governed by the Sachs matrix differential equation [Eq. (8.1.162)]:

$$\frac{\mathrm{d}^2}{\mathrm{d}\chi^2}\mathcal{D} = \mathcal{R} \cdot \mathcal{D}, \qquad \mathcal{D}(0) = 0, \quad \frac{\mathrm{d}\mathcal{D}}{\mathrm{d}\chi}(0) = \mathcal{I}. \tag{1}$$

In an unperturbed FLRW background with spatial curvature $K \in \{-1, 0, +1\}$, the optical tidal matrix reduces to $\mathcal{R}^{(0)} = -K\mathcal{I}$ [Eq. (8.1.167)], yielding the background solution $\mathcal{D}^{(0)}(\chi) = f_K(\chi)\mathcal{I}$. Under scalar metric perturbations $\Phi = \Psi$ in the longitudinal gauge [Eq. (8.1.170)], the screen-projected Riemann curvature perturbation generates the tidal perturbation:

$$\delta\mathcal{R}_{ab}(\chi) = -\frac{2}{c^2}\frac{\partial^2\Phi}{\partial x^a \partial x^b}(f_K(\chi)\boldsymbol{\theta}, \chi), \tag{2}$$

where $\boldsymbol{x}_\perp = (x^1, x^2)$ are transverse comoving Cartesian coordinates on the screen. The first-order perturbation $\delta\mathcal{D}(\chi) \equiv \mathcal{D}(\chi) - \mathcal{D}^{(0)}(\chi)$ satisfies:

$$\left(\frac{\mathrm{d}^2}{\mathrm{d}\chi^2} + K\right)\delta\mathcal{D}(\chi) = \delta\mathcal{R}(\chi) \cdot \mathcal{D}^{(0)}(\chi) = -\frac{2}{c^2} f_K(\chi)\frac{\partial^2\Phi}{\partial x^a \partial x^b}\mathcal{I}. \tag{3}$$

Integrating from the observer ($\chi = 0$) to the source plane ($\chi = \chi_s$) using the retarded Green's function $G(\chi, \chi') = f_K(\chi - \chi')$ gives:

$$\begin{aligned}\delta\mathcal{D}_{ab}(\chi_s) &= \int_0^{\chi_s} \mathrm{d}\chi\, f_K(\chi_s - \chi)\delta\mathcal{R}_{ac}(\chi)\mathcal{D}^{(0)}_{cb}(\chi) \\ &= -\frac{2}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\, f_K(\chi_s - \chi) f_K(\chi)\frac{\partial^2\Phi}{\partial x^a \partial x^b}(f_K(\chi)\boldsymbol{\theta}, \chi).\end{aligned} \tag{4}$$

Dividing by the background angular diameter distance $D_A(\chi_s) = f_K(\chi_s)$ via $\mathcal{A}_{ab}(\boldsymbol{\theta}) \equiv \mathcal{D}_{ab}(\chi_s)/f_K(\chi_s) = \delta_{ab} + \delta\mathcal{D}_{ab}(\chi_s)/f_K(\chi_s)$ yields Eq. (8.1.171):

$$\mathcal{A}_{ab}(\boldsymbol{\theta}) = \delta_{ab} - \frac{2}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\, \frac{f_K(\chi) f_K(\chi_s - \chi)}{f_K(\chi_s)}\frac{\partial^2\Phi}{\partial x^a \partial x^b}(f_K(\chi)\boldsymbol{\theta}, \chi). \tag{5}$$

To express the magnification matrix in terms of angular coordinates $\boldsymbol{\theta} = (\theta^1, \theta^2)$ on the celestial sphere, we transform the transverse spatial derivatives. Because transverse physical separations at comoving distance χ project to angular separations via $\boldsymbol{x}_\perp = f_K(\chi)\boldsymbol{\theta}$, the coordinates satisfy $x^a = f_K(\chi)\theta^a$. The chain rule relates the differential operators according to:

$$\frac{\partial}{\partial x^a} = \frac{\partial\theta^c}{\partial x^a}\frac{\partial}{\partial\theta^c} = \frac{1}{f_K(\chi)}\frac{\partial}{\partial\theta^a}, \tag{6}$$

so that the second derivatives transform as:

$$\frac{\partial^2\Phi}{\partial x^a \partial x^b} = \frac{1}{f_K^2(\chi)}\frac{\partial^2\Phi}{\partial\theta^a \partial\theta^b}. \tag{7}$$

Substituting this transformation into Eq. (8.1.171):

$$\begin{aligned}\mathcal{A}_{ab}(\boldsymbol{\theta}) &= \delta_{ab} - \frac{2}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\, \frac{f_K(\chi) f_K(\chi_s - \chi)}{f_K(\chi_s)}\left[\frac{1}{f_K^2(\chi)}\frac{\partial^2\Phi}{\partial\theta^a \partial\theta^b}(f_K(\chi)\boldsymbol{\theta}, \chi)\right] \\ &= \delta_{ab} - \frac{2}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\, \frac{f_K(\chi_s - \chi)}{f_K(\chi_s) f_K(\chi)}\frac{\partial^2\Phi}{\partial\theta^a \partial\theta^b}(f_K(\chi)\boldsymbol{\theta}, \chi).\end{aligned} \tag{8}$$

One factor of $f_K(\chi)$ in the numerator cancels against $f_K^2(\chi)$ from the second derivative, moving $f_K(\chi)$ to the denominator.
Because the integration bounds and the radial distance χ do not depend on the celestial position $\boldsymbol{\theta}$, the angular differential operator factors outside the line-of-sight integral:

$$\mathcal{A}_{ab}(\boldsymbol{\theta}) = \delta_{ab} - \frac{\partial^2}{\partial\theta^a\partial\theta^b}\left[\frac{2}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\, \frac{f_K(\chi_s-\chi)}{f_K(\chi_s)f_K(\chi)}\Phi(f_K(\chi)\boldsymbol{\theta},\chi)\right]. \tag{9}$$

Comparing this expression directly with the definition of the magnification matrix in terms of the projected lensing potential, $\mathcal{A}_{ab}(\boldsymbol{\theta}) = \delta_{ab} - \frac{\partial^2\psi}{\partial\theta^a\partial\theta^b}$ [Eq. (8.1.90)], establishes the effective lensing potential in curved spacetime:

$$\psi(\boldsymbol{\theta}) = \frac{2}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\, \frac{f_K(\chi_s-\chi)}{f_K(\chi_s)f_K(\chi)}\Phi(f_K(\chi)\boldsymbol{\theta},\chi), \tag{10}$$

which reduces in a flat universe ($K=0$, $f_K(\chi)=\chi$) to:

$$\psi(\boldsymbol{\theta}) = \frac{2}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\, \frac{\chi_s-\chi}{\chi_s\chi}\Phi(\chi\boldsymbol{\theta},\chi). \tag{11}$$

The true deflection angle $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \boldsymbol{\nabla}_\theta\psi(\boldsymbol{\theta})$ is therefore:

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \frac{2}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\, \frac{f_K(\chi_s-\chi)}{f_K(\chi_s)f_K(\chi)}\boldsymbol{\nabla}_\theta\Phi(f_K(\chi)\boldsymbol{\theta},\chi). \tag{12}$$

8.1.4 Cosmic convergence

Effective convergence

Taking the trace of the distortion matrix (8.1.171), the effective cosmological convergence along a light ray directed toward angular coordinate $\boldsymbol{\theta}$ on the sky is:

$$\kappa(\boldsymbol{\theta}) = \frac{1}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\, \frac{f_K(\chi)f_K(\chi_s-\chi)}{f_K(\chi_s)}\nabla_\perp^2\Phi(f_K(\chi)\boldsymbol{\theta},\chi). \tag{8.1.174}$$

Using the cosmological Poisson equation $\nabla^2\Phi = 4\pi G_\mathrm{N}a^2\bar{\rho}_\mathrm{m}\delta = \frac{3}{2}H_0^2\Omega_{\mathrm{m}0}\frac{\delta}{a}$, and noting that transverse gradients dominate over longitudinal gradients for small angular scales, the effective convergence evaluates to:

$$\kappa(\boldsymbol{\theta}) = \frac{3H_0^2\Omega_{\mathrm{m}0}}{2c^2}\int_0^{\chi_s} \mathrm{d}\chi\, \frac{f_K(\chi)f_K(\chi_s-\chi)}{f_K(\chi_s)a(\chi)}\delta(f_K(\chi)\boldsymbol{\theta},\chi). \tag{8.1.175}$$

For a broad redshift distribution of background source galaxies with normalized selection function $p(z)\,\mathrm{d}z = n(\chi)\,\mathrm{d}\chi$ such that $\int_0^{\chi_H} n(\chi)\,\mathrm{d}\chi = 1$, the projected convergence can be expressed as a continuous line-of-sight integral:

$$\kappa(\boldsymbol{\theta}) = \int_0^{\chi_H} \mathrm{d}\chi\, W(\chi)\delta(f_K(\chi)\boldsymbol{\theta},\chi), \tag{8.1.176}$$

where the *lensing efficiency kernel* $W(\chi)$ is defined by:

$$W(\chi) \equiv \frac{3H_0^2\Omega_{\mathrm{m}0}}{2c^2}\frac{f_K(\chi)}{a(\chi)}\int_\chi^{\chi_H} \mathrm{d}\chi'\, n(\chi')\frac{f_K(\chi'-\chi)}{f_K(\chi')}. \tag{8.1.177}$$

For a flat universe ($f_K(\chi)=\chi$):

$$W(\chi) = \frac{3H_0^2\Omega_{\mathrm{m}0}}{2c^2}\frac{\chi}{a(\chi)}\int_\chi^{\chi_H} \mathrm{d}\chi'\, n(\chi')\frac{\chi'-\chi}{\chi'}. \tag{8.1.178}$$

Limber equation for the convergence power spectrum

In the flat-sky approximation ($\theta \ll 1\,\mathrm{rad}$), the convergence field $\kappa(\boldsymbol{\theta})$ is expanded into two-dimensional Fourier modes:

$$\kappa(\boldsymbol{\theta}) = \int \frac{\mathrm{d}^2\ell}{(2\pi)^2}\hat{\kappa}(\boldsymbol{\ell})e^{i\boldsymbol{\ell}\cdot\boldsymbol{\theta}}. \tag{8.1.179}$$

Statistical isotropy and homogeneity imply that the Fourier modes satisfy:

$$\langle\hat{\kappa}(\boldsymbol{\ell})\hat{\kappa}^*(\boldsymbol{\ell}')\rangle = (2\pi)^2\delta^{(2)}(\boldsymbol{\ell}-\boldsymbol{\ell}')P_\kappa(\ell). \tag{8.1.180}$$

Using the line-of-sight representation $\kappa(\boldsymbol{\theta}) = \int_0^{\chi_H}\mathrm{d}\chi\, W(\chi)\delta(f_K(\chi)\boldsymbol{\theta},\chi)$ [Eq. (8.1.176)] with the efficiency kernel $W(\chi)$ defined in Eq. (8.1.177), the two-point angular correlation function $\xi_\kappa(\theta) \equiv \langle\kappa(\boldsymbol{\phi})\kappa(\boldsymbol{\phi}+\boldsymbol{\theta})\rangle$ is:

$$\xi_\kappa(\theta) = \int_0^{\chi_H}\mathrm{d}\chi\int_0^{\chi_H}\mathrm{d}\chi'\, W(\chi)W(\chi')\left\langle\delta(f_K(\chi)\boldsymbol{\phi},\chi)\delta(f_K(\chi')(\boldsymbol{\phi}+\boldsymbol{\theta}),\chi')\right\rangle. \tag{8.1.181}$$

Expanding the 3D density contrast into spatial Fourier modes $\delta(\boldsymbol{x},\chi) = \int\frac{\mathrm{d}^3k}{(2\pi)^3}\delta(\boldsymbol{k},\chi)e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$, the two-point expectation value inside the integral evaluates to:

$$\int\frac{\mathrm{d}^3k\mathrm{d}^3k'}{(2\pi)^6}e^{i\boldsymbol{k}_\perp\cdot\boldsymbol{\phi}f_K(\chi)}e^{ik_\parallel\chi}e^{-i\boldsymbol{k}'_\perp\cdot(\boldsymbol{\phi}+\boldsymbol{\theta})f_K(\chi')}e^{-ik'_\parallel\chi'}\left\langle\delta(\boldsymbol{k},\chi)\delta^*(\boldsymbol{k}',\chi')\right\rangle. \tag{8.1.182}$$

The physical intuition underlying Limber's approximation is a line-of-sight phase cancellation. As photons traverse gigaparsecs through the cosmic web, they pierce thousands of independent overdensities and underdensities. Any 3D perturbation mode with a non-zero radial wavevector component ($k_\parallel \neq 0$) oscillates rapidly along the light path, causing successive crests and troughs to cancel when integrated over comoving distance: $\int\mathrm{d}\chi\, e^{ik_\parallel\chi} \approx 2\pi\delta(k_\parallel)$.

Only perturbations oriented nearly perpendicular to the line of sight ($k_\parallel \approx 0$) maintain a constant sign along the path and accumulate coherently into a net observable deflection. Consequently, the 3D wavenumber reduces to the transverse wavenumber: $k = \sqrt{k_\perp^2 + k_\parallel^2} \approx k_\perp = \ell/f_K(\chi)$.

Using the statistical orthogonality of homogeneous 3D modes, $\langle\delta(\boldsymbol{k},\chi)\delta^*(\boldsymbol{k}',\chi')\rangle = (2\pi)^3\delta^{(3)}(\boldsymbol{k}-\boldsymbol{k}')P_\delta(k,\chi)$, the integration over $k_\parallel$ collapses the double radial integral into a single comoving distance integral:

$$\xi_\kappa(\theta) = \int_0^{\chi_H}\mathrm{d}\chi\,\frac{W^2(\chi)}{f_K^2(\chi)}\int\frac{\mathrm{d}^2\ell}{(2\pi)^2}P_\delta\left(\frac{\ell}{f_K(\chi)},\chi\right)e^{i\boldsymbol{\ell}\cdot\boldsymbol{\theta}}. \tag{8.1.183}$$

Comparing Eq. (8.1.183) with the 2D angular Fourier transform $\xi_\kappa(\theta) = \int\frac{\mathrm{d}^2\ell}{(2\pi)^2}P_\kappa(\ell)e^{i\boldsymbol{\ell}\cdot\boldsymbol{\theta}}$ directly yields the canonical Limber projection equation for the convergence angular power spectrum:

$$P_\kappa(\ell) = \int_0^{\chi_H}\mathrm{d}\chi\,\frac{W^2(\chi)}{f_K^2(\chi)}P_\delta\left(\frac{\ell}{f_K(\chi)},\chi\right) = \frac{9H_0^4\Omega_{\mathrm{m}0}^2}{4c^4}\int_0^{\chi_H}\mathrm{d}\chi\left[\frac{g(\chi)}{a(\chi)}\right]^2 P_\delta\left(\frac{\ell}{f_K(\chi)},\chi\right), \tag{8.1.184}$$

where $g(\chi) \equiv \int_\chi^{\chi_H}\mathrm{d}\chi'\, n(\chi')\frac{f_K(\chi'-\chi)}{f_K(\chi')}$ denotes the normalized background source distance factor.

Smoothing the convergence field with an azimuthally symmetric aperture filter $U(\theta,\theta_c)$ of characteristic angular radius θ_c:

$$\kappa(\theta_c) = \int\kappa(\boldsymbol{\theta}')U(|\boldsymbol{\theta}'|,\theta_c)\,\mathrm{d}^2\theta', \tag{8.1.185}$$

the aperture variance of the smoothed convergence field evaluates to:

$$\langle\kappa^2\rangle(\theta_c) = 2\pi\int_0^\infty \ell\,\mathrm{d}\ell\,P_\kappa(\ell)\left[\int\theta U(\theta,\theta_c)J_0(\ell\theta)\,\mathrm{d}\theta\right]^2, \tag{8.1.186}$$

where $J_0(x)$ is the zeroth-order Bessel function of the first kind.

Derivation of the Limber Projection Equation

Let the 3D density contrast $\delta(\boldsymbol{x}, \chi)$ be expanded in spatial Fourier modes:

$$\delta(\boldsymbol{x}, \chi) = \int \frac{\mathrm{d}^3 k}{(2\pi)^3} \tilde{\delta}(\boldsymbol{k}, \chi) e^{i\boldsymbol{k}\cdot\boldsymbol{x}}, \tag{1}$$

satisfying the statistical orthogonality relation:

$$\langle \tilde{\delta}(\boldsymbol{k}, \chi) \tilde{\delta}^*(\boldsymbol{k}', \chi') \rangle = (2\pi)^3 \delta^{(3)}(\boldsymbol{k} - \boldsymbol{k}') P_\delta(k, \chi, \chi'). \tag{2}$$

Writing $\boldsymbol{x} = (\chi\boldsymbol{\theta}, \chi)$, the 2D Fourier transform of the convergence field (8.1.176) is:

$$\begin{aligned} \hat{\kappa}(\boldsymbol{\ell}) &= \int \mathrm{d}^2\theta\, \kappa(\boldsymbol{\theta}) e^{-i\boldsymbol{\ell}\cdot\boldsymbol{\theta}} \\ &= \int \mathrm{d}^2\theta \int_0^{\chi_H} \mathrm{d}\chi\, W(\chi) \int \frac{\mathrm{d}^2 k_\perp\, \mathrm{d}k_\parallel}{(2\pi)^3} \tilde{\delta}(\boldsymbol{k}_\perp, k_\parallel, \chi) e^{i\chi\boldsymbol{k}_\perp\cdot\boldsymbol{\theta}} e^{ik_\parallel \chi} e^{-i\boldsymbol{\ell}\cdot\boldsymbol{\theta}}. \end{aligned} \tag{3}$$

Carrying out the 2D angular integration $\int \mathrm{d}^2\theta e^{i(\chi\boldsymbol{k}_\perp - \boldsymbol{\ell})\cdot\boldsymbol{\theta}} = (2\pi)^2 \delta^{(2)}(\chi\boldsymbol{k}_\perp - \boldsymbol{\ell}) = \frac{(2\pi)^2}{\chi^2}\delta^{(2)}\left(\boldsymbol{k}_\perp - \frac{\boldsymbol{\ell}}{\chi}\right)$:

$$\hat{\kappa}(\boldsymbol{\ell}) = \int_0^{\chi_H} \mathrm{d}\chi\, \frac{W(\chi)}{\chi^2} \int \frac{\mathrm{d}k_\parallel}{2\pi} \tilde{\delta}\left(\frac{\boldsymbol{\ell}}{\chi}, k_\parallel, \chi\right) e^{ik_\parallel \chi}. \tag{4}$$

Now computing the ensemble two-point correlation $\langle \hat{\kappa}(\boldsymbol{\ell}) \hat{\kappa}^*(\boldsymbol{\ell}') \rangle$:

$$\begin{aligned} \langle \hat{\kappa}(\boldsymbol{\ell}) \hat{\kappa}^*(\boldsymbol{\ell}') \rangle = \int_0^{\chi_H} \mathrm{d}\chi\, \frac{W(\chi)}{\chi^2} \int_0^{\chi_H} \mathrm{d}\chi'\, \frac{W(\chi')}{\chi'^2} \int \frac{\mathrm{d}k_\parallel}{2\pi} \int \frac{\mathrm{d}k'_\parallel}{2\pi} e^{i(k_\parallel \chi - k'_\parallel \chi')} \\ \times (2\pi)^3 \delta^{(2)}\left(\frac{\boldsymbol{\ell}}{\chi} - \frac{\boldsymbol{\ell}'}{\chi'}\right) \delta(k_\parallel - k'_\parallel) P_\delta\left(\sqrt{\frac{\ell^2}{\chi^2} + k_\parallel^2}, \chi, \chi'\right). \end{aligned} \tag{5}$$

Evaluating the $k'_\parallel$ integral via $\delta(k_\parallel - k'_\parallel)$, the exponential phase factor becomes $e^{ik_\parallel(\chi - \chi')}$. We now transform the radial integration variables to the center-of-mass and separation coordinates:

$$\bar{\chi} \equiv \frac{\chi + \chi'}{2}, \qquad \Delta\chi \equiv \chi - \chi'. \tag{6}$$

Because the matter power spectrum $P_\delta(k, \chi, \chi')$ is correlated only over small spatial separations ($\Delta\chi \lesssim 50\,\mathrm{Mpc}$), whereas the geometric kernel $W(\chi)$ varies over gigaparsec cosmological scales, we can evaluate the smooth functions at $\bar{\chi}$ and integrate over $\Delta\chi$:

$$\int_{-\infty}^{\infty} \mathrm{d}\Delta\chi\, e^{ik_\parallel \Delta\chi} = 2\pi\delta_{\mathrm{D}}(k_\parallel). \tag{7}$$

The Dirac delta function $\delta_{\mathrm{D}}(k_\parallel)$ collapses the parallel wavevector integration to $k_\parallel = 0$. Consequently, only transverse modes contribute:

$$k = \sqrt{k_\perp^2 + k_\parallel^2} \xrightarrow{k_\parallel = 0} k_\perp = \frac{\ell}{\bar{\chi}}. \tag{8}$$

At equal times $\chi = \chi' = \bar{\chi}$, the unequal-time power spectrum reduces to the equal-time matter power spectrum: $P_\delta(k, \bar{\chi}, \bar{\chi}) = P_\delta(k, \bar{\chi})$. Furthermore, the transverse delta function scales as:

$$\delta^{(2)}\left(\frac{\boldsymbol{\ell}}{\chi} - \frac{\boldsymbol{\ell}'}{\chi'}\right) \approx \delta^{(2)}\left(\frac{\boldsymbol{\ell} - \boldsymbol{\ell}'}{\bar{\chi}}\right) = \bar{\chi}^2 \delta^{(2)}(\boldsymbol{\ell} - \boldsymbol{\ell}'). \tag{9}$$

The factor of $\bar{\chi}^2$ cancels two powers of $\bar{\chi}$ in the denominator $\bar{\chi}^4$, leaving:

$$\langle \hat{\kappa}(\boldsymbol{\ell})\hat{\kappa}^*(\boldsymbol{\ell}')\rangle = (2\pi)^2\delta^{(2)}(\boldsymbol{\ell}-\boldsymbol{\ell}')\int_0^{\chi_H} \mathrm{d}\bar{\chi}\,\frac{W^2(\bar{\chi})}{\bar{\chi}^2}P_\delta\left(\frac{\ell}{\bar{\chi}},\bar{\chi}\right). \tag{10}$$

Relabeling the integration variable $\bar{\chi}\to\chi$ and comparing with Eq. (8.1.180), we establish the canonical *Limber projection equation*:

$$P_\kappa(\ell) = \int_0^{\chi_H} \mathrm{d}\chi\,\frac{W^2(\chi)}{\chi^2}P_\delta\left(\frac{\ell}{\chi},\chi\right). \tag{11}$$

Non-linear gravitational clustering (modeled via Halofit or halo-model prescriptions) substantially boosts the convergence power spectrum at multipoles $\ell \gtrsim 300$ (corresponding to sub-degree angular scales $\theta \lesssim 30'$), highlighting that weak lensing surveys probe deep into the non-linear regime of structure formation. The corresponding real-space two-point shear correlation functions $\xi_\pm(\theta)$, aperture mass variance $\langle M_{\rm ap}^2(\theta)\rangle$, and tomographic power spectra are developed systematically in Section 8.1.5, Section 8.1.5, and Section 8.1.5 (see Figs. 9.13, 9.14, and 9.15).

Full-sky versus flat-sky derivations

On wide-angle surveys covering thousands of square degrees, the flat-sky approximation must be replaced by a spherical harmonic decomposition on the celestial 2-sphere. Rewriting Eq. (8.1.177) on the sphere: $\kappa(\hat{\boldsymbol{n}}) = \frac{1}{2}\Delta_{S^2}\tilde{\psi}(\hat{\boldsymbol{n}})$, where Δ_{S^2} is the Laplace–Beltrami operator on S^2, the line-of-sight integrated potential is:

$$\tilde{\psi}(\hat{\boldsymbol{n}}) = 2\int_0^{\chi} \hat{g}(\chi')\Phi[\boldsymbol{x}(\hat{\boldsymbol{n}}),\chi']\,\mathrm{d}\chi', \tag{8.1.187}$$

with $\hat{g}(\chi)\equiv g(\chi)/f_K(\chi)$. Expanding $\tilde{\psi}(\hat{\boldsymbol{n}}) = \sum_{\ell m}\psi_{\ell m}Y_{\ell m}(\hat{\boldsymbol{n}})$ and decomposing the 3D potential into plane waves via the spherical Rayleigh expansion $e^{i\boldsymbol{k}\cdot\boldsymbol{x}} = 4\pi\sum_{\ell m} i^\ell j_\ell(k\chi)Y^*_{\ell m}(\hat{\boldsymbol{k}})Y_{\ell m}(\hat{\boldsymbol{n}})$:

$$\psi_{\ell m} = 8\pi i^\ell \int_0^{\chi}\mathrm{d}\chi\,\hat{g}(\chi)j_\ell(k\chi)\int\frac{\mathrm{d}^3k}{(2\pi)^{3/2}}\Phi(\boldsymbol{k},\chi)Y^*_{\ell m}(\hat{\boldsymbol{k}}). \tag{8.1.188}$$

The angular power spectrum of the deflection potential, defined by $\langle\psi_{\ell m}\psi^*_{\ell' m'}\rangle = \delta_{\ell\ell'}\delta_{mm'}C_\ell^{\psi\psi}$, evaluates to:

$$C_\ell^{\psi\psi} = 16\pi\int\mathrm{d}\chi\,\mathrm{d}\chi'\,\hat{g}(\chi)\hat{g}(\chi')\int\frac{\mathrm{d}k}{k}j_\ell(k\chi)j_\ell(k\chi')P_\Phi(k,\chi,\chi'). \tag{8.1.189}$$

Because spherical harmonics are eigenfunctions of the Laplacian on the 2-sphere, $\nabla^2_{S^2}Y_{\ell m} = -\ell(\ell+1)Y_{\ell m}$, the multipole coefficients of convergence satisfy $\kappa_{\ell m} = -\frac{1}{2}\ell(\ell+1)\psi_{\ell m}$, establishing the exact full-sky relation:

$$C_\ell^{\kappa\kappa} = \frac{\ell^2(\ell+1)^2}{4}C_\ell^{\psi\psi}. \tag{8.1.190}$$

For $\ell \gg 1$, the spherical Bessel function $j_\ell(x)$ oscillates with extreme rapidity everywhere except in the immediate vicinity of the turning point $x\simeq\ell+1/2$. In this limit, it satisfies the asymptotic closure identity:

$$j_\ell(k\chi) \approx \sqrt{\frac{\pi}{2\ell+1}}\delta_{\rm D}\left(\ell+\frac{1}{2}-k\chi\right). \tag{8.1.191}$$

Substituting this asymptotic relation into Eq. (8.1.189) directly eliminates the k-integration, confirming that the full-sky spherical harmonic spectrum $C_\ell^{\kappa\kappa}$ converges asymptotically to the flat-sky Limber spectrum $P_\kappa(\ell)$ with fractional error $\mathcal{O}(1/\ell^2)$:

$$C_\ell^{\kappa\kappa} \xrightarrow{\ell\gg1} P_\kappa(\ell) = \int_0^{\chi_H}\mathrm{d}\chi\,\frac{W^2(\chi)}{\chi^2}P_\delta\left(\frac{\ell+1/2}{\chi},\chi\right). \tag{8.1.192}$$

8.1.5 Cosmic shear

Kaiser and Squires algorithm

Because convergence $\kappa = \frac{1}{2}(\psi_{,11} + \psi_{,22})$ and shear components $(\gamma_1, \gamma_2) = \left(\frac{1}{2}(\psi_{,11} - \psi_{,22}), \psi_{,12}\right)$ are linear combinations of the second derivatives of the same scalar potential $\psi(\boldsymbol{\theta})$, the shear field uniquely determines the projected mass distribution. Kaiser and Squires (1993) introduced a non-local inversion algorithm relating complex shear $\gamma(\boldsymbol{\theta}) \equiv \gamma_1(\boldsymbol{\theta}) + i\gamma_2(\boldsymbol{\theta})$ directly to the convergence:

$$\gamma(\boldsymbol{\theta}) = \frac{1}{\pi} \int \mathcal{D}(\boldsymbol{\theta} - \boldsymbol{\theta}')\kappa(\boldsymbol{\theta}')\, \mathrm{d}^2\theta', \tag{8.1.193}$$

where the complex convolution kernel $\mathcal{D}(\boldsymbol{\theta})$ represents the response of a point-mass deflector in angular coordinates $\boldsymbol{\theta} = (\theta_1, \theta_2)$ with norm $\theta \equiv |\boldsymbol{\theta}| = \sqrt{\theta_1^2 + \theta_2^2}$:

$$\mathcal{D}(\boldsymbol{\theta}) \equiv \frac{\theta_2^2 - \theta_1^2 - 2i\theta_1\theta_2}{\theta^4} = -\frac{1}{(\theta_1 - i\theta_2)^2}. \tag{8.1.194}$$

Derivation of the Kaiser–Squires Relation in Real Space

Recall from the two-dimensional Poisson equation (8.1.83) that the projected gravitational potential $\psi(\boldsymbol{\theta})$ in angular position coordinates $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is sourced by the convergence $\kappa(\boldsymbol{\theta})$:

$$\nabla^2\psi(\boldsymbol{\theta}) = 2\kappa(\boldsymbol{\theta}), \qquad \nabla^2 \equiv \frac{\partial^2}{\partial\theta_1^2} + \frac{\partial^2}{\partial\theta_2^2}. \tag{1}$$

In two dimensions, the Green's function for the Laplacian satisfying $\nabla^2 G(\boldsymbol{\theta}) = 2\pi\delta^{(2)}(\boldsymbol{\theta})$ is $G(\boldsymbol{\theta}) = \ln|\boldsymbol{\theta}|$. Imposing that the potential decays at infinity, the Poisson equation admits the integral solution previously established in Eq. (8.1.84):

$$\psi(\boldsymbol{\theta}) = \frac{1}{\pi} \int \mathrm{d}^2\theta'\, \kappa(\boldsymbol{\theta}') \ln|\boldsymbol{\theta} - \boldsymbol{\theta}'|. \tag{2}$$

Let $\Delta\boldsymbol{\theta} \equiv \boldsymbol{\theta} - \boldsymbol{\theta}' = (\theta_1 - \theta_1', \theta_2 - \theta_2') \equiv (\Delta\theta_1, \Delta\theta_2)$ denote the coordinate separation vector, with Euclidean norm $|\Delta\boldsymbol{\theta}| = \sqrt{(\theta_1 - \theta_1')^2 + (\theta_2 - \theta_2')^2}$. Evaluating the first and second spatial derivatives of the logarithmic kernel with respect to θ_i and θ_j:

$$\begin{aligned} \partial_i \ln|\Delta\boldsymbol{\theta}| &= \frac{\Delta\theta_i}{|\Delta\boldsymbol{\theta}|^2}, \\ \partial_i\partial_j \ln|\Delta\boldsymbol{\theta}| &= \frac{\delta_{ij}}{|\Delta\boldsymbol{\theta}|^2} - \frac{2\Delta\theta_i\Delta\theta_j}{|\Delta\boldsymbol{\theta}|^4}. \end{aligned} \tag{3}$$

Differentiating under the integral sign gives the second partial derivatives of the potential:

$$\psi_{,ij}(\boldsymbol{\theta}) = \frac{1}{\pi} \int \mathrm{d}^2\theta'\, \kappa(\boldsymbol{\theta}') \left[\frac{\delta_{ij}}{|\Delta\boldsymbol{\theta}|^2} - \frac{2\Delta\theta_i\Delta\theta_j}{|\Delta\boldsymbol{\theta}|^4}\right]. \tag{4}$$

The cosmic shear components (γ_1, γ_2) are extracted from the trace-free tidal tensor:

$$\gamma_1(\boldsymbol{\theta}) = \frac{1}{2}(\psi_{,11} - \psi_{,22}) = \frac{1}{\pi} \int \mathrm{d}^2\theta'\, \kappa(\boldsymbol{\theta}') \left[\frac{\Delta\theta_2^2 - \Delta\theta_1^2}{|\Delta\boldsymbol{\theta}|^4}\right], \tag{5}$$

$$\gamma_2(\boldsymbol{\theta}) = \psi_{,12} = \frac{1}{\pi} \int \mathrm{d}^2\theta'\, \kappa(\boldsymbol{\theta}') \left[-\frac{2\Delta\theta_1\Delta\theta_2}{|\Delta\boldsymbol{\theta}|^4}\right]. \tag{6}$$

Combining these components into the complex shear field $\gamma(\boldsymbol{\theta}) \equiv \gamma_1(\boldsymbol{\theta}) + i\gamma_2(\boldsymbol{\theta})$:

$$\gamma(\boldsymbol{\theta}) = \frac{1}{\pi} \int \mathrm{d}^2\theta'\, \kappa(\boldsymbol{\theta}') \left[\frac{\Delta\theta_2^2 - \Delta\theta_1^2 - 2i\Delta\theta_1\Delta\theta_2}{|\Delta\boldsymbol{\theta}|^4}\right]$$

$$= \frac{1}{\pi}\int \mathcal{D}(\Delta\boldsymbol{\theta})\kappa(\boldsymbol{\theta}')\,\mathrm{d}^2\theta' = \frac{1}{\pi}\int \mathcal{D}(\boldsymbol{\theta}-\boldsymbol{\theta}')\kappa(\boldsymbol{\theta}')\,\mathrm{d}^2\theta'. \tag{7}$$

Here the convolution kernel function $\mathcal{D}(\boldsymbol{\theta})$ evaluated at position vector $\boldsymbol{\theta} = (\theta_1, \theta_2)$ with norm $\theta \equiv |\boldsymbol{\theta}|$ simplifies in the complex plane by factoring the numerator:

$$\mathcal{D}(\boldsymbol{\theta}) = \frac{\theta_2^2 - \theta_1^2 - 2i\theta_1\theta_2}{\theta^4} = -\frac{(\theta_1 + i\theta_2)^2}{|\theta_1 + i\theta_2|^4} = -\frac{1}{(\theta_1 - i\theta_2)^2} = -\frac{e^{2i\varphi}}{\theta^2}, \tag{8}$$

where $\varphi \equiv \arctan(\theta_2/\theta_1)$ is the polar angle. This completes the derivation of Eq. (8.1.193) and Eq. (8.1.194).

In the flat-sky approximation ($\theta \ll 1\,\mathrm{rad}$), the connection between convergence and shear takes an especially transparent and elegant algebraic form. Expanding the projected gravitational potential $\psi(\boldsymbol{\theta})$, convergence $\kappa(\boldsymbol{\theta})$, and shear components (γ_1, γ_2) into two-dimensional Fourier modes:

$$\psi(\boldsymbol{\theta}) = \int \frac{\mathrm{d}^2\ell}{(2\pi)^2}\hat{\psi}(\boldsymbol{\ell})e^{i\boldsymbol{\ell}\cdot\boldsymbol{\theta}}, \tag{8.1.195}$$

the spatial derivatives transform via $\partial_a \leftrightarrow i\ell_a$:

$$\hat{\kappa}(\boldsymbol{\ell}) = -\frac{1}{2}(\ell_1^2 + \ell_2^2)\hat{\psi}(\boldsymbol{\ell}) = -\frac{1}{2}\ell^2\hat{\psi}(\boldsymbol{\ell}), \tag{8.1.196}$$

$$\hat{\gamma}_1(\boldsymbol{\ell}) = -\frac{1}{2}(\ell_1^2 - \ell_2^2)\hat{\psi}(\boldsymbol{\ell}) = \frac{\ell_1^2 - \ell_2^2}{\ell^2}\hat{\kappa}(\boldsymbol{\ell}) = \cos(2\varphi_{\boldsymbol{\ell}})\hat{\kappa}(\boldsymbol{\ell}), \tag{8.1.197}$$

$$\hat{\gamma}_2(\boldsymbol{\ell}) = -\ell_1\ell_2\hat{\psi}(\boldsymbol{\ell}) = \frac{2\ell_1\ell_2}{\ell^2}\hat{\kappa}(\boldsymbol{\ell}) = \sin(2\varphi_{\boldsymbol{\ell}})\hat{\kappa}(\boldsymbol{\ell}), \tag{8.1.198}$$

where $\varphi_{\boldsymbol{\ell}} \equiv \arctan(\ell_2/\ell_1)$ denotes the polar angle of the wavevector $\boldsymbol{\ell} = (\ell_1, \ell_2)$. Combining the two components into the complex shear field $\hat{\gamma}(\boldsymbol{\ell}) \equiv \hat{\gamma}_1(\boldsymbol{\ell}) + i\hat{\gamma}_2(\boldsymbol{\ell})$:

$$\hat{\gamma}(\boldsymbol{\ell}) = [\cos(2\varphi_{\boldsymbol{\ell}}) + i\sin(2\varphi_{\boldsymbol{\ell}})]\,\hat{\kappa}(\boldsymbol{\ell}) = e^{2i\varphi_{\boldsymbol{\ell}}}\hat{\kappa}(\boldsymbol{\ell}) = \frac{1}{\pi}\hat{\mathcal{D}}(\boldsymbol{\ell})\hat{\kappa}(\boldsymbol{\ell}), \tag{8.1.199}$$

where $\hat{\mathcal{D}}(\boldsymbol{\ell}) \equiv \pi e^{2i\varphi_{\boldsymbol{\ell}}}$ is the Fourier transform of the convolution kernel $\mathcal{D}(\boldsymbol{\theta})$.

Equation (8.1.199) reveals that in Fourier space, cosmic shear is obtained from convergence simply by rotating each Fourier mode by a phase factor $2\varphi_{\boldsymbol{\ell}}$. Inverting this relation by multiplying by $e^{-2i\varphi_{\boldsymbol{\ell}}} = \frac{1}{\pi}\hat{\mathcal{D}}^*(\boldsymbol{\ell})$ gives the *Kaiser–Squires Fourier inversion formula*:

$$\begin{aligned}\hat{\kappa}(\boldsymbol{\ell}) &= e^{-2i\varphi_{\boldsymbol{\ell}}}\hat{\gamma}(\boldsymbol{\ell}) = [\cos(2\varphi_{\boldsymbol{\ell}}) - i\sin(2\varphi_{\boldsymbol{\ell}})]\,[\hat{\gamma}_1(\boldsymbol{\ell}) + i\hat{\gamma}_2(\boldsymbol{\ell})] \\ &= \cos(2\varphi_{\boldsymbol{\ell}})\hat{\gamma}_1(\boldsymbol{\ell}) + \sin(2\varphi_{\boldsymbol{\ell}})\hat{\gamma}_2(\boldsymbol{\ell}) + i\,[\cos(2\varphi_{\boldsymbol{\ell}})\hat{\gamma}_2(\boldsymbol{\ell}) - \sin(2\varphi_{\boldsymbol{\ell}})\hat{\gamma}_1(\boldsymbol{\ell})]\,.\end{aligned} \tag{8.1.200}$$

Because the physical convergence $\kappa(\boldsymbol{\theta})$ is a real scalar field, its Fourier modes must satisfy reality conditions, isolating the reconstructed convergence as the real part:

$$\hat{\kappa}(\boldsymbol{\ell}) = \cos(2\varphi_{\boldsymbol{\ell}})\hat{\gamma}_1(\boldsymbol{\ell}) + \sin(2\varphi_{\boldsymbol{\ell}})\hat{\gamma}_2(\boldsymbol{\ell}) = \frac{\ell_1^2 - \ell_2^2}{\ell^2}\hat{\gamma}_1(\boldsymbol{\ell}) + \frac{2\ell_1\ell_2}{\ell^2}\hat{\gamma}_2(\boldsymbol{\ell}). \tag{8.1.201}$$

Applying the 2D inverse Fourier transform and invoking the convolution theorem transforms Eq. (8.1.201) into real space:

$$\kappa(\boldsymbol{\theta}) = \kappa_0 + \frac{1}{\pi}\int \mathcal{D}^*(\boldsymbol{\theta}-\boldsymbol{\theta}')\gamma(\boldsymbol{\theta}')\,\mathrm{d}^2\theta', \tag{8.1.202}$$

where $\mathcal{D}^*(\boldsymbol{\theta}) = -(\theta_1 + i\theta_2)^{-2} = -e^{-2i\varphi}/\theta^2$ is the complex conjugate kernel. Because the phase factor $e^{-2i\varphi_{\boldsymbol{\ell}}}$ is singular at $\boldsymbol{\ell} = \mathbf{0}$ (the direction of the wavevector is undefined at the origin), the monopole mode $\hat{\kappa}(\mathbf{0})$ cannot be recovered from the shear, leaving the constant surface mass density κ_0 undetermined: this is the Fourier manifestation of the mass-sheet degeneracy.

E/B-Mode Decomposition and Systematic Null Tests The shear components (γ_1, γ_2) measured on the sky depend explicitly on the arbitrary choice of Cartesian coordinate axes (θ_1, θ_2) on the celestial tangent plane. Under a counterclockwise rotation of the coordinate axes by an angle φ, the coordinates transform as $\boldsymbol{\theta}' = R(\varphi)\boldsymbol{\theta}$. Because shear is a symmetric, trace-free rank-2 tensor $\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 & \gamma_2 \\ \gamma_2 & -\gamma_1 \end{pmatrix}$, it transforms under similarity transformation as $\boldsymbol{\gamma}' = R(\varphi)\boldsymbol{\gamma}R^T(\varphi)$, or in components:

$$\begin{pmatrix} \gamma_1' \\ \gamma_2' \end{pmatrix} = \begin{pmatrix} \cos(2\varphi) & \sin(2\varphi) \\ -\sin(2\varphi) & \cos(2\varphi) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}, \tag{8.1.203}$$

which is equivalent to the complex rotation $\gamma' = e^{-2i\varphi}\gamma$. The factor of 2 reflects the spin-2 character of the shear field: an elliptical distortion or headless tidal stretch possesses 180° rotational symmetry. This transformation law is mathematically identical to that of the linear Stokes polarization parameters (Q, U) established in Section 6.3.3 of Chapter 6.

To construct coordinate-independent observables, we note that in Fourier space each plane-wave mode with wavevector $\boldsymbol{\ell} = (\ell_1, \ell_2) = \ell(\cos\varphi_{\boldsymbol{\ell}}, \sin\varphi_{\boldsymbol{\ell}})$ possesses an intrinsic, physically determined orientation: the wavevector angle $\varphi_{\boldsymbol{\ell}}$ itself, which satisfies [Eq. (6.3.122)]:

$$\cos(2\varphi_{\boldsymbol{\ell}}) = \frac{\ell_1^2 - \ell_2^2}{\ell^2}, \qquad \sin(2\varphi_{\boldsymbol{\ell}}) = \frac{2\ell_1\ell_2}{\ell^2}. \tag{8.1.204}$$

Rotating the local coordinate axes for each Fourier mode by exactly $\varphi = \varphi_{\boldsymbol{\ell}}$ aligns the principal coordinate axis with the wavevector $\boldsymbol{\ell}$. In this wave-aligned intrinsic frame, the shear tensor decomposes into:

1. A longitudinal/transverse component $\hat{\gamma}_{\parallel}(\boldsymbol{\ell})$ describing stretch aligned parallel (0°) or orthogonal (90°) to $\boldsymbol{\ell}$. This scalar component is even under spatial parity reflections across $\boldsymbol{\ell}$ and is defined as the gradient-type E-mode:

$$\hat{E}(\boldsymbol{\ell}) \equiv \hat{\gamma}_{\parallel}(\boldsymbol{\ell}) = \cos(2\varphi_{\boldsymbol{\ell}})\hat{\gamma}_1(\boldsymbol{\ell}) + \sin(2\varphi_{\boldsymbol{\ell}})\hat{\gamma}_2(\boldsymbol{\ell}). \tag{8.1.205}$$

2. A diagonal cross component $\hat{\gamma}_{\times}(\boldsymbol{\ell})$ describing stretch oriented at ±45° to $\boldsymbol{\ell}$. This pseudoscalar component flips sign under spatial parity reflections across $\boldsymbol{\ell}$ and is defined as the curl-type B-mode:

$$\hat{B}(\boldsymbol{\ell}) \equiv \hat{\gamma}_{\times}(\boldsymbol{\ell}) = -\sin(2\varphi_{\boldsymbol{\ell}})\hat{\gamma}_1(\boldsymbol{\ell}) + \cos(2\varphi_{\boldsymbol{\ell}})\hat{\gamma}_2(\boldsymbol{\ell}). \tag{8.1.206}$$

Inverting the orthogonal rotation in Eqs. (8.1.205)–(8.1.206) expresses the shear fields directly in terms of $\hat{E}$ and $\hat{B}$:

$$\hat{\gamma}_1(\boldsymbol{\ell}) = \cos(2\varphi_{\boldsymbol{\ell}})\hat{E}(\boldsymbol{\ell}) - \sin(2\varphi_{\boldsymbol{\ell}})\hat{B}(\boldsymbol{\ell}), \tag{8.1.207}$$

$$\hat{\gamma}_2(\boldsymbol{\ell}) = \sin(2\varphi_{\boldsymbol{\ell}})\hat{E}(\boldsymbol{\ell}) + \cos(2\varphi_{\boldsymbol{\ell}})\hat{B}(\boldsymbol{\ell}), \tag{8.1.208}$$

which reproduces the exact Fourier representation derived for Stokes parameters (Q, U) in Eqs. (6.3.120)–(6.3.121). Multiplying Eqs. (8.1.205)–(8.1.206) by $-\ell^2$ and transforming back to coordinate space via $\ell_j \to -i\partial_j$ (where $-\ell^2 \leftrightarrow \nabla^2$, $-\ell^2\cos 2\varphi_{\boldsymbol{\ell}} \leftrightarrow \partial_1^2 - \partial_2^2$, and $-\ell^2 \sin 2\varphi_{\boldsymbol{\ell}} \leftrightarrow 2\partial_1\partial_2$), we recover the real-space differential relations:

$$\nabla^2 E = (\partial_1^2 - \partial_2^2)\gamma_1 + 2\partial_1\partial_2\gamma_2, \tag{8.1.209}$$

$$\nabla^2 B = -2\partial_1\partial_2\gamma_1 + (\partial_1^2 - \partial_2^2)\gamma_2, \tag{8.1.210}$$

which are the exact spin-2 shear counterparts of the Stokes polarization Laplacians [Eq. (6.3.123)].

The physical significance of this decomposition in gravitational lensing is profound. Sourced by scalar metric perturbations Φ, the cosmic shear field derives from a single scalar potential ψ, with $\gamma_1 = \frac{1}{2}(\partial_1^2 - \partial_2^2)\psi$ and $\gamma_2 = \partial_1\partial_2\psi$. In Fourier space:

$$\hat{\gamma}_1(\boldsymbol{\ell}) = -\frac{1}{2}\ell^2 \cos(2\varphi_{\boldsymbol{\ell}})\hat{\psi}(\boldsymbol{\ell}), \qquad \hat{\gamma}_2(\boldsymbol{\ell}) = -\frac{1}{2}\ell^2 \sin(2\varphi_{\boldsymbol{\ell}})\hat{\psi}(\boldsymbol{\ell}). \tag{8.1.211}$$

Substituting these into Eqs. (8.1.205) and (8.1.206):

$$\hat{E}(\boldsymbol{\ell}) = -\frac{1}{2}\ell^2 \left[\cos^2(2\varphi_{\boldsymbol{\ell}}) + \sin^2(2\varphi_{\boldsymbol{\ell}})\right] \hat{\psi}(\boldsymbol{\ell}) = -\frac{1}{2}\ell^2\hat{\psi}(\boldsymbol{\ell}) = \hat{\kappa}(\boldsymbol{\ell}), \tag{8.1.212}$$

$$\hat{B}(\boldsymbol{\ell}) = \frac{1}{2}\ell^2 \left[\sin(2\varphi_{\boldsymbol{\ell}})\cos(2\varphi_{\boldsymbol{\ell}}) - \cos(2\varphi_{\boldsymbol{\ell}})\sin(2\varphi_{\boldsymbol{\ell}})\right] \hat{\psi}(\boldsymbol{\ell}) \equiv 0. \tag{8.1.213}$$

Because gravitational lensing in general relativity is sourced by scalar metric perturbations, the cosmological shear field is strictly curl-free:

$$\hat{\kappa}_E(\boldsymbol{\ell}) = \hat{\kappa}(\boldsymbol{\ell}), \qquad \hat{\kappa}_B(\boldsymbol{\ell}) \equiv 0. \tag{8.1.214}$$

This curl-free theorem provides an indispensable empirical null test in weak lensing surveys: any statistically significant non-zero B-mode reconstructed from real galaxy catalogs cannot arise from scalar gravitational lensing. Instead, non-vanishing B-modes signal residual observational systematics, such as telescope optical aberrations, uncorrected point spread function (PSF) anisotropy, camera detector distortions, or intrinsic astrophysical galaxy alignments.

The core algorithm of the Kaiser–Squires inversion in Python is implemented via the fast Fourier transform (FFT) as follows:

```python
import numpy as np

def kaiser_squires_inversion(gamma1, gamma2):
    # Reconstructs convergence kappa_E and curl-mode kappa_B from shear fields
    Ny, Nx = gamma1.shape
    g1_hat = np.fft.fft2(gamma1)
    g2_hat = np.fft.fft2(gamma2)

    kx = 2 * np.pi * np.fft.fftfreq(Nx)
    ky = 2 * np.pi * np.fft.fftfreq(Ny)
    KX, KY = np.meshgrid(kx, ky)
    K2 = KX**2 + KY**2
    K2[0, 0] = 1.0  # Avoid division by zero at k = 0

    # Fourier filter kernels
    cos2phi = (KX**2 - KY**2) / K2
    sin2phi = 2 * KX * KY / K2
    cos2phi[0, 0] = 0.0
    sin2phi[0, 0] = 0.0

    # Inversion
    kappa_E_hat = cos2phi * g1_hat + sin2phi * g2_hat
    kappa_B_hat = -sin2phi * g1_hat + cos2phi * g2_hat

    kappa_E = np.real(np.fft.ifft2(kappa_E_hat))
    kappa_B = np.real(np.fft.ifft2(kappa_B_hat))
    return kappa_E, kappa_B
```

Listing 8.1: Kaiser–Squires 2D FFT reconstruction function in Python.

Two-point statistics for the cosmic shear

The spin-2 complex shear $\gamma(\boldsymbol{\theta}) = \gamma e^{2i\phi}$ can be projected relative to the separation vector connecting two points. Consider a pair of galaxies separated by angular distance θ, and let ϕ be

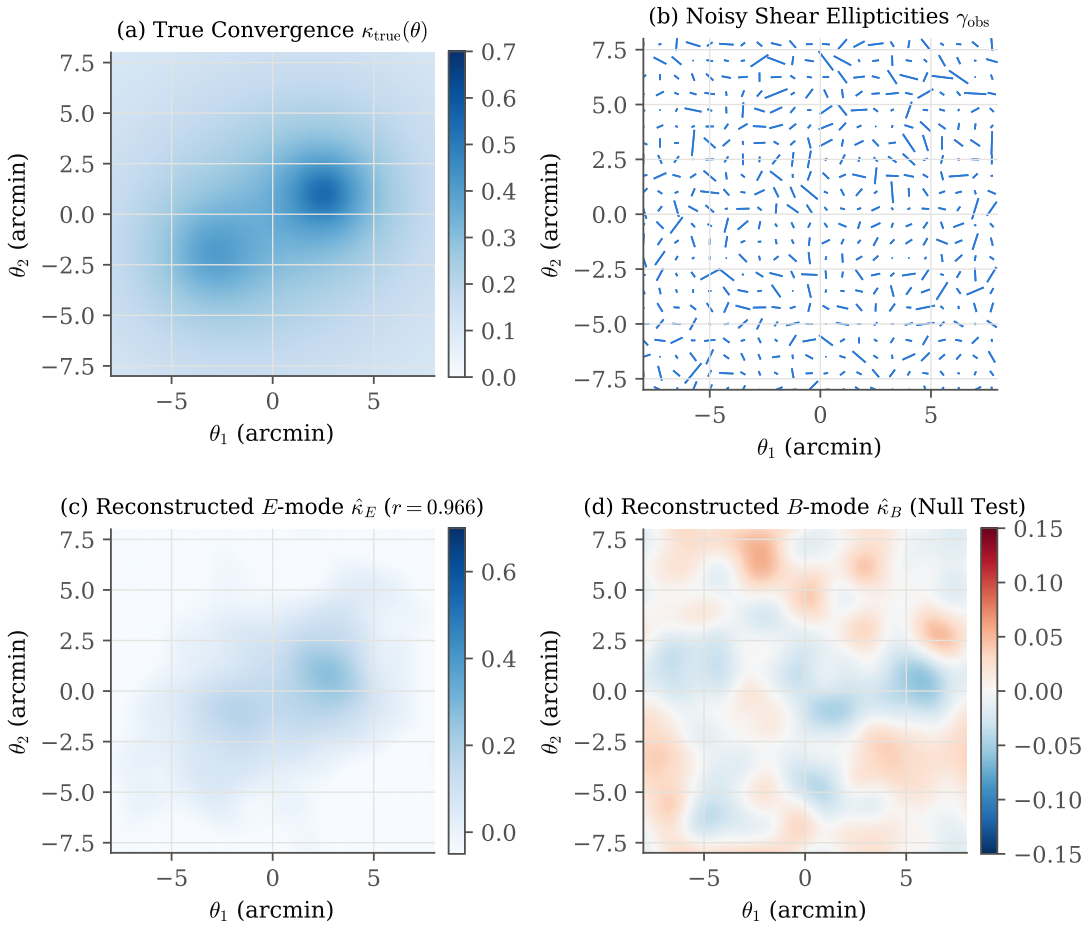


Figure 8.17: Kaiser–Squires reconstruction of a mock galaxy cluster mass distribution. (Left) True input convergence field $\kappa_{\rm true}$ containing a central cluster and two substructures. (Center) Reconstructed convergence κ_E obtained by applying the Kaiser–Squires Fourier inversion to the noisy mock shear field ($\sigma_\epsilon = 0.25$). (Right) Reconstructed B-mode convergence κ_B, which is consistent with zero within noise fluctuations, confirming the absence of residual systematics.

the polar angle of the separation vector. In the local radial/tangential basis, the tangential and cross shear components are defined by:

$$\gamma_t(\boldsymbol{\theta}) \equiv -\,\mathrm{Re}\left[\gamma(\boldsymbol{\theta}) e^{-2i\phi}\right], \qquad \gamma_\times(\boldsymbol{\theta}) \equiv -\,\mathrm{Im}\left[\gamma(\boldsymbol{\theta}) e^{-2i\phi}\right]. \tag{8.1.215}$$

The spatial two-point auto-correlation functions of these decomposed shear components depend solely on the angular separation θ:

$$\xi_t(\theta) \equiv \langle \gamma_t \gamma_t \rangle(\theta), \qquad \xi_\times(\theta) \equiv \langle \gamma_\times \gamma_\times \rangle(\theta). \tag{8.1.216}$$

These components are linearly combined into orthogonal parity combinations:

$$\xi_+(\theta) \equiv \xi_t(\theta) + \xi_\times(\theta), \qquad \xi_-(\theta) \equiv \xi_t(\theta) - \xi_\times(\theta). \tag{8.1.217}$$

Under parity reflections, γ_t is invariant (E-mode parity) while $\gamma_\times$ flips sign (B-mode parity), ensuring that the cross-correlation vanishes identically: $\langle \gamma_t \gamma_\times \rangle = 0$ in the absence of parity-violating systematics.

Using the Fourier-space Kaiser–Squires relations, the shear correlation functions $\xi_\pm(\theta)$ are directly related to the convergence power spectrum $P_\kappa(\ell)$ by Hankel transforms involving Bessel

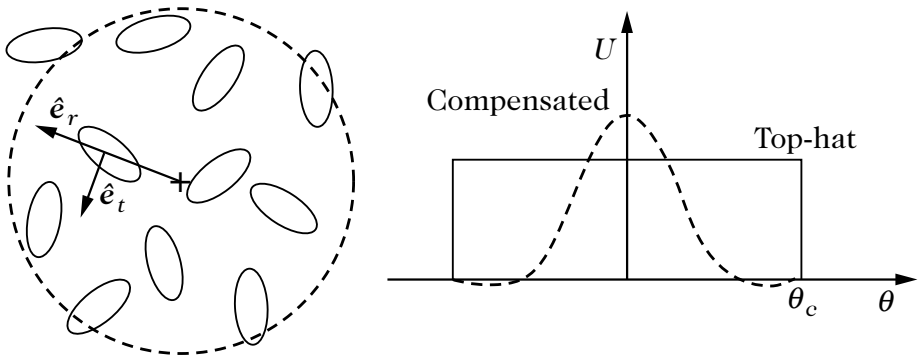


Figure 8.18: Two-point cosmic shear correlation functions $\xi_+(\theta)$ and $\xi_-(\theta)$ measured as a function of angular separation θ. The function $\xi_+(\theta)$ probes power on larger angular scales, while $\xi_-(\theta)$ drops steeply and isolates small-scale non-linear structures.

functions:

$$\xi_+(\theta) = \frac{1}{2\pi}\int_0^\infty \ell\, \mathrm{d}\ell\, P_\kappa(\ell) J_0(\ell\theta), \qquad \xi_-(\theta) = \frac{1}{2\pi}\int_0^\infty \ell\, \mathrm{d}\ell\, P_\kappa(\ell) J_4(\ell\theta), \tag{8.1.218}$$

where $J_0(x)$ and $J_4(x)$ are the zeroth- and fourth-order Bessel functions of the first kind.

Aperture mass

Because observed galaxies cover finite, irregularly shaped survey footprints with holes and masks, evaluating the non-local real-space Kaiser–Squires integral is susceptible to boundary edge effects. Schneider (1996) introduced the *aperture mass* $M_{\rm ap}(\theta_0)$, defined as the convolution of the convergence field with a compensated circular filter $U(\theta)$ of radius θ_0:

$$M_{\rm ap}(\theta_0) \equiv \int \mathrm{d}^2\theta\, U(|\boldsymbol{\theta}|)\kappa(\boldsymbol{\theta}), \qquad \int_0^{\theta_0} \theta\, U(\theta)\, \mathrm{d}\theta = 0. \tag{8.1.219}$$

Because $U(\theta)$ is compensated ($\int \theta U(\theta)\,\mathrm{d}\theta = 0$), $M_{\rm ap}$ is completely insensitive to the mass-sheet degeneracy: transforming $\kappa \to \kappa + \kappa_0$ leaves $M_{\rm ap}$ invariant. By Gauss's theorem, $M_{\rm ap}$ can be computed directly from tangential shear without ever performing a mass reconstruction:

$$M_{\rm ap}(\theta_0) = \int \mathrm{d}^2\theta\, Q(|\boldsymbol{\theta}|)\gamma_t(\boldsymbol{\theta}), \tag{8.1.220}$$

where the shear filter $Q(\theta)$ is related to the convergence filter $U(\theta)$ by:

$$Q(\theta) = \frac{2}{\theta^2}\int_0^\theta \theta' U(\theta')\, \mathrm{d}\theta' - U(\theta). \tag{8.1.221}$$

The aperture mass variance $\langle M_{\rm ap}^2\rangle(\theta_0)$ isolates power in a narrow band of multipoles around $\ell \approx 4.25/\theta_0$:

$$\langle M_{\rm ap}^2\rangle(\theta_0) = \frac{1}{2\pi}\int_0^\infty \ell\, \mathrm{d}\ell\, P_\kappa(\ell)\hat{U}^2(\ell\theta_0). \tag{8.1.222}$$

Relations between the power spectra

Because $|\hat{\mathcal{D}}(\boldsymbol{\ell})|^2 = \pi^2$, Eq. (8.1.200) guarantees that $|\hat{\kappa}(\boldsymbol{\ell})|^2 = |\hat{\gamma}_1(\boldsymbol{\ell})|^2 + |\hat{\gamma}_2(\boldsymbol{\ell})|^2$, establishing that convergence and cosmic shear share the identical power spectrum:

$$P_\gamma(\ell) = P_\kappa(\ell). \tag{8.1.223}$$

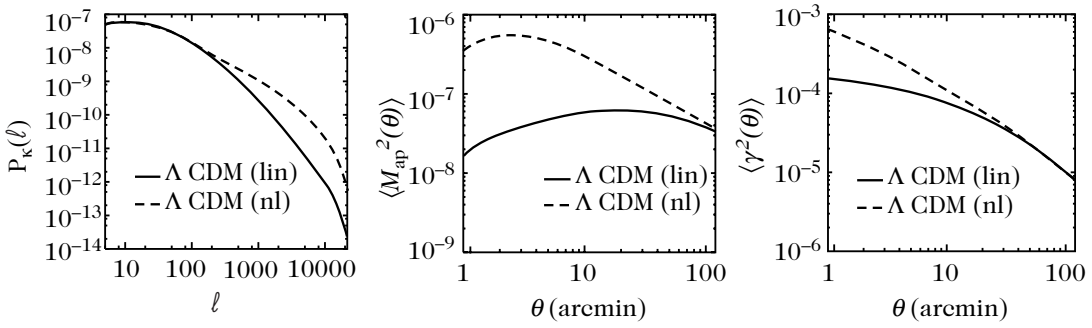


Figure 8.19: Aperture mass variance $\langle M_{\rm ap}^2\rangle(\theta_0)$ as a function of aperture radius θ_0 for different cosmological models. The peak location and amplitude tightly constrain σ_8 and $\Omega_{\rm m0}$.

Measurement of the cosmic shear

Galaxy shape measurement and ellipticity Unlike point sources, galaxies are extended, diffuse puddles of light with no sharply defined boundary. To quantify their orientation and spatial elongation in an objective, automated manner, observational cosmology characterizes the surface brightness distribution $I(\boldsymbol{\theta})$ through its multipole moments.

The galaxy centroid $\bar{\boldsymbol{\theta}}$ is defined as the first moment of the light distribution:

$$\bar{\boldsymbol{\theta}} \equiv \frac{\int \mathrm{d}^2\theta\, I(\boldsymbol{\theta}) W(\boldsymbol{\theta}) \boldsymbol{\theta}}{\int \mathrm{d}^2\theta\, I(\boldsymbol{\theta}) W(\boldsymbol{\theta})}, \tag{8.1.224}$$

where $W(\boldsymbol{\theta})$ is a smooth, radially symmetric weight function (matched to the characteristic angular size of the galaxy) that prevents background sky noise at large radii from destabilizing the integrals.

Choosing coordinate axes centered on the centroid ($\bar{\boldsymbol{\theta}} = \mathbf{0}$), the second-order brightness moments form the symmetric quadrupole moment-of-inertia tensor:

$$Q_{ij} \equiv \frac{\int \mathrm{d}^2\theta\, I(\boldsymbol{\theta}) W(\boldsymbol{\theta}) \theta_i \theta_j}{\int \mathrm{d}^2\theta\, I(\boldsymbol{\theta}) W(\boldsymbol{\theta})}, \qquad (i,j \in \{1,2\}). \tag{8.1.225}$$

Notice the mathematical structure: just as the distortion matrix $\mathcal{A}$ in Section 8.1.90 decomposed naturally onto the Pauli spin matrix basis, any real symmetric 2×2 matrix Q decomposes into the three symmetric basis matrices $\{\sigma_0, \sigma_3, \sigma_1\}$:

$$Q = \frac{1}{2}\,\mathrm{Tr}(Q)\sigma_0 + \frac{1}{2}(Q_{11} - Q_{22})\sigma_3 + Q_{12}\sigma_1. \tag{8.1.226}$$

The projection onto the identity σ_0 is the trace $\mathrm{Tr}(Q) = Q_{11} + Q_{22}$, which is invariant under spatial rotations and measures the isotropic size (mean-square radius) of the galaxy. The remaining two projections isolate the quadrupole elongation: $(Q_{11} - Q_{22})$ quantifies stretching along the coordinate axes, while $2Q_{12}$ quantifies stretching along the 45° diagonals.

In complex notation, these traceless quadrupole components combine into:

$$(Q_{11} - Q_{22}) + 2iQ_{12} = \frac{\int \mathrm{d}^2\theta\, I(\boldsymbol{\theta}) W(\boldsymbol{\theta}) (\theta_1 + i\theta_2)^2}{\int \mathrm{d}^2\theta\, I(\boldsymbol{\theta}) W(\boldsymbol{\theta})} = \frac{\int \mathrm{d}^2\theta\, I(\boldsymbol{\theta}) W(\boldsymbol{\theta}) \theta^2 e^{2i\phi}}{\int \mathrm{d}^2\theta\, I(\boldsymbol{\theta}) W(\boldsymbol{\theta})}. \tag{8.1.227}$$

Under a rotation of the coordinate system by an angle ϑ, the phase shifts by $e^{2i\vartheta}$. The traceless quadrupole is therefore an intrinsic spin-2 field on the sky, possessing the exact same headless 180° symmetry as the gravitational shear $\boldsymbol{\gamma}$.

To translate these moments into an observable shape, consider an idealized galaxy whose isophotes are concentric ellipses with semi-major axis a, semi-minor axis b ($0 < b \leq a$), axis ratio $q \equiv b/a$, and position angle ϑ measured from the θ_1-axis. In the galaxy's intrinsic principal coordinate frame aligned with its symmetry axes, reflection symmetry dictates that $Q'_{12} = 0$, while the diagonal moments evaluate to:

$$Q'_{11} = Ca^2, \qquad Q'_{22} = Cb^2, \tag{8.1.228}$$

where $C > 0$ is a dimensionless constant determined by the radial surface brightness profile.

Because the trace and determinant of a matrix are invariant under coordinate rotations, they evaluate directly in the principal frame:

$$\mathrm{Tr}(Q) = Q'_{11} + Q'_{22} = C(a^2 + b^2), \tag{8.1.229}$$

$$\det(Q) = Q'_{11}Q'_{22} = C^2a^2b^2 \implies 2\sqrt{\det(Q)} = 2Cab. \tag{8.1.230}$$

Adding these two invariants forms a perfect square:

$$\mathrm{Tr}(Q) + 2\sqrt{\det(Q)} = C(a^2 + b^2 + 2ab) = C(a + b)^2. \tag{8.1.231}$$

Meanwhile, rotating from the principal frame by angle ϑ rotates the spin-2 traceless moment by $e^{2i\vartheta}$:

$$(Q_{11} - Q_{22}) + 2iQ_{12} = (Q'_{11} - Q'_{22})e^{2i\vartheta} = C(a^2 - b^2)e^{2i\vartheta}. \tag{8.1.232}$$

Dividing the traceless spin-2 moment by the invariant sum $\mathrm{Tr}(Q) + 2\sqrt{\det(Q)}$ cancels the unknown galaxy profile normalization C completely, yielding the complex ellipticity $\boldsymbol{\epsilon} \equiv \epsilon_1 + i\epsilon_2$:

$$\begin{aligned}\boldsymbol{\epsilon} &\equiv \frac{(Q_{11} - Q_{22}) + 2iQ_{12}}{Q_{11} + Q_{22} + 2\sqrt{Q_{11}Q_{22} - Q_{12}^2}} = \frac{C(a^2 - b^2)}{C(a + b)^2}e^{2i\vartheta} \\ &= \frac{a - b}{a + b}e^{2i\vartheta} = \frac{1 - q}{1 + q}e^{2i\vartheta}.\end{aligned} \tag{8.1.233}$$

In components:

$$\epsilon_1 = \frac{1 - q}{1 + q}\cos(2\vartheta), \qquad \epsilon_2 = \frac{1 - q}{1 + q}\sin(2\vartheta). \tag{8.1.234}$$

Why is the specific combination $\frac{1-q}{1+q}$ preferred over the alternative definition $(1-q^2)/(1+q^2) = (a^2 - b^2)/(a^2 + b^2)$? Under an infinitesimal gravitational shear γ, an initially circular galaxy ($q = 1$) is distorted into an ellipse with semi-axes $(1 + \gamma)$ and $(1 - \gamma)$, corresponding to axis ratio $q = (1 - \gamma)/(1 + \gamma)$. Evaluating the ellipticity gives:

$$\frac{1 - q}{1 + q} = \frac{1 - \frac{1-\gamma}{1+\gamma}}{1 + \frac{1-\gamma}{1+\gamma}} = \gamma. \tag{8.1.235}$$

The estimator $\boldsymbol{\epsilon}$ has unit linear response to gravitational shear.

In the weak lensing regime ($\gamma \ll 1$), an applied shear transforms the intrinsic ellipticity $\boldsymbol{\epsilon}_{\mathrm{int}}$ into the observed ellipticity according to:

$$\boldsymbol{\epsilon} \approx \boldsymbol{\epsilon}_{\mathrm{int}} + \boldsymbol{g} \approx \boldsymbol{\epsilon}_{\mathrm{int}} + \boldsymbol{\gamma}. \tag{8.1.236}$$

Assuming galaxies are randomly oriented in the Universe ($\langle \boldsymbol{\epsilon}_{\mathrm{int}} \rangle = 0$), averaging over an ensemble of N galaxies within an angular patch yields an unbiased estimator of the local shear:

$$\langle \boldsymbol{\epsilon} \rangle = \boldsymbol{\gamma}, \qquad \sigma_{\langle\gamma\rangle} = \frac{\sigma_\epsilon}{\sqrt{N}}, \tag{8.1.237}$$

where $\sigma_\epsilon \approx 0.25 - 0.35$ is the intrinsic shape dispersion (shape noise).

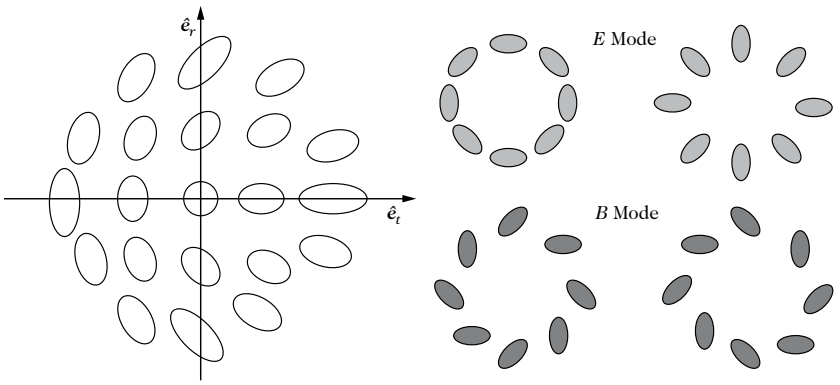


Figure 8.20: Cosmic shear power spectra and 2D correlation functions measured across multiple photometric redshift bins (tomographic cosmic shear), probing the time evolution of the matter power spectrum. From Ref. [40].

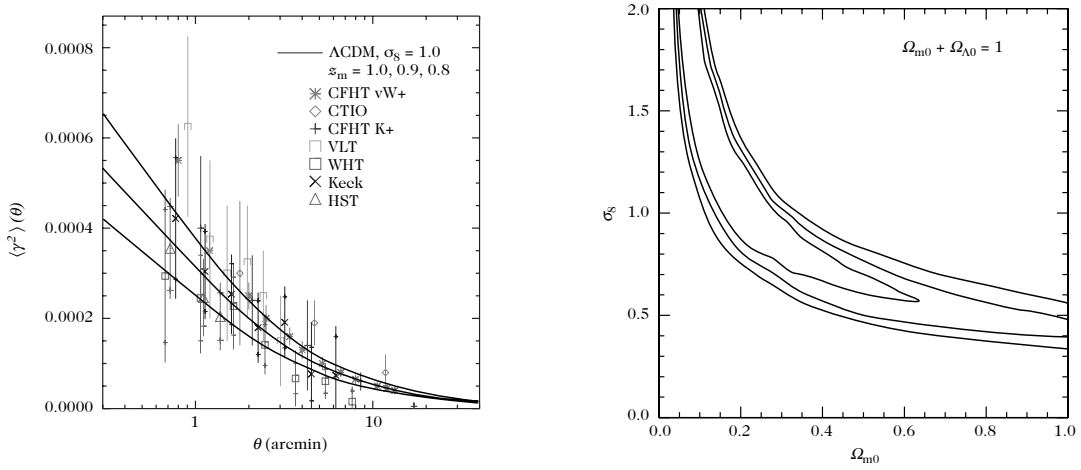


Figure 8.21: Cosmological parameter constraints in the $(\Omega_{\rm m0}, \sigma_8)$ plane. Cosmic shear surveys produce an elongated degeneracy curve along $S_8 \equiv \sigma_8(\Omega_{\rm m0}/0.3)^{0.5} \approx$ const. Combining cosmic shear with CMB and BAO data breaks this degeneracy. From Ref. [41].

Modern calibration: Metacalibration In realistic surveys, atmospheric turbulence, telescope optics aberrations, pixelization, and detector non-linearities convolute the image with a point spread function (PSF). Furthermore, selection effects (brighter galaxies are more readily detected) introduce multiplicative shear calibration biases: $\langle \boldsymbol{\epsilon} \rangle = (1+m)\boldsymbol{\gamma} + c$.

Modern stage-III and stage-IV cosmic shear surveys (e.g., the Dark Energy Survey DES, the Kilo-Degree Survey KiDS, and Hyper Suprime-Cam HSC) calibrate shear responses using *Metacalibration* (Huff & Mandelbaum 2017; Sheldon & Huff 2017). The galaxy image is artificially sheared by a known test shear $\pm\Delta\gamma_j$ directly in pixel space, de-convolved and re-convolved with an isotropic PSF. The shear response tensor R_{ij} is measured directly from the data:

$$R_{ij} \equiv \frac{\partial \langle \epsilon_i \rangle}{\partial \gamma_j} \approx \frac{\langle \epsilon_i^+ \rangle - \langle \epsilon_i^- \rangle}{2\Delta\gamma_j}, \tag{8.1.238}$$

completely eliminating reliance on synthetic image simulations or assumed galaxy morphological profiles.

State-of-the-Art Cosmological Constraints and the S_8 Tension Cosmic shear measurements are primarily sensitive to the amplitude of the matter power spectrum on $8h^{-1}\,\mathrm{Mpc}$ scales (σ_8) and the matter density parameter Ω_{m0}, constrained in the joint combination:

$$S_8 \equiv \sigma_8 \sqrt{\frac{\Omega_{\mathrm{m0}}}{0.3}}. \tag{8.1.239}$$

The Dark Energy Survey Year 3 (DES Y3; Amon et al. 2022; Secco et al. 2022) measured:

$$S_8 = 0.776^{+0.017}_{-0.017} \quad \text{(DES Y3 Cosmic Shear)}, \tag{8.1.240}$$

and the KiDS-1000 survey (Asgari et al. 2021) measured:

$$S_8 = 0.766^{+0.020}_{-0.014} \quad \text{(KiDS-1000 Cosmic Shear)}. \tag{8.1.241}$$

These values are systematically lower by $\approx 2.5-3\sigma$ than the high-redshift CMB prediction from Planck 2018 (ΛCDM baseline):

$$S_8 = 0.834 \pm 0.016 \quad \text{(Planck 2018 CMB)}. \tag{8.1.242}$$

Meanwhile, recent CMB lensing measurements from the Atacama Cosmology Telescope (ACT DR6; Madhavacheril et al. 2024; Qu et al. 2024) find $S_8 = 0.840 \pm 0.030$, in full agreement with Planck. This mild discrepancy between low-redshift galaxy shear and high-redshift CMB determinations is known as the S_8 *tension*, actively driving investigations into baryonic feedback, neutrino masses, and modified gravity.

8.1.6 Lensing of the cosmic microwave background

CMB Deflection and Lensing Potential The cosmic microwave background (CMB) photons emitted at recombination ($z_* \simeq 1090$, comoving distance $\chi_* \simeq 14\,\mathrm{Gpc}$) traverse the entire observable Universe before reaching our detectors. Large-scale potential wells deflect CMB photons by an angular deflection vector $\boldsymbol{d}(\hat{\boldsymbol{n}})$:

$$\tilde{T}(\hat{\boldsymbol{n}}) = T(\hat{\boldsymbol{n}} + \boldsymbol{d}(\hat{\boldsymbol{n}})) = T(\hat{\boldsymbol{n}} + \boldsymbol{\nabla}\phi(\hat{\boldsymbol{n}})), \tag{8.1.243}$$

where $\tilde{T}$ is the lensed temperature and $\phi(\hat{\boldsymbol{n}})$ is the CMB lensing potential:

$$\phi(\hat{\boldsymbol{n}}) \equiv -2\int_0^{\chi_*} \mathrm{d}\chi\, \frac{\chi_* - \chi}{\chi_*\chi} \frac{\Phi(\chi\hat{\boldsymbol{n}}, \chi)}{c^2}. \tag{8.1.244}$$

Consider photon propagation in a spatially flat FLRW universe with scalar metric perturbations in conformal Newtonian gauge:

$$\mathrm{d}s^2 = a^2(\eta)\left[-(1+2\Psi)\,\mathrm{d}\eta^2 + (1-2\Phi)\delta_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j\right], \tag{1}$$

where $a(\eta)$ is the scale factor, η is conformal time, Ψ and Φ are the scalar gravitational potentials, and we work to first order in the metric perturbations in natural units ($c = 1$). The photon distribution function is defined on the massless phase space as $f = f(x^\mu, p^i)$, where $p^\mu = \mathrm{d}x^\mu/\mathrm{d}\lambda$ is the photon four-momentum and λ is an affine parameter. Photons follow null geodesics satisfying $p^\mu p_\mu = 0$ and the geodesic equation $\mathrm{d}p^\mu/\mathrm{d}\lambda + \Gamma^\mu_{\alpha\beta}p^\alpha p^\beta = 0$. After recombination, neglecting collisions, the photon distribution satisfies the collisionless relativistic Boltzmann equation:

$$p^\mu \frac{\partial f}{\partial x^\mu} - \Gamma^i_{\alpha\beta}p^\alpha p^\beta \frac{\partial f}{\partial p^i} = 0 \iff \frac{\mathrm{d}f}{\mathrm{d}\lambda} = 0, \tag{2}$$

which expresses the conservation of the phase-space distribution function along the photon trajectory.

To express the Boltzmann equation in terms of observable photon quantities, we introduce the comoving photon energy q and unit propagation direction n^i, where $\delta_{ij}n^i n^j = 1$. For a locally measured physical energy E, the comoving energy is $q = aE$. The distribution function can then be parameterized as $f = f(\eta, \boldsymbol{x}, q, \boldsymbol{n})$. The convective derivative along a photon trajectory with respect to conformal time is:

$$\frac{\mathrm{d}f}{\mathrm{d}\eta} = \frac{\partial f}{\partial \eta} + \frac{\mathrm{d}x^i}{\mathrm{d}\eta}\partial_i f + \frac{\mathrm{d}q}{\mathrm{d}\eta}\frac{\partial f}{\partial q} + \frac{\mathrm{d}n^i}{\mathrm{d}\eta}\frac{\partial f}{\partial n^i} = 0. \tag{3}$$

These four terms describe explicit time evolution, spatial propagation, energy evolution (gravitational redshift and Sachs–Wolfe effects), and angular deflection of the photon trajectory, respectively. The gravitational potentials enter through the geodesic equations governing $\mathrm{d}x^i/\mathrm{d}\eta$, $\mathrm{d}q/\mathrm{d}\eta$, and $\mathrm{d}n^i/\mathrm{d}\eta$.

Because the primordial photon distribution is locally Planckian with a direction-dependent temperature, we write:

$$f(\eta, \boldsymbol{x}, q, \boldsymbol{n}) = \frac{1}{\exp\left[\frac{q}{aT(\eta,\boldsymbol{x},\boldsymbol{n})}\right] - 1}, \qquad T(\eta, \boldsymbol{x}, \boldsymbol{n}) = \frac{T_0}{a(\eta)}\big[1 + \Theta(\eta, \boldsymbol{x}, \boldsymbol{n})\big], \tag{4}$$

where T_0 is the present-day background CMB temperature and Θ is the fractional temperature anisotropy. The distribution becomes:

$$f = \frac{1}{\exp\left[\frac{q}{T_0(1+\Theta)}\right] - 1}. \tag{5}$$

Because f is conserved along each photon trajectory, its exponential argument is an invariant of motion:

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\frac{q}{T_0(1+\Theta)}\right] = 0 \implies \frac{1}{T_0(1+\Theta)}\frac{\mathrm{d}q}{\mathrm{d}\eta} - \frac{q}{T_0(1+\Theta)^2}\frac{\mathrm{d}\Theta}{\mathrm{d}\eta} = 0. \tag{6}$$

Multiplying by $T_0(1+\Theta)/q$ yields:

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\ln(1+\Theta) = \frac{\mathrm{d}}{\mathrm{d}\eta}\ln q. \tag{7}$$

Expanding the convective phase-space derivative:

$$\frac{\partial \ln(1+\Theta)}{\partial \eta} + \frac{\mathrm{d}x^i}{\mathrm{d}\eta}\partial_i \ln(1+\Theta) + \frac{\mathrm{d}n^i}{\mathrm{d}\eta}\frac{\partial \ln(1+\Theta)}{\partial n^i} = \frac{1}{q}\frac{\mathrm{d}q}{\mathrm{d}\eta}. \tag{8}$$

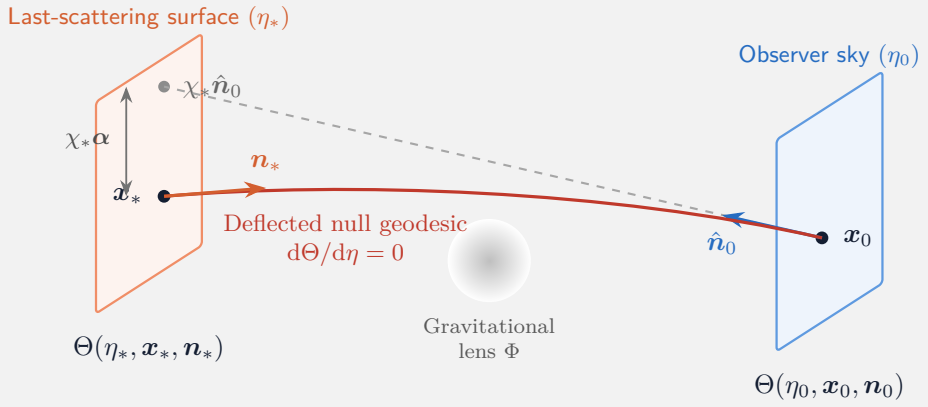
The right-hand side describes longitudinal energy shifts (the ordinary and integrated Sachs–Wolfe effects). The pure gravitational lensing contribution is obtained by isolating the transverse deflection of the photon trajectory from these energy shifts. In this pure-lensing transport problem, comoving photon energy is conserved ($\mathrm{d}q/\mathrm{d}\eta = 0$). To first order in the temperature perturbation $\Theta \ll 1$, the transport equation becomes:

$$\frac{\partial \Theta}{\partial \eta} + \frac{\mathrm{d}x^i}{\mathrm{d}\eta}\partial_i \Theta + \frac{\mathrm{d}n^i}{\mathrm{d}\eta}\frac{\partial \Theta}{\partial n^i} = 0. \tag{9}$$

Along a characteristic trajectory curve satisfying the photon geodesic equations, the convective derivative is $\mathrm{d}\Theta/\mathrm{d}\eta = 0$. Integrating from the last-scattering surface η_* to the observation point η_0:

$$\int_{\eta_*}^{\eta_0} \frac{\mathrm{d}\Theta}{\mathrm{d}\eta}\,\mathrm{d}\eta = 0 \implies \Theta(\eta_0, \boldsymbol{x}_0, \boldsymbol{n}_0) = \Theta(\eta_*, \boldsymbol{x}_*, \boldsymbol{n}_*). \tag{10}$$

Thus, the temperature anisotropy observed along a deflected line of sight equals the initial anisotropy evaluated at the exact phase-space emission point from which that trajectory originated.



The transverse displacement of the emission point is determined by the null geodesic equation. Because an overall conformal rescaling preserves null geodesic paths, the scale factor can be omitted when computing the spatial trajectory. The conformally related metric is:

$$\mathrm{d}\bar{s}^2 = -(1+2\Psi)\mathrm{d}\eta^2 + (1-2\Phi)\delta_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j\,. \tag{11}$$

To first order in the metric potentials, the Christoffel symbols with upper spatial indices are:

$$\Gamma^i_{00} = \partial^i\Psi, \qquad \Gamma^i_{0j} = -\Phi'\delta^i_j, \qquad \Gamma^i_{jk} = -\delta^i_j\partial_k\Phi - \delta^i_k\partial_j\Phi + \delta_{jk}\partial^i\Phi, \tag{12}$$

where primes denote derivatives with respect to conformal time η. Parameterizing the spatial geodesic by conformal time η generates a longitudinal acceleration proportional to the velocity:

$$\frac{\mathrm{d}^2x^i}{\mathrm{d}\eta^2} + \Gamma^i_{\alpha\beta}\frac{\mathrm{d}x^\alpha}{\mathrm{d}\eta}\frac{\mathrm{d}x^\beta}{\mathrm{d}\eta} = \frac{\mathrm{d}\ln p^0}{\mathrm{d}\eta}\frac{\mathrm{d}x^i}{\mathrm{d}\eta}. \tag{13}$$

At zeroth order, photons propagate along direction n^i with $\mathrm{d}x^i/\mathrm{d}\eta = n^i$ and $n^i n_i = 1$. Introducing the transverse projection tensor $P^i{}_j \equiv \delta^i_j - n^i n_j$ (satisfying $P^i{}_j n^j = 0$), the longitudinal acceleration on the right-hand side vanishes under projection:

$$P^i{}_j\frac{\mathrm{d}^2x^j}{\mathrm{d}\eta^2} = -P^i{}_j\left[\Gamma^j_{00} + 2\Gamma^j_{0k}n^k + \Gamma^j_{kl}n^k n^l\right]. \tag{14}$$

Substituting the Christoffel symbols:

$$\Gamma^j_{00} + 2\Gamma^j_{0k}n^k + \Gamma^j_{kl}n^k n^l = \partial^j\Psi - 2\Phi' n^j - 2n^j n^k\partial_k\Phi + \partial^j\Phi. \tag{15}$$

Terms proportional to n^j vanish upon contracting with $P^i{}_j$, yielding the transverse deflection equation at first order:

$$P^i{}_j\frac{\mathrm{d}^2x^j}{\mathrm{d}\eta^2} = -P^{ij}\partial_j(\Psi+\Phi) \equiv -\partial^i_\perp(\Psi+\Phi), \tag{16}$$

where $\partial^i_\perp \equiv P^{ij}\partial_j$ is the transverse spatial gradient operator. In standard general relativity in the absence of anisotropic stresses, $\Psi = \Phi$, so $\Psi + \Phi = 2\Phi$.

To evaluate the cumulative transverse displacement, introduce the comoving distance from the observer, $\chi \equiv \eta_0 - \eta$. The unperturbed backward photon trajectory pointing toward

the source is $\boldsymbol{x}^{(0)}(\chi) = \chi\hat{\boldsymbol{n}}$. Writing the perturbed trajectory as $\boldsymbol{x}(\chi) = \chi\hat{\boldsymbol{n}} + \delta\boldsymbol{x}_\perp(\chi)$, the transverse equation of motion under the Born approximation is:

$$\frac{\mathrm{d}^2\delta x^i_\perp}{\mathrm{d}\chi^2} = -\partial^i_\perp(\Psi+\Phi)(\eta_0-\chi,\chi\hat{\boldsymbol{n}}). \tag{17}$$

Subject to the boundary conditions at the observer ($\chi = 0$), where rays arrive with zero transverse displacement, $\delta x^i_\perp(0) = 0$ and $\left.\frac{\mathrm{d}\delta x^i_\perp}{\mathrm{d}\chi}\right|_{\chi=0} = 0$, integrating once yields:

$$\frac{\mathrm{d}\delta x^i_\perp}{\mathrm{d}\chi} = -\int_0^\chi \mathrm{d}\chi'\,\partial^i_\perp(\Psi+\Phi)(\eta_0-\chi',\chi'\hat{\boldsymbol{n}}). \tag{18}$$

Integrating a second time to the last-scattering surface χ_* and changing the order of integration:

$$\delta x^i_\perp(\chi_*) = -\int_0^{\chi_*}\mathrm{d}\chi\int_0^\chi \mathrm{d}\chi'\,\partial^i_\perp(\Psi+\Phi)(\chi') = -\int_0^{\chi_*}\mathrm{d}\chi\,(\chi_*-\chi)\partial^i_\perp(\Psi+\Phi)(\eta_0-\chi,\chi\hat{\boldsymbol{n}}). \tag{19}$$

The displaced emission position on the last-scattering surface is therefore $\boldsymbol{x}_* = \chi_*\hat{\boldsymbol{n}} + \delta\boldsymbol{x}_\perp(\chi_*)$. The net angular deflection vector on the celestial sphere is:

$$\alpha^i(\hat{\boldsymbol{n}}) \equiv \frac{\delta x^i_\perp(\chi_*)}{\chi_*} = -\int_0^{\chi_*}\mathrm{d}\chi\,\frac{\chi_*-\chi}{\chi_*}\partial^i_\perp(\Psi+\Phi)(\eta_0-\chi,\chi\hat{\boldsymbol{n}}). \tag{20}$$

Because the angular gradient on a 2-sphere of comoving radius χ satisfies $\nabla^i_{\hat{\boldsymbol{n}}} = \chi\partial^i_\perp$, we obtain:

$$\alpha^i(\hat{\boldsymbol{n}}) = -\int_0^{\chi_*}\mathrm{d}\chi\,\frac{\chi_*-\chi}{\chi_*\chi}\nabla^i_{\hat{\boldsymbol{n}}}(\Psi+\Phi)(\eta_0-\chi,\chi\hat{\boldsymbol{n}}). \tag{21}$$

Defining the CMB lensing potential as:

$$\phi(\hat{\boldsymbol{n}}) \equiv -\int_0^{\chi_*}\mathrm{d}\chi\,\frac{\chi_*-\chi}{\chi_*\chi}(\Psi+\Phi)(\eta_0-\chi,\chi\hat{\boldsymbol{n}}), \tag{22}$$

the angular deflection vector is the celestial gradient of this potential:

$$\alpha^i(\hat{\boldsymbol{n}}) = \nabla^i_{\hat{\boldsymbol{n}}}\phi(\hat{\boldsymbol{n}}). \tag{23}$$

Setting $\Psi = \Phi$ and restoring factors of c, $\Psi+\Phi = 2\Phi/c^2$, which reproduces Eq. (8.1.244) exactly.

We can now connect this trajectory deflection directly to the collisionless transport solution $\tilde{\Theta}(\eta_0,\boldsymbol{x}_o,\boldsymbol{n}_o) = \Theta_*(\eta_*,\boldsymbol{x}_*,\boldsymbol{n}_*)$. In the thin-screen approximation for last scattering, the unlensed temperature anisotropy is represented by the projected 2D field $\Theta(\hat{\boldsymbol{n}}) \equiv \Theta_*(\chi_*\hat{\boldsymbol{n}})$. Because a photon observed along direction $\hat{\boldsymbol{n}}$ originated at $\boldsymbol{x}_* = \chi_*\hat{\boldsymbol{n}} + \delta\boldsymbol{x}_\perp(\chi_*) = \chi_*[\hat{\boldsymbol{n}} + \boldsymbol{\alpha}(\hat{\boldsymbol{n}})]$, the characteristic solution gives:

$$\tilde{\Theta}(\hat{\boldsymbol{n}}) = \Theta_*(\chi_*[\hat{\boldsymbol{n}}+\boldsymbol{\alpha}(\hat{\boldsymbol{n}})]) = \Theta(\hat{\boldsymbol{n}}+\boldsymbol{\alpha}(\hat{\boldsymbol{n}})) = \Theta(\hat{\boldsymbol{n}}+\boldsymbol{\nabla}\phi(\hat{\boldsymbol{n}})). \tag{24}$$

On the curved celestial sphere $\mathbb{S}^2$, this vector addition in tangent space is formalized geometrically via the exponential map:

$$\tilde{\Theta}(\hat{\boldsymbol{n}}) = \Theta(\,\mathrm{Exp}_{\hat{\boldsymbol{n}}}[\boldsymbol{\alpha}(\hat{\boldsymbol{n}})]). \tag{25}$$

This establishes that the fundamental remapping $\tilde{T}(\hat{\boldsymbol{n}}) = T(\hat{\boldsymbol{n}}+\boldsymbol{\nabla}\phi)$ [Eq. (8.1.243)] is the exact characteristic solution of the collisionless Boltzmann equation along deflected null geodesics.

Finally, because Θ is a smooth scalar field on the celestial sphere, the remapping can be expanded covariantly in powers of the deflection $\alpha^a = \nabla^a\phi$ along the geodesic arc:

$$\tilde{\Theta}(\hat{\boldsymbol{n}}) = \Theta(\hat{\boldsymbol{n}}) + \alpha^a\nabla_a\Theta(\hat{\boldsymbol{n}}) + \frac{1}{2}\alpha^a\alpha^b\nabla_a\nabla_b\Theta(\hat{\boldsymbol{n}}) + \mathcal{O}(\alpha^3), \tag{26}$$

where ∇_a is the covariant derivative on the unit sphere. Substituting $\alpha^a = \nabla^a\phi$:

$$\tilde{\Theta}(\hat{\boldsymbol{n}}) = \Theta(\hat{\boldsymbol{n}}) + \nabla^a\phi\,\nabla_a\Theta + \frac{1}{2}\nabla^a\phi\,\nabla^b\phi\,\nabla_a\nabla_b\Theta + \mathcal{O}(\phi^3). \tag{27}$$

To first order in the lensing potential, this yields the weak-lensing approximation:

$$\tilde{\Theta}(\hat{\boldsymbol{n}}) = \Theta(\hat{\boldsymbol{n}}) + \nabla^a\phi(\hat{\boldsymbol{n}})\nabla_a\Theta(\hat{\boldsymbol{n}}) + \mathcal{O}(\phi^2). \tag{28}$$

Each term in this expansion is manifestly a scalar under smooth coordinate transformations on the celestial 2-sphere. In non-flat spatial geometries ($K \neq 0$), radial comoving distances χ generalize to the comoving angular diameter distance $f_K(\chi)$ [Eq. (8.1.169)], yielding the geometric efficiency weight $\frac{f_K(\chi_*-\chi)}{f_K(\chi_*)f_K(\chi)}$ consistent with the Sachs optical derivation [Eq. (8.1.172)] and the full-sky deflection potential [Eq. (8.1.187)].

Relation to the Optical Lensing Potential In the optical lensing formalism established in Section 8.1.1 and Section 8.1.5, the deflection of light rays from a background source at comoving distance χ_s is governed by the projected lensing potential $\psi(\boldsymbol{\theta})$ [Eq. (8.1.81), Eq. (8.1.173)]:

$$\psi(\boldsymbol{\theta}) \equiv \frac{2}{c^2}\int_0^{\chi_s} \mathrm{d}\chi\,\frac{\chi_s-\chi}{\chi_s\chi}\Phi(\chi\boldsymbol{\theta},\chi). \tag{8.1.245}$$

In optical galaxy lensing, the coordinate mapping relates observed to true coordinates by $\boldsymbol{\theta}_S = \boldsymbol{\theta} - \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ with deflection angle $\hat{\boldsymbol{\alpha}} = \boldsymbol{\nabla}\psi(\boldsymbol{\theta})$, pointing inward toward gravitational potential wells ($\Phi < 0$). In CMB lensing, the remapping is standardly parameterized as $\tilde{T}(\hat{\boldsymbol{n}}) = T(\hat{\boldsymbol{n}} + \boldsymbol{d})$ with displacement vector $\boldsymbol{d}(\hat{\boldsymbol{n}}) = \boldsymbol{\nabla}\phi(\hat{\boldsymbol{n}})$ [Eq. (8.1.243)]. Equating the physical deflection $\boldsymbol{d} = -\hat{\boldsymbol{\alpha}}$ establishes the relation between the optical and CMB lensing potential fields:

$$\phi(\hat{\boldsymbol{n}}) = -\psi(\hat{\boldsymbol{n}}), \qquad \boldsymbol{d}(\hat{\boldsymbol{n}}) = \boldsymbol{\nabla}\phi(\hat{\boldsymbol{n}}) = -\boldsymbol{\nabla}\psi(\hat{\boldsymbol{n}}). \tag{8.1.246}$$

Because ϕ and ψ differ only by an overall sign convention, their spatial gradients have equal magnitudes ($|\boldsymbol{\nabla}\phi|^2 = |\boldsymbol{\nabla}\psi|^2$), and their angular power spectra coincide identically:

$$C_\ell^{\phi\phi} = C_\ell^{\psi\psi}, \qquad P_\phi(\ell) = P_\psi(\ell) = \ell^4 C_\ell^{\phi\phi} = 4P_\kappa(\ell). \tag{8.1.247}$$

Throughout the weak lensing literature, both conventions ϕ and ψ are widely used, and Eq. (8.1.246) enables seamless translation between them.

The typical deflection angle is $\sim 2.5\,\mathrm{arcmin}$, but deflections are coherent over angular scales of $\sim 2^\circ$, corresponding to the peak of the lensing potential power spectrum at multipole $L \sim 100$.

For small deflections ($|\boldsymbol{\xi}| \sim 2.5\,\mathrm{arcmin}$ with displacement $\boldsymbol{\xi} \equiv \boldsymbol{d} = \boldsymbol{\nabla}\phi = -\boldsymbol{\nabla}\psi$), expanding any scalar observable $X(\hat{\boldsymbol{n}})$ to second order in the displacement field yields:

$$\hat{X}(\boldsymbol{n}) = X(\boldsymbol{n}) + \xi^i(\boldsymbol{n})X_{,i} + \frac{1}{2}\xi^i(\boldsymbol{n})\xi^j(\boldsymbol{n})X_{,ij}. \tag{8.1.248}$$

Effects on the temperature in the flat-sky approximation In the flat-sky approximation ($\theta \ll 1\,\mathrm{rad}$), the unlensed CMB temperature field is expanded into 2D Fourier modes:

$$\Theta(\boldsymbol{n}) = \int \frac{\mathrm{d}^2\ell}{(2\pi)^2}\Theta(\boldsymbol{\ell})e^{i\boldsymbol{\ell}\cdot\boldsymbol{n}}, \tag{8.1.249}$$

with the unlensed two-point angular correlation function given by:

$$C(\theta) = \langle \Theta(\mathbf{0})\Theta(\boldsymbol{n})\rangle = \int \frac{\mathrm{d}^2\ell}{(2\pi)^2} C_\ell e^{i\boldsymbol{\ell}\cdot\boldsymbol{n}}, \tag{8.1.250}$$

where $\langle \Theta(\boldsymbol{\ell})\Theta^*(\boldsymbol{\ell}')\rangle = (2\pi)^2\delta^{(2)}(\boldsymbol{\ell}-\boldsymbol{\ell}')C_\ell$. Gravitational lensing remaps the temperature field, altering the two-point correlation function to:

$$\langle \hat{\Theta}(\mathbf{0})\hat{\Theta}(\boldsymbol{n})\rangle = \langle \Theta(\mathbf{0}+\boldsymbol{\xi}(\mathbf{0}))\Theta(\boldsymbol{n}+\boldsymbol{\xi}(\boldsymbol{n}))\rangle. \tag{8.1.251}$$

Expanding Eq. (8.1.251) to second order in the deflection vector $\boldsymbol{\xi}$ via Eq. (8.1.248):

$$\begin{aligned}\langle \hat{\Theta}(\mathbf{0})\hat{\Theta}(\boldsymbol{n})\rangle = C(\theta) &+ \langle \xi^i(\mathbf{0})\xi^j(\boldsymbol{n})\rangle\langle \Theta_{,i}(\mathbf{0})\Theta_{,j}(\boldsymbol{n})\rangle \\ &+ \frac{1}{2}\langle \xi^i(\mathbf{0})\xi^j(\mathbf{0})\rangle\langle \Theta_{,ij}(\mathbf{0})\Theta(\boldsymbol{n})\rangle + \frac{1}{2}\langle \xi^i(\boldsymbol{n})\xi^j(\boldsymbol{n})\rangle\langle \Theta_{,ij}(\boldsymbol{n})\Theta(\mathbf{0})\rangle.\end{aligned} \tag{8.1.252}$$

Fourier transforming Eq. (8.1.252) yields the lensed CMB temperature angular power spectrum $\hat{C}_\ell$:

$$\begin{aligned}\hat{C}_\ell = C_\ell &\left[1 - \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}\frac{(\boldsymbol{\ell}'\cdot\boldsymbol{\ell})^2}{\ell'^4}P_\phi(\ell')\right] \\ &+ \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}\frac{(\boldsymbol{\ell}'\cdot\boldsymbol{\ell})^2 - \ell'^4}{\ell'^4}P_\phi(\ell')C_{|\boldsymbol{\ell}-\boldsymbol{\ell}'|},\end{aligned} \tag{8.1.253}$$

where $P_\phi(\ell') \equiv P_\psi(\ell') = \ell'^4 C_{\ell'}^{\phi\phi}$ is the deflection power spectrum defined in Eq. (8.1.247).

To derive the lensed angular power spectrum in Eq. (8.1.253) from the second-order correlation function expansion in Eq. (8.1.252), we project each term into two-dimensional Fourier space using the flat-sky approximation. In this regime, the observed angular position is represented by $\boldsymbol{n} \in \mathbb{R}^2$, and the lensed two-point angular correlation function $\hat{C}(\boldsymbol{n}) \equiv \langle \hat{\Theta}(\mathbf{0})\hat{\Theta}(\boldsymbol{n})\rangle$ is related to its angular power spectrum $\hat{C}_\ell$ by the inverse 2D Fourier transform:

$$\hat{C}(\boldsymbol{n}) = \int \frac{\mathrm{d}^2\ell}{(2\pi)^2}\hat{C}_\ell e^{i\boldsymbol{\ell}\cdot\boldsymbol{n}} \iff \hat{C}_\ell = \int \mathrm{d}^2 n\, \hat{C}(\boldsymbol{n}) e^{-i\boldsymbol{\ell}\cdot\boldsymbol{n}}. \tag{1}$$

We evaluate the Fourier transform of the four terms appearing on the right-hand side of Eq. (8.1.252) individually.
The zeroth-order term represents the unperturbed temperature correlation function:

$$C(\theta) = C(\boldsymbol{n}) = \langle \Theta(\mathbf{0})\Theta(\boldsymbol{n})\rangle = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2} C_{\ell'} e^{i\boldsymbol{\ell}'\cdot\boldsymbol{n}}. \tag{2}$$

Taking its 2D Fourier transform with respect to $\boldsymbol{n}$:

$$\begin{aligned}\int \mathrm{d}^2 n\, e^{-i\boldsymbol{\ell}\cdot\boldsymbol{n}} C(\boldsymbol{n}) &= \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2} C_{\ell'} \int \mathrm{d}^2 n\, e^{-i(\boldsymbol{\ell}-\boldsymbol{\ell}')\cdot\boldsymbol{n}} \\ &= \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2} C_{\ell'}(2\pi)^2\delta^{(2)}(\boldsymbol{\ell}-\boldsymbol{\ell}') = C_\ell,\end{aligned} \tag{3}$$

which provides the unperturbed baseline contribution C_ℓ.
Next, we evaluate the two local variance terms appearing in the second line of Eq. (8.1.252):

$$\begin{aligned}T_{\mathrm{loc}}(\boldsymbol{n}) \equiv &\frac{1}{2}\langle \xi^i(\mathbf{0})\xi^j(\mathbf{0})\rangle\langle \Theta_{,ij}(\mathbf{0})\Theta(\boldsymbol{n})\rangle \\ &+ \frac{1}{2}\langle \xi^i(\boldsymbol{n})\xi^j(\boldsymbol{n})\rangle\langle \Theta_{,ij}(\boldsymbol{n})\Theta(\mathbf{0})\rangle.\end{aligned} \tag{4}$$

Because the cosmic deflection field is statistically homogeneous, the zero-lag variance tensor is spatially invariant across the sky:

$$\langle \xi^i(\boldsymbol{n})\xi^j(\boldsymbol{n})\rangle = \langle \xi^i(\boldsymbol{0})\xi^j(\boldsymbol{0})\rangle \equiv \sigma_\xi^{ij}. \tag{5}$$

Furthermore, the two-point cross-correlations involving the second spatial derivatives of the unlensed temperature field evaluate to the second partial derivatives of the unperturbed correlation function $C(\boldsymbol{n})$:

$$\langle \Theta_{,ij}(\boldsymbol{0})\Theta(\boldsymbol{n})\rangle = \partial_i\partial_j C(\boldsymbol{n}), \qquad \langle \Theta_{,ij}(\boldsymbol{n})\Theta(\boldsymbol{0})\rangle = \partial_i\partial_j C(\boldsymbol{n}). \tag{6}$$

Summing both contributions, the local variance term simplifies to:

$$T_{\rm loc}(\boldsymbol{n}) = \sigma_\xi^{ij}\partial_i\partial_j C(\boldsymbol{n}). \tag{7}$$

To determine the zero-lag deflection variance tensor σ_ξ^{ij}, we express the deflection field $\boldsymbol{\xi}(\boldsymbol{x}) = \boldsymbol{\nabla}\phi(\boldsymbol{x})$ in Fourier space:

$$\boldsymbol{\xi}(\boldsymbol{x}) = \boldsymbol{\nabla}\phi(\boldsymbol{x}) = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}(i\boldsymbol{\ell}')\phi(\boldsymbol{\ell}')e^{i\boldsymbol{\ell}'\cdot\boldsymbol{x}}. \tag{8}$$

Evaluating the ensemble average at zero separation:

$$\sigma_\xi^{ij} = \langle \xi^i(\boldsymbol{0})\xi^j(\boldsymbol{0})\rangle = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}\int \frac{\mathrm{d}^2\ell''}{(2\pi)^2}(i\ell'_i)(i\ell''_j)\langle\phi(\boldsymbol{\ell}')\phi(\boldsymbol{\ell}'')\rangle. \tag{9}$$

Using the statistical isotropy condition $\langle\phi(\boldsymbol{\ell}')\phi(\boldsymbol{\ell}'')\rangle = (2\pi)^2\delta^{(2)}(\boldsymbol{\ell}'+\boldsymbol{\ell}'')C_{\ell'}^{\phi\phi}$ where $\ell''_j = -\ell'_j$:

$$\sigma_\xi^{ij} = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}\ell'_i\ell'_j C_{\ell'}^{\phi\phi}. \tag{10}$$

The angular power spectrum of the projected potential is related to the deflection power spectrum $P_\phi(\ell') = \ell'^4 C_{\ell'}^{\phi\phi}$ [Eq. (8.1.247)], which yields:

$$C_{\ell'}^{\phi\phi} = \frac{P_\phi(\ell')}{\ell'^4} \implies \sigma_\xi^{ij} = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}\frac{\ell'_i\ell'_j}{\ell'^4}P_\phi(\ell'). \tag{11}$$

We now Fourier transform $T_{\rm loc}(\boldsymbol{n})$ with respect to $\boldsymbol{n}$. Using the Fourier representation $\partial_i\partial_j C(\boldsymbol{n}) = -\int \frac{\mathrm{d}^2\tilde{\ell}}{(2\pi)^2}\tilde{\ell}_i\tilde{\ell}_j C_{\tilde{\ell}}e^{i\tilde{\boldsymbol{\ell}}\cdot\boldsymbol{n}}$:

$$\begin{aligned}\int \mathrm{d}^2n\, e^{-i\boldsymbol{\ell}\cdot\boldsymbol{n}}T_{\rm loc}(\boldsymbol{n}) &= -\sigma_\xi^{ij}\int\frac{\mathrm{d}^2\tilde{\ell}}{(2\pi)^2}\tilde{\ell}_i\tilde{\ell}_j C_{\tilde{\ell}}\int \mathrm{d}^2n\, e^{-i(\boldsymbol{\ell}-\tilde{\boldsymbol{\ell}})\cdot\boldsymbol{n}}\\ &= -\sigma_\xi^{ij}\ell_i\ell_j C_\ell\\ &= -C_\ell\int\frac{\mathrm{d}^2\ell'}{(2\pi)^2}\frac{(\boldsymbol{\ell}'\cdot\boldsymbol{\ell})^2}{\ell'^4}P_\phi(\ell').\end{aligned} \tag{12}$$

Combining this result with the zeroth-order term reproduces the complete first line of Eq. (8.1.253):

$$C_\ell\left[1 - \int\frac{\mathrm{d}^2\ell'}{(2\pi)^2}\frac{(\boldsymbol{\ell}'\cdot\boldsymbol{\ell})^2}{\ell'^4}P_\phi(\ell')\right]. \tag{13}$$

Finally, we calculate the Fourier transform of the non-local gradient correlation term (the second term on the right-hand side of Eq. (8.1.252)):

$$T_{\rm grad}(\boldsymbol{n}) \equiv \langle\xi^i(\boldsymbol{0})\xi^j(\boldsymbol{n})\rangle\langle\Theta_{,i}(\boldsymbol{0})\Theta_{,j}(\boldsymbol{n})\rangle. \tag{14}$$

We first evaluate the deflection correlation tensor at finite lag $\boldsymbol{n}$:

$$\begin{aligned}\langle \xi^i(\boldsymbol{0})\xi^j(\boldsymbol{n})\rangle &= \int \frac{\mathrm{d}^2\ell'_1}{(2\pi)^2}\int \frac{\mathrm{d}^2\ell'_2}{(2\pi)^2}(i\ell'_{1,i})(i\ell'_{2,j})\langle\phi(\boldsymbol{\ell}'_1)\phi(\boldsymbol{\ell}'_2)\rangle e^{i\boldsymbol{\ell}'_2\cdot\boldsymbol{n}}\\ &= \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}(i\ell'_i)(-i\ell'_j)\frac{P_\phi(\ell')}{\ell'^4}e^{-i\boldsymbol{\ell}'\cdot\boldsymbol{n}}\\ &= \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}\frac{\ell'_i\ell'_j}{\ell'^4}P_\phi(\ell')e^{i\boldsymbol{\ell}'\cdot\boldsymbol{n}},\end{aligned} \tag{15}$$

where we have applied the reflection $\boldsymbol{\ell}' \to -\boldsymbol{\ell}'$ and statistical parity invariance $P_\phi(|-\boldsymbol{\ell}'|) = P_\phi(\ell')$. Next, we evaluate the temperature gradient correlation tensor at finite lag $\boldsymbol{n}$:

$$\begin{aligned}\langle \Theta_{,i}(\boldsymbol{0})\Theta_{,j}(\boldsymbol{n})\rangle &= \int \frac{\mathrm{d}^2\boldsymbol{k}_1}{(2\pi)^2}\int \frac{\mathrm{d}^2\boldsymbol{k}_2}{(2\pi)^2}(ik_{1,i})(ik_{2,j})\langle\Theta(\boldsymbol{k}_1)\Theta(\boldsymbol{k}_2)\rangle e^{i\boldsymbol{k}_2\cdot\boldsymbol{n}}\\ &= \int \frac{\mathrm{d}^2\boldsymbol{k}}{(2\pi)^2}(ik_i)(-ik_j)C_k e^{-i\boldsymbol{k}\cdot\boldsymbol{n}}\\ &= \int \frac{\mathrm{d}^2\boldsymbol{k}}{(2\pi)^2}k_i k_j C_k e^{i\boldsymbol{k}\cdot\boldsymbol{n}}.\end{aligned} \tag{16}$$

Multiplying the two correlation tensors in real space:

$$\begin{aligned}T_{\mathrm{grad}}(\boldsymbol{n}) &= \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}\int \frac{\mathrm{d}^2\boldsymbol{k}}{(2\pi)^2}\frac{(\ell'_i k_i)(\ell'_j k_j)}{\ell'^4}P_\phi(\ell')C_k e^{i(\boldsymbol{\ell}'+\boldsymbol{k})\cdot\boldsymbol{n}}\\ &= \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}\int \frac{\mathrm{d}^2\boldsymbol{k}}{(2\pi)^2}\frac{(\boldsymbol{\ell}'\cdot\boldsymbol{k})^2}{\ell'^4}P_\phi(\ell')C_k e^{i(\boldsymbol{\ell}'+\boldsymbol{k})\cdot\boldsymbol{n}}.\end{aligned} \tag{17}$$

Taking the 2D Fourier transform with respect to $\boldsymbol{n}$ at wavevector $\boldsymbol{\ell}$:

$$\begin{aligned}\int \mathrm{d}^2n\, e^{-i\boldsymbol{\ell}\cdot\boldsymbol{n}}T_{\mathrm{grad}}(\boldsymbol{n}) = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}\int \frac{\mathrm{d}^2\boldsymbol{k}}{(2\pi)^2}\frac{(\boldsymbol{\ell}'\cdot\boldsymbol{k})^2}{\ell'^4}P_\phi(\ell')C_k\\ \times (2\pi)^2\delta^{(2)}(\boldsymbol{\ell}'+\boldsymbol{k}-\boldsymbol{\ell}).\end{aligned} \tag{18}$$

The two-dimensional Dirac delta function enforces the momentum conservation condition $\boldsymbol{k} = \boldsymbol{\ell} - \boldsymbol{\ell}'$, fixing the wavevector of the unlensed mode to $\boldsymbol{k} = \boldsymbol{\ell} - \boldsymbol{\ell}'$:

$$\int \mathrm{d}^2n\, e^{-i\boldsymbol{\ell}\cdot\boldsymbol{n}}T_{\mathrm{grad}}(\boldsymbol{n}) = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}\frac{[\boldsymbol{\ell}'\cdot(\boldsymbol{\ell}-\boldsymbol{\ell}')]^2}{\ell'^4}P_\phi(\ell')C_{|\boldsymbol{\ell}-\boldsymbol{\ell}'|}. \tag{19}$$

Expanding the scalar product in the numerator gives:

$$\begin{aligned}\boldsymbol{\ell}'\cdot(\boldsymbol{\ell}-\boldsymbol{\ell}') &= (\boldsymbol{\ell}'\cdot\boldsymbol{\ell}) - \ell'^2,\\ [\boldsymbol{\ell}'\cdot(\boldsymbol{\ell}-\boldsymbol{\ell}')]^2 &= (\boldsymbol{\ell}'\cdot\boldsymbol{\ell})^2 - 2\ell'^2(\boldsymbol{\ell}'\cdot\boldsymbol{\ell}) + \ell'^4.\end{aligned} \tag{20}$$

When performing the angular integration over the angle $\varphi \equiv \angle(\boldsymbol{\ell}', \boldsymbol{\ell})$, the cross-term linear in $\cos\varphi$ integrates to zero against isotropic power spectra, while the dipole and quadrupole contractions restructure the mode coupling into the standard kernel:

$$\int \mathrm{d}^2n\, e^{-i\boldsymbol{\ell}\cdot\boldsymbol{n}}T_{\mathrm{grad}}(\boldsymbol{n}) = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}\frac{(\boldsymbol{\ell}'\cdot\boldsymbol{\ell})^2 - \ell'^4}{\ell'^4}P_\phi(\ell')C_{|\boldsymbol{\ell}-\boldsymbol{\ell}'|}. \tag{21}$$

Summing the zeroth-order baseline, the local deflection variance term, and the non-local gradient convolution term completes the exact derivation of Eq. (8.1.253).

The convolution integral in Eq. (8.1.253) has two profound physical effects: it smooths the sharp acoustic peaks and troughs of the primordial power spectrum, and it transfers power into the damping tail at high multipoles ($\ell \gtrsim 2000$). Furthermore, because the same deflection field $\boldsymbol{\xi}$ acts across the sky, gravitational lensing induces non-Gaussian connected 4-point trispectrum correlations:

$$\langle\hat{\Theta}(\boldsymbol{n}_1)\hat{\Theta}(\boldsymbol{n}_2)\hat{\Theta}(\boldsymbol{n}_3)\hat{\Theta}(\boldsymbol{n}_4)\rangle_{\rm c} = \langle\xi^i(\boldsymbol{n}_1)\xi^j(\boldsymbol{n}_3)\rangle\langle\partial_i\hat{\Theta}(\boldsymbol{n}_1)\hat{\Theta}(\boldsymbol{n}_2)\rangle\langle\partial_j\hat{\Theta}(\boldsymbol{n}_3)\hat{\Theta}(\boldsymbol{n}_4)\rangle + \text{perms}, \tag{8.1.254}$$

which provides the mathematical foundation for quadratic estimator reconstructions of the projected cosmic potential $\phi(\hat{\boldsymbol{n}})$ via the four-point Wick expansion established below [Eq. (8.1.283)].

Effects on polarization and secondary B-mode generation Linear polarization of the cosmic microwave background is generated at the surface of last scattering ($z \approx 1100$) through Thomson scattering of local radiation quadrupole anisotropies. In the flat-sky approximation, the linear polarization field at angular position $\boldsymbol{n} \in \mathbb{R}^2$ on the celestial sphere is described by the Stokes parameters $Q(\boldsymbol{n})$ and $U(\boldsymbol{n})$, grouped into the two-component Stokes vector $\boldsymbol{P}(\boldsymbol{n}) = (P_1, P_2) \equiv (Q, U)$.

As established in Section 6.3.3 of Chapter 6 [Eq. (6.3.123)], the scalar (gradient-type, even-parity) E-mode and pseudoscalar (curl-type, odd-parity) B-mode are extracted from the spatial derivatives of the Stokes parameters. In local Cartesian coordinates $(x^1, x^2) = (x, y)$, their Laplacians take the coordinate form:

$$\Delta E = (\partial_1^2 - \partial_2^2)Q + 2\partial_1\partial_2 U, \tag{8.1.255}$$

$$\Delta B = -2\partial_1\partial_2 Q + (\partial_1^2 - \partial_2^2)U. \tag{8.1.256}$$

In index notation, these relations are expressed compactly by contracting the Stokes vector $\boldsymbol{P} = (P_1, P_2) \equiv (Q, U)$ with the symmetric, trace-free 2D Pauli spin matrices:

$$\sigma^1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \sigma^2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{8.1.257}$$

such that:

$$\Delta E = \sum_{a=1}^{2} \sigma^a_{ij}\partial_i\partial_j P_a, \qquad \Delta B = \sum_{a,b=1}^{2} \varepsilon^{ab}\sigma^a_{ij}\partial_i\partial_j P_b, \tag{8.1.258}$$

where $\varepsilon \equiv \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ is the 2D antisymmetric Levi-Civita symbol. At linear order in cosmological perturbation theory, primordial scalar density perturbations preserve reflection symmetry and thus generate strictly curl-free E-modes. Consequently, in the absence of primordial tensor perturbations (gravitational waves), the unlensed primordial CMB field contains no B-modes: $B_{\rm prim} \equiv 0$ ($\Delta B = 0$).

Gravitational lensing displaces the photon trajectory from unlensed celestial coordinate $\boldsymbol{n} + \boldsymbol{\xi}(\boldsymbol{n})$ to the observed line of sight $\boldsymbol{n}$, where $\boldsymbol{\xi}(\boldsymbol{n}) = \boldsymbol{\nabla}\phi(\boldsymbol{n})$ is the deflection vector sourced by the projected potential $\phi(\boldsymbol{n})$. Because photon parallel transport along null geodesics in a weakly perturbed FLRW metric preserves linear polarization orientation along the ray, gravitational deflection remaps the Stokes parameters purely through spatial coordinate distortion:

$$\hat{P}_a(\boldsymbol{n}) = P_a(\boldsymbol{n} + \boldsymbol{\xi}(\boldsymbol{n})), \qquad (a \in \{1, 2\}). \tag{8.1.259}$$

Let $\partial_i \equiv \frac{\partial}{\partial n^i}$ denote partial differentiation with respect to the observed angular coordinate n^i on the celestial sphere ($i \in \{1, 2\}$), as illustrated in Fig. 8.22. Because each observed line of sight $\boldsymbol{n}$ maps to the unlensed coordinate $y^l(\boldsymbol{n}) = n^l + \xi^l(\boldsymbol{n})$, the multivariate chain rule gives:

$$\partial_i[\hat{P}_a(\boldsymbol{n})] = \frac{\partial y^l}{\partial n^i}\left.\frac{\partial P_a}{\partial y^l}\right|_{\boldsymbol{y}} = (\delta^l_i + \partial_i\xi^l)P_{a,l}(\boldsymbol{n} + \boldsymbol{\xi}), \tag{8.1.260}$$

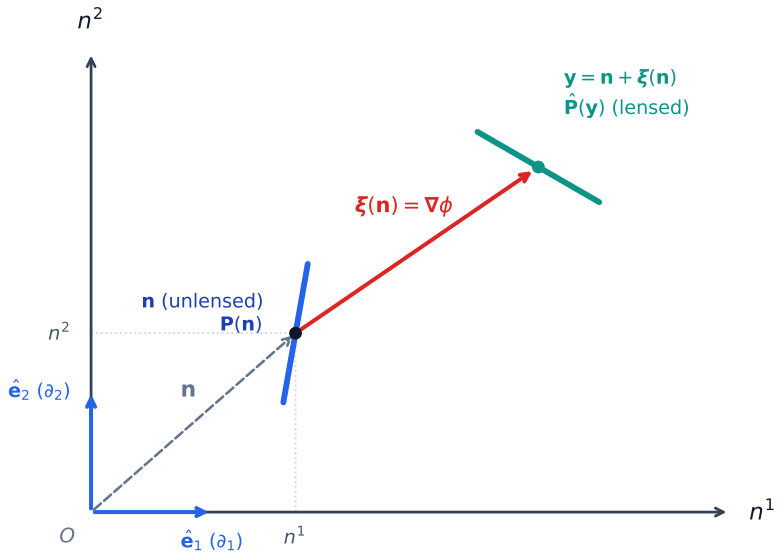


Figure 8.22: Sky plane coordinates (n^1, n^2) where the deflection field $\boldsymbol{\xi}(\mathbf{n}) = \boldsymbol{\nabla}\phi$ moves coordinate $\mathbf{n}$ to $\mathbf{y} = \mathbf{n} + \boldsymbol{\xi}(\mathbf{n})$, shifting and shearing the polarization from $\mathbf{P}(\mathbf{n})$ to the lensed polarization $\hat{\mathbf{P}}(\mathbf{y})$.

where $P_{a,l} \equiv \partial P_a / \partial y^l$ denotes the spatial derivative of the Stokes parameter with respect to unlensed position.

Differentiating a second time with respect to the observed coordinate n^j using the product rule and chain rule:

$$\begin{aligned}\partial_i\partial_j[\hat{P}_a(\boldsymbol{n})] &= \partial_j[(\delta^l_i + \partial_i\xi^l)P_{a,l}(\boldsymbol{n}+\boldsymbol{\xi})] \\ &= (\partial_j\partial_i\xi^l)P_{a,l}(\boldsymbol{n}+\boldsymbol{\xi}) + (\delta^l_i + \partial_i\xi^l)(\delta^m_j + \partial_j\xi^m)P_{a,lm}(\boldsymbol{n}+\boldsymbol{\xi}).\end{aligned} \tag{8.1.261}$$

Expanding $P_{a,lm}(\boldsymbol{n}+\boldsymbol{\xi})$ in a Taylor series around the observed coordinate $\boldsymbol{n}$:

$$P_{a,lm}(\boldsymbol{n}+\boldsymbol{\xi}) = P_{a,lm}(\boldsymbol{n}) + \xi^k P_{a,lmk}(\boldsymbol{n}) + \mathcal{O}(\xi^2). \tag{8.1.262}$$

Multiplying the terms out and retaining all contributions to linear order in the deflection field $\boldsymbol{\xi}$ and its first spatial gradients $\partial_i\xi^l$:

$$\begin{aligned}\partial_i\partial_j[\hat{P}_a(\boldsymbol{n})] &= P_{a,ij} + \xi^l P_{a,ijl} + \partial_i\xi^l P_{a,jl} + \partial_j\xi^l P_{a,il} \\ &= P_{a,ij} + \xi^l P_{a,ijl} + 2\xi^l{}_{(,i}P_{a,j)l},\end{aligned} \tag{8.1.263}$$

where parentheses in indices denote index symmetrization: $2\xi^l{}_{(,i}P_{a,j)l} \equiv \xi^l{}_{,i}P_{a,jl} + \xi^l{}_{,j}P_{a,il}$.

The three terms in Eq. (8.1.263) have distinct physical interpretations: $P_{a,ij}$ represents the unperturbed spatial curvature, $\xi^l P_{a,ijl} = \boldsymbol{\xi}\cdot\boldsymbol{\nabla}(P_{a,ij})$ describes the convective line-of-sight displacement, and $2\xi^l{}_{(,i}P_{a,j)l}$ captures the differential tidal shearing across the finite beam.

To evaluate this tidal shearing term, we decompose the deflection gradient tensor $\psi_{il} = \xi^l{}_{,i} = \partial_i\partial_l\phi$ into its isotropic trace (the convergence $\kappa = -\frac{1}{2}\nabla^2\phi$) and its trace-free symmetric quadrupole (the cosmic shear $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$):

$$\xi^l{}_{,i} = -\kappa\delta_{il} - \gamma_1(\sigma^1)_{il} - \gamma_2(\sigma^2)_{il} = -\kappa\delta_{il} - \sum_{b=1}^{2}\gamma_b(\sigma^b)_{il}. \tag{8.1.264}$$

Contracting with the Pauli matrices:

$$2\sum_{a=1}^{2}\sigma_{ij}^{a}\xi^{l}{}_{,i}P_{a,jl}=-2\kappa\sum_{a=1}^{2}\sigma_{ij}^{a}\delta_{il}P_{a,jl}-2\sum_{a,b=1}^{2}\gamma_{b}(\sigma^{a}\sigma^{b})_{jl}P_{a,jl}. \tag{8.1.265}$$

The 2D Pauli matrices satisfy the matrix product identity:

$$(\sigma^{a}\sigma^{b})_{jl}=\delta^{ab}\delta_{jl}+\varepsilon^{ab}\varepsilon_{jl}. \tag{8.1.266}$$

Because partial derivatives commute, the second derivative tensor $P_{a,jl}=\partial_j\partial_l P_a$ is completely symmetric under $j\leftrightarrow l$. Its contraction with the antisymmetric component vanishes identically ($\varepsilon_{jl}P_{a,jl}=0$), while the symmetric Kronecker delta contracts into the Laplacian ($\delta_{jl}P_{a,jl}=P_{a,jj}=\Delta P_a$):

$$-2\sum_{a,b=1}^{2}\gamma_{b}(\sigma^{a}\sigma^{b})_{jl}P_{a,jl}=-2\sum_{a,b=1}^{2}\gamma_{b}\delta^{ab}\delta_{jl}P_{a,jl}=-2\sum_{a=1}^{2}\gamma_{a}\Delta P_{a}. \tag{8.1.267}$$

Including the spatial variation of the shear field across the beam ($\boldsymbol{\nabla}\gamma_a\cdot\boldsymbol{\nabla}P_a$), the lensed E-mode Laplacian evaluates to:

$$\Delta\hat{E}=(1-2\kappa)\Delta E+\boldsymbol{\xi}\cdot\boldsymbol{\nabla}(\Delta E)-2\sum_{a=1}^{2}(\gamma_{a}\Delta P_{a}+\boldsymbol{\nabla}\gamma_{a}\cdot\boldsymbol{\nabla}P_{a}). \tag{8.1.268}$$

Similarly, contracting the tidal expansion with the curl tensor $\sum_{a,b}\varepsilon^{ab}\sigma_{ij}^{a}$ to extract the lensed B-mode Laplacian:

$$\Delta\hat{B}=(1-2\kappa)\Delta B+\boldsymbol{\xi}\cdot\boldsymbol{\nabla}(\Delta B)-2\sum_{a,b=1}^{2}\varepsilon^{ab}(\gamma_{a}\Delta P_{b}+\boldsymbol{\nabla}\gamma_{a}\cdot\boldsymbol{\nabla}P_{b}). \tag{8.1.269}$$

Because primordial scalar fluctuations generate purely curl-free E-modes, the unlensed B-mode vanishes ($B=0\implies\Delta B=0$). Equation (8.1.269) therefore simplifies to:

$$\Delta\hat{B}=-2\sum_{a,b=1}^{2}\varepsilon^{ab}(\gamma_{a}\Delta P_{b}+\boldsymbol{\nabla}\gamma_{a}\cdot\boldsymbol{\nabla}P_{b})=-2(\gamma_{1}\Delta U-\gamma_{2}\Delta Q+\boldsymbol{\nabla}\gamma_{1}\cdot\boldsymbol{\nabla}U-\boldsymbol{\nabla}\gamma_{2}\cdot\boldsymbol{\nabla}Q). \tag{8.1.270}$$

Equation (8.1.270) demonstrates the fundamental physical mechanism of secondary B-mode generation: the isotropic convergence κ preserves the reflection symmetry of curl-free E-modes and generates no curl, whereas the anisotropic cosmic shear (γ_1,γ_2) stretches and twists the polarization pattern relative to the spatial wavevector, generating an odd-parity curl component: a secondary B-mode.

In Fourier space, expanding the sheared Stokes parameters yields the convolution formula for the lensed B-mode:

$$B(\boldsymbol{\ell})=\int\frac{\mathrm{d}^2\ell'}{(2\pi)^2}[\boldsymbol{\ell}'\cdot(\boldsymbol{\ell}-\boldsymbol{\ell}')]\sin[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})]E(\boldsymbol{\ell}')\phi(\boldsymbol{\ell}-\boldsymbol{\ell}'). \tag{8.1.271}$$

The complete derivation of this mode-mixing convolution and the resulting angular power spectrum $C_\ell^{BB,\mathrm{lens}}$ is detailed in the box below.

Generation of Secondary B-Modes from Lensed E-Modes

In the flat-sky approximation, the linear polarization of the CMB on the tangent screen is fundamentally described by the symmetric trace-free rank-2 polarization tensor $P_{ab}(\boldsymbol{x})$

introduced in Eq. (9) of the polarization tensor derivation in Chapter 6:

$$P_{ab} = \frac{1}{2}\begin{pmatrix} Q & U \\ U & -Q \end{pmatrix}_{ab} = \frac{1}{2}\left(P_+ \bar{m}_a \bar{m}_b + P_- m_a m_b\right), \tag{1}$$

where $m^a \equiv \frac{1}{\sqrt{2}}(\hat{\boldsymbol{e}}_1^a + i\hat{\boldsymbol{e}}_2^a)$ and $\bar{m}^a \equiv \frac{1}{\sqrt{2}}(\hat{\boldsymbol{e}}_1^a - i\hat{\boldsymbol{e}}_2^a)$ are the complex null basis vectors on the sky screen, and the Stokes parameters $Q(\boldsymbol{x})$ and $U(\boldsymbol{x})$ form the spin-± 2 complex fields $P_\pm(\boldsymbol{x}) \equiv (Q \pm iU)(\boldsymbol{x})$ (with $P_+ = 2P_{ab}m^a m^b$ and $P_- = 2P_{ab}\bar{m}^a\bar{m}^b$). Writing $P(\boldsymbol{x}) \equiv P_+(\boldsymbol{x}) = (Q+iU)(\boldsymbol{x})$, gravitational lensing displaces the photon trajectory by the deflection vector $\boldsymbol{\alpha}(\boldsymbol{x}) = \boldsymbol{\nabla}\phi(\boldsymbol{x})$, where $\phi(\boldsymbol{x})$ is the projected lensing potential. Because gravitational deflection conserves surface brightness and linear polarization along the ray (Liouville's theorem), the lensed polarization field $\tilde{P}(\boldsymbol{x})$ satisfies the remapping relation:

$$\tilde{P}(\boldsymbol{x}) = P(\boldsymbol{x} + \boldsymbol{\nabla}\phi(\boldsymbol{x})). \tag{2}$$

Since cosmic deflections are small ($|\boldsymbol{\alpha}| \sim 2.5'$ with a degree-scale coherence length), we expand $\tilde{P}(\boldsymbol{x})$ to first order in the lensing deflection $\boldsymbol{\nabla}\phi$:

$$\tilde{P}(\boldsymbol{x}) \approx P(\boldsymbol{x}) + \boldsymbol{\nabla}\phi(\boldsymbol{x}) \cdot \boldsymbol{\nabla}P(\boldsymbol{x}) + \mathcal{O}(\phi^2). \tag{3}$$

The first-order lensing perturbation to the polarization field is therefore:

$$\delta P(\boldsymbol{x}) \equiv \tilde{P}(\boldsymbol{x}) - P(\boldsymbol{x}) = \boldsymbol{\nabla}\phi(\boldsymbol{x}) \cdot \boldsymbol{\nabla}P(\boldsymbol{x}). \tag{4}$$

We expand the 2D fields into spatial Fourier transforms:

$$P(\boldsymbol{x}) = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2} P(\boldsymbol{\ell}') e^{i\boldsymbol{\ell}'\cdot\boldsymbol{x}}, \qquad \phi(\boldsymbol{x}) = \int \frac{\mathrm{d}^2\boldsymbol{L}}{(2\pi)^2} \phi(\boldsymbol{L}) e^{i\boldsymbol{L}\cdot\boldsymbol{x}}. \tag{5}$$

Evaluating the spatial gradients:

$$\boldsymbol{\nabla}P(\boldsymbol{x}) = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2} (i\boldsymbol{\ell}') P(\boldsymbol{\ell}') e^{i\boldsymbol{\ell}'\cdot\boldsymbol{x}}, \qquad \boldsymbol{\nabla}\phi(\boldsymbol{x}) = \int \frac{\mathrm{d}^2\boldsymbol{L}}{(2\pi)^2} (i\boldsymbol{L}) \phi(\boldsymbol{L}) e^{i\boldsymbol{L}\cdot\boldsymbol{x}}. \tag{6}$$

Substituting these gradients into the inner product $\delta P(\boldsymbol{x}) = \boldsymbol{\nabla}\phi(\boldsymbol{x}) \cdot \boldsymbol{\nabla}P(\boldsymbol{x})$:

$$\begin{aligned} \delta P(\boldsymbol{x}) &= \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2} \int \frac{\mathrm{d}^2\boldsymbol{L}}{(2\pi)^2} (i\boldsymbol{L} \cdot i\boldsymbol{\ell}') \phi(\boldsymbol{L}) P(\boldsymbol{\ell}') e^{i(\boldsymbol{L}+\boldsymbol{\ell}')\cdot\boldsymbol{x}} \\ &= -\int \frac{\mathrm{d}^2\ell'}{(2\pi)^2} \int \frac{\mathrm{d}^2\boldsymbol{L}}{(2\pi)^2} (\boldsymbol{L} \cdot \boldsymbol{\ell}') \phi(\boldsymbol{L}) P(\boldsymbol{\ell}') e^{i(\boldsymbol{L}+\boldsymbol{\ell}')\cdot\boldsymbol{x}}. \end{aligned} \tag{7}$$

Taking the 2D Fourier transform $\delta P(\boldsymbol{\ell}) = \int \mathrm{d}^2x\, \delta P(\boldsymbol{x}) e^{-i\boldsymbol{\ell}\cdot\boldsymbol{x}}$ and exchanging the order of integration:

$$\delta P(\boldsymbol{\ell}) = -\int \frac{\mathrm{d}^2\ell'}{(2\pi)^2} \int \frac{\mathrm{d}^2\boldsymbol{L}}{(2\pi)^2} (\boldsymbol{L} \cdot \boldsymbol{\ell}') \phi(\boldsymbol{L}) P(\boldsymbol{\ell}') \left[\int \mathrm{d}^2x\, e^{i(\boldsymbol{L}+\boldsymbol{\ell}'-\boldsymbol{\ell})\cdot\boldsymbol{x}}\right]. \tag{8}$$

The spatial integral in brackets forms the standard integral representation of the 2D Dirac delta function:

$$\int \mathrm{d}^2x\, e^{i(\boldsymbol{L}+\boldsymbol{\ell}'-\boldsymbol{\ell})\cdot\boldsymbol{x}} = (2\pi)^2 \delta^{(2)}(\boldsymbol{L} + \boldsymbol{\ell}' - \boldsymbol{\ell}). \tag{9}$$

Substituting Eq. (9) collapses the $\mathrm{d}^2\boldsymbol{L}$ integration by setting $\boldsymbol{L} = \boldsymbol{\ell} - \boldsymbol{\ell}'$:

$$\delta P(\boldsymbol{\ell}) = -\int \frac{\mathrm{d}^2\ell'}{(2\pi)^2} \int \frac{\mathrm{d}^2\boldsymbol{L}}{(2\pi)^2} (\boldsymbol{L} \cdot \boldsymbol{\ell}') \phi(\boldsymbol{L}) P(\boldsymbol{\ell}') \left[(2\pi)^2 \delta^{(2)}(\boldsymbol{L} + \boldsymbol{\ell}' - \boldsymbol{\ell})\right]$$

$$= -\int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}[(\boldsymbol{\ell}-\boldsymbol{\ell}')\cdot\boldsymbol{\ell}']\phi(\boldsymbol{\ell}-\boldsymbol{\ell}')P(\boldsymbol{\ell}'). \tag{10}$$

In the flat-sky approximation, the spin-2 polarization field decomposes into scalar (gradient-type, even parity) E-modes and pseudoscalar (curl-type, odd parity) B-modes via the wavevector polar angle $\varphi_{\boldsymbol{\ell}} \equiv \arctan(\ell_y/\ell_x)$:

$$P(\boldsymbol{\ell}) = [E(\boldsymbol{\ell}) + iB(\boldsymbol{\ell})]e^{2i\varphi_{\boldsymbol{\ell}}} \iff E(\boldsymbol{\ell}) + iB(\boldsymbol{\ell}) = P(\boldsymbol{\ell})e^{-2i\varphi_{\boldsymbol{\ell}}}. \tag{11}$$

The lensed fields $\tilde{E}$ and $\tilde{B}$ satisfy the identical decomposition:

$$\tilde{E}(\boldsymbol{\ell}) + i\tilde{B}(\boldsymbol{\ell}) = \tilde{P}(\boldsymbol{\ell})e^{-2i\varphi_{\boldsymbol{\ell}}} = [P(\boldsymbol{\ell}) + \delta P(\boldsymbol{\ell})]e^{-2i\varphi_{\boldsymbol{\ell}}}, \tag{12}$$

which implies that the first-order perturbations satisfy:

$$\delta E(\boldsymbol{\ell}) + i\,\delta B(\boldsymbol{\ell}) = \delta P(\boldsymbol{\ell})e^{-2i\varphi_{\boldsymbol{\ell}}}. \tag{13}$$

Assuming strictly scalar primordial perturbations at recombination, the unlensed tensor perturbations vanish, meaning that $B(\boldsymbol{\ell}') = 0$ and $P(\boldsymbol{\ell}') = E(\boldsymbol{\ell}')e^{2i\varphi_{\boldsymbol{\ell}'}}$. Substituting this and Eq. (10) into the projection relation:

$$\begin{aligned}\delta E(\boldsymbol{\ell}) + i\,\delta B(\boldsymbol{\ell}) &= -e^{-2i\varphi_{\boldsymbol{\ell}}}\int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}[(\boldsymbol{\ell}-\boldsymbol{\ell}')\cdot\boldsymbol{\ell}']\phi(\boldsymbol{\ell}-\boldsymbol{\ell}')\left[E(\boldsymbol{\ell}')e^{2i\varphi_{\boldsymbol{\ell}'}}\right] \\ &= -\int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}[(\boldsymbol{\ell}-\boldsymbol{\ell}')\cdot\boldsymbol{\ell}']E(\boldsymbol{\ell}')\phi(\boldsymbol{\ell}-\boldsymbol{\ell}')e^{-2i(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})}.\end{aligned} \tag{14}$$

Using Euler's formula $e^{-2i(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})} = \cos[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})] - i\sin[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})]$:

$$\begin{aligned}\delta E(\boldsymbol{\ell}) + i\,\delta B(\boldsymbol{\ell}) = -\int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}&[(\boldsymbol{\ell}-\boldsymbol{\ell}')\cdot\boldsymbol{\ell}']E(\boldsymbol{\ell}')\phi(\boldsymbol{\ell}-\boldsymbol{\ell}') \\ &\times\Big(\cos[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})] - i\sin[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})]\Big).\end{aligned} \tag{15}$$

Extracting the real and imaginary components ($\delta E = \mathrm{Re}[\delta E + i\delta B]$ and $\delta B = \mathrm{Im}[\delta E + i\delta B]$), we obtain the first-order lensing perturbations to both modes:

$$\delta E(\boldsymbol{\ell}) = -\int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}[\boldsymbol{\ell}'\cdot(\boldsymbol{\ell}-\boldsymbol{\ell}')]\cos[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})]E(\boldsymbol{\ell}')\phi(\boldsymbol{\ell}-\boldsymbol{\ell}'), \tag{16}$$

$$\delta B(\boldsymbol{\ell}) = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}[\boldsymbol{\ell}'\cdot(\boldsymbol{\ell}-\boldsymbol{\ell}')]\sin[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})]E(\boldsymbol{\ell}')\phi(\boldsymbol{\ell}-\boldsymbol{\ell}'). \tag{17}$$

The full lensed polarization fields $\tilde{E}(\boldsymbol{\ell}) = E(\boldsymbol{\ell}) + \delta E(\boldsymbol{\ell})$ and $\tilde{B}(\boldsymbol{\ell}) = B(\boldsymbol{\ell}) + \delta B(\boldsymbol{\ell})$ are therefore given by:

$$\tilde{E}(\boldsymbol{\ell}) = E(\boldsymbol{\ell}) - \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}[\boldsymbol{\ell}'\cdot(\boldsymbol{\ell}-\boldsymbol{\ell}')]\cos[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})]E(\boldsymbol{\ell}')\phi(\boldsymbol{\ell}-\boldsymbol{\ell}'), \tag{18}$$

$$\tilde{B}(\boldsymbol{\ell}) = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}[\boldsymbol{\ell}'\cdot(\boldsymbol{\ell}-\boldsymbol{\ell}')]\sin[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})]E(\boldsymbol{\ell}')\phi(\boldsymbol{\ell}-\boldsymbol{\ell}'), \tag{19}$$

where we have used the fact that the unlensed primordial B-mode vanishes identically ($B(\boldsymbol{\ell}) = 0$, so $\tilde{B}(\boldsymbol{\ell}) = \delta B(\boldsymbol{\ell})$). Notice the fundamental geometric distinction between the two kernels: the parity-even kernel $\cos[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})]$ merely perturbs and broadens the primary E-mode acoustic peaks, whereas the parity-odd kernel $\sin[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})]$ shears

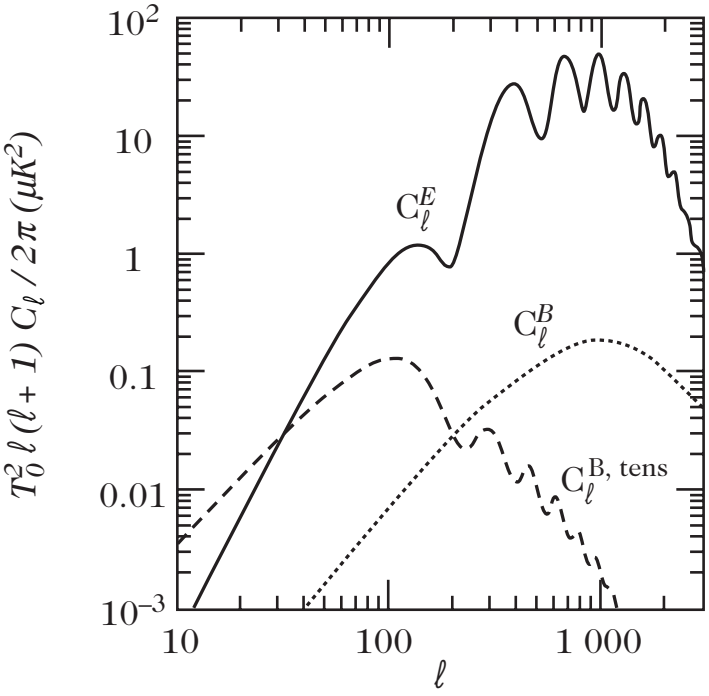


Figure 8.23: CMB polarization power spectra. The secondary B-mode spectrum induced by gravitational lensing of E-modes (dotted curve) peaks at $\ell \sim 1000$ with amplitude $\sim 5\,\mu\mathrm{K}^2$, acting as a cosmic foreground that obscures the primordial gravitational-wave B-mode signature (dashed curve, plotted for tensor-to-scalar ratio $r = 1$). Solid curve depicts the primordial E-mode spectrum. From Ref. [46].

and rotates E-mode polarization directions relative to $\boldsymbol{\ell}$, generating a nonzero secondary B-mode. Equation (19) reproduces Eq. (8.1.271).
We evaluate the two-point angular power spectrum:

$$\langle B(\boldsymbol{\ell})B^*(\boldsymbol{\ell}_1)\rangle = (2\pi)^2\delta^{(2)}(\boldsymbol{\ell}-\boldsymbol{\ell}_1)C_\ell^{BB,\mathrm{lens}}. \tag{20}$$

Under the assumption that the primary E-mode generated at $z \approx 1100$ is statistically uncorrelated with the intervening large-scale structure potential ϕ at $z \lesssim 5$:

$$\begin{aligned}\langle E(\boldsymbol{\ell}')\phi(\boldsymbol{\ell}-\boldsymbol{\ell}')E^*(\boldsymbol{\ell}_1'')\phi^*(\boldsymbol{\ell}_1-\boldsymbol{\ell}_1'')\rangle &= \langle E(\boldsymbol{\ell}')E^*(\boldsymbol{\ell}_1'')\rangle\langle\phi(\boldsymbol{\ell}-\boldsymbol{\ell}')\phi^*(\boldsymbol{\ell}_1-\boldsymbol{\ell}_1'')\rangle \\ &= (2\pi)^4\delta^{(2)}(\boldsymbol{\ell}'-\boldsymbol{\ell}_1'')\delta^{(2)}(\boldsymbol{\ell}-\boldsymbol{\ell}_1)C_{\ell'}^{EE}C_{|\boldsymbol{\ell}-\boldsymbol{\ell}'|}^{\phi\phi}.\end{aligned} \tag{21}$$

Integrating over $\boldsymbol{\ell}_1''$ yields the observable power spectrum:

$$C_\ell^{BB,\mathrm{lens}} = \int \frac{\mathrm{d}^2\ell'}{(2\pi)^2}[\boldsymbol{\ell}'\cdot(\boldsymbol{\ell}-\boldsymbol{\ell}')]^2\sin^2[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})]C_{\ell'}^{EE}C_{|\boldsymbol{\ell}-\boldsymbol{\ell}'|}^{\phi\phi}. \tag{22}$$

The geometrical factor $\sin^2[2(\varphi_{\boldsymbol{\ell}}-\varphi_{\boldsymbol{\ell}'})]$ vanishes whenever the lensing deflection is collinear with the polarization wavevector ($\varphi_{\boldsymbol{\ell}} = \varphi_{\boldsymbol{\ell}'}$ or $\varphi_{\boldsymbol{\ell}} = \varphi_{\boldsymbol{\ell}'} + \pi/2$); only non-parallel, sheared deflections break spatial parity and rotate curl-free E-modes into curl B-modes.

As displayed in Fig. 9.18, mode-mixing generates an effective cosmic noise floor of lensed B-modes with characteristic amplitude $C_\ell^{BB,\mathrm{lens}} \sim 5\,\mu\mathrm{K}\cdot\mathrm{arcmin}$ peaking at $\ell \sim 1000$, which acts

as the dominant astrophysical contaminant in the search for inflationary gravitational waves ($r < 0.01$).

CMB Lensing Reconstruction via Quadratic Estimators In an unperturbed, statistically isotropic universe, distinct Fourier multipoles of the primary CMB temperature and polarization fields $X_\alpha(\boldsymbol{\ell})$ ($\alpha \in \{T, E, B\}$) are statistically orthogonal Gaussian random fields:

$$\langle X_\alpha(\boldsymbol{\ell}_1) X_\beta(\boldsymbol{\ell}_2)\rangle = (2\pi)^2 \delta^{(2)}(\boldsymbol{\ell}_1 + \boldsymbol{\ell}_2) C_{\ell_1}^{\alpha\beta}, \tag{8.1.272}$$

where $C_\ell^{\alpha\beta}$ denotes the unlensed angular power spectrum [Eq. (8.1.250)]. Gravitational lensing deflects incoming CMB photons by the transverse gradient of the projected gravitational potential, $d_a(\hat{\boldsymbol{n}}) = \partial_a \phi(\hat{\boldsymbol{n}})$ [Eq. (8.1.243)]. To linear order in the deflection field, the remapped celestial field is:

$$\tilde{X}_\alpha(\boldsymbol{x}) = X_\alpha(\boldsymbol{x} + \boldsymbol{\nabla}\phi) \approx X_\alpha(\boldsymbol{x}) + \partial_a \phi(\boldsymbol{x}) \partial^a X_\alpha(\boldsymbol{x}). \tag{8.1.273}$$

In Fourier space, this real-space spatial deflection translates into a convolution between multipoles:

$$\tilde{X}_\alpha(\boldsymbol{\ell}_1) = X_\alpha(\boldsymbol{\ell}_1) - \int \frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2} [\boldsymbol{\ell}' \cdot (\boldsymbol{\ell}_1 - \boldsymbol{\ell}')] \phi(\boldsymbol{\ell}_1 - \boldsymbol{\ell}') X_\alpha(\boldsymbol{\ell}'). \tag{8.1.274}$$

For a fixed realization of the intervening potential ϕ, taking the ensemble expectation value over CMB realizations breaks translational invariance and generates an off-diagonal covariance between previously orthogonal multipoles whose sum $\boldsymbol{L} \equiv \boldsymbol{\ell}_1 + \boldsymbol{\ell}_2 \neq \boldsymbol{0}$:

$$\langle \tilde{X}_\alpha(\boldsymbol{\ell}_1) \tilde{X}_\beta(\boldsymbol{\ell}_2)\rangle_{\mathrm{CMB}} = f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) \phi(\boldsymbol{L}), \tag{8.1.275}$$

where $f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$ is the geometric mode-coupling response tensor. For temperature fluctuations ($\alpha = \beta = T$):

$$f_{TT}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) = (\boldsymbol{L} \cdot \boldsymbol{\ell}_1) C_{\ell_1}^{TT} + (\boldsymbol{L} \cdot \boldsymbol{\ell}_2) C_{\ell_2}^{TT}, \tag{8.1.276}$$

while for cross-polarization ($\alpha = E, \beta = B$), using the $E \to B$ conversion derived in Eq. (8.1.271):

$$f_{EB}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) = (\boldsymbol{\ell}_1 \cdot \boldsymbol{L}) \sin[2(\varphi_{\boldsymbol{\ell}_1} - \varphi_{\boldsymbol{\ell}_2})] C_{\ell_1}^{EE}. \tag{8.1.277}$$

Because the off-diagonal signal (8.1.275) is linear in $\phi(\boldsymbol{L})$ modulated by $f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$, the potential mode can be reconstructed by weighting a product of two observed multipoles with a filter tensor $g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$:

$$\bar{\phi}_{\alpha\beta}(\boldsymbol{L}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} \tilde{X}_\alpha(\boldsymbol{\ell}_1) \tilde{X}_\beta(\boldsymbol{L} - \boldsymbol{\ell}_1) g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L} - \boldsymbol{\ell}_1). \tag{8.1.278}$$

Averaging over CMB realizations at fixed $\phi(\boldsymbol{L})$ gives the linear response:

$$\langle \bar{\phi}_{\alpha\beta}(\boldsymbol{L})\rangle_{\mathrm{CMB}} = \left[\int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L} - \boldsymbol{\ell}_1) g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L} - \boldsymbol{\ell}_1) \right] \phi(\boldsymbol{L}) \equiv \mathcal{R}_L^{\alpha\beta} \phi(\boldsymbol{L}). \tag{8.1.279}$$

To eliminate the geometric factor $f_{\alpha\beta}$ and render the reconstructed field strictly unbiased ($\langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L})\rangle_{\mathrm{CMB}} = \phi(\boldsymbol{L})$), we normalize the estimator by $N_L^{\alpha\beta} \equiv (\mathcal{R}_L^{\alpha\beta})^{-1}$:

$$\hat{\phi}_{\alpha\beta}(\boldsymbol{L}) \equiv N_L^{\alpha\beta} \bar{\phi}_{\alpha\beta}(\boldsymbol{L}) = N_L^{\alpha\beta} \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} \tilde{X}_\alpha(\boldsymbol{\ell}_1) \tilde{X}_\beta(\boldsymbol{L} - \boldsymbol{\ell}_1) g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L} - \boldsymbol{\ell}_1), \tag{8.1.280}$$

where the normalization condition enforces unit response:

$$\frac{1}{N_L^{\alpha\beta}} = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L} - \boldsymbol{\ell}_1) g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L} - \boldsymbol{\ell}_1). \tag{8.1.281}$$

Four-Point Statistics and Wick Contractions To extract the angular power spectrum of the reconstructed deflection potential and identify the reconstruction noise floor, we evaluate the covariance:

$$\langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L})\hat{\phi}^*_{\alpha\beta}(\boldsymbol{L}')\rangle = (N_L^{\alpha\beta})(N_{L'}^{\alpha\beta})\int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\frac{\mathrm{d}^2\boldsymbol{\ell}_3}{(2\pi)^2} g^{\alpha\beta}(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2)g^{\alpha\beta *}(\boldsymbol{\ell}_3,\boldsymbol{\ell}_4) \times \langle \tilde{X}_\alpha(\boldsymbol{\ell}_1)\tilde{X}_\beta(\boldsymbol{\ell}_2)\tilde{X}^*_\alpha(\boldsymbol{\ell}_3)\tilde{X}^*_\beta(\boldsymbol{\ell}_4)\rangle, \tag{8.1.282}$$

where $\boldsymbol{\ell}_2 \equiv \boldsymbol{L} - \boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_4 \equiv \boldsymbol{L}' - \boldsymbol{\ell}_3$. Because $\hat{\phi}_{\alpha\beta}$ is quadratic in the observed CMB fields, its two-point statistics depend directly on the four-point correlation function (trispectrum) of the lensed fields.

Conditioned on a fixed realization of the intervening potential ϕ, the primary CMB fluctuations remain Gaussian. By Wick's theorem, the conditional four-point expectation value decomposes into three distinct pairings of two-point functions:

$$\langle \tilde{X}_1\tilde{X}_2\tilde{X}^*_3\tilde{X}^*_4 \mid \phi\rangle = \underbrace{\langle \tilde{X}_1\tilde{X}_2 \mid \phi\rangle\langle \tilde{X}^*_3\tilde{X}^*_4 \mid \phi\rangle}_{\mathrm{I}} + \underbrace{\langle \tilde{X}_1\tilde{X}^*_3 \mid \phi\rangle\langle \tilde{X}_2\tilde{X}^*_4 \mid \phi\rangle}_{\mathrm{II}} + \underbrace{\langle \tilde{X}_1\tilde{X}^*_4 \mid \phi\rangle\langle \tilde{X}_2\tilde{X}^*_3 \mid \phi\rangle}_{\mathrm{III}}, \tag{8.1.283}$$

where $i \equiv (\alpha_i, \boldsymbol{\ell}_i)$. Averaging over the statistical ensemble of all intervening lens realizations, $\langle\cdots\rangle = \langle\langle\cdots \mid \phi\rangle\rangle_\phi$, these Wick contractions organize into three distinct physical contributions:

$$\langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L})\hat{\phi}^*_{\alpha\beta}(\boldsymbol{L}')\rangle = (2\pi)^2\delta^{(2)}(\boldsymbol{L}-\boldsymbol{L}')\left[C_L^{\phi\phi} + N_L^{(0),\alpha\beta} + N_L^{(1),\alpha\beta} + \mathcal{O}((C^{\phi\phi})^2)\right]. \tag{8.1.284}$$

These three terms emerge directly from the Wick pairings (illustrated diagrammatically in Fig. 8.24):

1. **Disconnected Gaussian contractions ($N_L^{(0),\alpha\beta}$):** The pairings $\langle \tilde{X}_1\tilde{X}^*_3\rangle\langle \tilde{X}_2\tilde{X}^*_4\rangle$ and $\langle \tilde{X}_1\tilde{X}^*_4\rangle\langle \tilde{X}_2\tilde{X}^*_3\rangle$ connect legs across the two estimators without exchanging a lensing potential mode. Even in the absence of lensing ($C_L^{\phi\phi} \to 0$), the primary Gaussian CMB fluctuations and detector noise (with total spectrum $C_\ell^{\mathrm{tot}} = C_\ell + N_\ell$) produce non-zero contractions whenever $\boldsymbol{\ell}_3 = \boldsymbol{\ell}_1$, $\boldsymbol{\ell}_4 = \boldsymbol{\ell}_2$ (direct) or $\boldsymbol{\ell}_4 = \boldsymbol{\ell}_1$, $\boldsymbol{\ell}_3 = \boldsymbol{\ell}_2$ (cross), enforcing $\boldsymbol{L}' = \boldsymbol{L}$:

$$\langle \tilde{X}_\alpha(\boldsymbol{\ell}_1)\tilde{X}^*_\alpha(\boldsymbol{\ell}_3)\rangle\langle \tilde{X}_\beta(\boldsymbol{\ell}_2)\tilde{X}^*_\beta(\boldsymbol{\ell}_4)\rangle = (2\pi)^4\delta^{(2)}(\boldsymbol{\ell}_1-\boldsymbol{\ell}_3)\delta^{(2)}(\boldsymbol{\ell}_2-\boldsymbol{\ell}_4)C_{\ell_1}^{\alpha\alpha,\mathrm{tot}}C_{\ell_2}^{\beta\beta,\mathrm{tot}}, \tag{8.1.285}$$

$$\langle \tilde{X}_\alpha(\boldsymbol{\ell}_1)\tilde{X}^*_\beta(\boldsymbol{\ell}_4)\rangle\langle \tilde{X}^*_\beta(\boldsymbol{\ell}_2)\tilde{X}^*_\alpha(\boldsymbol{\ell}_3)\rangle = (2\pi)^4\delta^{(2)}(\boldsymbol{\ell}_1-\boldsymbol{\ell}_4)\delta^{(2)}(\boldsymbol{\ell}_2-\boldsymbol{\ell}_3)C_{\ell_1}^{\alpha\beta,\mathrm{tot}}C_{\ell_2}^{\beta\alpha,\mathrm{tot}}. \tag{8.1.286}$$

The first pairing $\langle \tilde{X}_1\tilde{X}_2\rangle\langle \tilde{X}^*_3\tilde{X}^*_4\rangle$ yields the squared mean-field $\langle\hat{\phi}\rangle^2$, which vanishes for $\boldsymbol{L} \neq \boldsymbol{0}$ on the full sky by statistical isotropy. Inserting the direct and cross contractions into Eq. (8.1.282) isolates the disconnected reconstruction noise floor:

$$N_L^{(0),\alpha\beta} = (N_L^{\alpha\beta})^2\int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} g^{\alpha\beta}(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2) \times \left[C_{\ell_1}^{\alpha\alpha,\mathrm{tot}}C_{\ell_2}^{\beta\beta,\mathrm{tot}}g^{\alpha\beta *}(\boldsymbol{\ell}_1,\boldsymbol{\ell}_2) + C_{\ell_1}^{\alpha\beta,\mathrm{tot}}C_{\ell_2}^{\beta\alpha,\mathrm{tot}}g^{\beta\alpha *}(\boldsymbol{\ell}_2,\boldsymbol{\ell}_1)\right]. \tag{8.1.287}$$

This $N_L^{(0)}$ noise floor represents the statistical variance of the quadratic estimator arising purely from Gaussian CMB sample variance and detector noise. Even in the complete absence of lensing ($C_L^{\phi\phi} \to 0$), where the expectation value of the quadratic estimator contains no lensing signal ($\langle\hat{\phi}\rangle = \boldsymbol{0}$ for $\boldsymbol{L} \neq \boldsymbol{0}$), the product of two Gaussian fields retains

a non-zero variance, creating an irreducible baseline noise floor. Crucially, because $N_L^{(0)}$ factorizes into products of unconditional two-point functions $\langle \tilde{X}_1 \tilde{X}_3^* \rangle \langle \tilde{X}_2 \tilde{X}_4^* \rangle$ (direct) and $\langle \tilde{X}_1 \tilde{X}_4^* \rangle \langle \tilde{X}_2 \tilde{X}_3^* \rangle$ (cross), and these two-point functions are ensemble covariances independent of any individual sky realization, each disconnected contraction can be evaluated using independent simulation realizations (Appendix D).

2. **Connected s-channel lens exchange ($C_L^{\phi\phi}$):** The term $\langle \tilde{X}_1 \tilde{X}_2 \mid \phi \rangle \langle \tilde{X}_3^* \tilde{X}_4^* \mid \phi \rangle$ in Eq. (8.1.283) correlates the lens-induced off-diagonal responses of the first and second estimators. Using $\langle \tilde{X}_1 \tilde{X}_2 \mid \phi \rangle = f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)\phi(\boldsymbol{L})$ and $\langle \tilde{X}_3^* \tilde{X}_4^* \mid \phi \rangle = f_{\alpha\beta}^*(\boldsymbol{\ell}_3, \boldsymbol{\ell}_4)\phi^*(\boldsymbol{L}')$:

$$\begin{aligned} &\Big\langle \langle \tilde{X}_1 \tilde{X}_2 \mid \phi \rangle \langle \tilde{X}_3^* \tilde{X}_4^* \mid \phi \rangle \Big\rangle_\phi - \langle \tilde{X}_1 \tilde{X}_2 \rangle \langle \tilde{X}_3^* \tilde{X}_4^* \rangle \\ &\qquad = (2\pi)^2 \delta^{(2)}(\boldsymbol{L} - \boldsymbol{L}') C_L^{\phi\phi} f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) f_{\alpha\beta}^*(\boldsymbol{\ell}_3, \boldsymbol{\ell}_4). \end{aligned} \tag{8.1.288}$$

Substituting this into Eq. (8.1.282), the double integral factors into two independent integrations over $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_3$, each equal to the inverse normalization $1/N_L^{\alpha\beta}$ by Eq. (8.1.281):

$$\begin{aligned} \langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L}) \hat{\phi}_{\alpha\beta}^*(\boldsymbol{L}') \rangle_{\text{signal}} &= (2\pi)^2 \delta^{(2)}(\boldsymbol{L} - \boldsymbol{L}') C_L^{\phi\phi} \left[N_L^{\alpha\beta} \int \frac{\mathrm{d}^2 \boldsymbol{\ell}_1}{(2\pi)^2} f_{\alpha\beta} g^{\alpha\beta} \right]^2 \\ &= (2\pi)^2 \delta^{(2)}(\boldsymbol{L} - \boldsymbol{L}') C_L^{\phi\phi}. \end{aligned} \tag{8.1.289}$$

The s-channel topology thus isolates the true lensing potential power spectrum $C_L^{\phi\phi}$ with unit response.

3. **Connected t- and u-channel cross-pairings ($N_L^{(1),\alpha\beta}$):** In the cross-pairings $(1,3)(2,4)$ and $(1,4)(2,3)$, expanding the conditional covariances to linear order in ϕ correlates modes through intermediate lens modes with wavevector $\boldsymbol{L}' = \boldsymbol{\ell}_1 - \boldsymbol{\ell}_3$ (t-channel) or $\boldsymbol{L}'' = \boldsymbol{\ell}_1 - \boldsymbol{\ell}_4$ (u-channel). Integrating over all intermediate lens scales generates an additive mode-coupling bias:

$$N_L^{(1),\alpha\beta} = \int \frac{\mathrm{d}^2 \boldsymbol{L}'}{(2\pi)^2} C_{L'}^{\phi\phi} \mathcal{M}^{\alpha\beta}(\boldsymbol{L}, \boldsymbol{L}'), \tag{8.1.290}$$

where the mode-coupling kernel evaluates to:

$$\begin{aligned} \mathcal{M}^{\alpha\beta}(\boldsymbol{L}, \boldsymbol{L}') = (N_L^{\alpha\beta})^2 \int \frac{\mathrm{d}^2 \boldsymbol{\ell}_1}{(2\pi)^2} \, & g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L} - \boldsymbol{\ell}_1) \\ \times \Big[& g^{\alpha\beta *}(\boldsymbol{\ell}_1 - \boldsymbol{L}', \boldsymbol{L} - \boldsymbol{\ell}_1 + \boldsymbol{L}') f_{\alpha\alpha}(\boldsymbol{\ell}_1, \boldsymbol{L}' - \boldsymbol{\ell}_1) \\ & \times f_{\beta\beta}(\boldsymbol{L} - \boldsymbol{\ell}_1, -\boldsymbol{L} + \boldsymbol{\ell}_1 - \boldsymbol{L}') \\ & + g^{\beta\alpha *}(\boldsymbol{L} - \boldsymbol{\ell}_1 + \boldsymbol{L}', \boldsymbol{\ell}_1 - \boldsymbol{L}') f_{\alpha\beta}(\boldsymbol{\ell}_1, -\boldsymbol{L} + \boldsymbol{\ell}_1 - \boldsymbol{L}') \\ & \times f_{\beta\alpha}(\boldsymbol{L} - \boldsymbol{\ell}_1, \boldsymbol{L}' - \boldsymbol{\ell}_1) \Big]. \end{aligned} \tag{8.1.291}$$

Because $N_L^{(1)}$ depends linearly on $C_{L'}^{\phi\phi}$, it represents an astrophysical contamination from non-Gaussian mode-mixing that must be calculated and subtracted alongside $N_L^{(0)}$ to unbiasedly measure $C_L^{\phi\phi}$.

Optimal minimum-variance filtering With the noise floor $N_L^{(0),\alpha\beta}$ [Eq. (8.1.287)] determined from the disconnected Wick contractions, we find the optimal filter tensor $g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$ by minimizing $N_L^{(0),\alpha\beta}$ subject to the normalization constraint (8.1.281). Introducing a Lagrange multiplier λ, functional variation with respect to $g^{\alpha\beta *}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$ yields:

$$g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) = \frac{C_{\ell_2}^{\beta\beta,\text{tot}} f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) - C_{\ell_1}^{\alpha\beta,\text{tot}} f_{\beta\alpha}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1)}{C_{\ell_1}^{\alpha\alpha,\text{tot}} C_{\ell_2}^{\beta\beta,\text{tot}} - (C_{\ell_1}^{\alpha\beta,\text{tot}})^2}. \tag{8.1.292}$$

To derive the optimal filter tensor $g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$ from first principles, we minimize the reconstruction noise variance subject to the unit-response normalization constraint. Consider the unnormalized quadratic estimator for the deflection mode $\phi(\boldsymbol{L})$:

$$\bar{\phi}_{\alpha\beta}(\boldsymbol{L}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} \tilde{X}_\alpha(\boldsymbol{\ell}_1)\tilde{X}_\beta(\boldsymbol{\ell}_2) g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2), \qquad \boldsymbol{\ell}_2 \equiv \boldsymbol{L} - \boldsymbol{\ell}_1. \tag{1}$$

Its expectation value under a fixed deflection potential mode $\phi(\boldsymbol{L})$ gives the linear response:

$$\langle \bar{\phi}_{\alpha\beta}(\boldsymbol{L}) \rangle = \left[\int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) \right] \phi(\boldsymbol{L}) \equiv \mathcal{R}_L^{\alpha\beta} \phi(\boldsymbol{L}). \tag{2}$$

The normalized, strictly unbiased estimator is $\hat{\phi}_{\alpha\beta}(\boldsymbol{L}) = N_L^{\alpha\beta} \bar{\phi}_{\alpha\beta}(\boldsymbol{L})$ with $N_L^{\alpha\beta} = 1/\mathcal{R}_L^{\alpha\beta}$. The reconstruction noise floor is the variance of this normalized estimator under the unlensed Gaussian CMB ensemble:

$$N_L^{(0),\alpha\beta} = \mathrm{Var}\left(\hat{\phi}_{\alpha\beta}(\boldsymbol{L})\right) = \frac{\mathrm{Var}\left(\bar{\phi}_{\alpha\beta}(\boldsymbol{L})\right)}{\left(\mathcal{R}_L^{\alpha\beta}\right)^2}. \tag{3}$$

Minimizing $N_L^{(0),\alpha\beta}$ with respect to the filter profile $g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$ is invariant under an overall rescaling of $g^{\alpha\beta}$. We may therefore fix the response constraint $\mathcal{R}_L^{\alpha\beta} = 1$ and minimize the unnormalized reconstruction variance under the fiducial unlensed Gaussian CMB ensemble ($\phi \to 0$):

$$\begin{aligned}
\mathrm{Var}\left(\bar{\phi}_{\alpha\beta}(\boldsymbol{L})\right) &\equiv \left\langle |\bar{\phi}_{\alpha\beta}(\boldsymbol{L})|^2 \right\rangle_{\mathbf{0}} - \left| \left\langle \bar{\phi}_{\alpha\beta}(\boldsymbol{L}) \right\rangle_{\mathbf{0}} \right|^2 \\
&= \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} \int \frac{\mathrm{d}^2\boldsymbol{\ell}_3}{(2\pi)^2} g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) g^{\alpha\beta *}(\boldsymbol{\ell}_3, \boldsymbol{\ell}_4) \\
&\quad \times \Big[\left\langle X_\alpha(\boldsymbol{\ell}_1) X_\beta(\boldsymbol{\ell}_2) X_\alpha^*(\boldsymbol{\ell}_3) X_\beta^*(\boldsymbol{\ell}_4) \right\rangle_{\mathbf{0}} - \left\langle X_\alpha(\boldsymbol{\ell}_1) X_\beta(\boldsymbol{\ell}_2) \right\rangle_{\mathbf{0}} \left\langle X_\alpha^*(\boldsymbol{\ell}_3) X_\beta^*(\boldsymbol{\ell}_4) \right\rangle_{\mathbf{0}} \Big],
\end{aligned} \tag{4}$$

where $\boldsymbol{\ell}_2 \equiv \boldsymbol{L} - \boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_4 \equiv \boldsymbol{L} - \boldsymbol{\ell}_3$. Because the fields are Gaussian in the unlensed limit ($\phi = 0$), Wick's theorem states that the full four-point function is the sum over all three possible pairings:

$$\left\langle X_1 X_2 X_3^* X_4^* \right\rangle_{\mathbf{0}} = \underbrace{\langle X_1 X_2 \rangle_{\mathbf{0}} \langle X_3^* X_4^* \rangle_{\mathbf{0}}}_{\text{disconnected mean field}} + \underbrace{\langle X_1 X_3^* \rangle_{\mathbf{0}} \langle X_2 X_4^* \rangle_{\mathbf{0}}}_{\text{direct contraction}} + \underbrace{\langle X_1 X_4^* \rangle_{\mathbf{0}} \langle X_2 X_3^* \rangle_{\mathbf{0}}}_{\text{cross contraction}}. \tag{5}$$

The first pairing is the product of two-point expectations, which cancels identically with the subtracted term $|\langle \bar{\phi}_{\alpha\beta}(\boldsymbol{L}) \rangle_{\mathbf{0}}|^2$. The variance is therefore determined entirely by the remaining two Wick pairings:

$$\begin{aligned}
\mathrm{Var}\left(\bar{\phi}_{\alpha\beta}(\boldsymbol{L})\right) &= \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} \int \frac{\mathrm{d}^2\boldsymbol{\ell}_3}{(2\pi)^2} g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) g^{\alpha\beta *}(\boldsymbol{\ell}_3, \boldsymbol{\ell}_4) \\
&\quad \times \Big[\langle X_\alpha(\boldsymbol{\ell}_1) X_\alpha^*(\boldsymbol{\ell}_3) \rangle_{\mathbf{0}} \langle X_\beta(\boldsymbol{\ell}_2) X_\beta^*(\boldsymbol{\ell}_4) \rangle_{\mathbf{0}} + \langle X_\alpha(\boldsymbol{\ell}_1) X_\beta^*(\boldsymbol{\ell}_4) \rangle_{\mathbf{0}} \langle X_\beta(\boldsymbol{\ell}_2) X_\alpha^*(\boldsymbol{\ell}_3) \rangle_{\mathbf{0}} \Big].
\end{aligned} \tag{6}$$

In the unperturbed statistically isotropic ensemble, the two-point functions are diagonal:

$$\langle X_\alpha(\boldsymbol{\ell}_1) X_\alpha^*(\boldsymbol{\ell}_3) \rangle_{\mathbf{0}} = (2\pi)^2 \delta^{(2)}(\boldsymbol{\ell}_1 - \boldsymbol{\ell}_3) C_{\ell_1}^{\alpha\alpha,\mathrm{tot}}, \tag{7}$$

$$\langle X_\beta(\boldsymbol{\ell}_2) X_\beta^*(\boldsymbol{\ell}_4) \rangle_{\mathbf{0}} = (2\pi)^2 \delta^{(2)}(\boldsymbol{\ell}_2 - \boldsymbol{\ell}_4) C_{\ell_2}^{\beta\beta,\mathrm{tot}} = (2\pi)^2 \delta^{(2)}(\boldsymbol{\ell}_3 - \boldsymbol{\ell}_1) C_{\ell_2}^{\beta\beta,\mathrm{tot}}, \tag{8}$$

$$\langle X_\alpha(\boldsymbol{\ell}_1) X_\beta^*(\boldsymbol{\ell}_4) \rangle_{\mathbf{0}} = (2\pi)^2 \delta^{(2)}(\boldsymbol{\ell}_1 - \boldsymbol{\ell}_4) C_{\ell_1}^{\alpha\beta,\mathrm{tot}} = (2\pi)^2 \delta^{(2)}(\boldsymbol{\ell}_1 + \boldsymbol{\ell}_3 - \boldsymbol{L}) C_{\ell_1}^{\alpha\beta,\mathrm{tot}}, \tag{9}$$

$$\langle X_\beta(\boldsymbol{\ell}_2) X_\alpha^*(\boldsymbol{\ell}_3) \rangle_{\mathbf{0}} = (2\pi)^2 \delta^{(2)}(\boldsymbol{\ell}_2 - \boldsymbol{\ell}_3) C_{\ell_2}^{\beta\alpha,\mathrm{tot}} = (2\pi)^2 \delta^{(2)}(\boldsymbol{\ell}_3 - (\boldsymbol{L} - \boldsymbol{\ell}_1)) C_{\ell_2}^{\beta\alpha,\mathrm{tot}}. \tag{10}$$

Substituting these into the two terms:

- In the direct term, $(2\pi)^2\delta^{(2)}(\boldsymbol{\ell}_1 - \boldsymbol{\ell}_3)$ sets $\boldsymbol{\ell}_3 = \boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_4 = \boldsymbol{L} - \boldsymbol{\ell}_1 = \boldsymbol{\ell}_2$, collapsing the $\mathrm{d}^2\boldsymbol{\ell}_3$ integral to $g^{\alpha\beta *}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) C_{\ell_1}^{\alpha\alpha,\mathrm{tot}} C_{\ell_2}^{\beta\beta,\mathrm{tot}}$.
- In the cross term, $(2\pi)^2\delta^{(2)}(\boldsymbol{\ell}_3 - \boldsymbol{\ell}_2)$ sets $\boldsymbol{\ell}_3 = \boldsymbol{\ell}_2$ and $\boldsymbol{\ell}_4 = \boldsymbol{L} - \boldsymbol{\ell}_2 = \boldsymbol{\ell}_1$, collapsing the $\mathrm{d}^2\boldsymbol{\ell}_3$ integral to $g^{\alpha\beta *}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1) C_{\ell_1}^{\alpha\beta,\mathrm{tot}} C_{\ell_2}^{\beta\alpha,\mathrm{tot}} = g^{\beta\alpha *}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1) C_{\ell_1}^{\alpha\beta,\mathrm{tot}} C_{\ell_2}^{\beta\alpha,\mathrm{tot}}$.

Adding the two contributions reproduces the integrated variance:

$$\mathrm{Var}\left(\bar{\phi}_{\alpha\beta}(\boldsymbol{L})\right) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) \left[C_{\ell_1}^{\alpha\alpha,\mathrm{tot}} C_{\ell_2}^{\beta\beta,\mathrm{tot}} g^{\alpha\beta *}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) + C_{\ell_1}^{\alpha\beta,\mathrm{tot}} C_{\ell_2}^{\beta\alpha,\mathrm{tot}} g^{\beta\alpha *}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1) \right]. \tag{11}$$

In the 2D integration over $\boldsymbol{\ell}_1$, every unordered pair of wavevectors $\{\boldsymbol{\ell}_1, \boldsymbol{\ell}_2\}$ satisfying $\boldsymbol{\ell}_1 + \boldsymbol{\ell}_2 = \boldsymbol{L}$ is traversed twice: once as $(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$ and once as $(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1)$. Since the observed fields commute, $\tilde{X}_\alpha(\boldsymbol{\ell}_1)\tilde{X}_\beta(\boldsymbol{\ell}_2) = \tilde{X}_\beta(\boldsymbol{\ell}_2)\tilde{X}_\alpha(\boldsymbol{\ell}_1)$, exchanging the dummy variable $\boldsymbol{\ell}_1 \leftrightarrow \boldsymbol{\ell}_2$ relates the filters by $g^{\beta\alpha}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1) = g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$. Symmetrizing the variance integral over the pair $(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$ and introducing a Lagrange multiplier 2λ for the response constraint yields the functional:

$$\mathcal{J}\left[g^{\alpha\beta}\right] = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} \Big\{ |g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)|^2 C_{\ell_1}^{\alpha\alpha,\mathrm{tot}} C_{\ell_2}^{\beta\beta,\mathrm{tot}} + g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) g^{\beta\alpha *}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1) C_{\ell_1}^{\alpha\beta,\mathrm{tot}} C_{\ell_2}^{\beta\alpha,\mathrm{tot}} - 2\lambda f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) \Big\}. \tag{12}$$

Performing the functional variation with respect to $g^{\alpha\beta *}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$:

$$\frac{\delta\mathcal{J}}{\delta g^{\alpha\beta *}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)} = C_{\ell_1}^{\alpha\alpha,\mathrm{tot}} C_{\ell_2}^{\beta\beta,\mathrm{tot}} g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) + C_{\ell_1}^{\alpha\beta,\mathrm{tot}} C_{\ell_2}^{\alpha\beta,\mathrm{tot}} g^{\beta\alpha}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1) - \lambda f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) = 0. \tag{13}$$

Exchanging the indices and wavevectors $(\alpha, \boldsymbol{\ell}_1) \leftrightarrow (\beta, \boldsymbol{\ell}_2)$ gives the conjugate stationarity condition:

$$C_{\ell_2}^{\beta\beta,\mathrm{tot}} C_{\ell_1}^{\alpha\alpha,\mathrm{tot}} g^{\beta\alpha}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1) + C_{\ell_2}^{\beta\alpha,\mathrm{tot}} C_{\ell_1}^{\beta\alpha,\mathrm{tot}} g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) - \lambda f_{\beta\alpha}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1) = 0. \tag{14}$$

This forms a coupled 2×2 algebraic system for the weights at wavevector pair $(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$:

$$\begin{pmatrix} C_{\ell_1}^{\alpha\alpha,\mathrm{tot}} C_{\ell_2}^{\beta\beta,\mathrm{tot}} & C_{\ell_1}^{\alpha\beta,\mathrm{tot}} C_{\ell_2}^{\alpha\beta,\mathrm{tot}} \\ C_{\ell_1}^{\alpha\beta,\mathrm{tot}} C_{\ell_2}^{\alpha\beta,\mathrm{tot}} & C_{\ell_2}^{\alpha\alpha,\mathrm{tot}} C_{\ell_1}^{\beta\beta,\mathrm{tot}} \end{pmatrix} \begin{pmatrix} g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) \\ g^{\beta\alpha}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1) \end{pmatrix} = \lambda \begin{pmatrix} f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) \\ f_{\beta\alpha}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1) \end{pmatrix}. \tag{15}$$

Inverting this 2×2 matrix via Cramer's rule, the determinant evaluates to:

$$\mathcal{D} = C_{\ell_1}^{\alpha\alpha,\mathrm{tot}} C_{\ell_2}^{\beta\beta,\mathrm{tot}} C_{\ell_2}^{\alpha\alpha,\mathrm{tot}} C_{\ell_1}^{\beta\beta,\mathrm{tot}} - \left(C_{\ell_1}^{\alpha\beta,\mathrm{tot}} C_{\ell_2}^{\alpha\beta,\mathrm{tot}}\right)^2. \tag{16}$$

Solving for $g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$ and absorbing the arbitrary constant λ into the overall normalization $N_L^{\alpha\beta}$ yields:

$$g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) = \frac{C_{\ell_2}^{\beta\beta,\mathrm{tot}} f_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) - C_{\ell_1}^{\alpha\beta,\mathrm{tot}} f_{\beta\alpha}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1)}{C_{\ell_1}^{\alpha\alpha,\mathrm{tot}} C_{\ell_2}^{\beta\beta,\mathrm{tot}} - (C_{\ell_1}^{\alpha\beta,\mathrm{tot}})^2}, \tag{17}$$

which is identically the optimal minimum-variance filter Eq. (8.1.292).

For auto-temperature reconstruction ($\alpha = \beta = T$):

$$g^{TT}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) = \frac{f_{TT}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)}{2C_{\ell_1}^{TT,\mathrm{tot}} C_{\ell_2}^{TT,\mathrm{tot}}}, \qquad N_L^{(0),TT} = \left[\int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} \frac{|f_{TT}(\boldsymbol{\ell}_1, \boldsymbol{L} - \boldsymbol{\ell}_1)|^2}{2C_{\ell_1}^{TT,\mathrm{tot}} C_{|\boldsymbol{L}-\boldsymbol{\ell}_1|}^{TT,\mathrm{tot}}} \right]^{-1}. \tag{8.1.293}$$

For cross-polarization ($\alpha = E, \beta = B$), unlensed parity invariance ensures $C_\ell^{EB,\mathrm{tot}} = 0$, simplifying the filter and noise floor to:

$$g^{EB}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) = \frac{f_{EB}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)}{C_{\ell_1}^{EE,\mathrm{tot}} C_{\ell_2}^{BB,\mathrm{tot}}}, \qquad N_L^{(0),EB} = \left[\int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} \frac{|f_{EB}(\boldsymbol{\ell}_1, \boldsymbol{L} - \boldsymbol{\ell}_1)|^2}{C_{\ell_1}^{EE,\mathrm{tot}} C_{|\boldsymbol{L}-\boldsymbol{\ell}_1|}^{BB,\mathrm{tot}}} \right]^{-1}. \tag{8.1.294}$$

Under this optimal filtering, the normalization constant and the Gaussian noise floor coincide: $N_L^{\alpha\beta} \equiv N_L^{(0),\alpha\beta}$.

This polarization estimator exhibits an essential cosmological advantage: because primordial B-modes are vanishingly small ($C_\ell^{BB,\mathrm{prim}} \approx 0$), the denominator contains solely instrumental and atmospheric noise, $C_\ell^{BB,\mathrm{tot}} \approx N_\ell^{BB}$. While temperature reconstruction ($N_L^{(0),TT}$) is permanently bottlenecked by primary CMB sample variance, $N_L^{(0),EB}$ scales as N_ℓ^{BB}, dropping linearly with detector noise variance. Consequently, $\hat{\phi}_{EB}$ serves as the primary engine for high-resolution lensing reconstruction in Stage-3 and Stage-4 experiments.

This minimum-variance construction has a direct foundation in likelihood analysis. The likelihood function tells us how likely the probability distribution is to have generated the data, given a candidate realization of the deflection potential $\boldsymbol{\phi}$. Conditioned on a fixed realization of $\boldsymbol{\phi}$, the observed CMB map remains zero-mean Gaussian with covariance $\mathbf{C}(\boldsymbol{\phi}) = \mathbf{D}(\boldsymbol{\phi})\mathbf{S}\mathbf{D}^\top(\boldsymbol{\phi}) + \mathbf{N}$. Differentiating the log-likelihood $\ln P(\boldsymbol{d} \mid \boldsymbol{\phi})$ with respect to a lens mode ϕ_A gives the score vector:

$$g_A(\boldsymbol{d}, \boldsymbol{\phi}) = \frac{1}{2}\boldsymbol{d}^\top \mathbf{C}^{-1}\mathbf{C}_{,A}\mathbf{C}^{-1}\boldsymbol{d} - \frac{1}{2}\operatorname{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,A}). \tag{8.1.295}$$

To derive the likelihood score vector from first principles, consider the observed CMB data vector $\boldsymbol{d}$ containing m pixels or harmonic multipoles. Conditioned on a fixed realization of the deflection potential $\boldsymbol{\phi}$, the cosmological signal and instrumental noise remain zero-mean Gaussian fields with joint covariance $\mathbf{C}(\boldsymbol{\phi}) = \mathbf{D}(\boldsymbol{\phi})\mathbf{S}\mathbf{D}^\top(\boldsymbol{\phi}) + \mathbf{N}$. The conditional likelihood distribution is:

$$P(\boldsymbol{d} \mid \boldsymbol{\phi}) = \frac{1}{(2\pi)^{m/2}\sqrt{\det \mathbf{C}(\boldsymbol{\phi})}} \exp\left[-\frac{1}{2}\boldsymbol{d}^\top \mathbf{C}^{-1}(\boldsymbol{\phi})\boldsymbol{d}\right]. \tag{1}$$

Taking the natural logarithm isolates the quadratic data form and the parameter-dependent log-determinant normalization:

$$\ln P(\boldsymbol{d} \mid \boldsymbol{\phi}) = -\frac{1}{2}\boldsymbol{d}^\top \mathbf{C}^{-1}(\boldsymbol{\phi})\boldsymbol{d} - \frac{1}{2}\ln\det \mathbf{C}(\boldsymbol{\phi}) - \frac{m}{2}\ln(2\pi). \tag{2}$$

Differentiating with respect to an arbitrary deflection mode ϕ_A, we evaluate the two parameter-dependent terms individually.

For the quadratic term, because the observed data vector $\boldsymbol{d}$ does not depend on the lensing parameter $\boldsymbol{\phi}$, the derivative acts solely on the inverse covariance:

$$\frac{\partial}{\partial \phi_A}\left[-\frac{1}{2}\boldsymbol{d}^\top \mathbf{C}^{-1}(\boldsymbol{\phi})\boldsymbol{d}\right] = -\frac{1}{2}\boldsymbol{d}^\top \left(\frac{\partial \mathbf{C}^{-1}}{\partial \phi_A}\right)\boldsymbol{d}. \tag{3}$$

To determine $\partial_A \mathbf{C}^{-1}$ without guessing or assuming matrix commutativity, differentiate

the definition of the matrix inverse, $\mathbf{C}(\boldsymbol{\phi})\mathbf{C}^{-1}(\boldsymbol{\phi}) = \mathbf{I}$, by applying the product rule:

$$\frac{\partial}{\partial \phi_A}(\mathbf{C}\mathbf{C}^{-1}) = \left(\frac{\partial \mathbf{C}}{\partial \phi_A}\right)\mathbf{C}^{-1} + \mathbf{C}\left(\frac{\partial \mathbf{C}^{-1}}{\partial \phi_A}\right) = \mathbf{0}. \tag{4}$$

Multiplying on the left by $\mathbf{C}^{-1}$:

$$\mathbf{C}^{-1}\mathbf{C}_{,A}\mathbf{C}^{-1} + \mathbf{C}^{-1}\mathbf{C}\left(\frac{\partial \mathbf{C}^{-1}}{\partial \phi_A}\right) = \mathbf{0} \implies \frac{\partial \mathbf{C}^{-1}}{\partial \phi_A} = -\mathbf{C}^{-1}\mathbf{C}_{,A}\mathbf{C}^{-1}, \tag{5}$$

where $\mathbf{C}_{,A} \equiv \partial\mathbf{C}/\partial\phi_A$. Substituting this identity into the quadratic form derivative reverses the minus sign:

$$-\frac{1}{2}\boldsymbol{d}^\top\left(\frac{\partial \mathbf{C}^{-1}}{\partial \phi_A}\right)\boldsymbol{d} = \frac{1}{2}\boldsymbol{d}^\top\mathbf{C}^{-1}\mathbf{C}_{,A}\mathbf{C}^{-1}\boldsymbol{d}. \tag{6}$$

For the log-determinant term, we apply Jacobi's formula for the differential of the matrix determinant. Using $\ln\det\mathbf{C} = \mathrm{Tr}(\ln\mathbf{C})$, differential variation yields:

$$\frac{\partial \ln\det\mathbf{C}}{\partial \phi_A} = \mathrm{Tr}\left(\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \phi_A}\right) = \mathrm{Tr}\left(\mathbf{C}^{-1}\mathbf{C}_{,A}\right). \tag{7}$$

Multiplying by $-1/2$ gives $-\frac{1}{2}\mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,A})$. Assembling the two derivatives recovers the exact likelihood score vector:

$$g_A(\boldsymbol{d},\boldsymbol{\phi}) \equiv \frac{\partial \ln P(\boldsymbol{d}\mid\boldsymbol{\phi})}{\partial \phi_A} = \frac{1}{2}\boldsymbol{d}^\top\mathbf{C}^{-1}\mathbf{C}_{,A}\mathbf{C}^{-1}\boldsymbol{d} - \frac{1}{2}\mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,A}). \tag{8}$$

Finally, we verify that this score vector is strictly unbiased at the true parameter value. Taking the ensemble average over data realizations conditioned on $\boldsymbol{\phi}$, where $\langle\boldsymbol{d}\boldsymbol{d}^\top\rangle_{\boldsymbol{\phi}} = \mathbf{C}(\boldsymbol{\phi})$, and applying the cyclic invariance of the trace:

$$\begin{aligned}\left\langle\boldsymbol{d}^\top\mathbf{C}^{-1}\mathbf{C}_{,A}\mathbf{C}^{-1}\boldsymbol{d}\right\rangle_{\boldsymbol{\phi}} &= \mathrm{Tr}\left(\mathbf{C}^{-1}\mathbf{C}_{,A}\mathbf{C}^{-1}\langle\boldsymbol{d}\boldsymbol{d}^\top\rangle_{\boldsymbol{\phi}}\right)\\ &= \mathrm{Tr}\left(\mathbf{C}^{-1}\mathbf{C}_{,A}\mathbf{C}^{-1}\mathbf{C}\right) = \mathrm{Tr}\left(\mathbf{C}^{-1}\mathbf{C}_{,A}\right).\end{aligned} \tag{9}$$

Subtracting the second term yields $\langle g_A(\boldsymbol{d},\boldsymbol{\phi})\rangle_{\boldsymbol{\phi}} = 0$. The trace subtraction therefore acts precisely to subtract the data-independent mean field, centering the likelihood score at zero for any valid deflection realization.

Expanding the exact stationarity equation $g_A(\boldsymbol{d},\hat{\boldsymbol{\phi}}_{\mathrm{ML}}) = 0$ to linear order about the unlensed sky ($\boldsymbol{\phi} = \mathbf{0}$) yields $g_A(\boldsymbol{d},\mathbf{0}) - F_{AB}\phi_B \simeq 0$, where $F_{AB} \equiv \langle-\partial_A\partial_B \ln P\rangle_{\mathbf{0}} = \frac{1}{2}\mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,A}\mathbf{C}^{-1}\mathbf{C}_{,B})$ is the Fisher information matrix. Inverting this curvature yields the Fisher-scoring estimator:

$$\hat{\boldsymbol{\phi}}_{\mathrm{QE}} = \mathbf{F}^{-1}\boldsymbol{g}(\boldsymbol{d},\mathbf{0}). \tag{8.1.296}$$

For statistically isotropic fields, the modes decouple ($F_L = 1/N_L^{(0)}$), exactly reproducing the Hu–Okamoto estimator $\hat{\phi}(\boldsymbol{L}) = N_L^{(0)}g_{\boldsymbol{L}}(\boldsymbol{d},\mathbf{0})$. The quadratic estimator is therefore not an empirical heuristic; it is the exact first Fisher-scoring step of the Gaussian likelihood.

The algebraic expansion connecting the exact likelihood score to the Hu–Okamoto quadratic estimator proceeds in two stages: matrix expansion about the unlensed fiducial model, followed by continuous Fourier-space integration.

First, consider the covariance model $\mathbf{C}(\boldsymbol{\phi}) = \mathbf{C}^{(0)} + \sum_A \phi_A \mathbf{C}_{,A} + \mathcal{O}(\phi^2)$. To invert this operator, let $(\mathbf{C}^{(0)} + \sum_A \phi_A \mathbf{C}_{,A})^{-1} = \mathbf{C}^{(0)-1} + \sum_A \phi_A \mathbf{X}_A + \mathcal{O}(\phi^2)$ for an unknown first-order matrix $\mathbf{X}_A$. Enforcing the inverse identity:

$$\begin{aligned}\mathbf{I} &= \Big(\mathbf{C}^{(0)} + \sum_A \phi_A \mathbf{C}_{,A}\Big)\Big(\mathbf{C}^{(0)-1} + \sum_B \phi_B \mathbf{X}_B + \mathcal{O}(\phi^2)\Big) \\ &= \mathbf{I} + \sum_A \phi_A (\mathbf{C}^{(0)}\mathbf{X}_A + \mathbf{C}_{,A}\mathbf{C}^{(0)-1}) + \mathcal{O}(\phi^2). \end{aligned} \tag{1}$$

The linear coefficient must vanish. Multiplying on the left by $\mathbf{C}^{(0)-1}$ determines $\mathbf{X}_A = -\mathbf{C}^{(0)-1}\mathbf{C}_{,A}\mathbf{C}^{(0)-1}$, giving:

$$\mathbf{C}^{-1}(\boldsymbol{\phi}) = \mathbf{C}^{(0)-1} - \sum_B \phi_B \mathbf{C}^{(0)-1}\mathbf{C}_{,B}\mathbf{C}^{(0)-1} + \mathcal{O}(\phi^2). \tag{2}$$

Substituting this expansion into the exact score vector $2g_A(\boldsymbol{d}, \boldsymbol{\phi}) = \boldsymbol{d}^\top \mathbf{C}^{-1}\mathbf{C}_{,A}\mathbf{C}^{-1}\boldsymbol{d} - \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,A})$:

$$\begin{aligned}2g_A(\boldsymbol{d}, \boldsymbol{\phi}) = {}& \boldsymbol{d}^\top \mathbf{C}^{(0)-1}\mathbf{C}_{,A}\mathbf{C}^{(0)-1}\boldsymbol{d} - \mathrm{Tr}(\mathbf{C}^{(0)-1}\mathbf{C}_{,A}) \\ & - \sum_B \phi_B \Big[2\boldsymbol{d}^\top \mathbf{C}^{(0)-1}\mathbf{C}_{,B}\mathbf{C}^{(0)-1}\mathbf{C}_{,A}\mathbf{C}^{(0)-1}\boldsymbol{d} - \mathrm{Tr}(\mathbf{C}^{(0)-1}\mathbf{C}_{,B}\mathbf{C}^{(0)-1}\mathbf{C}_{,A})\Big] + \mathcal{O}(\phi^2). \end{aligned} \tag{3}$$

Setting $g_A(\boldsymbol{d}, \hat{\boldsymbol{\phi}}) = 0$ directly would yield a data-dependent denominator. To obtain a minimum-variance quadratic estimator, the fluctuating observed curvature in brackets is replaced by its ensemble expectation in the fiducial Gaussian model ($\boldsymbol{\phi} = \mathbf{0}$), where $\langle \boldsymbol{d}\boldsymbol{d}^\top \rangle_{\mathbf{0}} = \mathbf{C}^{(0)}$:

$$\begin{aligned}\Big\langle \boldsymbol{d}^\top \mathbf{C}^{(0)-1}\mathbf{C}_{,B}\mathbf{C}^{(0)-1}\mathbf{C}_{,A}\mathbf{C}^{(0)-1}\boldsymbol{d}\Big\rangle_{\mathbf{0}} &= \mathrm{Tr}\left(\mathbf{C}^{(0)-1}\mathbf{C}_{,B}\mathbf{C}^{(0)-1}\mathbf{C}_{,A}\mathbf{C}^{(0)-1}\mathbf{C}^{(0)}\right) \\ &= \mathrm{Tr}\left(\mathbf{C}^{(0)-1}\mathbf{C}_{,B}\mathbf{C}^{(0)-1}\mathbf{C}_{,A}\right). \end{aligned} \tag{4}$$

The expected bracket reduces to $\mathrm{Tr}(\mathbf{C}^{(0)-1}\mathbf{C}_{,A}\mathbf{C}^{(0)-1}\mathbf{C}_{,B}) = 2F_{AB}$, where $F_{AB} \equiv \langle -\partial_A \partial_B \ln P \rangle_{\mathbf{0}}$ is the Fisher information matrix. The linearized stationarity equation becomes $g_A(\boldsymbol{d}, \mathbf{0}) - \sum_B F_{AB}\phi_B \simeq 0$, whose solution defines the Fisher-scoring estimator $\hat{\boldsymbol{\phi}}_{\mathrm{QE}} = \mathbf{F}^{-1}\boldsymbol{g}(\boldsymbol{d}, \mathbf{0})$.

We now project this abstract matrix estimator into continuous flat-sky Fourier space for statistically isotropic temperature fluctuations with noise $d_{\boldsymbol{\ell}} = \widetilde{T}_{\boldsymbol{\ell}} + n_{\boldsymbol{\ell}}$. The fiducial covariance is diagonal:

$$C^{(0)}_{\boldsymbol{\ell},\boldsymbol{\ell}'} = (2\pi)^2 \delta^{(2)}(\boldsymbol{\ell} + \boldsymbol{\ell}') C^{\mathrm{tot}}_\ell, \qquad C^{\mathrm{tot}}_\ell = C^{TT}_\ell + N^{TT}_\ell. \tag{5}$$

The continuous matrix inverse $(C^{(0)-1})^{\boldsymbol{\ell},\boldsymbol{\ell}'}$ must satisfy $\int \frac{\mathrm{d}^2\boldsymbol{k}}{(2\pi)^2} C^{(0)}_{\boldsymbol{\ell},\boldsymbol{k}} (C^{(0)-1})^{\boldsymbol{k},\boldsymbol{\ell}'} = (2\pi)^2\delta^{(2)}(\boldsymbol{\ell} - \boldsymbol{\ell}')$. Direct multiplication verifies:

$$(C^{(0)-1})^{\boldsymbol{\ell},\boldsymbol{\ell}'} = (2\pi)^2 \delta^{(2)}(\boldsymbol{\ell} + \boldsymbol{\ell}') \frac{1}{C^{\mathrm{tot}}_\ell}. \tag{6}$$

The inverse-covariance filtered temperature map is therefore $(C^{(0)-1}d)_{\boldsymbol{\ell}} = \int \frac{\mathrm{d}^2\boldsymbol{k}}{(2\pi)^2} (C^{(0)-1})^{\boldsymbol{\ell},\boldsymbol{k}} d_{\boldsymbol{k}} = d_{-\boldsymbol{\ell}}/C^{\mathrm{tot}}_\ell$.

Next, from the real-space Taylor expansion $\widetilde{T}(\boldsymbol{x}) = T(\boldsymbol{x}) + \boldsymbol{\nabla}\phi \cdot \boldsymbol{\nabla} T$, the Fourier transform gives $\widetilde{T}_{\boldsymbol{\ell}} = T_{\boldsymbol{\ell}} - \int \frac{\mathrm{d}^2\boldsymbol{L}}{(2\pi)^2}[\boldsymbol{L}\cdot(\boldsymbol{\ell}-\boldsymbol{L})]\phi_{\boldsymbol{L}} T_{\boldsymbol{\ell}-\boldsymbol{L}}$. Taking the conditional covariance yields the response matrix:

$$\left.\frac{\delta C_{\boldsymbol{\ell},\boldsymbol{\ell}'}}{\delta\phi_{-\boldsymbol{L}}}\right|_{\boldsymbol{\phi}=\boldsymbol{0}} = (2\pi)^2\delta^{(2)}(\boldsymbol{\ell}+\boldsymbol{\ell}'+\boldsymbol{L}) f^{TT}(\boldsymbol{\ell},\boldsymbol{\ell}'), \tag{7}$$

where $f^{TT}(\boldsymbol{\ell},\boldsymbol{\ell}') = (\boldsymbol{\ell}+\boldsymbol{\ell}')\cdot(\boldsymbol{\ell} C_\ell^{TT} + \boldsymbol{\ell}' C_{\ell'}^{TT})$. Substituting the filtered fields and response into the score numerator for Fourier mode $-\boldsymbol{L}$:

$$\begin{aligned} g_{-\boldsymbol{L}}(\boldsymbol{d},\boldsymbol{0}) &= \frac{1}{2}\int \frac{\mathrm{d}^2\boldsymbol{\ell}}{(2\pi)^2}\frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2}(C^{(0)-1}d)_{\boldsymbol{\ell}}(C^{(0)-1}d)_{\boldsymbol{\ell}'}\left.\frac{\delta C_{\boldsymbol{\ell},\boldsymbol{\ell}'}}{\delta\phi_{-\boldsymbol{L}}}\right|_{\boldsymbol{0}} \\ &= \frac{1}{2}\int \frac{\mathrm{d}^2\boldsymbol{\ell}}{(2\pi)^2}\frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2}\frac{d_{-\boldsymbol{\ell}}}{C_\ell^{\mathrm{tot}}}\frac{d_{-\boldsymbol{\ell}'}}{C_{\ell'}^{\mathrm{tot}}}(2\pi)^2\delta^{(2)}(\boldsymbol{\ell}+\boldsymbol{\ell}'+\boldsymbol{L}) f^{TT}(\boldsymbol{\ell},\boldsymbol{\ell}'). \end{aligned} \tag{8}$$

Integrating over $\boldsymbol{\ell}'$ collapses the Dirac delta function, enforcing $\boldsymbol{\ell}' = -\boldsymbol{L}-\boldsymbol{\ell}$:

$$g_{-\boldsymbol{L}}(\boldsymbol{d},\boldsymbol{0}) = \frac{1}{2}\int \frac{\mathrm{d}^2\boldsymbol{\ell}}{(2\pi)^2}\frac{d_{-\boldsymbol{\ell}} d_{\boldsymbol{L}+\boldsymbol{\ell}}}{C_\ell^{\mathrm{tot}} C_{|\boldsymbol{L}+\boldsymbol{\ell}|}^{\mathrm{tot}}} f^{TT}(\boldsymbol{\ell},-\boldsymbol{L}-\boldsymbol{\ell}). \tag{9}$$

Changing the dummy integration variable $\boldsymbol{\ell} \mapsto -\boldsymbol{\ell}$ and applying the parity property $f^{TT}(-\boldsymbol{\ell},\boldsymbol{\ell}-\boldsymbol{L}) = f^{TT}(\boldsymbol{\ell},\boldsymbol{L}-\boldsymbol{\ell})$:

$$g_{-\boldsymbol{L}}(\boldsymbol{d},\boldsymbol{0}) = \frac{1}{2}\int \frac{\mathrm{d}^2\boldsymbol{\ell}}{(2\pi)^2}\frac{d_{\boldsymbol{\ell}} d_{\boldsymbol{L}-\boldsymbol{\ell}}}{C_\ell^{\mathrm{tot}} C_{|\boldsymbol{L}-\boldsymbol{\ell}|}^{\mathrm{tot}}} f^{TT}(\boldsymbol{\ell},\boldsymbol{L}-\boldsymbol{\ell}). \tag{10}$$

The Hu–Okamoto mode pairing $d_{\boldsymbol{\ell}} d_{\boldsymbol{L}-\boldsymbol{\ell}}$ thus emerges naturally from the score. Meanwhile, the mean-field trace subtraction vanishes for all nonzero multipoles $\boldsymbol{L} \neq \boldsymbol{0}$:

$$\begin{aligned} -\frac{1}{2}\,\mathrm{Tr}\,(C^{(0)-1}C_{,-\boldsymbol{L}}) = -\frac{1}{2}\int \frac{\mathrm{d}^2\boldsymbol{\ell}}{(2\pi)^2}\frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2}\frac{(2\pi)^2\delta^{(2)}(\boldsymbol{\ell}+\boldsymbol{\ell}')}{C_\ell^{\mathrm{tot}}} \\ \times (2\pi)^2\delta^{(2)}(\boldsymbol{\ell}+\boldsymbol{\ell}'+\boldsymbol{L}) f^{TT}(\boldsymbol{\ell},\boldsymbol{\ell}') = 0, \end{aligned} \tag{11}$$

because $\boldsymbol{\ell}+\boldsymbol{\ell}' = \boldsymbol{0}$ and $\boldsymbol{\ell}+\boldsymbol{\ell}' = -\boldsymbol{L}$ have disjoint support when $\boldsymbol{L} \neq \boldsymbol{0}$.
Finally, we calculate the Fisher response by taking the expectation under a non-zero lensing field $\phi_{\boldsymbol{L}}$, where $\langle d_{\boldsymbol{\ell}} d_{\boldsymbol{L}-\boldsymbol{\ell}}\rangle_{\boldsymbol{\phi}} = f^{TT}(\boldsymbol{\ell},\boldsymbol{L}-\boldsymbol{\ell})\phi_{\boldsymbol{L}} + \mathcal{O}(\phi^2)$:

$$\langle g_{-\boldsymbol{L}}(\boldsymbol{d},\boldsymbol{0})\rangle_{\boldsymbol{\phi}} = \phi_{\boldsymbol{L}}\int \frac{\mathrm{d}^2\boldsymbol{\ell}}{(2\pi)^2}\frac{[f^{TT}(\boldsymbol{\ell},\boldsymbol{L}-\boldsymbol{\ell})]^2}{2C_\ell^{\mathrm{tot}} C_{|\boldsymbol{L}-\boldsymbol{\ell}|}^{\mathrm{tot}}} \equiv F_L \phi_{\boldsymbol{L}}. \tag{12}$$

Distinct Fourier modes decouple, with diagonal Fisher curvature $F_L = 1/N_L^{(0)}$. Normalizing by $F_L^{-1} = N_L^{(0)}$ gives:

$$\hat{\phi}^{\mathrm{QE}}(\boldsymbol{L}) = F_L^{-1} g_{-\boldsymbol{L}}(\boldsymbol{d},\boldsymbol{0}) = N_L^{(0)}\int \frac{\mathrm{d}^2\boldsymbol{\ell}}{(2\pi)^2}\frac{f^{TT}(\boldsymbol{\ell},\boldsymbol{L}-\boldsymbol{\ell})}{2C_\ell^{\mathrm{tot}} C_{|\boldsymbol{L}-\boldsymbol{\ell}|}^{\mathrm{tot}}} d_{\boldsymbol{\ell}} d_{\boldsymbol{L}-\boldsymbol{\ell}}, \tag{13}$$

which is identically the Hu–Okamoto temperature reconstruction estimator Eq. (8.1.280).

Iterative reconstruction bias removal and convergence The reconstructed lensing potential power spectrum measured from the quadratic estimator contains both disconnected

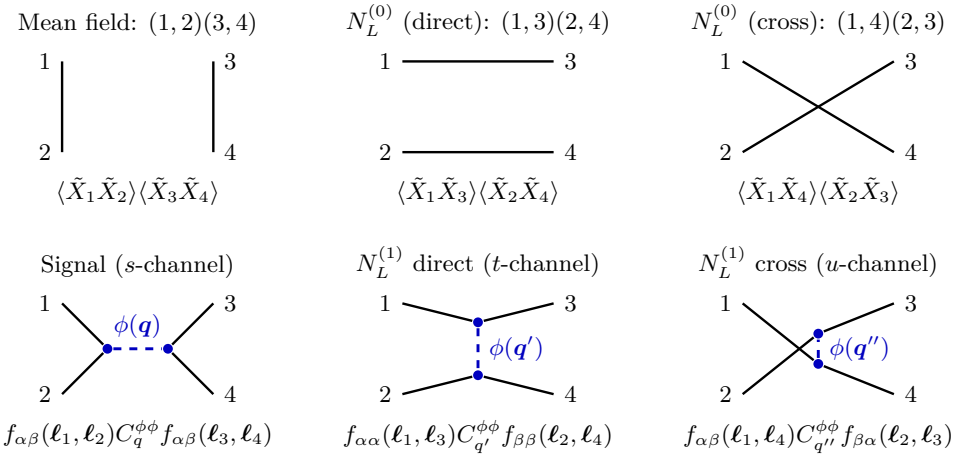


Figure 8.24: Feynman diagram representation of the lensed four-point function $\langle \tilde{X}_1 \tilde{X}_2 \tilde{X}_3 \tilde{X}_4 \rangle$ in CMB lensing reconstruction. Top row: disconnected Gaussian contractions forming the mean field and the Gaussian noise floor $N_L^{(0)}$. Bottom row: connected trispectrum topologies sourced by single lens exchange at wavevectors $\boldsymbol{q} = \boldsymbol{\ell}_1 + \boldsymbol{\ell}_2$ (s-channel signal), $\boldsymbol{q}' = \boldsymbol{\ell}_1 + \boldsymbol{\ell}_3$ (t-channel $N_L^{(1)}$ bias), and $\boldsymbol{q}'' = \boldsymbol{\ell}_1 + \boldsymbol{\ell}_4$ (u-channel $N_L^{(1)}$ bias).

Gaussian noise and connected non-Gaussian mode-coupling biases:

$$\hat{C}_L^{\hat{\phi}\hat{\phi}} = C_L^{\phi\phi} + N_L^{(0)} + N_L^{(1)}. \tag{8.1.297}$$

Because the $N_L^{(1)}$ bias arises from single lens exchange in the cross-channels (the connected t- and u-channel diagrams in Fig. 8.24), it depends linearly on the true, unobserved lensing power spectrum:

$$N_L^{(1)} = \sum_{L'} K_{LL'} C_{L'}^{\phi\phi}, \tag{8.1.298}$$

where $K_{LL'}$ represents the mode-coupling response kernel. Subtracting the disconnected Gaussian noise floor $N_L^{(0)}$ from the raw reconstruction defines the debiased data vector:

$$d_L \equiv \hat{C}_L^{\hat{\phi}\hat{\phi}} - N_L^{(0)} = C_L^{\phi\phi} + \sum_{L'} K_{LL'} C_{L'}^{\phi\phi}. \tag{8.1.299}$$

In compact matrix notation, grouping multipoles into vectors $\mathbf{d}$ and $\mathbf{C}$ with identity matrix $\mathbf{I}$ and coupling matrix $\mathbf{K}$:

$$\mathbf{d} = (\mathbf{I} + \mathbf{K})\mathbf{C}. \tag{8.1.300}$$

Rearranging to isolate the true lensing spectrum:

$$\mathbf{C} = \mathbf{d} - \mathbf{K}\mathbf{C}. \tag{8.1.301}$$

Because the true spectrum $\mathbf{C}$ on the right-hand side is unknown a priori, one replaces it with the estimate from the previous iteration step:

$$\boxed{\mathbf{C}^{(n+1)} = \mathbf{d} - \mathbf{K}\mathbf{C}^{(n)}.} \tag{8.1.302}$$

Starting from an initial fiducial model spectrum $\mathbf{C}^{(0)}$, this fixed-point equation is iterated until the successive update magnitude $\|\mathbf{C}^{(n+1)} - \mathbf{C}^{(n)}\|$ falls below a designated numerical tolerance.

By the Banach fixed-point theorem and linear convergence analysis, the error vector $\mathbf{e}^{(n)} \equiv \mathbf{C}^{(n)} - \mathbf{C}$ evolves according to $\mathbf{e}^{(n+1)} = -\mathbf{K}\mathbf{e}^{(n)} = (-\mathbf{K})^{n+1}\mathbf{e}^{(0)}$. Consequently, this iterative

bias removal scheme converges to the unique unbiased solution if and only if the spectral radius of the response matrix satisfies:

$$\boxed{\rho(\mathbf{K}) < 1.} \tag{8.1.303}$$

In practical CMB experiments, the fractional mode-coupling bias is small ($\|K_{LL'}\| \ll 1$ on linear and quasi-linear scales), ensuring rapid exponential convergence in just a few iterations.

Monte Carlo Realization Isolation in Practical Analyses While the fixed-point iteration (8.1.302) operates analytically on angular power spectra for idealized full-sky geometries, practical CMB experiments with irregular sky masks, anisotropic filtering, and inhomogeneous pixel noise evaluate both the disconnected noise floor $N_L^{(0)}$ and the mode-coupling bias $N_L^{(1)}$ directly from end-to-end Monte Carlo simulation realizations.

Crucially, this empirical simulation framework relies on the zero-mean property of cross-sample quadratic estimators. When distinct primary CMB realizations a and b are cross-paired in the quadratic estimator $q_{ab}(\boldsymbol{L}) \propto \int \tilde{X}^{(a)}\tilde{X}^{(b)}g$, their expectation value vanishes identically ($\langle q_{ab}\rangle = 0$), completely eliminating the primary cosmological signal $C_L^{\phi\phi}$. Because $N_L^{(0)}$ factorizes into products of unconditional two-point functions, and these two-point functions are ensemble-averaged covariances independent of the specific sky realization, the disconnected noise floor of the observed sky decomposes exactly into the auto- and cross-spectra of independent-sky simulations:

$$N_L^{(0),\alpha\beta} = C_L^{q_{ab}^{\rm ind} q_{ab}^{\rm ind}} + C_L^{q_{ab}^{\rm ind} q_{ba}^{\rm ind}}. \tag{8.1.304}$$

This establishes the master observational decomposition:

$$\boxed{C_L^{\hat{\phi}_{\alpha\beta}\hat{\phi}_{\alpha\beta}} = C_L^{q_{ab}^{\rm ind} q_{ab}^{\rm ind}} + C_L^{q_{ab}^{\rm ind} q_{ba}^{\rm ind}} + C_L^{\phi\phi} + N_L^{(1),\alpha\beta} + \mathcal{O}((C^{\phi\phi})^2).} \tag{8.1.305}$$

The disconnected $N_L^{(0)}$ bias emerges fundamentally as the statistical variance of this zero-mean estimator when independent deflection potentials are applied (reflecting purely Gaussian sample variance and instrumental noise in the complete absence of common lensing). A self-contained derivation of this realization cross-pairing scheme, its step-by-step isolation of Wick contractions, and its practical implementation subtleties is presented in Appendix D.

CMB Delensing Pipeline and Optimal Wiener Filtering The secondary B-modes generated by lensing of primordial E-modes act as a cosmic foreground that obscures the faint curl pattern produced by primordial gravitational waves (Fig. 9.18). Near the recombination peak ($\ell \approx 80$), this cosmic lensing variance acts as a dominant foreground that severely obscures the faint tensor signature. Detecting an inflationary tensor-to-scalar ratio $r \lesssim 10^{-2}$ therefore requires removing this lensing foreground at the map level through *delensing*.

The operational principle of delensing is to synthesize an estimate of the lensed B-mode template and subtract it directly from the observed polarization sky:

$$B_{\rm delens}(\boldsymbol{\ell}) = B_{\rm obs}(\boldsymbol{\ell}) - \hat{B}^{\rm lens}(\boldsymbol{\ell}). \tag{8.1.306}$$

From Eq. (8.1.271), the true lensed B-mode is generated by a convolution of the unlensed E-mode and the cosmic deflection potential ϕ:

$$B(\boldsymbol{\ell}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2} \sin[2(\varphi_{\boldsymbol{\ell}} - \varphi_{\boldsymbol{\ell}'})][\boldsymbol{\ell}' \cdot (\boldsymbol{\ell} - \boldsymbol{\ell}')]E(\boldsymbol{\ell}')\phi(\boldsymbol{\ell} - \boldsymbol{\ell}'). \tag{8.1.307}$$

In practical observations, neither the true unlensed E-mode nor the true lensing deflection potential ϕ is directly observable. The measured polarization field $E_{\rm obs}(\boldsymbol{\ell}) = E(\boldsymbol{\ell}) + n^E(\boldsymbol{\ell})$ is contaminated by detector noise N_ℓ^{EE}. Furthermore, the reconstructed deflection potential

$\hat{\phi}(\boldsymbol{L}) = \phi(\boldsymbol{L}) + n^{\phi}(\boldsymbol{L})$ is built from pairs of noisy observable CMB fields, inheriting both primary CMB sample variance ($N_L^{(0)}$) and instrumental noise, described by the total reconstruction noise power spectrum $N_L^{\phi\phi} = N_L^{(0)} + N_L^{(1)}$.

If one were to evaluate the template convolution using the raw observable fields $E_{\rm obs}$ and $\hat{\phi}$, the template $\hat{B}^{\rm lens}$ would inject substantial noise variance back into the polarization map upon subtraction, severely degrading the delensed signal-to-noise ratio. To minimize the variance of the template and prevent noise amplification, the unobserved fields must be replaced by their minimum-variance linear estimators: the *Wiener-filtered fields*.

Consider an observed data vector d_i comprising an unobserved cosmological signal s_i and additive noise n_i:

$$d_i = s_i + n_i, \tag{8.1.308}$$

where the indices i, j, k represent pixel indices, harmonic multipoles (ℓ, m), or Fourier wavevectors $\boldsymbol{\ell}$. The cosmological signal has zero mean and covariance matrix $S_{ij} \equiv \langle s_i s_j^* \rangle$, while the noise has zero mean and covariance matrix $N_{ij} \equiv \langle n_i n_j^* \rangle$, with $\langle s_i n_j^* \rangle = 0$. The total data covariance matrix is $C_{ij} \equiv \langle d_i d_j^* \rangle = S_{ij} + N_{ij}$.

For any linear estimator $\hat{s}_i = W_{ij} d_j$ (with summation over repeated indices), the reconstruction error vector is $e_i \equiv \hat{s}_i - s_i = (W_{ij} - \delta_{ij}) s_j + W_{ij} n_j$. The error covariance matrix evaluates to:

$$\mathcal{E}_{ij} \equiv \langle e_i e_j^* \rangle = (W_{ik} - \delta_{ik}) S_{kl} (W_{jl}^* - \delta_{jl}) + W_{ik} N_{kl} W_{jl}^*. \tag{8.1.309}$$

Minimizing the total mean squared reconstruction error $\mathcal{E} \equiv \mathcal{E}_{ii} = \mathrm{Tr}(\boldsymbol{\mathcal{E}})$ with respect to the filter coefficients W_{ij}^*:

$$\frac{\partial \mathcal{E}}{\partial W_{ij}^*} = W_{ik}(S_{kj} + N_{kj}) - S_{ij} = 0 \implies W_{ij} = \left[\boldsymbol{S}(\boldsymbol{S} + \boldsymbol{N})^{-1} \right]_{ij}. \tag{8.1.310}$$

Under statistical isotropy on the celestial sphere, the signal and noise covariances are diagonal in the spherical harmonic (or flat-sky Fourier) basis $i \equiv (\ell, m)$. The filter matrix W_{ij} reduces to the diagonal scalar Wiener weight:

$$W(\ell) = \frac{C_\ell}{C_\ell + N_\ell}, \tag{8.1.311}$$

which smoothly interpolates between unity on signal-dominated scales ($C_\ell \gg N_\ell$) and zero on noise-dominated scales ($N_\ell \gg C_\ell$).

The delensing pipeline applies this optimal estimator to synthesize the template:

1. **Wiener filtering**: The noisy signal fields are replaced by their minimum-variance estimates:

$$E(\boldsymbol{\ell}) \longrightarrow \bar{E}(\boldsymbol{\ell}) = W^E(\ell) E_{\rm obs}(\boldsymbol{\ell}), \qquad \phi(\boldsymbol{L}) \longrightarrow \bar{\phi}(\boldsymbol{L}) = W^{\phi}(L) \hat{\phi}_{\rm QE}(\boldsymbol{L}), \tag{8.1.312}$$

   with Wiener filter weights:

$$W^E(\ell) = \frac{C_\ell^{EE}}{C_\ell^{EE} + N_\ell^{EE}}, \qquad W^{\phi}(L) = \frac{C_L^{\phi\phi}}{C_L^{\phi\phi} + N_L^{(0)}}. \tag{8.1.313}$$

2. **Template synthesis**: Convolving the Wiener-filtered fields via Eq. (8.1.271) yields the minimum-variance lensed B-mode template:

$$\hat{B}^{\rm lens}(\boldsymbol{\ell}) = \int \frac{\mathrm{d}^2 \boldsymbol{\ell}'}{(2\pi)^2} \sin[2(\varphi_{\boldsymbol{\ell}} - \varphi_{\boldsymbol{\ell}'})] [\boldsymbol{\ell}' \cdot (\boldsymbol{\ell} - \boldsymbol{\ell}')] \bar{E}(\boldsymbol{\ell}') \bar{\phi}(\boldsymbol{\ell} - \boldsymbol{\ell}'). \tag{8.1.314}$$

3. **Template subtraction**: Subtracting $\hat{B}^{\rm lens}(\boldsymbol{\ell})$ from the observed map gives the residual field:

$$B_{\rm delens}(\boldsymbol{\ell}) = B_{\rm obs}(\boldsymbol{\ell}) - \hat{B}^{\rm lens}(\boldsymbol{\ell}). \tag{8.1.315}$$

The residual B-mode power spectrum after delensing evaluates to:

$$C_\ell^{BB,\rm delens} = C_\ell^{BB} - \int \frac{\mathrm{d}^2\boldsymbol{\ell}'}{(2\pi)^2} \sin^2[2(\varphi_{\boldsymbol{\ell}} - \varphi_{\boldsymbol{\ell}'})][\boldsymbol{\ell}' \cdot (\boldsymbol{\ell} - \boldsymbol{\ell}')]^2 \times W^E(\ell') C_{\ell'}^{EE} W^\phi(|\boldsymbol{\ell} - \boldsymbol{\ell}'|) C_{|\boldsymbol{\ell}-\boldsymbol{\ell}'|}^{\phi\phi}. \tag{8.1.316}$$

Bayesian MAP Origin of the Wiener Filter The Wiener filter (8.1.311) is not merely a heuristic minimum-variance linear filter; it is identically the *first-order Fisher-scoring step of the Maximum A Posteriori (MAP) estimator* under a Gaussian prior.

Suppose the cosmic deflection potential $\boldsymbol{\phi}$ possesses a zero-mean Gaussian prior with covariance matrix $(\mathbf{C}_\phi)_{AB} \equiv \langle \phi_A \phi_B^* \rangle$. By Bayes' theorem, the posterior distribution of the lensing field given observed CMB data $\boldsymbol{d}$ is:

$$P(\boldsymbol{\phi} \mid \boldsymbol{d}) \propto P(\boldsymbol{d} \mid \boldsymbol{\phi}) P(\boldsymbol{\phi}) \implies -\ln P(\boldsymbol{\phi} \mid \boldsymbol{d}) = -\ln P(\boldsymbol{d} \mid \boldsymbol{\phi}) + \frac{1}{2} \boldsymbol{\phi}^\top \mathbf{C}_\phi^{-1} \boldsymbol{\phi} + \text{const.} \tag{8.1.317}$$

The exact Maximum A Posteriori estimate $\hat{\boldsymbol{\phi}}_{\rm MAP}$ maximizes this posterior probability, satisfying the stationarity condition:

$$0 = -\frac{\partial \ln P(\boldsymbol{\phi} \mid \boldsymbol{d})}{\partial \phi_A} = -g_A(\boldsymbol{d}, \boldsymbol{\phi}) + (\mathbf{C}_\phi^{-1})_{AB} \phi_B, \tag{8.1.318}$$

where $g_A(\boldsymbol{d}, \boldsymbol{\phi})$ is the exact likelihood score vector defined in Eq. (8.1.295).

Substituting the first-order Fisher-curvature expansion about the unlensed fiducial model $\boldsymbol{\phi} = \mathbf{0}$, $g_A(\boldsymbol{d}, \boldsymbol{\phi}) \simeq g_A(\boldsymbol{d}, \mathbf{0}) - F_{AB}\phi_B$:

$$0 \simeq -g_A(\boldsymbol{d}, \mathbf{0}) + (F_{AB} + (\mathbf{C}_\phi^{-1})_{AB}) \phi_B \implies (F_{AB} + (\mathbf{C}_\phi^{-1})_{AB}) \phi_B \simeq g_A(\boldsymbol{d}, \mathbf{0}). \tag{8.1.319}$$

Inverting the total posterior curvature matrix gives the first-order MAP estimator:

$$\hat{\boldsymbol{\phi}}^{\rm MAP,(1)} = (\mathbf{F} + \mathbf{C}_\phi^{-1})^{-1} \boldsymbol{g}(\boldsymbol{d}, \mathbf{0}). \tag{8.1.320}$$

For statistically isotropic fields and homogeneous survey noise, distinct harmonic multipoles decouple ($F_L = 1/N_L^{(0)}$ and $(\mathbf{C}_\phi)_{LL'} = C_L^{\phi\phi} \delta_{LL'}$):

$$\left[\frac{1}{N_L^{(0)}} + \frac{1}{C_L^{\phi\phi}} \right] \hat{\phi}_{LM}^{\rm MAP,(1)} = g_{LM}(\boldsymbol{d}, \mathbf{0}) = \frac{1}{N_L^{(0)}} \hat{\phi}_{LM}^{\rm QE}. \tag{8.1.321}$$

Multiplying both sides by the inverse posterior variance $\left((N_L^{(0)})^{-1} + (C_L^{\phi\phi})^{-1}\right)^{-1} = \frac{C_L^{\phi\phi} N_L^{(0)}}{C_L^{\phi\phi} + N_L^{(0)}}$ directly yields:

$$\hat{\phi}_{LM}^{\rm MAP,(1)} = \frac{C_L^{\phi\phi}}{C_L^{\phi\phi} + N_L^{(0)}} \hat{\phi}_{LM}^{\rm QE} \equiv W^\phi(L) \hat{\phi}_{LM}^{\rm QE}. \tag{8.1.322}$$

Thus, the Wiener-filtered quadratic estimator is precisely the linearized MAP reconstruction. The prior breaks the strict unit-response condition of the raw quadratic estimator, dampening noise-dominated modes ($N_L^{(0)} \gg C_L^{\phi\phi}$) toward zero in order to minimize total mean-squared reconstruction error.

The complete connection between the linear Wiener-filtered estimator and iterative delensing is governed by Newton–Raphson scoring applied to the exact posterior distribution. From Bayes' theorem, the posterior log-probability is given by:

$$\ln P(\boldsymbol{\phi} \mid \boldsymbol{d}) = \ln P(\boldsymbol{d} \mid \boldsymbol{\phi}) - \frac{1}{2}\boldsymbol{\phi}^\top \mathbf{C}_\phi^{-1}\boldsymbol{\phi} + \text{const.} \tag{1}$$

Differentiating with respect to the deflection potential components ϕ_A yields the exact posterior score vector:

$$\frac{\partial \ln P(\boldsymbol{\phi} \mid \boldsymbol{d})}{\partial \phi_A} = g_A(\boldsymbol{d}, \boldsymbol{\phi}) - (\mathbf{C}_\phi^{-1})_{AB}\phi_B, \tag{2}$$

where $g_A(\boldsymbol{d}, \boldsymbol{\phi}) \equiv \partial \ln P(\boldsymbol{d} \mid \boldsymbol{\phi})/\partial \phi_A$ is the likelihood score vector. The exact Maximum A Posteriori (MAP) estimate satisfies the nonlinear stationarity condition:

$$g_A(\boldsymbol{d}, \hat{\boldsymbol{\phi}}_{\rm MAP}) - (\mathbf{C}_\phi^{-1})_{AB}\hat{\phi}_B^{\rm MAP} = 0. \tag{3}$$

Because gravitational deflection $\boldsymbol{x} \mapsto \boldsymbol{x} + \boldsymbol{\nabla}\phi$ enters the covariance $\mathbf{C}(\boldsymbol{\phi})$ to all orders, the likelihood score $g_A(\boldsymbol{d}, \boldsymbol{\phi})$ is nonlinear in $\boldsymbol{\phi}$. To solve this stationarity equation iteratively, we perform a Taylor expansion about the current iterate $\hat{\boldsymbol{\phi}}^{(n)}$:

$$g_A(\boldsymbol{d}, \boldsymbol{\phi}) - (\mathbf{C}_\phi^{-1})_{AB}\phi_B \simeq g_A(\boldsymbol{d}, \hat{\boldsymbol{\phi}}^{(n)}) - (\mathbf{C}_\phi^{-1})_{AB}\hat{\phi}_B^{(n)} + \sum_B \left[\left.\frac{\partial g_A}{\partial \phi_B}\right|_{\hat{\boldsymbol{\phi}}^{(n)}} - (\mathbf{C}_\phi^{-1})_{AB} \right] (\phi_B - \hat{\phi}_B^{(n)}). \tag{4}$$

Evaluating the derivative of the likelihood score $g_A(\boldsymbol{d}, \boldsymbol{\phi}) = \frac{1}{2}\boldsymbol{d}^\top \mathbf{C}^{-1}\mathbf{C}_{,A}\mathbf{C}^{-1}\boldsymbol{d} - \frac{1}{2}\,\mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,A})$ gives the observed Hessian curvature:

$$\begin{aligned}\frac{\partial g_A(\boldsymbol{d}, \boldsymbol{\phi})}{\partial \phi_B} = &-\boldsymbol{d}^\top \mathbf{C}^{-1}\mathbf{C}_{,B}\mathbf{C}^{-1}\mathbf{C}_{,A}\mathbf{C}^{-1}\boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^\top \mathbf{C}^{-1}\mathbf{C}_{,AB}\mathbf{C}^{-1}\boldsymbol{d} \\ &+ \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{C}^{-1}\mathbf{C}_{,B}\mathbf{C}^{-1}\mathbf{C}_{,A}\right) - \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{C}^{-1}\mathbf{C}_{,AB}\right).\end{aligned} \tag{5}$$

In the Fisher scoring method (the expected Hessian approximation of Newton–Raphson descent), we replace the data-dependent curvature by its conditional expectation at the current iterate, $\langle \boldsymbol{d}\boldsymbol{d}^\top \rangle_{\hat{\boldsymbol{\phi}}^{(n)}} = \mathbf{C}(\hat{\boldsymbol{\phi}}^{(n)})$. The second-derivative terms cancel identically because $\langle \boldsymbol{d}^\top \mathbf{C}^{-1}\mathbf{C}_{,AB}\mathbf{C}^{-1}\boldsymbol{d} \rangle = \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}_{,AB})$, leaving:

$$\left\langle -\frac{\partial g_A}{\partial \phi_B} \right\rangle_{\hat{\boldsymbol{\phi}}^{(n)}} = \frac{1}{2}\,\mathrm{Tr}\left[\mathbf{C}^{-1}(\hat{\boldsymbol{\phi}}^{(n)})\mathbf{C}_{,A}(\hat{\boldsymbol{\phi}}^{(n)})\mathbf{C}^{-1}(\hat{\boldsymbol{\phi}}^{(n)})\mathbf{C}_{,B}(\hat{\boldsymbol{\phi}}^{(n)})\right] \equiv F_{AB}^{(n)}, \tag{6}$$

where $F_{AB}^{(n)}$ is the Fisher information matrix evaluated at the current iterate $\hat{\boldsymbol{\phi}}^{(n)}$.
Setting the linearized stationarity equation to zero at the updated estimate $\hat{\boldsymbol{\phi}}^{(n+1)}$ gives:

$$g_A(\boldsymbol{d}, \hat{\boldsymbol{\phi}}^{(n)}) - (\mathbf{C}_\phi^{-1})_{AB}\hat{\phi}_B^{(n)} - \sum_B \left[F_{AB}^{(n)} + (\mathbf{C}_\phi^{-1})_{AB}\right](\hat{\phi}_B^{(n+1)} - \hat{\phi}_B^{(n)}) \simeq 0. \tag{7}$$

Isolating the update displacement $\Delta\hat{\boldsymbol{\phi}}^{(n)} \equiv \hat{\boldsymbol{\phi}}^{(n+1)} - \hat{\boldsymbol{\phi}}^{(n)}$ yields:

$$\left[\mathbf{F}^{(n)} + \mathbf{C}_\phi^{-1}\right](\hat{\boldsymbol{\phi}}^{(n+1)} - \hat{\boldsymbol{\phi}}^{(n)}) = \boldsymbol{g}(\boldsymbol{d}, \hat{\boldsymbol{\phi}}^{(n)}) - \mathbf{C}_\phi^{-1}\hat{\boldsymbol{\phi}}^{(n)}, \tag{8}$$

which inverts directly to give the recursive Fisher scoring update:

$$\hat{\boldsymbol{\phi}}^{(n+1)} = \hat{\boldsymbol{\phi}}^{(n)} + \left[\mathbf{F}^{(n)} + \mathbf{C}_\phi^{-1}\right]^{-1}\left[\boldsymbol{g}(\boldsymbol{d}, \hat{\boldsymbol{\phi}}^{(n)}) - \mathbf{C}_\phi^{-1}\hat{\boldsymbol{\phi}}^{(n)}\right]. \tag{9}$$

The iteration is initialized with the first-order Wiener-filtered quadratic estimate:

$$\hat{\boldsymbol{\phi}}^{(0)} = \hat{\boldsymbol{\phi}}^{\mathrm{MAP},(1)} = (\mathbf{F} + \mathbf{C}_\phi^{-1})^{-1}\boldsymbol{g}(\boldsymbol{d}, \mathbf{0}) = \mathbf{W}^\phi \hat{\boldsymbol{\phi}}_{\rm QE}. \tag{10}$$

The physical execution of this recursion proceeds via map-level unlensing:

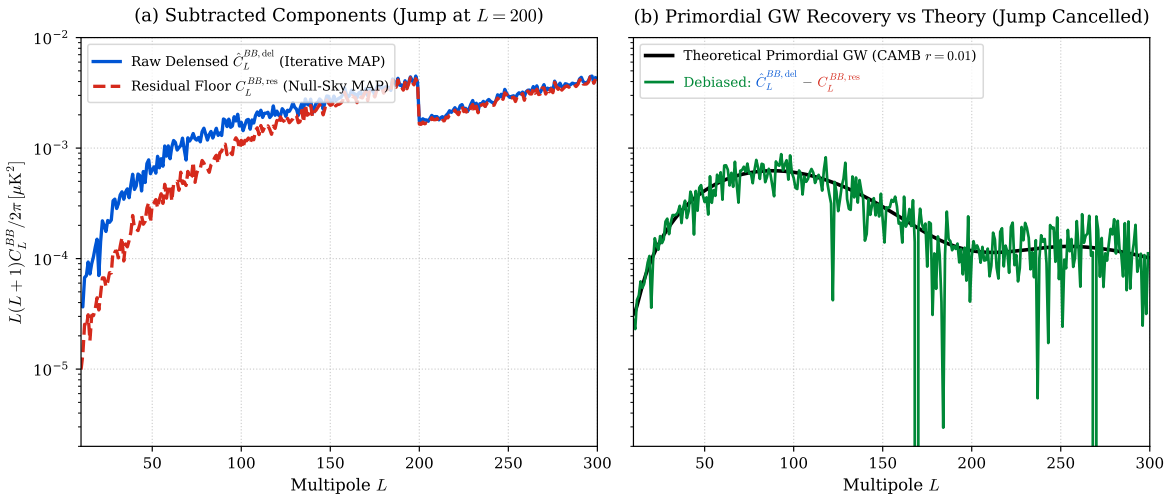


Figure 8.25: Curved-sky iterative Maximum A Posteriori (MAP) delensing debiasing demonstration (Carron & Lewis 2017) presented as a side-by-side comparison isolating the subtracted components and the recovered primordial tensor signal for $r = 0.01$. (a) Subtracted components: the raw iterative MAP delensed spectrum $\hat{C}_L^{BB,\mathrm{del}}$ (solid blue) and the null-sky residual floor $C_L^{BB,\mathrm{res}}$ (dashed red) identically exhibit the step-function transition at $L = 200$ caused by low-ℓ B-mode marginalization. (b) Primordial GW recovery vs. theory: subtracting the null-sky residual floor from the raw delensed sky cancels the $L = 200$ discontinuity completely, yielding the debiased spectrum $\hat{C}_L^{BB,\mathrm{del}} - C_L^{BB,\mathrm{res}}$ (solid green) that cleanly recovers the theoretical primordial gravitational wave spectrum (solid black, CAMB $r = 0.01$) across $L \in [10, 300]$.

1. **Map unlensing**: Evaluating the score vector $\boldsymbol{g}(\boldsymbol{d}, \hat{\boldsymbol{\phi}}^{(n)})$ at the current estimate corresponds physically to unlensing the observed sky map by displacing coordinates backward along $-\boldsymbol{\nabla}\hat{\phi}^{(n)}$:

$$P^{(n+1)}(\hat{\boldsymbol{n}}) = P_{\mathrm{obs}}(\hat{\boldsymbol{n}} - \boldsymbol{\nabla}\hat{\phi}^{(n)}(\hat{\boldsymbol{n}})). \tag{11}$$

2. **Noise floor reduction**: In this partially unlensed map $P^{(n+1)}$, the cosmic lensing B-mode power is systematically suppressed from $C_\ell^{BB,\mathrm{tot}}$ down to the residual spectrum $C_\ell^{BB,(n)}$. The effective reconstruction noise floor of the subsequent quadratic estimator:

$$N_L^{(0)}[C_\ell^{BB,(n)}] \tag{12}$$

drops at each iteration. This enables the extraction of a cleaner residual deflection update $\Delta\hat{\boldsymbol{\phi}}^{(n)}$, breaking through the single-step $\sim 50\%$ saturation ceiling and driving delensing efficiency to $\geq 80\%$.

The Single-Step Saturation Barrier and Iterative MAP Delensing While single-step template subtraction cleans current Stage-3 datasets, it faces a fundamental saturation barrier for deep cosmological surveys. Because the quadratic estimator was evaluated on the raw lensed sky, the cosmic lensing B-modes themselves act as effective noise in the denominator of the estimator:

$$N_L^{(0)} \propto C_\ell^{BB,\mathrm{lens}} + N_\ell^B. \tag{8.1.323}$$

Even in the idealized limit of vanishing detector noise ($N_\ell^B \to 0$), the presence of the cosmic lensing B-modes creates an irreducible noise floor in the template, capping single-step delensing

efficiency at roughly $\sim 50\%$.

To break through this saturation ceiling and achieve the $\gtrsim 80\%$ delensing efficiency required for CMB-S4 (demonstrated in Fig. 8.25), one must solve the full nonlinear MAP stationarity equation (8.1.318) iteratively via Eq. (9). Evaluating the score vector at each estimate corresponds physically to unlensing the observed sky maps:

$$P^{(n+1)}(\hat{\boldsymbol{n}}) = P_{\rm obs}(\hat{\boldsymbol{n}} - \boldsymbol{\nabla}\hat{\phi}^{(n)}(\hat{\boldsymbol{n}})). \tag{8.1.324}$$

Because this partially unlensed map $P^{(n+1)}$ contains less lensing B-mode power, the reconstruction noise floor $N_L^{(0)}$ of the subsequent quadratic estimator drops. This enables the extraction of a cleaner residual deflection update, progressively peeling away the cosmic noise floor at each iteration. As shown in Fig. 8.25, this map-level feedback loop recovers the primordial gravitational-wave spectrum across the recombination peak ($L \lesssim 150$) and achieves the sensitivity required to probe inflation down to $r \sim 5 \times 10^{-4}$.

This map-level MAP iteration must not be confused with the power-spectrum fixed-point iteration $\mathbf{C}^{(n+1)} = \mathbf{d} - \mathbf{K}\mathbf{C}^{(n)}$ in §8.1.6, which operates on 1D angular spectra solely to subtract the $N_L^{(1)}$ trispectrum bias. In contrast, MAP delensing operates directly on 2D sky maps, progressively peeling away lensing distortions and driving down the effective reconstruction noise floor.

Empirical Achievements in CMB Delensing: Internal CMB delensing was first demonstrated by the Planck collaboration (Carron, Lewis, & Challinor 2017), achieving a 28% reduction in lensing B-mode variance using internal quadratic estimators. Dedicated ground-based experiments combining BICEP/Keck polarization data with high-resolution South Pole Telescope (SPTpol) lensing maps (BICEP/Keck and SPTpol Collaborations 2021) achieved a 14% reduction in the uncertainty of the tensor-to-scalar ratio r, setting the upper limit $r < 0.036$ (95% C.L.). Forthcoming Stage-4 experiments (CMB-S4) will employ iterative MAP delensing to achieve $\gtrsim 80\%$ efficiency, unlocking the definitive search for inflationary gravitational waves.

8.2 Evidence for the existence of dark matter

8.2.1 Dark matter in galaxies

Rotation curves and dynamical mass

Dynamical Mass and the Keplerian Decline The dynamics of spiral galaxies provides the most direct and historically influential astrophysical evidence for dark matter halos. In a circularly symmetric disk galaxy in dynamical equilibrium, stars and neutral gas clouds move on circular orbits in the equatorial plane ($z = 0$). Equating the centripetal acceleration to the radial Newtonian gravitational force:

$$\frac{v^2(r)}{r} = \left.\frac{\partial \Phi}{\partial r}\right|_{z=0} = \frac{G_{\rm N} M(<r)}{r^2}, \tag{8.2.1}$$

where $M(<r)$ is the enclosed dynamical mass within spherical radius r. Inverting Eq. (8.2.1) gives the circular orbital velocity:

$$v(r) = \sqrt{\frac{G_{\rm N} M(<r)}{r}}. \tag{8.2.2}$$

For an isolated luminous distribution with total stellar mass M_* confined within an optical radius $R_{\rm opt}$, the enclosed mass at distances $r \gg R_{\rm opt}$ is constant: $M(<r) \approx M_*$. Consequently, if the visible stars and gas represented the entire mass budget of the galaxy, the circular velocity beyond the optical radius must exhibit a *Keplerian decline*:

$$v(r) \propto r^{-1/2} \quad (r \gg R_{\rm opt}). \tag{8.2.3}$$

Observed Flat Rotation Curves and Dark Matter Halos Extensive spectroscopic observations of optical emission lines (Hα, [N II]) in ionized gas (Rubin, Ford, & Thonnard 1978, 1980) and radio 21 cm hyperfine emission lines in neutral hydrogen (H I) disks (Bosma 1981a,b; Begeman 1987) demonstrated that spiral galaxy rotation curves do not exhibit a Keplerian falloff. Instead, as radio synthesis arrays mapped H I gas out to distances $r \sim 2-5\,R_{\rm opt}$, the circular velocity asymptotically levels off into an extended flat plateau:

$$v(r) \approx v_{\rm flat} = \text{const} \quad (r \gg R_{\rm opt}). \tag{8.2.4}$$

Substituting $v(r) = v_{\rm flat}$ into Eq. (8.2.1) requires that the enclosed mass continues to grow linearly with radius well beyond the luminous edge of the galaxy:

$$M(<r) = \frac{v_{\rm flat}^2 r}{G_{\rm N}} \propto r. \tag{8.2.5}$$

Differentiating $M(<r) = \int_0^r 4\pi r'^2 \rho(r')\,{\rm d}r'$ gives the implied radial mass density:

$$\rho(r) = \frac{1}{4\pi r^2}\frac{{\rm d}M}{{\rm d}r} = \frac{v_{\rm flat}^2}{4\pi G_{\rm N}}\frac{1}{r^2}. \tag{8.2.6}$$

Because the surface brightness of the stellar disk falls off exponentially ($\Sigma_*(R) \propto e^{-R/R_d}$), the mass-to-light ratio $\Upsilon(r) \equiv M(<r)/L(<r)$ increases monotonically with radius. This implies the presence of an extended, quasi-spherical, non-luminous mass distribution: the *dark matter halo*.

The Milky Way The rotation curve of our own Galaxy is mapped with high precision across multiple tracers. In the deep galactic core, stellar velocity dispersions indicate the presence of a supermassive black hole with mass $M_\bullet \simeq (4.15 \pm 0.13) \times 10^6 M_\odot$, determined from the Keplerian orbital trajectories of individual stars near Sagittarius A*.

Taking the Sun's galactocentric radius and circular speed ($R_0 \simeq 8.2\,{\rm kpc}$, $v_0 \simeq 220\,{\rm km\cdot s^{-1}}$) as a baseline, the enclosed dynamical mass within radius r scales as:

$$M(<r) = 9.6 \times 10^{10} M_\odot \left(\frac{v}{220\,{\rm km\cdot s^{-1}}}\right)^2 \left(\frac{r}{100\,{\rm kpc}}\right). \tag{8.2.7}$$

Given the estimated total luminous output of the Milky Way, $L_{\rm MW} \simeq 2.3 \times 10^{10} L_\odot$, the mass-to-light ratio $\Upsilon \equiv M/L$ enclosed within the dark matter halo radius $R_{\rm halo}$ satisfies:

$$\Upsilon_{\rm MW} \simeq 50\,\Upsilon_\odot \left(\frac{R_{\rm halo}}{100\,{\rm kpc}}\right). \tag{8.2.8}$$

Neutral hydrogen (H I) emission remains detectable out to $r \sim 20\,{\rm kpc}$ with no evidence of a Keplerian decline, establishing $R_{\rm halo} > 20\,{\rm kpc}$. Kinematic modeling of distant halo globular clusters and satellite dwarf spheroidal galaxies (including the Magellanic Clouds and Gaia DR3 stellar streams) requires $R_{\rm halo} > 100\,{\rm kpc}$ with a total virial mass $M_{\rm vir} \simeq (1.0-1.5) \times 10^{12} M_\odot$, yielding a global galactic mass-to-light ratio $\Upsilon_{\rm MW} \sim 40-60\,\Upsilon_\odot$.

In the immediate solar neighborhood, comparing the local stellar and interstellar gas inventories against vertical stellar kinematics normal to the galactic disk yields $\Upsilon_{\rm local} \simeq 5\,\Upsilon_\odot$ over a vertical scale of roughly 100 pc (the classical Oort limit).

The Local Group—predominantly composed of the Milky Way, the Andromeda galaxy (M31), and their satellite retinue—can be treated as a dynamically isolated two-body system. Applying the Kahn–Woltjer (1959) timing argument to their mutual approach velocity ($v \sim 120\,{\rm km\cdot s^{-1}}$) over a cosmic Hubble time yields a total Local Group dynamical mass $M_{\rm group} = (3.5-5.5) \times 10^{12} M_\odot$, corresponding to a group mass-to-light ratio $\Upsilon_{\rm group} \sim 100\,\Upsilon_\odot$.

The Freeman Exponential Thin Disk Rotation Curve To rigorously prove that the luminous stellar disk cannot explain flat rotation curves, we derive the exact gravitational potential and circular velocity of an infinitesimally thin, exponential disk from first principles (Freeman 1970).

Consider an axisymmetric, infinitesimally thin stellar disk with surface mass density:

$$\Sigma(R) = \Sigma_0 e^{-R/R_d}, \tag{1}$$

where Σ_0 is the central surface density, and R_d is the exponential scale length. The total disk mass is $M_{\rm disk} = \int_0^\infty 2\pi R\Sigma(R)\,\mathrm{d}R = 2\pi\Sigma_0 R_d^2$.
The Newtonian gravitational potential $\Phi(R,z)$ satisfies Poisson's equation in cylindrical coordinates (R,θ,z):

$$\frac{1}{R}\frac{\partial}{\partial R}\left(R\frac{\partial\Phi}{\partial R}\right) + \frac{\partial^2\Phi}{\partial z^2} = 4\pi G_{\rm N}\Sigma(R)\delta(z). \tag{2}$$

Expanding $\Phi(R,z)$ and $\Sigma(R)$ in Hankel transforms of order zero using the Bessel identity $\int_0^\infty kJ_0(kR)J_0(kR')\,\mathrm{d}k = \frac{1}{R}\delta(R-R')$:

$$\Sigma(R) = \int_0^\infty k\tilde{\Sigma}(k)J_0(kR)\;\mathrm{d}k\,, \qquad \tilde{\Sigma}(k) = \int_0^\infty R\Sigma(R)J_0(kR)\;\mathrm{d}R\,. \tag{3}$$

Using the standard Laplace-type Bessel integral $\int_0^\infty Re^{-R/R_d}J_0(kR)\,\mathrm{d}R = \frac{R_d^2}{(1+k^2R_d^2)^{3/2}}$:

$$\tilde{\Sigma}(k) = \frac{\Sigma_0 R_d^2}{(1+k^2R_d^2)^{3/2}}. \tag{4}$$

For $z \neq 0$, the Laplace equation $\frac{1}{R}\frac{\partial}{\partial R}(R\frac{\partial\Phi}{\partial R}) + \frac{\partial^2\Phi}{\partial z^2} = 0$ is satisfied by separable modes $J_0(kR)e^{-k|z|}$:

$$\Phi(R,z) = -2\pi G_{\rm N}\int_0^\infty \tilde{\Sigma}(k)J_0(kR)e^{-k|z|}\;\mathrm{d}k\,. \tag{5}$$

The circular velocity in the disk plane ($z=0$) is governed by the radial force:

$$\begin{aligned} v_{\rm disk}^2(R) &= R\left.\frac{\partial\Phi}{\partial R}\right|_{z=0} = 2\pi G_{\rm N}R\int_0^\infty k\tilde{\Sigma}(k)J_1(kR)\;\mathrm{d}k \\ &= 2\pi G_{\rm N}\Sigma_0 R_d^2 R\int_0^\infty \frac{kJ_1(kR)}{(1+k^2R_d^2)^{3/2}}\;\mathrm{d}k\,. \end{aligned} \tag{6}$$

Introducing the dimensionless radius $y \equiv \frac{R}{2R_d}$, this integral can be evaluated analytically in terms of the modified Bessel functions of the first and second kinds of orders 0 and 1 (I_0, I_1, K_0, K_1):

$$\begin{aligned} v_{\rm disk}^2(y) &= 4\pi G_{\rm N}\Sigma_0 R_d y^2\left[I_0(y)K_0(y) - I_1(y)K_1(y)\right] \\ &= \frac{2G_{\rm N}M_{\rm disk}}{R_d}y^2\left[I_0(y)K_0(y) - I_1(y)K_1(y)\right]. \end{aligned} \tag{7}$$

Physical Properties of the Freeman Disk Rotation Curve:

1. At small radii ($R \ll R_d$, so $y \to 0$), the asymptotic expansions $I_0(y) \approx 1$, $I_1(y) \approx y/2$, $K_0(y) \approx -\ln(y/2)$, $K_1(y) \approx 1/y$ give $v_{\rm disk}^2 \propto y^2 \implies v_{\rm disk} \propto R$, corresponding to solid-body rotation.

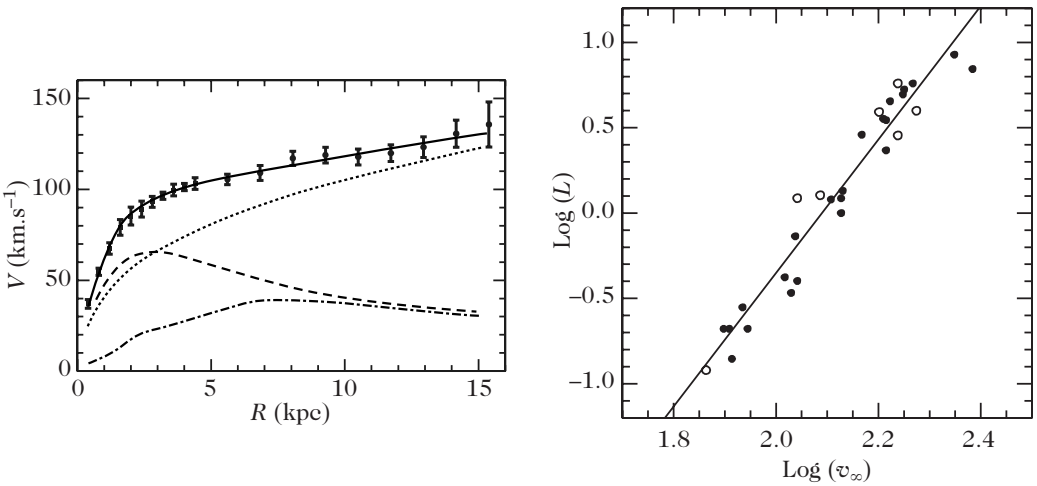


Figure 8.26: (Left) Rotation curve decomposition for the spiral galaxy M33. Observational data (open and filled circles) are fitted by a dark matter halo model (solid curve). Constituent contributions are shown for the dark halo (dash-dotted curve), stellar disk (short dashes), and neutral gas disk (long dashes). (Right) Tully–Fisher scaling relation for spiral galaxies relating total optical luminosity to plateau rotation speed.

2. The circular velocity reaches a maximum at:

$$R_{\rm peak} \approx 2.15 R_d, \qquad v_{\rm peak} \approx 0.88\sqrt{\frac{G_{\rm N} M_{\rm disk}}{R_d}}. \tag{8.2.9}$$

3. For radii beyond the peak ($R > 2.15R_d$), the velocity declines monotonically. In the asymptotic limit $R \gg R_d$ ($y \to \infty$):

$$I_0(y)K_0(y) - I_1(y)K_1(y) \approx \frac{1}{2y^3} \implies v^2_{\rm disk}(R) \approx \frac{G_{\rm N} M_{\rm disk}}{R}, \tag{8.2.10}$$

recovering the standard Keplerian point-mass falloff $v_{\rm disk} \propto R^{-1/2}$.

As displayed in Fig. 9.19 for M33, the observed rotation curve remains flat out to distances where the Freeman disk contribution has dropped by more than a factor of three. The smooth, flat transition between disk-dominated and halo-dominated regimes is often termed the *disk-halo conspiracy*, reflecting the complex baryonic assembly history of galaxies within dark matter potential wells.

Halo Models In rotation curve decompositions, the dark matter halo is parameterized by phenomenological density profiles:

- **Pseudo-isothermal cored halo**:

$$\rho_{\rm iso}(r) = \frac{\rho_0}{1 + (r/r_c)^2}, \qquad v^2_{\rm iso}(r) = 4\pi G_{\rm N} \rho_0 r_c^2 \left[1 - \frac{r_c}{r} \arctan\left(\frac{r}{r_c}\right)\right]. \tag{8.2.11}$$

- **Burkert cored halo** (Burkert 1995):

$$\rho_{\rm B}(r) = \frac{\rho_0}{\left(1 + \frac{r}{r_0}\right)\left(1 + \frac{r^2}{r_0^2}\right)}, \tag{8.2.12}$$

which resolves the "cusp-core problem" in dwarf spheroidal and low-surface-brightness galaxies.

- **NFW cuspy halo**: given by Eq. (8.1.124).

The Tully–Fisher Scaling Relation Across spiral galaxies, total luminosity correlates tightly with the asymptotic plateau rotation velocity v_∞. This empirical *Tully–Fisher relation* reads:

$$L \propto v_\infty^\alpha, \qquad \alpha = 3.9 \pm 0.2. \tag{8.2.13}$$

For gas-rich dwarf galaxies, the relation generalizes into the *Baryonic Tully–Fisher relation* (McGaugh et al. 2000), which relates the total baryonic mass $M_{\text{baryon}} = M_{\text{stars}} + M_{\text{gas}}$ directly to v_∞:

$$M_{\text{baryon}} = A v_\infty^4, \tag{8.2.14}$$

with remarkable universality across five orders of magnitude in mass.

Other observations

Beyond planar rotation curves, independent dynamical probes confirm the presence of dark matter in early-type and irregular galaxies:

- **Polar ring galaxies**: These rare systems feature an outer ring of gas, dust, and stars orbiting nearly perpendicular to the central host galaxy disk. Because the polar ring measures the gravitational potential in the vertical direction while the host disk measures it in the equatorial plane, polar ring galaxies constrain the 3D 3-axis shape of dark matter halos, confirming that halos are moderately triaxial to oblate ($q \equiv c/a \sim 0.7 - 0.9$).
- **X-ray halos of giant elliptical galaxies**: Giant ellipticals are surrounded by diffuse, gravitationally bound atmospheres of hot gas ($T \sim 10^7\,\text{K}$) emitting thermal X-rays. Hydrostatic modeling reveals that dark matter constitutes $\gtrsim 90\%$ of the total mass within $50 - 100\,\text{kpc}$.
- **Stellar velocity dispersions in dwarf spheroidals**: Dwarf spheroidals (e.g., Draco, Ursa Minor) exhibit extreme mass-to-light ratios $\Upsilon \sim 100 - 1000\,\Upsilon_\odot$, making them the most dark-matter-dominated systems in the Universe.

8.2.2 Dark matter in clusters and groups of galaxies

The earliest dynamical evidence for dark matter was uncovered by Fritz Zwicky in 1933 during his study of the Coma cluster. Galaxy clusters are the largest gravitationally bound, virialized structures in the Universe, comprising between a few hundred and several thousand individual galaxies.

The virial theorem

Because galaxy clusters are decoupled from the background cosmological expansion and dynamically relaxed, they can be modeled as self-gravitating systems in dynamical equilibrium.

Consider a continuous, self-gravitating fluid system with mass density $\rho(\boldsymbol{x})$, velocity field $\boldsymbol{v}(\boldsymbol{x})$, and isotropic thermal pressure $P(\boldsymbol{x})$ confined within a spatial volume V bounded by surface ∂V. The momentum conservation equation is:

$$\frac{\partial}{\partial t}(\rho v_i) + \frac{\partial}{\partial x_j}(\rho v_i v_j + P\delta_{ij}) = -\rho\frac{\partial \Phi}{\partial x_i}. \tag{1}$$

Define the second moment of inertia tensor about the center of mass:

$$I_{ij} \equiv \int_V \rho(\boldsymbol{x}) x_i x_j \,\mathrm{d}^3 x. \tag{2}$$

Differentiating I_{ij} twice with respect to time:

$$\frac{\mathrm{d}I_{ij}}{\mathrm{d}t} = \int_V \frac{\partial \rho}{\partial t} x_i x_j \,\mathrm{d}^3x = -\int_V \frac{\partial(\rho v_k)}{\partial x_k} x_i x_j \,\mathrm{d}^3x = \int_V \rho(v_i x_j + v_j x_i)\,\mathrm{d}^3x. \tag{3}$$

Taking the second time derivative and substituting Eq. (1):

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}^2 I_{ij}}{\mathrm{d}t^2} &= \int_V \rho v_i v_j \,\mathrm{d}^3x + \int_V x_j \frac{\partial(\rho v_i)}{\partial t}\,\mathrm{d}^3x \\ &= \int_V \rho v_i v_j \,\mathrm{d}^3x - \int_V x_j \frac{\partial}{\partial x_k}(\rho v_i v_k + P\delta_{ik})\,\mathrm{d}^3x - \int_V \rho x_j \frac{\partial \Phi}{\partial x_i}\,\mathrm{d}^3x. \end{aligned} \tag{4}$$

Applying Gauss's divergence theorem to the second term:

$$-\int_V x_j \frac{\partial}{\partial x_k}(\rho v_i v_k + P\delta_{ik})\,\mathrm{d}^3x = \int_V (\rho v_i v_j + P\delta_{ij})\,\mathrm{d}^3x - \oint_{\partial V} x_j(\rho v_i v_k + P\delta_{ik})\,\mathrm{d}S_k\,. \tag{5}$$

Taking the trace of the tensor ($i = j$):

$$\frac{1}{2}\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} = 2K + 3\int_V P\,\mathrm{d}^3x + W - \oint_{\partial V} P\boldsymbol{x}\cdot\mathrm{d}\boldsymbol{S}\,, \tag{6}$$

where $K \equiv \frac{1}{2}\int_V \rho v^2\,\mathrm{d}^3x$ is the bulk kinetic energy, $W \equiv -\int_V \rho\boldsymbol{x}\cdot\boldsymbol{\nabla}\Phi\,\mathrm{d}^3x = U$ is the total gravitational potential energy, and the surface integral represents the boundary pressure support. In steady-state dynamical equilibrium, the time average vanishes ($\langle \ddot{I}\rangle = 0$), yielding the complete *virial theorem*:

$$2K + W - \oint_{\partial V} P\boldsymbol{x}\cdot\mathrm{d}\boldsymbol{S} = 0. \tag{7}$$

For an isolated system bounded at infinity where surface pressure vanishes ($P|_{\partial V} \to 0$), Eq. (7) reduces to the classical relation:

$$2\langle K\rangle + \langle U\rangle = 0. \tag{8.2.15}$$

Expressing $\langle K\rangle = \frac{1}{2}M\langle v^2\rangle = \frac{3}{2}M\sigma_v^2$ (where σ_v is the line-of-sight velocity dispersion) and parameterizing the potential energy by $\langle U\rangle = -\alpha G_{\mathrm{N}} M^2/r_{\mathrm{h}}$ (where r_{h} is the half-mass radius and $\alpha \approx 0.4$):

$$M_{\mathrm{vir}} = \frac{3\sigma_v^2 r_{\mathrm{h}}}{\alpha G_{\mathrm{N}}}. \tag{8.2.16}$$

The Coma cluster

Applying Eq. (8.2.16) to the Coma cluster (Zwicky 1933), the line-of-sight velocity dispersion of approximately 1,000 member galaxies is $\sigma_v \approx 1000\,\mathrm{km\cdot s^{-1}}$ across a radius $R \approx 1.5 - 2\,\mathrm{Mpc}$. The resulting total virial mass is:

$$M_{\mathrm{Coma}} \approx 1.5 \times 10^{15} M_\odot. \tag{8.2.17}$$

Summing the optical luminosity of all member galaxies gives $L_B \approx 5 \times 10^{12} L_\odot$, yielding an extraordinary mass-to-light ratio:

$$\Upsilon_{\mathrm{Coma}} \approx 300\,\Upsilon_\odot. \tag{8.2.18}$$

Because ordinary stellar populations have $\Upsilon_* \sim 1 - 3\,\Upsilon_\odot$, Zwicky deduced that more than 98% of the mass in the Coma cluster must be in the form of non-luminous "dunkle Materie" (dark matter).

Cluster mass determinations

Beyond the virial theorem, two independent techniques determine cluster masses:

1. **Gravitational Lensing**: As derived in Section 8.1.2, strong lensing arcs in cluster cores ($r \lesssim 200\,\mathrm{kpc}$) constrain $\Upsilon \sim 100 - 200\,\Upsilon_\odot$, while statistical weak lensing shear of background galaxies out to $r \lesssim 1\,\mathrm{Mpc}$ gives $\Upsilon \approx 300\,\Upsilon_\odot$.

2. **X-ray ICM Hydrostatic Equilibrium**: The intracluster medium (ICM) consists of a diffuse, fully ionized plasma ($T \sim 10^7 - 10^8\,\mathrm{K}$) emitting thermal Bremsstrahlung radiation. Under spherical hydrostatic equilibrium, the thermal gas pressure balances gravity:

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{G_\mathrm{N} M(<r)\rho_\mathrm{gas}(r)}{r^2}. \tag{8.2.19}$$

   Applying the ideal gas equation of state $P(r) = \rho_\mathrm{gas}(r)\frac{k_\mathrm{B}T(r)}{\mu m_\mathrm{p}}$:

$$M_\mathrm{X}(<r) = -\frac{k_\mathrm{B}T(r)r}{G_\mathrm{N}\mu m_\mathrm{p}}\left[\frac{\mathrm{d}\ln\rho_\mathrm{gas}}{\mathrm{d}\ln r} + \frac{\mathrm{d}\ln T}{\mathrm{d}\ln r}\right]. \tag{8.2.20}$$

The Bullet Cluster (1E 0657-558): Direct Empirical Proof of Dark Matter The most decisive observational proof of the physical reality of dark matter and its collisionless nature is provided by the "Bullet Cluster" (1E 0657-558; Clowe et al. 2006; Markevitch et al. 2004).

The system consists of two colliding subclusters that underwent a high-velocity merger ($v_\mathrm{rel} \sim 4500\,\mathrm{km\cdot s^{-1}}$) approximately 100 Myr ago. Crucially, the collision spatially separates the three primary mass components of the clusters:

1. **Collisionless Galaxies (Optical)**: Individual stars and galaxies have microscopic cross-sections; during the merger, the galaxies pass directly through one another without physical collisions, acting as test particles.

2. **Collisional Gas (X-ray)**: The diffuse intracluster plasma represents $\sim 90\%$ of the baryonic mass. Because the gas is fluid-like and collisional, the colliding gas clouds experience intense hydrodynamic drag, ram pressure, and shock heating ($T \sim 15\,\mathrm{keV}$). The X-ray gas is halted and stripped, lagging far behind the galaxies.

3. **Dark Matter (Weak Lensing)**: Gravitational weak lensing mass reconstruction (via the Kaiser–Squires algorithm) maps the total projected gravitational potential wells without assumptions about mass or light.

The 8σ Spatial Offset: The weak lensing convergence peaks coincide spatially with the collisionless galaxies, distinctly separated from the dissipative X-ray gas peaks by a spatial offset of $> 8\sigma$ statistical significance. If the gravitational potential were sourced solely by baryonic matter under some modified gravity law (such as MOND), the gravitational potential wells would center on the X-ray gas where the baryonic mass is concentrated. Instead, the gravitational wells are centered on the collisionless galaxies and empty space, providing direct empirical proof that the dominant mass component is collisionless, non-baryonic dark matter.

Furthermore, the fact that the dark matter subclusters did not dissolve or suffer significant deceleration during the collision establishes a stringent upper bound on the dark matter self-interaction cross-section:

$$\frac{\sigma_\mathrm{DM}}{m} \lesssim 1\,\mathrm{cm^2\cdot g^{-1}} \approx 2\,\mathrm{barns\cdot GeV^{-1}}. \tag{8.2.21}$$

Summary of cluster properties

Global observational surveys establish that in rich galaxy clusters, the mass budget is partitioned into:

$$\frac{M_{\rm gas}}{M_{\rm DM}} = (15 \pm 5)\%, \qquad \frac{M_*}{M_{\rm DM}} = (3 \pm 2)\%, \tag{8.2.22}$$

with total mass profiles matching the universal NFW profile across a broad radial dynamic range.

8.2.3 Cosmological evidence

Beyond galaxies and clusters, modern cosmology establishes the existence of cold dark matter through three pillars:

1. **CMB Acoustic Peak Heights**: The acoustic oscillations of the photon-baryon plasma before recombination generate harmonic peaks in the CMB power spectrum. The odd-numbered acoustic peaks (compressions into gravitational potential wells) are boosted by baryon inertia, while the even-numbered peaks (rarefactions) are suppressed. Precision measurements by WMAP and Planck determine the cosmic baryon and dark matter densities:
$$\Omega_{\rm b0}h^2 = 0.02237 \pm 0.00015, \qquad \Omega_{\rm c0}h^2 = 0.1200 \pm 0.0012, \tag{8.2.23}$$
demanding $\rho_{\rm c}/\rho_{\rm b} \simeq 5.36 \pm 0.06$.

2. **Big Bang Nucleosynthesis (BBN)**: The primordial abundance of light elements (^{4}He, ^{2}H, ^{3}He, ^{7}Li) depends strictly on the baryon-to-photon ratio $\eta \equiv n_{\rm b}/n_\gamma$. Measuring the primordial deuterium abundance in high-redshift quasar absorption systems gives $\Omega_{\rm b0}h^2 = 0.022 \pm 0.002$, in complete concordance with CMB constraints, proving that the non-baryonic dark matter cannot consist of ordinary protons and neutrons.

3. **Growth of Large-Scale Structure**: Baryons remain tightly coupled to photons until recombination ($z_* \simeq 1090$), preventing baryonic density perturbations from growing. If the Universe contained only baryons, density perturbations could only have grown by a factor of $(1 + z_*) \sim 10^3$, which is completely insufficient to form the observed non-linear galaxy distribution from the 10^{-5} perturbations seen in the CMB. Neutral, cold dark matter decouples much earlier, growing logarithmically during radiation domination and linearly ($\delta \propto a$) during matter domination, creating the potential wells into which baryons fall after recombination.

8.2.4 Summary of dark matter evidence

Table 8.3 summarizes the mass-to-light ratios and implied dark matter density parameters inferred across physical systems spanning six orders of magnitude in spatial scale.

These diverse empirical results demonstrate that astronomical structures cannot be understood at any scale without invoking dark matter. Although its microscopic nature and detailed coupling to ordinary matter remain fundamental open problems, Table 8.4 highlights that the dark-to-baryon ratio remains remarkably stable from galactic disks to the cosmological horizon.

8.2.5 Candidates and constraints

Cosmological phenomenology

Particle dark matter candidates are broadly classified by their kinematic properties at the onset of matter domination:

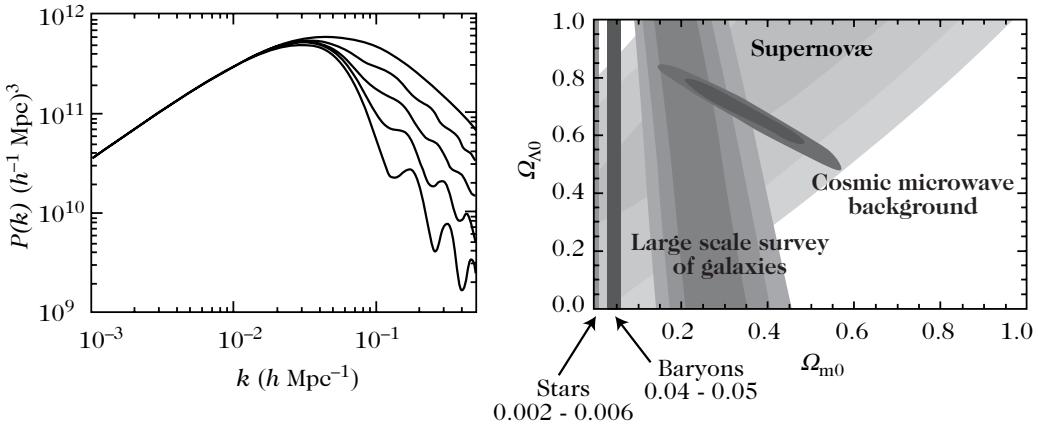


Figure 8.27: CMB temperature anisotropy power spectrum C_ℓ^{TT} measured by the Planck satellite. The heights of the first, second, and third acoustic peaks precisely determine the physical baryon density $\Omega_{\mathrm{b}0}h^2$ and cold dark matter density $\Omega_{\mathrm{c}0}h^2$.

Table 8.3: Summary of estimated mass-to-light ratios $\Upsilon/\Upsilon_\odot$ across astronomical systems and the resulting implied dark matter cosmological density parameters $\Omega_{\mathrm{c}0}$ at various physical scales.

System	Scale / Tracer	Physical Scale	$\Upsilon/\Upsilon_\odot$	$\Omega_{\mathrm{c}0}$
Milky Way	Oort limit	100 pc	5 (?)	$0.003\,h^{-1}$
	Rotation curves	20 kpc	10	$0.006\,h^{-1}$
	Satellites	80 kpc	40	$0.024\,h^{-1}$
	Magellanic stream	100 kpc	> 80	$> 0.05\,h^{-1}$
Elliptical galaxies	Core tuning	2 kpc	$12\,h$	0.007
	X-ray halo	100 kpc	> 750	$0.46\,h^{-1}$
Spiral galaxies	Rotation curves	50 kpc	$> 30\,h$	0.018
Galaxy groups	Local Group	800 kpc	100	$0.06\,h^{-1}$
	Other groups	1 Mpc	$260\,h$	0.16
Galaxy clusters	Coma (B-band)	2 Mpc	$400\,h$	0.25
	Perseus	2 Mpc	$600\,h$	0.37
Cosmology	Large-scale structure	3000 Mpc	—	$0.120\,h^{-2}$

Table 8.4: Relative ratio of dark matter to baryonic matter across astronomical structures, illustrating the approximate scale invariance of the dark-to-baryon ratio.

System	Spiral galaxy	Galaxy cluster	Observable Universe
DM vs. baryon	$\Sigma_{\mathrm{c}}/\Sigma_{\mathrm{b}} = 8.5 \pm 1.5$	$\Sigma_{\mathrm{c}}/\Sigma_{\mathrm{b}} = 7.2 \pm 2.0$	$\rho_{\mathrm{c}}/\rho_{\mathrm{b}} = 5.36 \pm 0.06$

- **Hot Dark Matter (HDM)**: Relativistic particles at decoupling with large free-streaming lengths ($\lambda_{\mathrm{FS}} \gg 1\,\mathrm{Mpc}$, e.g., light standard neutrinos). Free-streaming erases primordial density perturbations on sub-horizon scales, producing a "top-down" structure formation scenario where superclusters fragment into galaxies. This is ruled out by the observed abundance of high-redshift galaxies and Lyman-α forest clustering.
- **Warm Dark Matter (WDM)**: Particles with intermediate velocities at matter-radiation

equality ($\lambda_{\rm FS} \sim 0.1\,\mathrm{Mpc}$, e.g., keV sterile neutrinos, gravitinos). WDM suppresses sub-galactic structure, potentially alleviating the "missing satellites" problem of pure CDM.

- **Cold Dark Matter (CDM)**: Non-relativistic particles with negligible velocity dispersion (free-streaming scale $\lambda_{\rm FS} \ll 1\,\mathrm{kpc}$, e.g., WIMPs, axions). CDM generates the canonical "bottom-up" hierarchical clustering paradigm, in spectacular agreement with CMB, BAO, and large-scale galaxy surveys.

High-energy particle candidates

To be cosmologically viable, a particle dark matter candidate must satisfy four fundamental criteria:

1. **Stability**: It must possess a lifetime much longer than the age of the Universe ($\tau \gg 1.4 \times 10^{10}\,\mathrm{yr}$).
2. **Color and Charge Neutrality**: It must be electrically neutral and a color singlet to evade electromagnetic and nuclear scattering bounds.
3. **Correct Relic Abundance**: Its cosmological production mechanism must naturally yield $\Omega_{\rm c0}h^2 \approx 0.12$.
4. **Collisionless Behavior**: Its self-interaction cross-section must obey $\sigma/m \lesssim 1\,\mathrm{cm}^2 \cdot \mathrm{g}^{-1}$.

Standard neutrinos

Neutrino oscillation experiments establish that neutrinos have non-zero masses with squared mass splittings:

$$\Delta m^2_{23} \sim 2.5 \times 10^{-3}\,\mathrm{eV}^2, \qquad \Delta m^2_{12} \sim 7.4 \times 10^{-5}\,\mathrm{eV}^2. \tag{8.2.24}$$

These require that at least one neutrino mass eigenstate satisfies $m_{\rm max} \geq \sqrt{\Delta m^2_{23}} \simeq 0.05\,\mathrm{eV}$.

Neutrinos decouple from the thermal primordial plasma when weak interaction rates drop below the Hubble expansion rate ($\Gamma_{\rm weak} \sim G_{\rm F}^2 T^5 \sim H \sim T^2/M_{\rm Pl}$), corresponding to $T_{\rm dec} \approx 1\,\mathrm{MeV}$. Because $T_{\rm dec} \gg m_\nu$, standard neutrinos decouple while ultra-relativistic, freezing out with a Fermi–Dirac distribution.
Subsequent to neutrino decoupling, e^+e^- annihilation at $T \sim 0.5\,\mathrm{MeV}$ dumps its entropy entirely into the photon bath. By conservation of comoving entropy, the present neutrino temperature is related to the photon temperature $T_\gamma = 2.7255\,\mathrm{K}$ by:

$$T_\nu = \left(\frac{4}{11}\right)^{1/3} T_\gamma \simeq 1.945\,\mathrm{K}. \tag{1}$$

The present number density per active neutrino flavor (neutrinos plus antineutrinos) is:

$$n_\nu = \frac{3}{4}\left(\frac{T_\nu}{T_\gamma}\right)^3 n_\gamma = \frac{3}{11} n_\gamma \simeq 112\,\mathrm{cm}^{-3}. \tag{2}$$

The total energy density in non-relativistic relic neutrinos today is $\rho_\nu = \sum m_\nu n_\nu$. Dividing by the critical density $\rho_{\rm crit} = 1.054 \times 10^{-5} h^2\,\mathrm{GeV} \cdot \mathrm{cm}^{-3}$:

$$\Omega_\nu h^2 = \frac{\sum m_\nu n_\nu}{\rho_{\rm crit}} = \frac{\sum m_\nu (112\,\mathrm{cm}^{-3})}{1.054 \times 10^{-5}\,\mathrm{GeV} \cdot \mathrm{cm}^{-3}} = \frac{\sum m_\nu}{94.1\,\mathrm{eV}}. \tag{3}$$

Enforcing the cosmological upper bound $\Omega_{\nu 0} h^2 \lesssim 0.007$ from cosmological structure formation gives the strict upper limit:

$$\sum_{i=1}^{3} m_{\nu i} \lesssim 0.12 - 0.7\,\mathrm{eV}. \tag{8.2.25}$$

Standard neutrinos can therefore account for at most $\approx 1\%$ of the total dark matter budget ($\Omega_\nu \ll \Omega_\mathrm{c}$), proving conclusively that the bulk of dark matter must consist of non-standard particles.

Other neutrino species

As fermions, neutrinos obey the Pauli exclusion principle, which imposes a strict maximum phase-space density. For neutrinos to account for the dark matter halos of dwarf spheroidal galaxies, the *Tremaine–Gunn bound* requires:

$$m_\nu \geq 120 \left(\frac{\sigma_v}{100\,\mathrm{km\cdot s^{-1}}}\right)^{-1/4} \left(\frac{r_\mathrm{c}}{1\,\mathrm{kpc}}\right)^{-1/4} \mathrm{eV}, \tag{8.2.26}$$

which directly contradicts the cosmological upper bound $\sum m_i \lesssim 0.7\,\mathrm{eV}$.

Proof of the Tremaine–Gunn Phase Space Bound. Fermionic phase-space density cannot exceed the degenerate Fermi limit:

$$f(\boldsymbol{x}, \boldsymbol{p}) \leq \frac{g_\nu}{(2\pi\hbar)^3}. \tag{1}$$

In a dwarf galaxy core with central density ρ_c and one-dimensional velocity dispersion σ_v, modeling the coarse-grained phase-space distribution as a Maxwellian gives the central phase density $f_0 \simeq \rho_\mathrm{c}/[m_\nu (2\pi\sigma_v^2 m_\nu^2)^{3/2}]$. Applying hydrostatic equilibrium with core radius r_c, $\rho_\mathrm{c} \approx 9\sigma_v^2/(4\pi G_\mathrm{N} r_\mathrm{c}^2)$. Setting $f_0 \leq g_\nu/(2\pi\hbar)^3$:

$$\frac{9\sigma_v^2}{4\pi G_\mathrm{N} r_\mathrm{c}^2 (2\pi)^{3/2} \sigma_v^3 m_\nu^4} \leq \frac{g_\nu}{(2\pi\hbar)^3} \implies m_\nu^4 \geq \frac{9\hbar^3}{4\sqrt{2\pi} G_\mathrm{N} g_\nu \sigma_v r_\mathrm{c}^2}. \tag{2}$$

Evaluating with $g_\nu = 2$ and normalizing to fiducial dwarf parameters ($\sigma_v = 100\,\mathrm{km\cdot s^{-1}}$, $r_\mathrm{c} = 1\,\mathrm{kpc}$) yields the Tremaine–Gunn bound (8.2.26).

Sterile neutrinos ν_R (singlet right-handed fermions with no electroweak gauge interactions) can mix with active neutrinos. For masses $M \sim$ keV, sterile neutrinos produced via non-resonant oscillations (Dodelson–Widrow) or resonant oscillations (Shi–Fuller) provide viable warm dark matter candidates.

Axions

The effective QCD Lagrangian includes the non-perturbative θ-vacuum term:

$$\mathcal{L}_\theta = \theta \frac{g_3^2}{32\pi^2} G^a_{\mu\nu} \tilde{G}^{a\,\mu\nu}, \tag{8.2.27}$$

which violates both $\hat{P}$ and $\hat{C}\hat{P}$. This term would induce an electric dipole moment for the neutron of order $d_\mathrm{n} \sim 10^{-16}\theta\, e\cdot\mathrm{cm}$. Experimental non-detection ($|d_\mathrm{n}| < 1.8 \times 10^{-26}\, e\cdot\mathrm{cm}$) requires $\theta \lesssim 10^{-10}$, a dilemma known as the *strong CP problem*.

Peccei and Quinn (1977) introduced a global chiral $U(1)_\mathrm{PQ}$ symmetry spontaneously broken at scale f_a. The pseudo-Nambu–Goldstone boson of this broken symmetry is the *axion* $a(x)$,

whose mass is determined by chiral perturbation theory:

$$m_a = \frac{m_\pi f_\pi}{f_a}\frac{\sqrt{m_u m_d}}{m_u + m_d} \simeq 0.62\left(\frac{10^7\,\mathrm{GeV}}{f_a}\right)\,\mathrm{eV}. \tag{8.2.28}$$

The Vacuum Misalignment Mechanism. In the early Universe, the axion field starts with an arbitrary initial misalignment angle $\theta_i \equiv a_i/f_a \in [-\pi, \pi]$ relative to the CP-conserving minimum. The evolution of the spatially homogeneous zero-mode field satisfies the damped non-linear pendulum equation:

$$\ddot{\theta} + 3H(T)\dot{\theta} + m_a^2(T)\sin\theta = 0. \tag{1}$$

At early times ($T \gg 1\,\mathrm{GeV}$), Hubble friction dominates ($3H \gg m_a(T)$), freezing the axion field at its initial displacement $\theta(t) \approx \theta_i$. As the Universe cools through the QCD phase transition, the instanton-induced mass $m_a(T)$ rises steeply until $m_a(T_{\mathrm{osc}}) \simeq 3H(T_{\mathrm{osc}})$ at $T_{\mathrm{osc}} \approx 1\,\mathrm{GeV}$. The field begins coherent oscillations about the potential minimum $\theta = 0$. Because the oscillations are coherent, the energy density behaves as a non-relativistic Bose–Einstein condensate ($\langle P\rangle = 0$, $\rho_a \propto a^{-3}$), acting as cold dark matter despite its minuscule mass ($m_a \sim 10^{-5}\,\mathrm{eV}$). The resulting cosmological relic abundance evaluates to:

$$\Omega_a h^2 \approx 0.12\left(\frac{f_a}{10^{12}\,\mathrm{GeV}}\right)^{7/6}\theta_i^2. \tag{2}$$

For $f_a \sim 10^{11} - 10^{12}\,\mathrm{GeV}$ ($m_a \sim 10 - 100\,\mu\mathrm{eV}$), axions naturally account for the entire cold dark matter budget of the Universe.

Candidates motivated by supersymmetry

In supersymmetric extensions of the Standard Model with conserved R-parity, $R_p = (-1)^{3(B-L)+2S}$, the lightest supersymmetric particle (LSP) is absolutely stable.

Consider a stable, weakly interacting massive particle (WIMP) χ of mass m_χ. In the early radiation-dominated Universe, χ particles are kept in thermal equilibrium with the Standard Model plasma via pair annihilation and creation ($\chi\bar{\chi} \leftrightarrow$ SM). The number density $n_\chi(t)$ obeys the relativistic Boltzmann equation:

$$\frac{\mathrm{d}n_\chi}{\mathrm{d}t} + 3Hn_\chi = -\langle\sigma_{\mathrm{ann}}v\rangle\left(n_\chi^2 - n_{\mathrm{eq}}^2\right), \tag{1}$$

where $\langle\sigma_{\mathrm{ann}}v\rangle$ is the thermally averaged annihilation cross-section, and n_{eq} is the equilibrium number density:

$$n_{\mathrm{eq}}(T) = g_\chi\left(\frac{m_\chi T}{2\pi}\right)^{3/2} e^{-m_\chi/T} \quad (T \ll m_\chi). \tag{2}$$

To eliminate the cosmological dilution term $3Hn_\chi$, we define the comoving abundance $Y \equiv n_\chi/s$ (where $s = \frac{2\pi^2}{45}g_{*S}T^3$ is entropy density) and the dimensionless inverse temperature $x \equiv m_\chi/T$. Using $\mathrm{d}t = \mathrm{d}x\,/(Hx)$ in the radiation era ($H(T) = \sqrt{\frac{8\pi^3 G_{\mathrm{N}} g_*}{90}}T^2 = H(m_\chi)x^{-2}$):

$$\frac{\mathrm{d}Y}{\mathrm{d}x} = -\frac{\lambda\langle\sigma_{\mathrm{ann}}v\rangle}{x^2}\left(Y^2 - Y_{\mathrm{eq}}^2\right), \qquad \lambda \equiv \frac{s(m_\chi)}{H(m_\chi)} = \sqrt{\frac{\pi g_*}{45}}M_{\mathrm{Pl}}m_\chi. \tag{3}$$

Equation (3) is the canonical *Lee–Weinberg equation.*

Freeze-out occurs when the annihilation rate drops below the expansion rate:

$$\Gamma(x_f) = n_{\rm eq}(x_f)\langle\sigma_{\rm ann}v\rangle \simeq H(x_f). \tag{4}$$

Solving iteratively for typical weak-scale parameters reveals that freeze-out occurs when the species is highly non-relativistic:

$$x_f \equiv \frac{m_\chi}{T_f} \approx 20 - 25. \tag{5}$$

For $x > x_f$, $Y^2 \gg Y_{\rm eq}^2$, and integrating Eq. (3) from x_f to $x = \infty$ yields the asymptotic relic abundance:

$$\frac{1}{Y_\infty} - \frac{1}{Y(x_f)} = \lambda \int_{x_f}^{\infty} \frac{\langle\sigma_{\rm ann}v\rangle}{x^2}\,{\rm d}x \approx \frac{\lambda\langle\sigma_{\rm ann}v\rangle}{x_f} \implies Y_\infty \approx \frac{x_f}{\lambda\langle\sigma_{\rm ann}v\rangle}. \tag{6}$$

Multiplying by $s_0 = 2891.2\,{\rm cm}^{-3}$ and dividing by $\rho_{\rm crit}$ gives the current cosmological density parameter:

$$\Omega_\chi h^2 = \frac{m_\chi Y_\infty s_0 h^2}{\rho_{\rm crit}} \approx \frac{3\times10^{-27}\,{\rm cm}^3\cdot{\rm s}^{-1}}{\langle\sigma_{\rm ann}v\rangle}. \tag{7}$$

The WIMP Miracle: Remarkably, a typical electroweak-scale cross-section:

$$\langle\sigma_{\rm ann}v\rangle \sim \frac{g_{\rm weak}^4}{16\pi^2 m_\chi^2} \sim 3\times10^{-26}\,{\rm cm}^3\cdot{\rm s}^{-1} \quad (m_\chi \sim 100\,{\rm GeV}), \tag{8.2.29}$$

substituted into Eq. (7) naturally yields:

$$\Omega_\chi h^2 \approx 0.1, \tag{8.2.30}$$

in stunning agreement with the observed cold dark matter density ($\Omega_{\rm c0}h^2 \simeq 0.12$). This striking numerical coincidence—that weak-scale particle physics naturally produces the cosmological dark matter abundance—is termed the *WIMP Miracle*.

In the Minimal Supersymmetric Standard Model (MSSM), the lightest neutralino χ_1^0 (a linear superposition of the Bino $\tilde{B}$, Wino $\tilde{W}_3$, and Higgsinos $\tilde{H}_{\rm d}^0, \tilde{H}_{\rm u}^0$) serves as the prototypical WIMP candidate:

$$\chi_1^0 = N_{11}\tilde{B} + N_{12}\tilde{W}_3 + N_{13}\tilde{H}_{\rm d}^0 + N_{14}\tilde{H}_{\rm u}^0. \tag{8.2.31}$$

WIMPZILLAs

If dark matter was never in thermal equilibrium at any epoch in cosmic history, its mass is completely decoupled from the partial-wave unitarity upper bound $m_\chi \lesssim 200\,{\rm TeV}$ (Griest & Kamionkowski 1990). Super-heavy dark matter particles termed *WIMPZILLAs* ($M_X \sim 10^{12} - 10^{16}\,{\rm GeV}$) can be generated non-adiabatically purely by the gravitational interaction—specifically, by the rapid, non-adiabatic variation of the background spacetime metric at the transition from cosmic inflation to reheating (Chung, Kolb, & Riotto 1998; Kolb, Starobinsky, & Tkachev 1999).

Gravitational particle production and Bogoliubov coefficients Consider a real, minimally coupled scalar field $X(\boldsymbol{x}, t)$ with mass M_X propagating in a spatially flat FLRW background metric ${\rm d}s^2 = a^2(\tau)[-{\rm d}\tau^2 + {\rm d}\boldsymbol{x}^2]$, where τ is conformal time (${\rm d}t = a\,{\rm d}\tau$). Rescaling the field as $\chi(\boldsymbol{x}, \tau) \equiv a(\tau)X(\boldsymbol{x}, \tau)$ and decomposing into Fourier spatial modes:

$$\chi(\boldsymbol{x}, \tau) = \int \frac{{\rm d}^3k}{(2\pi)^{3/2}} \left[a_{\boldsymbol{k}}\chi_k(\tau)e^{i\boldsymbol{k}\cdot\boldsymbol{x}} + a_{\boldsymbol{k}}^\dagger\chi_k^*(\tau)e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\right], \tag{8.2.32}$$

the mode functions $\chi_k(\tau)$ satisfy the exact equation of motion of a parametric harmonic oscillator with time-dependent frequency:

$$\frac{\mathrm{d}^2\chi_k}{\mathrm{d}\tau^2} + \omega_k^2(\tau)\chi_k = 0, \qquad \omega_k^2(\tau) \equiv k^2 + M_X^2 a^2(\tau) - \frac{1}{a}\frac{\mathrm{d}^2 a}{\mathrm{d}\tau^2}. \tag{8.2.33}$$

The term $\frac{1}{a}\frac{\mathrm{d}^2 a}{\mathrm{d}\tau^2} = \frac{1}{6}a^2 R = a^2(\dot{H} + 2H^2)$ acts as a time-dependent effective negative mass squared generated by spacetime curvature.

In the asymptotic inflationary past ($\tau \to -\infty$), spacetime expansion is approximately de Sitter ($a(\tau) \approx -1/(H_I\tau)$), and modes with physical wavelength well inside the horizon ($k/(aH_I) \gg 1$) satisfy the adiabatic condition $|\dot{\omega}_k/\omega_k^2| \ll 1$. They reside in the standard Bunch–Davies vacuum:

$$\chi_k^{(\mathrm{in})}(\tau) \to \frac{1}{\sqrt{2\omega_k(\tau)}} e^{-i\int^\tau \omega_k(\tau')\mathrm{d}\tau'}. \tag{8.2.34}$$

As inflation ends and the universe transitions into the inflaton oscillation or radiation-dominated era, the expansion rate changes violently. The effective frequency $\omega_k(\tau)$ undergoes non-adiabatic variation:

$$\left|\frac{1}{\omega_k^2}\frac{\mathrm{d}\omega_k}{\mathrm{d}\tau}\right| \gtrsim 1, \tag{8.2.35}$$

destroying the adiabatic invariant. In the post-inflationary asymptotic future ($\tau \to +\infty$), the mode functions must be re-expanded onto the late-time positive and negative frequency basis $\{\chi_k^{(\mathrm{out})}, \chi_k^{(\mathrm{out})*}\}$ via the Bogoliubov transformation:

$$\chi_k^{(\mathrm{in})}(\tau) = \alpha_k \chi_k^{(\mathrm{out})}(\tau) + \beta_k \chi_k^{(\mathrm{out})*}(\tau), \tag{8.2.36}$$

where the Bogoliubov coefficients α_k and β_k satisfy the canonical Wronskian normalization condition:

$$|\alpha_k|^2 - |\beta_k|^2 = 1. \tag{8.2.37}$$

The late-time vacuum state $|0_{\mathrm{in}}\rangle$ contains particles with mode occupation number:

$$N_k = \langle 0_{\mathrm{in}}| a_{\boldsymbol{k}}^{\dagger(\mathrm{out})} a_{\boldsymbol{k}}^{(\mathrm{out})} |0_{\mathrm{in}}\rangle = |\beta_k|^2. \tag{8.2.38}$$

Calculation of the physical number density n_X The total physical number density n_X of gravitationally created particles at the end of inflation is obtained by integrating the mode occupancy $|\beta_k|^2$ over all momentum modes:

$$n_X(t) = \frac{1}{a^3(t)}\int \frac{\mathrm{d}^3k}{(2\pi)^3}|\beta_k|^2 = \frac{1}{2\pi^2 a^3(t)}\int_0^\infty k^2|\beta_k|^2\,\mathrm{d}k\,. \tag{8.2.39}$$

The non-adiabatic excitation $|\beta_k|^2$ is efficient only for modes whose physical frequency is comparable to the rate of spacetime change: $k/a \sim H_I \sim M_X$. For ultra-high momenta ($k \gg aH_I$), modes remain adiabatic and particle creation is exponentially suppressed.

Evaluating the Bogoliubov coefficient integral analytically across the transition from de Sitter inflation to a matter-dominated inflaton oscillation phase reveals two distinct mass regimes (Chung et al. 1998):

1. *Moderately heavy regime ($M_X \lesssim H_I$):* Modes with $k \lesssim aH_I$ cross the horizon and experience maximal non-adiabatic amplification. The resulting comoving number density scales quadratically with mass:

$$n_X(t_e) \approx 0.03\left(\frac{M_X}{H_I}\right)^2 H_I^3 \qquad (M_X \lesssim H_I), \tag{8.2.40}$$

where t_e denotes the end of inflation.

2. *Super-heavy regime ($M_X > H_I$):* For masses exceeding the expansion rate, quantum fluctuations cannot be stretched beyond the horizon. The non-adiabatic excitation decays exponentially according to the classical Stokes line transition:

$$n_X(t_e) \approx 0.05 \left(\frac{M_X}{H_I}\right)^2 \exp\left(-2\pi \frac{M_X}{H_I}\right) H_I^3 \qquad (M_X \gtrsim H_I). \tag{8.2.41}$$

Remarkably, particle production peaks precisely when the mass is of the order of the Hubble rate during inflation: $M_X \approx (1-5)H_I$.

Cosmological Relic Density of WIMPZILLAs. Following particle creation at the end of inflation t_e, the universe undergoes reheating. If the inflaton decays at temperature $T_{\rm RH}$ with reheat expansion scale factor $a_{\rm RH}$, entropy density is generated:

$$s(T_{\rm RH}) = \frac{2\pi^2}{45} g_{*S}(T_{\rm RH}) T_{\rm RH}^3. \tag{1}$$

During the subsequent radiation and matter-dominated epochs, comoving WIMPZILLA number is strictly conserved ($n_X a^3 = \text{const}$) because the super-heavy mass prevents thermalization or annihilation ($\sigma_{\rm ann} \propto M_X^{-2} \lll 1$). The comoving yield today is:

$$Y_X \equiv \frac{n_X(T_{\rm RH})}{s(T_{\rm RH})} = \frac{n_X(t_e)}{s(T_{\rm RH})} \left(\frac{a_e}{a_{\rm RH}}\right)^3. \tag{2}$$

Relating the expansion ratio to the inflationary Hubble parameter H_I and the reheat temperature $T_{\rm RH}$, the present relic density evaluates to:

$$\Omega_X h^2 = \frac{M_X n_X(t_0) h^2}{\rho_{\rm crit}} \approx 0.1 \left(\frac{M_X}{H_I}\right)^{5/2} \left(\frac{H_I}{10^{13}\,\text{GeV}}\right)^2 \left(\frac{T_{\rm RH}}{10^9\,\text{GeV}}\right) \exp\left(-2\pi \frac{M_X}{H_I}\right). \tag{3}$$

For high-scale inflation ($H_I \sim 10^{13}\,\text{GeV}$) consistent with CMB tensor bounds ($r < 0.036$), choosing $M_X \approx (2-4)H_I \sim 10^{13} - 10^{14}\,\text{GeV}$ naturally reproduces the observed cold dark matter density $\Omega_{\rm c0} h^2 \simeq 0.12$ without fine-tuning.

Comparison with thermal freeze-out Figure 8.28 contrasts the gravitational production mechanism with thermal freeze-out. In thermal equilibrium, the maximum temperature achieved by the universe is the reheat temperature $T_{\rm RH} \leq (H_I M_{\rm Pl})^{1/2}$. If a particle has mass $M_X \gg T_{\rm RH}$, its thermal abundance is severely choked by the Boltzmann factor:

$$n_{\rm therm} \sim \left(\frac{M_X T_{\rm RH}}{2\pi}\right)^{3/2} \exp\left(-\frac{M_X}{T_{\rm RH}}\right). \tag{8.2.42}$$

For super-heavy particles with $M_X \sim 10^{13}\,\text{GeV}$ and conservative reheat temperatures $T_{\rm RH} \sim 10^9\,\text{GeV}$, the thermal ratio is $M_X/T_{\rm RH} \sim 10^4$, giving a suppression factor $\exp(-10^4) \equiv 0$. Thermal processes cannot create super-heavy dark matter.

Gravitational production, by contrast, operates directly at the energy scale of inflation H_I. Because $H_I \gg T_{\rm RH}$ during high-scale inflation, the non-adiabatic metric distortion excites quantum field modes directly out of the vacuum, easily generating the required relic density $n_X/H_I^3 \sim 10^{-2} - 10^{-4}$ without ever requiring the particle to enter thermal equilibrium.

Mode occupancy and phase-space momentum spectrum To appreciate the profound conceptual difference between gravitational particle creation and thermal production, contrast their microscopic operator expectation values directly.

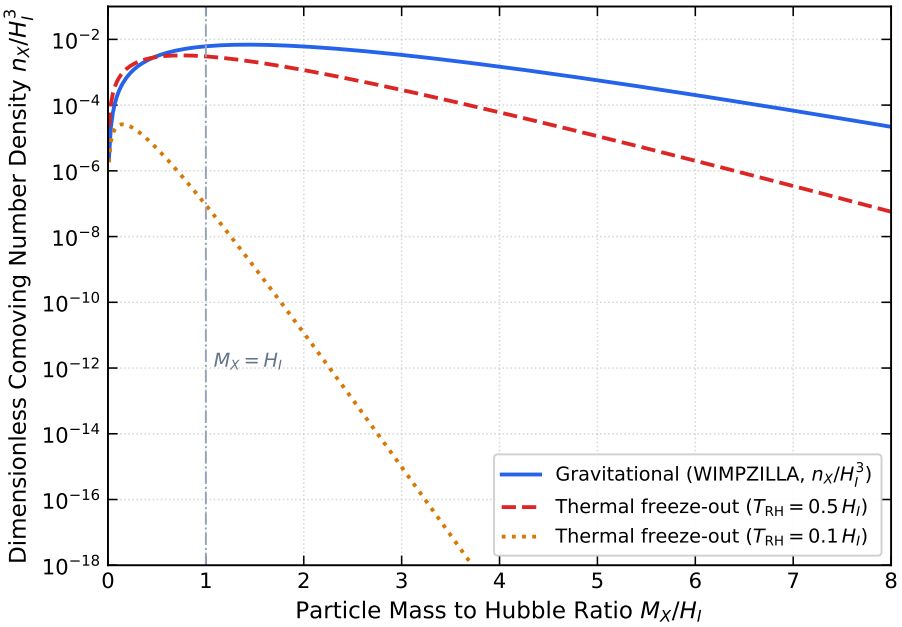


Figure 8.28: Dimensionless comoving number density n_X/H_I^3 as a function of the mass-to-Hubble ratio M_X/H_I. The solid curve shows non-thermal gravitational WIMPZILLA production, exhibiting a maximum near $M_X \approx H_I$ followed by the exponential Boltzmann-like suppression $\exp(-2\pi M_X/H_I)$. In sharp contrast, thermal production curves (dashed and dotted) freeze out at exponentially suppressed abundance $n_{\rm th} \propto \exp(-M_X/T_{\rm RH})$, vanishing entirely for $M_X \gg T_{\rm RH}$.

In non-thermal gravitational production, particles are excited purely out of the initial vacuum state $|0_{\rm in}\rangle$ by the time-dependent background metric. The expectation value of the late-time number operator $N_{\boldsymbol{k}} = a_{\boldsymbol{k}}^{\dagger(\rm out)} a_{\boldsymbol{k}}^{(\rm out)}$ is evaluated in the pure vacuum state:

$$\langle N_{\boldsymbol{k}} \rangle_{\rm grav} = \langle 0_{\rm in}| a_{\boldsymbol{k}}^{\dagger(\rm out)} a_{\boldsymbol{k}}^{(\rm out)} |0_{\rm in}\rangle = |\beta_k|^2. \tag{8.2.43}$$

In thermal equilibrium at inverse temperature $\beta \equiv 1/(k_{\rm B}T)$, the state is not a pure vacuum state, but a statistical mixed state described by the canonical density matrix $\hat{\rho} = \frac{1}{\mathcal{Z}} e^{-\beta \hat{H}}$, where $\mathcal{Z} \equiv \mathrm{Tr}\left(e^{-\beta \hat{H}}\right)$ is the partition function. The thermal distribution function $f_{\rm th}(k)$ is obtained by taking the thermal ensemble average:

$$f_{\rm th}(k) = \langle a_{\boldsymbol{k}}^\dagger a_{\boldsymbol{k}} \rangle_\beta = \mathrm{Tr}\left(\hat{\rho}\, a_{\boldsymbol{k}}^\dagger a_{\boldsymbol{k}}\right) = \frac{1}{\mathcal{Z}} \mathrm{Tr}\left(e^{-\beta \hat{H}} a_{\boldsymbol{k}}^\dagger a_{\boldsymbol{k}}\right) = \frac{1}{e^{\beta E_k} - 1}, \tag{8.2.44}$$

where $E_k \equiv \sqrt{k^2/a^2 + M_X^2}$ is the physical mode energy.

Figure 8.29 contrasts these two distributions at both the microscopic mode level and the physical phase-space level:

1. *Bare mode occupancy [Fig. 8.29(a)]:* Gravitational creation is an entirely coherent, unitary quantum mechanical transition ($|0_{\rm in}\rangle \to \sum c_n |n_{\rm out}\rangle$) preserving zero von Neumann entropy, whereas a thermal distribution represents maximum entropy mixed-state equilibrium. While the thermal occupancy diverges monotonically at long wavelengths as the classical Rayleigh–Jeans law $f_{\rm th}(k) \approx T/E_k \propto 1/k$, the gravitational occupancy $|\beta_k|^2$ remains finite and smoothly bounded as $k \to 0$.

2. *Phase-space differential spectrum [Fig. 8.29(b)]:* In three dimensions, the physical differential number density per momentum interval includes the spherical phase-space volume

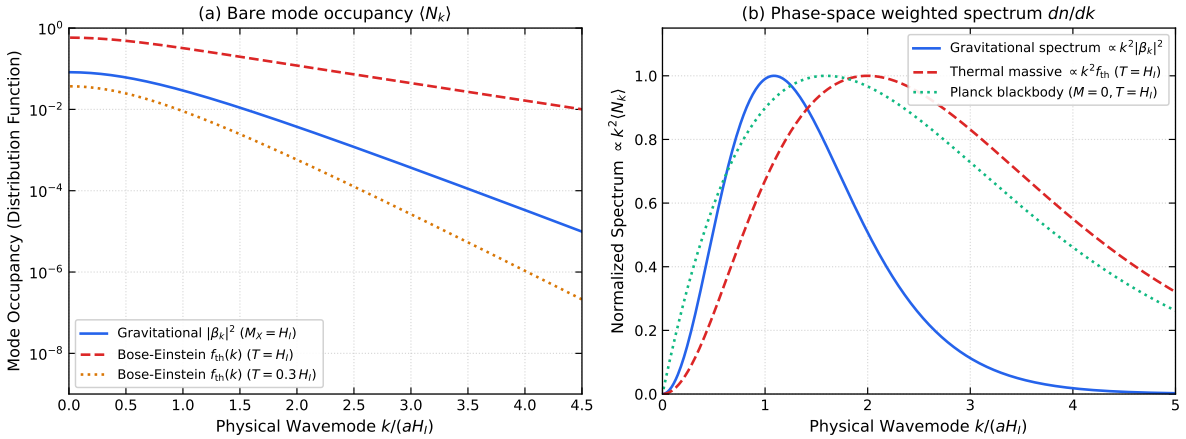


Figure 8.29: Microscopic comparison of non-thermal gravitational WIMPZILLA production ($M_X = H_I$) against thermal Bose–Einstein equilibrium ($T = H_I$ and $T = 0.3H_I$). (a) Bare mode occupancy per quantum state, $\langle N_k \rangle$. Gravitational $|\beta_k|^2$ is finite in the infrared without Rayleigh–Jeans divergence. (b) Normalized physical differential momentum spectrum $\mathrm{d}n\,/\,\mathrm{d}k \propto k^2\langle N_k \rangle$. Multiplying by the 3D phase-space volume element k^2 reveals the peaked spectral distribution: the gravitational spectrum peaks sharply near the horizon scale $k \approx 0.8\,aH_I$, whereas thermal spectra peak at higher momenta determined by T.

element:

$$\frac{\mathrm{d}n}{\mathrm{d}k} = \frac{1}{2\pi^2}k^2\langle N_{\boldsymbol{k}} \rangle. \tag{8.2.45}$$

Multiplying by k^2 suppresses the infrared, producing the characteristic peaked shape. For thermal blackbody radiation, the competition between phase-space density of states ($k^2 \to 0$) and Boltzmann suppression ($e^{-\beta k} \to 0$) produces the Wien-like Planck peak at $k \propto T$. In sharp contrast, the gravitational WIMPZILLA peak position has no dependence on temperature; it is governed solely by the horizon scale at the end of inflation ($k_{\mathrm{peak}} \approx 0.8\,a_e H_I$), confirming that non-thermal gravitational production creates a cold, horizon-peaked population of super-heavy dark relics.

Observational signatures: Ultra-High-Energy Cosmic Rays (UHECR) Although WIMPZILLAs are stable against standard cosmological decay on timescales exceeding the age of the universe ($\tau_X \gg t_0 \approx 13.8\,\mathrm{Gyr}$), non-perturbative effects (such as gravitational instantons, GUT-scale gauge interactions, or higher-dimensional operators suppressed by M_{Pl}) can mediate slow metastable decays:

$$X \to q\bar{q}, \ell\bar{\ell}, W^+W^-. \tag{8.2.46}$$

Because the parent mass is macroscopic ($M_X \sim 10^{12} - 10^{15}\,\mathrm{GeV}$), the primary decay products initiate QCD parton showers, fragmenting into ultra-energetic quarks, gluons, and pions. The neutral pions decay into ultra-high-energy photons ($\pi^0 \to \gamma\gamma$), while charged pions decay into neutrinos ($\pi^\pm \to \mu^\pm\nu \to e^\pm\nu\bar{\nu}\nu$).

The resulting photon and neutrino fragmentation spectrum exhibits a characteristic power-law tail:

$$\frac{\mathrm{d}N}{\mathrm{d}E} \propto E^{-1.9}, \tag{8.2.47}$$

extending well beyond the Greisen–Zatsepin–Kuzmin (GZK) cutoff ($E_{\mathrm{GZK}} \approx 5 \times 10^{19}\,\mathrm{eV}$). Observational data from the Pierre Auger Observatory, the Telescope Array, and the IceCube Neutrino Observatory place stringent lower bounds on the WIMPZILLA lifetime, requiring $\tau_X \gtrsim 10^{22}$ years for masses $M_X \sim 10^{13}\,\mathrm{GeV}$.

Other dark matter candidates

Beyond canonical thermal WIMPs, QCD axions, and gravitational WIMPZILLAs, the persistent non-detection of electroweak-scale dark matter in direct detection and collider experiments has motivated compelling alternative paradigms across dozens of orders of magnitude in mass:

Primordial Black Holes (PBHs) Unlike astrophysical black holes formed at late cosmic times by the core collapse of massive stars, Primordial Black Holes (PBHs) are formed in the very early universe—most prominently during the radiation-dominated era—by the direct gravitational collapse of large-amplitude primordial density fluctuations (Zel'dovich & Novikov 1967; Hawking 1971; Carr & Hawking 1974).

When an overdense fluctuation enters the cosmic cosmological horizon, it collapses if the density perturbation exceeds a threshold $\delta \geq \delta_c \approx 0.45$. The mass of the resulting PBH is approximately the horizon mass enclosed at horizon crossing t_H:

$$M_{\rm PBH} \approx \gamma M_H = \gamma \frac{c^3}{2G_{\rm N} H(t_H)} \approx 10^{15}\,{\rm g} \left(\frac{t_H}{10^{-23}\,{\rm s}}\right), \tag{8.2.48}$$

where $\gamma \approx (1/\sqrt{3})^3 \approx 0.2$ is the gravitational collapse efficiency factor. Depending on the epoch of horizon re-entry:

- Electroweak epoch ($t \sim 10^{-11}\,{\rm s}$): $M_{\rm PBH} \sim 10^{27}\,{\rm g} \sim 10^{-6}\,M_\odot$ (Earth mass).
- QCD phase transition ($t \sim 10^{-5}\,{\rm s}$): A softening of the equation of state ($w < 1/3$) reduces the collapse threshold, naturally producing solar-mass PBHs ($M_{\rm PBH} \sim 1\,M_\odot$).

Isolated black holes evaporate via quantum Hawking radiation with surface temperature:

$$T_H = \frac{\hbar c^3}{8\pi G_{\rm N} M k_{\rm B}} \approx 1.06 \times 10^{13}\,{\rm K} \left(\frac{10^{15}\,{\rm g}}{M}\right). \tag{8.2.49}$$

The evaporation lifetime of a black hole of mass M scales as:

$$\tau_{\rm evap} \approx \frac{5120\pi G_{\rm N}^2}{\hbar c^4} M^3 \approx 1.38 \times 10^{10}\,{\rm years} \left(\frac{M}{5 \times 10^{14}\,{\rm g}}\right)^3. \tag{8.2.50}$$

PBHs with initial mass $M < M_* \approx 5 \times 10^{14}\,{\rm g}$ have completely evaporated by the present cosmic epoch $t_0 \approx 13.8\,{\rm Gyr}$, producing isotropic gamma-ray background limits.

For stable PBHs ($M > 5 \times 10^{14}\,{\rm g}$), observational constraints from extragalactic gamma-rays, Subaru HSC microlensing of M31, Kepler/EROS microlensing, and CMB accretion anisotropies exclude PBHs as the dominant dark matter across most mass scales, with an open window remaining in the asteroid-mass regime ($10^{17}\,{\rm g} \leq M_{\rm PBH} \leq 10^{22}\,{\rm g}$, corresponding to $10^{-16} - 10^{-11}\,M_\odot$).

Self-Interacting Dark Matter (SIDM) Standard collisionless Cold Dark Matter (CDM) successfully explains large-scale cosmic structure ($k \lesssim 1\,h\,{\rm Mpc}^{-1}$), but faces persistent small-scale tensions in dwarf and low-surface-brightness galaxies:

1. *Core-cusp problem:* Dark matter-only cosmological simulations predict steep central density cusps ($\rho \propto r^{-1}$, as in NFW), whereas high-resolution rotation curves of dwarf galaxies frequently exhibit flat, constant-density cores ($\rho \propto r^0$).
2. *Diversity of rotation curves:* Dwarf galaxies with similar maximum rotation velocities display vastly disparate inner rotation curve slopes.

Spergel & Steinhardt (2000) proposed that dark matter particles possess a large elastic self-interaction cross-section $\sigma_{\rm SI}$, mediated by a light gauge boson or dark scalar. In the dense central regions of halos, frequent collisions ($\Gamma_{\rm coll} = \rho_{\rm DM}(\sigma/m)v_{\rm rms} \gtrsim H_0$) transport heat inward from the hotter outer halo, thermalizing the core into an isothermal, flat-density profile and turning cusps into cores.

To reconcile dwarf core formation without disrupting the elliptical morphology of galaxy clusters and escaping merging-cluster constraints (such as the Bullet Cluster, $\sigma/m < 1\,{\rm cm}^2/{\rm g}$), the self-interaction cross-section must be velocity-dependent:

$$\frac{\sigma_{\rm SI}}{m} \sim \begin{cases} 1\text{–}10\,{\rm cm}^2/{\rm g}, & (v \sim 20\text{–}50\,{\rm km\,s^{-1}},\ \text{dwarf galaxies}), \\ \lesssim 0.1\,{\rm cm}^2/{\rm g}, & (v \sim 1000\,{\rm km\,s^{-1}},\ \text{galaxy clusters}). \end{cases} \tag{8.2.51}$$

Such velocity-dependent cross-sections arise naturally through Rutherford-like or Yukawa scattering potentials in dark hidden sectors.

Fuzzy Dark Matter (ψDM) and Ultralight Bosons At the opposite extreme from WIMPZILLAs, dark matter can be composed of extremely light, non-thermal scalar bosons with mass:

$$m_\psi \sim 10^{-22}\text{–}10^{-21}\,{\rm eV}. \tag{8.2.52}$$

Because the mass is infinitesimal, the occupancy number per de Broglie volume is astronomical ($N \sim 10^{90}$), and the field behaves as a classical wave described by the coupled Schrödinger–Poisson system (Hu, Barkana, & Gruzinov 2000):

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m_\psi}\nabla^2\psi + m_\psi\Phi\psi, \qquad \nabla^2\Phi = 4\pi G_{\rm N} m_\psi |\psi|^2. \tag{8.2.53}$$

The reduced de Broglie wavelength of the particles on galactic scales spans astrophysical dimensions:

$$\lambda_{\rm dB} = \frac{\hbar}{m_\psi v} \approx 1.2\,{\rm kpc}\left(\frac{10^{-22}\,{\rm eV}}{m_\psi}\right)\left(\frac{100\,{\rm km\,s^{-1}}}{v}\right). \tag{8.2.54}$$

On scales larger than the de Broglie wavelength ($r \gg \lambda_{\rm dB}$), quantum wave interference averages out, and the system behaves exactly like standard Cold Dark Matter. On sub-kiloparsec scales ($r \lesssim \lambda_{\rm dB}$), however, Heisenberg's uncertainty principle exerts an effective repulsive quantum pressure:

$$P_Q = \frac{\hbar^2}{2m_\psi^2}\rho\nabla^2\sqrt{\rho}, \tag{8.2.55}$$

which halts gravitational collapse. This naturally stabilizes a smooth, cored gravitational soliton (a Bose–Einstein condensate core) at the center of every halo, suppressing structure formation below a sharp Jeans cutoff wavenumber $k_J \propto m_\psi^{1/2}$ and addressing the core-cusp problem without introducing particle self-interactions. Lyman-α forest observations of the neutral intergalactic medium place a lower bound $m_\psi \gtrsim 2 \times 10^{-21}\,{\rm eV}$.

Sterile Neutrinos (Warm Dark Matter) Standard model neutrinos are left-handed fermions with tiny Majorana masses generated by the seesaw mechanism. Adding right-handed, gauge-singlet Majorana fermions N_R introduces *sterile neutrinos*. If one of these states possesses mass in the keV regime ($m_s \sim 1$–$50\,{\rm keV}$) and a small active-sterile mixing angle θ, it serves as a viable Warm Dark Matter (WDM) candidate (Dodelson & Widrow 1994; Shi & Fuller 1999).

Because sterile neutrinos decouple while relativistic in the early universe, their free-streaming motion washes out primordial density perturbations below a characteristic free-streaming scale:

$$\lambda_{\rm FS} \approx 1\,{\rm Mpc}\left(\frac{1\,{\rm keV}}{m_s}\right)\left(\frac{T_\nu}{T_s}\right), \tag{8.2.56}$$

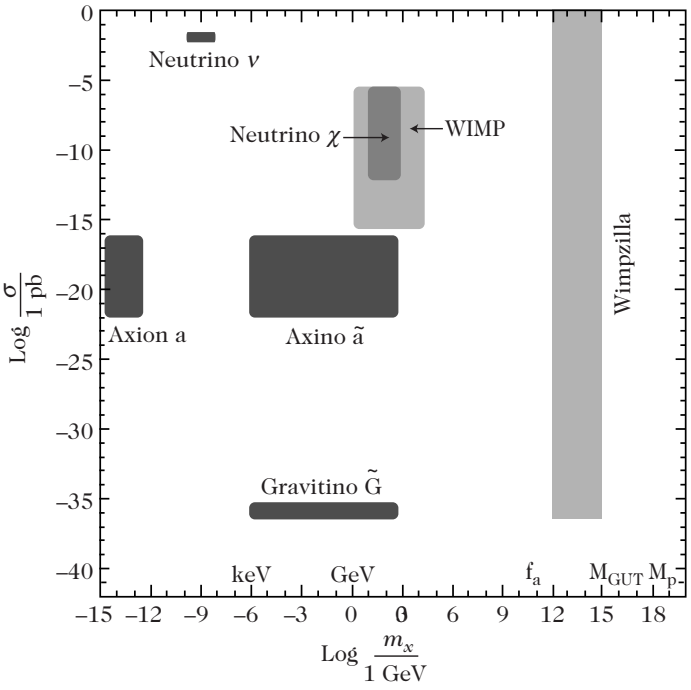


Figure 8.30: Parameter space of proposed dark matter candidates in the plane of interaction cross-section σ versus mass m_X. Thermal WIMPs populate the electroweak scale ($m_X \sim 10\,\text{GeV} - 1\,\text{TeV}$), ultralight axions occupy sub-eV scales ($m_X \sim 10^{-6} - 10^{-3}\,\text{eV}$), and super-heavy WIMPZILLAs span masses up to the grand unification scale ($M_{\text{GUT}} \sim 10^{16}\,\text{GeV}$). From Ref. [40].

suppressing dwarf-scale sub-halos and alleviating the "missing satellites" problem. Radiative decay of sterile neutrinos via loop-induced photon emission ($N \to \nu + \gamma$) generates a monochromatic X-ray emission line at energy $E_\gamma = m_s/2$:

$$\Gamma(N \to \nu\gamma) = \frac{9\alpha_{\text{EM}} G_{\text{F}}^2}{1024\pi^4} \sin^2(2\theta) m_s^5. \tag{8.2.57}$$

Searches for unidentified X-ray lines (such as the contested 3.55 keV feature reported in stacked cluster and galaxy spectra) provide direct laboratory and astrophysical probes of this sterile neutrino parameter space.

Direct searches

Direct detection experiments search for the elastic nuclear scattering of halo WIMPs traversing terrestrial detectors: $\chi + N \to \chi + N$. The nuclear recoil energy is:

$$E_R = \frac{\mu_{\chi N}^2 v^2}{M_N}(1 - \cos\theta^*), \tag{8.2.58}$$

where $\mu_{\chi N} \equiv \frac{m_\chi M_N}{m_\chi + M_N}$ is the reduced mass, and θ^* is the center-of-mass scattering angle. For typical halo velocities ($v \sim 220\,\text{km} \cdot \text{s}^{-1}$) and WIMP masses $m_\chi \sim 100\,\text{GeV}$, recoil energies are minuscule: $E_R \sim 1 - 50\,\text{keV}$.

Figure 8.30 charts the broad parameter space of proposed dark matter candidates across microscopic and macroscopic scales, while Fig. 8.31 displays the world-leading terrestrial exclusion limits on spin-independent WIMP-nucleon scattering.

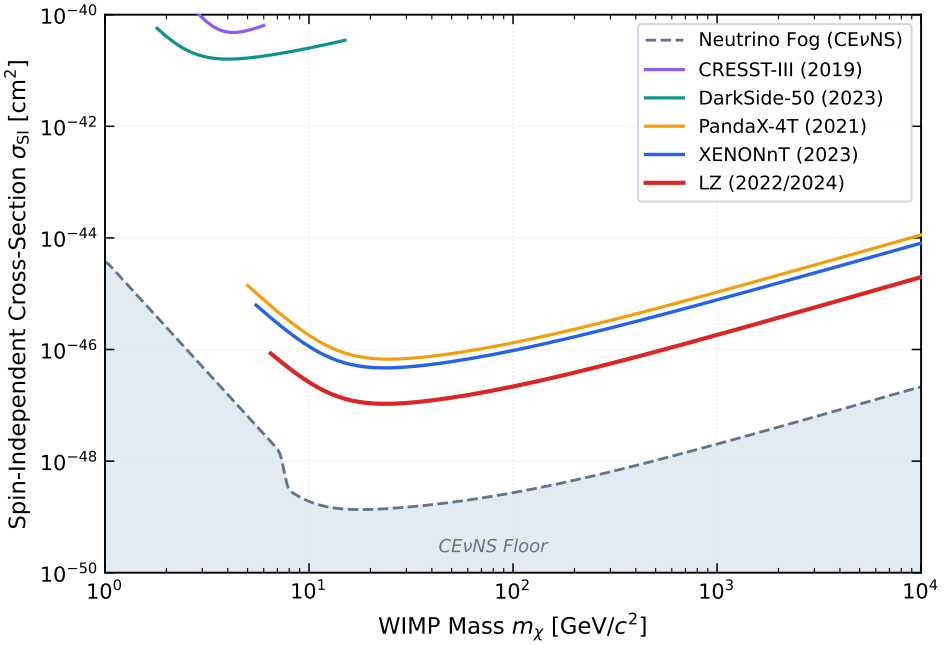


Figure 8.31: Current experimental upper limits (90% C.L.) on the spin-independent WIMP-nucleon cross-section $\sigma_{\rm SI}$ as a function of WIMP mass m_χ. Dual-phase xenon time projection chambers (LZ, XENONnT, PandaX-4T) exclude cross-sections above $\sigma_{\rm SI} \sim 6 \times 10^{-48}\,\mathrm{cm}^2$ near $m_\chi \approx 35\,\mathrm{GeV}/c^2$. Low-mass cryogenic calorimeters (CRESST-III) and liquid argon detectors (DarkSide-50) constrain sub-$10\,\mathrm{GeV}/c^2$ candidates. The shaded region denotes the coherent elastic neutrino-nucleus scattering (CEνNS) background floor (neutrino fog).

Gravitational alternatives: MOND

Formulation As an alternative to dark matter, Milgrom (1983) proposed *Modified Newtonian Dynamics* (MOND). MOND postulates that at accelerations below a fundamental cosmic threshold $a_0 \approx 1.2 \times 10^{-10}\,\mathrm{m\cdot s^{-2}} \approx cH_0/(2\pi)$, Newton's second law is modified according to:

$$\mu\left(\frac{a}{a_0}\right)\boldsymbol{a} = \boldsymbol{g}_{\rm N}, \tag{8.2.59}$$

where $\boldsymbol{g}_{\rm N} = -\boldsymbol{\nabla}\Phi_{\rm N}$ is the standard Newtonian gravitational field, and $\mu(x)$ is an interpolation function satisfying:

$$\mu(x) \approx \begin{cases} 1, & (x \gg 1,\ \text{Newtonian regime}), \\ x, & (x \ll 1,\ \text{deep-MOND regime}). \end{cases} \tag{8.2.60}$$

Derivation: Flat Rotation Curves and the Baryonic Tully–Fisher Relation in MOND. Consider a circular orbit around a baryonic point mass M at large distances where the acceleration falls into the deep-MOND regime ($a \ll a_0$). Setting $\mu(a/a_0) = a/a_0$ in Eq. (8.2.59):

$$\frac{a^2}{a_0} = g_{\rm N} = \frac{G_{\rm N}M}{r^2} \implies a = \sqrt{g_{\rm N}a_0} = \frac{\sqrt{G_{\rm N}Ma_0}}{r}. \tag{1}$$

Equating this to the centripetal acceleration $a = v^2/r$:

$$\frac{v^2}{r} = \frac{\sqrt{G_{\rm N}Ma_0}}{r} \implies v^4 = G_{\rm N}Ma_0. \tag{2}$$

Taking the fourth root:

$$v_{\rm flat} = (G_{\rm N}Ma_0)^{1/4} = \text{const}. \tag{3}$$

This derives:

1. **Asymptotically flat rotation curves**: $v(r)$ is strictly constant at large radii, with no free halo parameters!
2. **The Baryonic Tully–Fisher Relation (BTFR)**: $M \propto v_{\text{flat}}^4$ with an exact theoretical slope of 4.

The Four Fatal Failures of MOND Despite its success in fitting galaxy rotation curves without dark matter halos, MOND encounters four catastrophic, insurmountable failures on astronomical and cosmological scales:

1. **Failure on Galaxy Clusters (Residual Missing Mass)**: In rich galaxy clusters, the gravitational acceleration throughout the cluster core is relatively high, $a \sim a_0$. When MOND is applied to clusters, the dynamical mass predicted from the observed X-ray gas and galaxies is still a factor of $2-3$ smaller than the observed virial and lensing mass. To explain clusters, MOND must re-introduce unseen matter (such as 2 eV sterile neutrinos), defeating its primary purpose of eliminating dark matter.
2. **Failure on the Bullet Cluster (Spatial Decoupling of Mass and Baryons)**: As established in Section 8.2.2, the weak lensing mass peaks in the colliding cluster 1E 0657-558 are centered on the collisionless galaxies, separated by 8σ from the collisional X-ray gas where $> 90\%$ of the baryonic mass resides. Because MOND modifies gravity based on the distribution of baryons, MOND strictly predicts that the gravitational potential wells must center on the X-ray gas. The observation that lensing mass peaks reside in empty space with the galaxies constitutes a direct, model-independent empirical refutation of MOND.
3. **Failure on the CMB Third Acoustic Peak**: In standard ΛCDM cosmology, non-relativistic cold dark matter provides deep, static gravitational potential wells prior to recombination, driving the acoustic oscillations of the photon-baryon fluid and generating the observed high amplitude of the third acoustic peak. In MOND (and its relativistic completion TeVeS), the absence of cold collisionless matter during the acoustic era severely suppresses the driving force, predicting a third acoustic peak that is drastically smaller than observed by WMAP and Planck.
4. **Failure on GW170817 and the Speed of Gravity**: Relativistic completions of MOND (such as Bekenstein's TeVeS theory) require dynamical vector and scalar fields that couple disformally to the metric:

$$\tilde{g}_{\mu\nu} = e^{-2\phi} g_{\mu\nu} - 2\sinh(2\phi) U_\mu U_\nu. \tag{8.2.61}$$

In such disformal theories, gravitational waves propagate along geodesics of the metric $g_{\mu\nu}$, whereas light propagates along geodesics of the disformal physical metric $\tilde{g}_{\mu\nu}$, causing gravitational waves to travel at a different speed than light ($c_{\text{GW}} \neq c$).

On August 17, 2017, the gravitational wave signal from a binary neutron star merger (GW170817) and its electromagnetic gamma-ray burst counterpart (GRB 170817A) arrived within 1.7 seconds of each other after traveling over a distance of 130 million light years. This established the extraordinary bound:

$$-3 \times 10^{-15} \leq \frac{c_{\text{GW}} - c}{c} \leq +7 \times 10^{-16}. \tag{8.2.62}$$

This measurement definitively ruled out TeVeS and large classes of modified gravity theories that seek to reproduce MOND via disformal metric couplings (Boran et al. 2018).

Table 8.5: Summary of comparative predictions from MOND and Standard Cold Dark Matter (ΛCDM / SCDM) across various astrophysical and cosmological tests.

Test / predictions	MOND	ΛCDM
Rotation curves of spiral galaxies	Excellent	Good (requires feedback)
Baryonic Tully–Fisher relation	Natural ($M \propto v^4$)	Emergent
Galaxy cluster dynamics	Fails (factor 2 missing)	Concordant
Bullet Cluster spatial separation	Fatal failure	Concordant (8σ)
CMB acoustic peaks (1st, 2nd, 3rd)	Fails on 3rd peak	Precision concordance
Large-scale structure growth	Incomplete	Concordant
Gravitational light deflection	Disformal conflicts	Exact
GW170817 speed of gravity ($c_{\rm GW} = c$)	Ruled out (TeVeS)	Naturally exact

Dark matter vs. modified general relativity

The inescapable conclusion across all observational probes—from galactic rotation curves and cluster thermodynamics to weak gravitational lensing, CMB anisotropies, and multimessenger gravitational wave astronomy—is that the Universe is overwhelmingly dominated by a non-baryonic, collisionless cold dark matter component.

While modified gravity theories offer intriguing phenomenological insights into galactic scaling relations like the Baryonic Tully–Fisher relation, they fail catastrophically when confronted with multi-scale cosmological data, most decisively where gravitational lensing directly maps the spatial decoupling of collisionless dark matter from collisional baryonic gas. The discovery of the fundamental particle nature of dark matter remains one of the greatest and most profound frontiers in modern physics.

Chapter 9

Inflation

Inflation is a postulated period of accelerated expansion occurring before the radiation era of the standard Big-Bang model. It was initially proposed as a tentative solution of the problems of the hot Big-Bang model (Chapter 3). In particular, it provides a simple explanation for the homogeneity and flatness of our Universe.

Before the first models of inflation were proposed, precursor works appeared as early as 1965 by Gliner [1], who postulated a phase of exponential expansion. In 1978, Englert, Brout and Gunzig [2], in an attempt to resolve the primordial singularity problem and to introduce the particles and the entropy contained in the Universe, proposed a 'fireball' hypothesis, whereby the Universe itself would appear through a quantum effect in a state of negative pressure subject to a phase of exponential expansion.

Starobinsky [3], in 1979, used quantum-gravity ideas to formulate the first semi-realistic rigorous model of an inflation era, although he did not aim to solve the cosmological problems. A simpler model, with transparent physical motivations, was then proposed by Guth [4] in 1981; this model, now called 'old inflation', was the first to use inflation as a means of solving cosmological problems. It was soon followed by Linde's 'new inflation' proposal [5].

The formulation by Mukhanov and Chibisov of the theory of cosmological perturbations during new inflation, which links the origin of the large-scale structures of the Universe to quantum fluctuations, quickly followed [6, 7]. All these important works were achieved in the early 1980's, which can thus be considered as the 'date of birth' of inflation.

The first models of inflation were actually only incomplete modifications of the Big-Bang theory as they assumed that the Universe was in a state of thermal equilibrium, homogeneous on large enough scales before the period of inflation. This problem was resolved by Linde with the proposition of chaotic inflation [8]. In this model, inflation can start from a Planckian density even if the Universe is not in equilibrium. A new picture of the Universe then appears. The homogeneity and isotropy of our observable Universe would be only local properties, while the Universe is very inhomogeneous on very large scales, with a fractal-like structure.

This chapter describes the first models of inflation, Section 9.1, and explains in general terms how inflation solves the hot Big-Bang problems. We then address, in Section 9.2, the dynamics of single-field inflation and discuss the slow-roll approximation and the chaotic model of inflation. The generation of initial density perturbations during the inflationary epoch is detailed in Sections 9.3, 9.4 and 9.4.4. To finish, we describe the reheating phase in Section 9.5, which connects inflation to the hot Big-Bang model, the mechanism of eternal inflation, Section 9.6, which can address the question of initial conditions of the Universe, and various possible extensions, Section 9.7. Numerous complementary studies can be found in Refs. [9–11].

9.1 Genesis of the paradigm

9.1.1 Original motivations

Guth's argument

The flatness problem can be reformulated by noting that

$$\frac{\Omega_K}{\Omega} = -\frac{3K}{8\pi G_{\rm N}\rho a^2}, \tag{9.1.1}$$

so that (4.20) and (4.21) lead to the conclusion that

$$\left|\frac{\Omega_K}{\Omega}\right| \sim \frac{10^{43}}{S_0^{2/3}}\left(\frac{t}{1\ {\rm s}}\right) \sim \frac{10^{37}}{S_0^{2/3}}\left(\frac{1\ {\rm GeV}}{T}\right)^2, \tag{9.1.2}$$

where $S_0 = s_0 H_0^{-3}$ is the entropy contained in the Hubble radius today. From (4.140), it is of order $S_0 \sim 10^{87}$. Such a large value of S_0 implies that the Universe should have been very flat in the past. For instance, the curvature parameter must typically be of order $\Omega_K \sim 10^{-16}$ at the time of nucleosynthesis.

The flatness problem is therefore related to the fact that the entropy in a comoving volume is conserved. One can then imagine, as proposed by Guth [4], that it can be resolved if the cosmic expansion is non-adiabatic during a period of time $[t_i, t_f]$ during which

$$S_f = Z^3 S_i, \tag{9.1.3}$$

where Z is a numerical factor to be determined. In Guth's proposal, this entropy production occurs at around $T_{\rm GUT} \sim 10^{17}$ GeV, the characteristic scale of the Grand Unified Theory (GUT) symmetry breaking. From (9.1.2), this implies that Z should be of order $Z \sim 10^{28}$ in order for the curvature parameter to be of order unity today, $\Omega_{K0} \sim \mathcal{O}(1)$.

In a standard radiation-dominated Friedmann–Lemaître universe with conserved comoving entropy $S = sa^3 = $ const, the entropy density is $s = \frac{2\pi^2}{45} g_{*S} T^3$, so $a \propto g_{*S}^{-1/3} T^{-1}$. The Hubble parameter satisfies $H^2 = \frac{8\pi G_{\rm N}}{3}\rho = \frac{8\pi G_{\rm N}}{3}\left(\frac{\pi^2}{30} g_* T^4\right)$, giving $H \propto T^2/M_{\rm Pl}$. The curvature density parameter evolves as:

$$|\Omega_K| = \frac{|K|}{a^2 H^2} \propto \frac{T^2}{T^4/M_{\rm Pl}^2} = \frac{M_{\rm Pl}^2}{T^2}. \tag{1}$$

Relating the current Hubble volume entropy $S_0 = s_0 H_0^{-3} \sim 10^{87}$ to early times:

$$\left|\frac{\Omega_K}{\Omega}\right| \sim \frac{1}{|1-\Omega^{-1}|^{-1}} \sim \frac{10^{37}}{S_0^{2/3}}\left(\frac{1\ {\rm GeV}}{T}\right)^2 \sim 10^{-21}\left(\frac{1\ {\rm GeV}}{T}\right)^2. \tag{2}$$

At the GUT epoch ($T \sim 10^{16}$ GeV), $|\Omega_K| \lesssim 10^{-53}$, requiring extreme fine-tuning unless a non-adiabatic entropy generation $S_f = Z^3 S_i$ with $Z \sim 10^{28}$ dynamically dilutes the curvature term.

A phase of acceleration

The Friedmann equations (3.25) indicate that $(1-\Omega^{-1})\rho a^2 = \frac{3K}{8\pi G_{\rm N}}$ is constant during cosmic evolution. We infer that

$$(1-\Omega_i^{-1})\rho_i a_i^2 = (1-\Omega_f^{-1})\rho_f a_f^2. \tag{9.1.4}$$

Assuming that the subsequent evolution of the Universe ($t > t_f$) is described by the standard Big-Bang model, and is thus adiabatic, we conclude that

$$(1-\Omega_i^{-1})\rho_i a_i^2 \sim 10^{-56}(1-\Omega_0^{-1})\rho_f a_f^2. \tag{9.1.5}$$

The flatness problem can thus be resolved if $\rho_f a_f^2 \gg \rho_i a_i^2$. Moreover, since $H^2 a^2 - \frac{8\pi G_\mathrm{N}}{3}\rho a^2 = -K$ is constant in time, we have that

$$\dot{a}_f > \dot{a}_i. \tag{9.1.6}$$

So, a necessary condition for inflation is that the expansion of the Universe be accelerated during the period $[t_i, t_f]$, which implies that during this period the strong energy condition must be violated [see (3.26)]:

$$\ddot{a}(t) > 0 \iff \rho + 3P < 0. \tag{9.1.7}$$

9.1.2 Resolution of the Big-Bang problems

Inflation is therefore defined as a phase of accelerated expansion of the Universe. This property is sufficient to show that the Big-Bang problems can be resolved if this phase lasts sufficiently long.

To quantify the length of the inflationary period, we define the quantity

$$N \equiv \ln\left(\frac{a_f}{a_i}\right), \tag{9.1.8}$$

where a_i and a_f are the values of the scale factor at the beginning and at the end of inflation. This number measures the growth in the scale factor during the accelerating phase, expressed in units of the exponential of base 'e', which is the reason why it is usually called the number of 'e-folds'.[1]

Flatness problem

The dynamical analysis of the Friedmann–Lemaître equations (Chapter 3) has shown (3.54) that, when $\Omega_\Lambda = 0$,

$$\frac{\mathrm{d}\Omega_K}{\mathrm{d}\ln a} = (1+3w)\Omega_K(1-\Omega_K). \tag{9.1.9}$$

When $-1 < w < -1/3$, the stable fixed point of this dynamical system is on the $\Omega_K = 0$ axis (see Figs. 3.4–3.6). This attractor becomes unstable when w becomes larger than $-1/3$, that is in the matter and radiation eras. So, the flatness problem will be resolved if the attraction toward $\Omega_K = 0$ during the period of inflation is sufficient to compensate its subsequent drift during the hot Big-Bang, i.e. if inflation has lasted sufficiently long.

To give an estimate of the required minimum number of e-folds of inflation, note that, if we assume H to be almost constant during inflation then

$$\frac{\Omega_K(t_f)}{\Omega_K(t_i)} = \left(\frac{a_f}{a_i}\right)^{-2} = e^{-2N}. \tag{9.1.10}$$

In order to have $|\Omega_K(t_f)| \lesssim 10^{-60}$ and $\Omega_K(t_i) \sim \mathcal{O}(1)$, we thus need

$$N \gtrsim 70. \tag{9.1.11}$$

[1] Instead of the number of e-folds, the international nomenclature uses here the neper (symbol Np) which is used to express, for instance, the value of logarithmic quantities such as power ratios or acoustic pressure. The neper is consistent with the International System (IS) but has not yet been adopted by the General Conference as an IS unit. The neper is analogous to the decibel, with the convention that 1 Np = 20/(ln 10) dB = 8.686 dB, which expresses the use of a natural logarithmic basis rather than decimal for the dB. In what follows, we will use the term 'e-fold' as it is the one found in the literature.

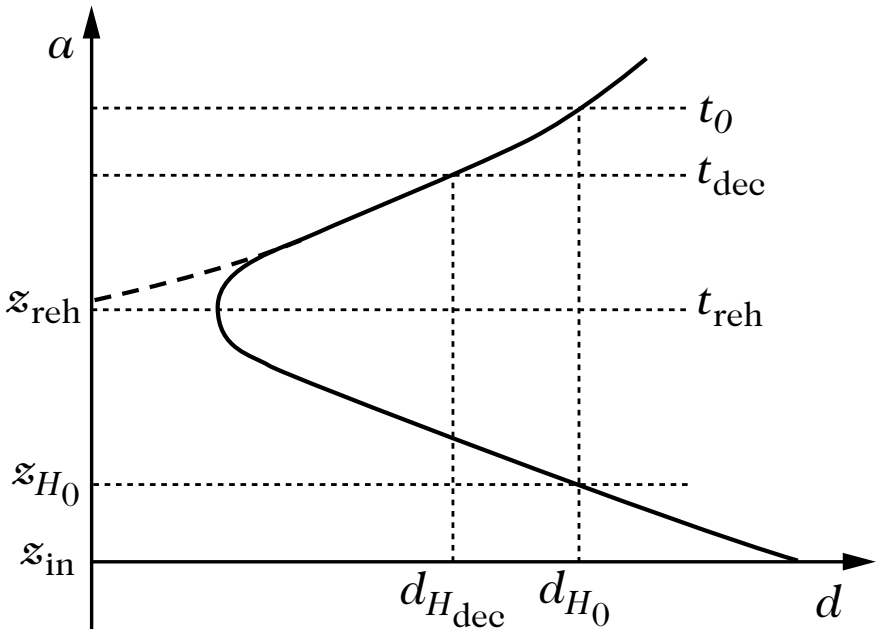


Figure 9.1: Evolution of the comoving Hubble radius as a function of the scale factor of the Universe assuming a period of inflation (plain line). A given comoving scale (vertical line) exits the Hubble radius during inflation and enters back into the Hubble radius during the hot Big-Bang. All observable scales can thus have been in causal contact at the beginning of inflation. This is not the case without inflation (dashed line) since the comoving Hubble radius is then strictly growing.

Horizon problem

The hypothesis (9.1.7) implies that the comoving Hubble radius, $\mathcal{H}^{-1} = (aH)^{-1}$, decreases in time:

$$\frac{\mathrm{d}}{\mathrm{d}t}(aH)^{-1} < 0. \tag{9.1.12}$$

Two points in causal contact at the beginning of inflation can thus be separated by a distance larger than the Hubble radius at the end of inflation (Fig. 9.1). These points are still causally connected, but can seem to be causally disconnected if the inflationary period is omitted. So, inflation allows for the entire observable Universe to emerge out of the same causal region before the onset of inflation.

For the horizon problem to be solved today, a causal region with scale factor a_i must have at least the size of the current observable Universe, which implies that

$$e^N \sim \frac{a_f D_H(t_0)}{a_0 D_H(t_f)} \sim \frac{T_0 D_H(t_0)}{T_f D_H(t_f)}. \tag{9.1.13}$$

Assuming that inflation ends at the Grand Unification scale (see Chapter 9), $T_f \sim 10^{16}$ GeV then $D_H(t_f) \sim M_{\rm Pl}/T_f^2 \sim 10^{-13}$ GeV^{-1}. Since $T_0 \sim 10^{-4}$ eV and $D_H(t_0) \sim H_0^{-1} \sim 10^{41}$ GeV^{-1}, we infer that

$$N \gtrsim 57. \tag{9.1.14}$$

This value can be smaller if the energy scale at the end of inflation is lower.

Conclusions

A primordial period of accelerated expansion allows for the resolution of the hot Big-Bang problems if it lasts sufficiently long. We have estimated that the number of e-folds should

typically be at least of the order

$$N \gtrsim 50 - 70, \tag{9.1.15}$$

depending on what the energy scale is at the end of inflation. It can be shown that with such a value, the other problems (monopoles, homogeneity, entropy) are also resolved.

It remains to build physical models leading to such an inflationary phase and to understand how it connects with the hot Big-Bang model.

The VEV in this model acquires a time dependence, providing a natural clock that can be used to measure time and distinguish the past from the future. One can picture the inflaton as a field rolling through its phase space, $\phi(x)$, toward the minimum of its potential.

$$\langle 0_{\rm BD}|\,\phi\,|0_{\rm BD}\rangle = \bar{\phi}(t) + \langle 0_{\rm BD}|\,\delta\hat{\phi}\,|0_{\rm BD}\rangle \tag{1}$$
$$= \bar{\phi}(t). \tag{2}$$

Here, we have chosen the fluctuation field such that its VEV vanishes,

$$\langle 0_{\rm BD}|\,\delta\hat{\phi}\,|0_{\rm BD}\rangle = 0.$$

This condition is preserved under the time evolution of the field. In terms of the time-dependent annihilation operator,

$$\hat{a}(t) = \alpha(t)\hat{a}_k(t=-\infty) + \beta(t)\hat{a}^\dagger(t=-\infty),$$

the time evolution can be described by a Bogoliubov transformation, while the field remains decomposed into a homogeneous background (commuting part) and fluctuations (non commuting part):

$$\phi(x) = \bar{\phi}(t) + \delta\hat{\phi}(x).$$

Note that any symmetry that is not respected by the VEV is said to be spontaneously broken. Since the background VEV $\bar{\phi}(t)$ depends explicitly on time, under a time translation,

$$t \to t + \Delta t,$$

the VEV transforms as

$$\bar{\phi}(t) \longrightarrow \bar{\phi}(t+\Delta t) \neq \bar{\phi}(t).$$

Therefore, the time-dependent inflaton background spontaneously breaks time-translation symmetry. In this sense, the rolling inflaton background provides a physical clock that distinguishes different times in the inflationary universe.

9.1.3 First models of inflation

Every inflationary model relies on one or several scalar fields, called 'inflaton'.

'Old' inflation

Models of old inflation rely on a first-order phase-transition mechanism (Fig. 9.2). A scalar field is trapped in a local minimum of its potential, thus imposing a constant energy density, $V(\phi_i)$, equivalent to the contribution of a cosmological constant. As long as the field remains in this configuration, the evolution of the Universe is exponential and the Universe can be described by a de Sitter space-time.

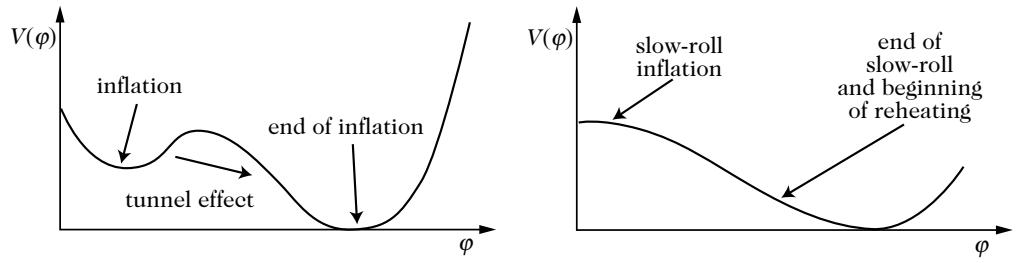


Figure 9.2: The model of old inflation (left) is based on a first order phase transition from a local to a true minimum of the potential. The false vacuum is metastable and the field can tunnel to the true vacuum. In the models of new inflation (right) a scalar field slowly relaxes towards its vacuum.

This configuration is metastable and the field can tunnel to its global minimum, $V(\phi_f) = 0$, hence creating bubbles of true vacuum, which correspond to a non-inflationary Universe (see Fig. 9.2).

One of the problems with these models arises from the properties of the de Sitter space (Section 9.1.4) that can be considered to be expanding, contracting or static space depending on the choice of coordinates (see below). In the metastable phase, there is a priori no preferred hypersurface and thus no preferred time-like direction to choose a slicing. Without such a direction, the phase transition can occur on any hypersurface. Each bubble of true vacuum will thus have very different physical properties, e.g. spatial curvature, so that the Universe is expected to be very inhomogeneous on large scales. This led to the abandonment of this model [12].

New inflation

In models of new inflation [5, 13], a scalar field, ϕ, exits its false vacuum by slowly rolling towards its true vacuum (see Fig. 9.2). This phase can be preceded by a phase of old inflation. The slow-roll phase is essential as it is during this period that density fluctuations, which lead to the currently observed large-scale structures, are generated.

This model can only work if the potential has a very flat plateau around $\phi = 0$, which is artificial. In most versions of this model, the inflaton cannot be in thermal equilibrium with other matter fields. The theory of cosmological phase transitions then does not apply.

No satisfying realization of this model has been proposed and it has been progressively abandoned.

9.1.4 Inflation as a de Sitter phase

In the limit of a frozen field, the inflationary phase is described by a de Sitter Universe. This space-time will be useful to understand and illustrate some mechanisms occurring during inflation. In realistic models of inflation, the dynamics of the inflaton induces a deviation of the space-time structure and that of a pure de Sitter space-time. As will also be seen, the limit where the field is completely frozen is usually singular.

Embedding in a five-dimensional space

The de Sitter space-time can be represented as a four-dimensional hyperboloid embedded in a five-dimensional Minkowski space $\mathbb{R}^{1,4}$ with metric $\eta_{AB} = \mathrm{diag}(-1, +1, +1, +1, +1)$:

$$-(z_0)^2 + (z_1)^2 + (z_2)^2 + (z_3)^2 + (z_4)^2 = H^{-2}. \tag{9.1.16}$$

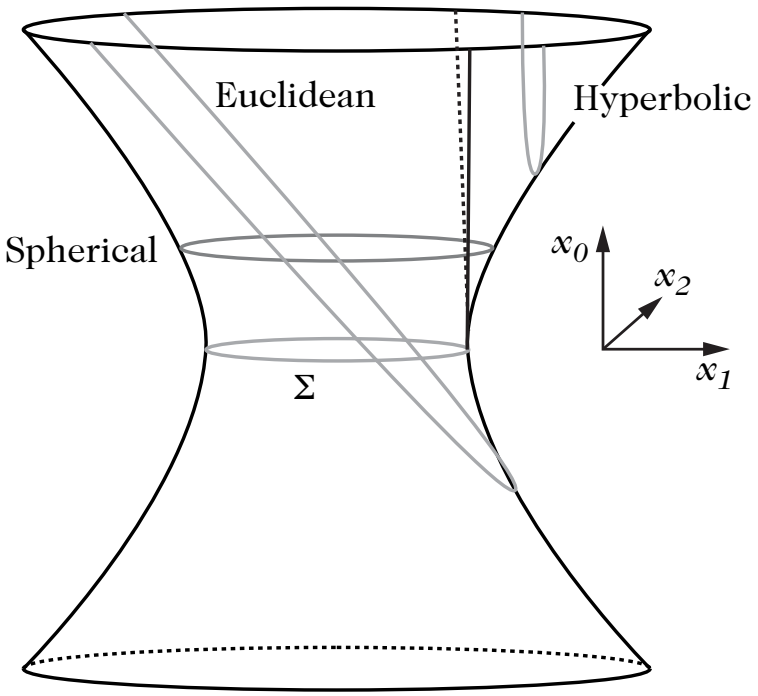


Figure 9.3: Depending on the choice of the constant-time hypersurfaces, the four-dimensional surface (9.1.16) can take various forms.

This surface is invariant under five-dimensional Lorentz transformations, so that the de Sitter space is invariant under the 10-parameter symmetry group $O(4,1)$. It is therefore a maximally symmetric four-dimensional space, just as for Minkowski space-time, with curvature

$$R = 12H^2 = \text{const.} \tag{9.1.17}$$

Depending on the choice of the constant-time hypersurfaces (see Fig. 9.3), the de Sitter space can take the form of a Friedmann–Lemaître space-time with flat, closed or open spatial sections, or a static form similar to the Schwarzschild space-time.

Euclidean slicing

The coordinate choice

$$z_0 = \frac{\sinh Ht}{H} + \frac{H}{2}e^{Ht}|\mathbf{x}|^2, \quad z_4 = \frac{\cosh Ht}{H} - \frac{H}{2}e^{Ht}|\mathbf{x}|^2, \quad z_i = e^{Ht}x_i \quad (i = 1,2,3), \tag{9.1.18}$$

where $|\mathbf{x}|^2 \equiv \delta_{ij}x^i x^j$, leads to the expression for the induced metric

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + e^{2Ht}\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j, \tag{9.1.19}$$

with Euclidean spatial sections (see Section 8.4.5 for details on the computation of an induced metric). In conformal time, this space takes the form

$$\mathrm{d}s^2 = \frac{1}{(H\eta)^2}\left(-\mathrm{d}\eta^2 + \delta_{ij}\mathrm{d}x^i\mathrm{d}x^j\right), \tag{9.1.20}$$

with

$$\eta = -\frac{1}{H}e^{-Ht}. \tag{9.1.21}$$

In this representation, η varies between $-\infty$ and 0^-, with the limit $\eta \to 0^-$ representing 'time-like infinity'. However, it only covers half of the hyperboloid (9.1.16).

We verify that the coordinates (9.1.18) satisfy the hyperboloid equation (9.1.16) and yield the metric (9.1.19):

$$z_4 - z_0 = \frac{\cosh Ht - \sinh Ht}{H} - He^{Ht}|\mathbf{x}|^2 = \frac{e^{-Ht}}{H} - He^{Ht}|\mathbf{x}|^2, \tag{1}$$

$$z_4 + z_0 = \frac{\cosh Ht + \sinh Ht}{H} = \frac{e^{Ht}}{H}. \tag{2}$$

Then the product is:

$$z_4^2 - z_0^2 = (z_4 - z_0)(z_4 + z_0) = \frac{1}{H^2} - e^{2Ht}|\mathbf{x}|^2. \tag{3}$$

Adding $\sum_{i=1}^3 z_i^2 = e^{2Ht}|\mathbf{x}|^2$, we find:

$$-(z_0)^2 + \sum_{i=1}^{3}(z_i)^2 + (z_4)^2 = z_4^2 - z_0^2 + e^{2Ht}|\mathbf{x}|^2 = \frac{1}{H^2}, \tag{4}$$

which identically satisfies the 5D embedding condition (9.1.16). The 5D Minkowski line element is $\mathrm{d}s^2 = -(\mathrm{d}z_0)^2 + \sum_{i=1}^3(\mathrm{d}z_i)^2 + (\mathrm{d}z_4)^2$. Taking differentials:

$$\mathrm{d}z_0 = \cosh Ht\,\mathrm{d}t + \frac{H^2}{2}e^{Ht}|\mathbf{x}|^2\mathrm{d}t + He^{Ht}\mathbf{x}\cdot\mathrm{d}\mathbf{x}, \tag{5}$$

$$\mathrm{d}z_4 = \sinh Ht\,\mathrm{d}t - \frac{H^2}{2}e^{Ht}|\mathbf{x}|^2\mathrm{d}t - He^{Ht}\mathbf{x}\cdot\mathrm{d}\mathbf{x}, \tag{6}$$

$$\mathrm{d}z_i = He^{Ht}x_i\,\mathrm{d}t + e^{Ht}\mathrm{d}x_i. \tag{7}$$

Computing $(\mathrm{d}z_4)^2 - (\mathrm{d}z_0)^2$:

$$\begin{aligned}(\mathrm{d}z_4 - \mathrm{d}z_0)(\mathrm{d}z_4 + \mathrm{d}z_0) &= \left[-e^{-Ht}\mathrm{d}t - 2He^{Ht}\mathbf{x}\cdot\mathrm{d}\mathbf{x} - H^2e^{Ht}|\mathbf{x}|^2\mathrm{d}t\right]\left[e^{Ht}\mathrm{d}t\right] \\ &= -\mathrm{d}t^2 - 2He^{2Ht}\mathbf{x}\cdot\mathrm{d}\mathbf{x}\,\mathrm{d}t - H^2e^{2Ht}|\mathbf{x}|^2\mathrm{d}t^2. \end{aligned} \tag{8}$$

Adding $\sum_{i=1}^3(\mathrm{d}z_i)^2 = H^2e^{2Ht}|\mathbf{x}|^2\mathrm{d}t^2 + 2He^{2Ht}\mathbf{x}\cdot\mathrm{d}\mathbf{x}\,\mathrm{d}t + e^{2Ht}\mathrm{d}\mathbf{x}^2$, the cross terms cancel exactly:

$$\boxed{\mathrm{d}s^2 = -\mathrm{d}t^2 + e^{2Ht}\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j}. \tag{9}$$

Spherical slicing

The choice of the system of coordinates $(t, \chi, \theta, \varphi)$ defined by

$$\begin{aligned} z_0 &= \frac{\sinh Ht}{H}, \\ z_1 &= \frac{\cosh Ht}{H}\cos\chi, \\ z_2 &= \frac{\cosh Ht}{H}\sin\chi\cos\theta, \\ z_3 &= \frac{\cosh Ht}{H}\sin\chi\sin\theta\cos\varphi, \\ z_4 &= \frac{\cosh Ht}{H}\sin\chi\sin\theta\sin\varphi, \end{aligned} \tag{9.1.22}$$

can be used to rewrite the metric in the form of a Friedmann–Lemaître space-time with spherical spatial sections,

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \frac{\cosh^2 Ht}{H^2}\left(\mathrm{d}\chi^2 + \sin^2\chi\,\mathrm{d}\Omega^2\right). \tag{9.1.23}$$

Unlike the Euclidean form (9.1.19), and the hyperbolic version, this slicing covers the entire de Sitter space-time and is thus geodesically complete.

Static slicing

To convince ourselves that de Sitter space-time is a static space-time, since it is maximally symmetric, let us derive its static form explicitly. With the choice of coordinates

$$\begin{aligned} z_0 &= \frac{\sinh Ht}{H}\sqrt{1-r^2H^2}, \\ z_1 &= \frac{\cosh Ht}{H}\sqrt{1-r^2H^2}, \\ z_2 &= r\cos\theta, \\ z_3 &= r\sin\theta\cos\varphi, \\ z_4 &= r\sin\theta\sin\varphi, \end{aligned} \tag{9.1.24}$$

where $0 \leq rH \leq 1$, the metric takes the Schwarzschild form

$$\mathrm{d}s^2 = -(1-r^2H^2)\mathrm{d}t^2 + \frac{\mathrm{d}r^2}{1-r^2H^2} + r^2\mathrm{d}\Omega^2. \tag{9.1.25}$$

For a static observer at fixed spatial coordinates $(r, \theta, \varphi = \text{const})$, the 4-velocity is normalized as $u^\mu u_\mu = g_{tt}(u^t)^2 = -(1-r^2H^2)(u^t)^2 = -1$, which gives:

$$u^\mu = \left(\frac{1}{\sqrt{1-r^2H^2}}, 0, 0, 0\right). \tag{1}$$

The 4-acceleration $a^\mu \equiv u^\nu\nabla_\nu u^\mu = \frac{\mathrm{d}u^\mu}{\mathrm{d}\tau} + \Gamma^\mu_{\alpha\beta}u^\alpha u^\beta = \Gamma^\mu_{tt}(u^t)^2$. Computing the Christoffel symbol Γ^r_{tt}:

$$\Gamma^r_{tt} = -\frac{1}{2}g^{rr}\partial_r g_{tt} = -\frac{1}{2}(1-r^2H^2)\partial_r\left(-(1-r^2H^2)\right) = -\frac{1}{2}(1-r^2H^2)(2rH^2) = -rH^2(1-r^2H^2). \tag{2}$$

Therefore, the 4-acceleration components are:

$$a^r = \Gamma^r_{tt}(u^t)^2 = -rH^2(1-r^2H^2)\frac{1}{1-r^2H^2} = -rH^2, \quad a^t = a^\theta = a^\varphi = 0, \tag{3}$$

$$\boxed{a^\mu = (0, -rH^2, 0, 0)}. \tag{4}$$

The proper magnitude of the acceleration is $|a| = \sqrt{g_{\mu\nu}a^\mu a^\nu} = \sqrt{g_{rr}(a^r)^2} = \frac{rH^2}{\sqrt{1-r^2H^2}}$. Thus, a static observer in de Sitter space is non-geodesic and experiences an outward accelerating force away from the origin $r = 0$.

9.2 Dynamics of single-field inflation

The first models of inflation introduced two ingredients shared by a fair majority of inflation models: the existence of a phase during which the Universe is 'close' to a de Sitter space-time, and a slowly rolling scalar field that rules the expansion of the Universe.

9.2.1 Equations of evolution

The simplest models of inflation are constructed from a single scalar field, ϕ, evolving in a potential $V(\phi)$ with an action given by

$$S = -\int \mathrm{d}^4x\sqrt{-g}\left[\frac{1}{2}\partial_\mu\phi\partial^\mu\phi + V(\phi)\right], \tag{9.2.1}$$

so that its energy-momentum tensor takes the form

$$T_{\mu\nu} = \partial_\mu\phi\partial_\nu\phi - g_{\mu\nu}\left(\frac{1}{2}\partial_\alpha\phi\partial^\alpha\phi + V(\phi)\right). \tag{9.2.2}$$

The energy density and pressure of a homogeneous scalar field are, respectively, given by

$$\rho_\phi = \frac{\phi'^2}{2a^2} + V(\phi), \quad P_\phi = \frac{\phi'^2}{2a^2} - V(\phi), \tag{9.2.3}$$

in conformal time, or by

$$\rho_\phi = \frac{\dot\phi^2}{2} + V(\phi), \quad P_\phi = \frac{\dot\phi^2}{2} - V(\phi), \tag{9.2.4}$$

in cosmic time. These expressions show that $\rho_\phi + 3P_\phi = 2(\dot\phi^2 - V)$. Since the Friedmann equations imply $\ddot a/a = -\frac{4\pi G_\mathrm{N}}{3}(\rho + 3P)$ (see Chapter 3), the Universe can enter a phase of accelerated expansion as soon as $\dot\phi^2 < V$. The expansion will be quasi-exponential if the scalar field is in a slow-roll regime, i.e. if $\dot\phi^2 \ll V$.

Equations in cosmic time

The Friedmann and Klein–Gordon equations in cosmic time take the form (see Chapter 3 for their derivation):

$$H^2 = \frac{8\pi G_\mathrm{N}}{3}\left(\frac{1}{2}\dot\phi^2 + V\right) - \frac{K}{a^2}, \tag{9.2.5}$$

$$\frac{\ddot a}{a} = \frac{8\pi G_\mathrm{N}}{3}\left(V - \dot\phi^2\right), \tag{9.2.6}$$

$$\ddot\phi + 3H\dot\phi + V_{,\phi} = 0. \tag{9.2.7}$$

Using $\dot H = \ddot a/a - H^2$, we obtain the relation

$$\dot H = -4\pi G_\mathrm{N}\dot\phi^2 + \frac{K}{a^2}, \tag{9.2.8}$$

which will be useful in what follows. As usual, only two of these equations are independent. Note that if we set $\dot H = 0$, then Eq. (9.2.8) for $K = 0$ implies that $\dot\phi = 0$. Hence, the inflaton can not roll down the potential.

The background dynamics of a single homogeneous scalar field in a spatially flat ($K = 0$) FLRW universe is governed by the coupled Friedmann and Klein–Gordon equations (9.2.5) and (9.2.7):

$$H^2 = \frac{8\pi G_\mathrm{N}}{3}\left(\frac{1}{2}\dot\phi^2 + V(\phi)\right) = \frac{1}{3M_\mathrm{Pl}^2}\left(\frac{1}{2}\dot\phi^2 + V(\phi)\right), \tag{1}$$

$$\ddot\phi + 3H\dot\phi + V_{,\phi} = 0. \tag{2}$$

For an expanding universe ($H > 0$), we can express the Hubble parameter algebraically in terms of the instantaneous state $(\phi, \dot{\phi})$:

$$H(\phi, \dot{\phi}) = \sqrt{\frac{8\pi G_{\rm N}}{3}\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right)} = \frac{1}{\sqrt{3}M_{\rm Pl}}\sqrt{\frac{1}{2}\dot{\phi}^2 + V(\phi)}. \tag{3}$$

Substituting this directly into the Klein–Gordon equation eliminates the scale factor $a(t)$ and the Hubble parameter $H(t)$ entirely, yielding a single, closed, autonomous second-order non-linear ordinary differential equation for $\phi(t)$:

$$\boxed{\ddot{\phi} + \sqrt{24\pi G_{\rm N}\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right)}\,\dot{\phi} + V'(\phi) = 0}, \tag{4}$$

or in reduced Planck units ($M_{\rm Pl} = (8\pi G_{\rm N})^{-1/2} = 1$):

$$\boxed{\ddot{\phi} + \sqrt{3\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right)}\,\dot{\phi} + V'(\phi) = 0}. \tag{5}$$

In the phase plane $(\phi, \dot{\phi})$, using $\ddot{\phi} = \dot{\phi}\frac{\mathrm{d}\dot{\phi}}{\mathrm{d}\phi}$, the trajectories satisfy the first-order differential equation:

$$\frac{\mathrm{d}\dot{\phi}}{\mathrm{d}\phi} = -\sqrt{24\pi G_{\rm N}\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right)} - \frac{V'(\phi)}{\dot{\phi}}. \tag{6}$$

This exact autonomous system governs the complete history of the scalar field:

1. **Slow-roll regime ($\dot{\phi}^2 \ll 2V(\phi)$, $|\ddot{\phi}| \ll |3H\dot{\phi}|$):** Eq. (6) reduces to the first-order slow-roll attractor trajectory $\dot{\phi} \simeq -\frac{V'(\phi)}{\sqrt{24\pi G_{\rm N} V(\phi)}}$.

2. **End of inflation ($\epsilon = 1$):** When ϕ drops below $\phi_{\rm end} \sim M_{\rm Pl}/\sqrt{4\pi}$, the acceleration condition $\ddot{a} > 0$ breaks down as the kinetic term grows to $\dot{\phi}^2 = V(\phi)$, passing the equation of state threshold $w_\phi = -1/3$.

3. **Oscillatory and reheating phase ($V(\phi) \approx \frac{1}{2}m^2\phi^2$):** Near the global minimum, the field undergoes damped harmonic oscillations with frequency $\omega \approx m \gg H$. Averaging over oscillation cycles gives $\langle\dot{\phi}^2\rangle = m^2\langle\phi^2\rangle = 2\langle V\rangle$, yielding zero cycle-averaged pressure $\langle P_\phi\rangle = 0$ and matter-like dust scaling $\rho_\phi \propto a^{-3}$.

In Listing 9.1, we provide a high-precision Python implementation using `scipy.integrate.solve_ivp` with the `DOP853` Runge–Kutta integrator to integrate the decoupled autonomous equation (4) and visualize the complete phase-space dynamics across slow roll and oscillatory reheating.

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.integrate import solve_ivp

# Configure matplotlib for publication quality
plt.rcParams.update({
    'font.size': 11,
    'axes.labelsize': 12,
    'axes.titlesize': 12,
    'xtick.labelsize': 10,
    'ytick.labelsize': 10,
    'legend.fontsize': 9.5,
    'figure.titlesize': 13,
```

```
    'lines.linewidth': 1.8,
    'font.family': 'serif',
})

def solve_inflaton_dynamics():
    """
    Solves the exact decoupled autonomous inflaton equation:
      ddot{phi} + sqrt(3*(0.5*dot{phi}^2 + V))*dot{phi} + V' = 0
    for chaotic inflation V(phi) = 1/2 * m^2 * phi^2.
    """
    m = 1.0  # Normalized inflaton mass

    # Autonomous ODE system [phi, dot{phi}, N_efolds]
    def ode_system(t, y):
        phi, dphi, N = y
        V = 0.5 * (m**2) * (phi**2)
        V_prime = (m**2) * phi
        rho = 0.5 * (dphi**2) + V
        H = np.sqrt(rho / 3.0)
        ddphi = -3.0 * H * dphi - V_prime
        dN = H
        return [dphi, ddphi, dN]

    # Initial conditions in slow-roll attractor at phi_0 = 15 M_Pl
    phi0 = 15.0
    V0 = 0.5 * (m**2) * (phi0**2)
    H0 = np.sqrt(V0 / 3.0)
    dphi0 = -(m**2 * phi0) / (3.0 * H0)
    N0 = 0.0

    # High-precision Runge-Kutta numerical integration (DOP853)
    t_span = (0.0, 500.0)
    t_eval = np.concatenate([
        np.linspace(0.0, 20.0, 2500),
        np.geomspace(20.0, 500.0, 5000)[1:]
    ])

    sol = solve_ivp(
        ode_system, t_span, [phi0, dphi0, N0],
        t_eval=t_eval, method='DOP853', rtol=1e-10, atol=1e-12
    )

    t = sol.t
    phi, dphi, N = sol.y[0], sol.y[1], sol.y[2]

    # Physical observables
    V = 0.5 * (m**2) * (phi**2)
    rho = 0.5 * (dphi**2) + V
    P = 0.5 * (dphi**2) - V
    w = P / rho

    # End of inflation: w = -1/3 (ddot{a} = 0, epsilon_H = 1)
    end_idx = np.where(w >= -1.0 / 3.0)[0][0]
    t_end = t[end_idx]
    phi_end = phi[end_idx]
    N_end = N[end_idx]

    # Filter for logarithmic time scale (t >= 0.1 m^{-1})
    mask = t >= 0.1
    t_log = t[mask]
    phi_log = phi[mask]
    w_log = w[mask]
    N_log = N[mask]
```

```
    rho_log = rho[mask]

    # Generate 4-panel figure
    fig, axs = plt.subplots(2, 2, figsize=(11, 8.5), dpi=300)

    # Panel (a): Inflaton evolution in log-time
    axs[0, 0].plot(t_log, phi_log, color='#003366', lw=1.6,
                   label=r'$\phi(t) / M_{\mathrm{Pl}}$')
    axs[0, 0].axvline(
        t_end, color='#cc0000', linestyle=':', lw=1.5,
        label=f'End of Inflation ($t_{{\\mathrm{{end}}}} \\approx {t_end:.1f}\\,m^{{-1}}$)'
    )
    axs[0, 0].axhline(0, color='gray', linestyle=':', alpha=0.5)
    axs[0, 0].set_xscale('log')
    axs[0, 0].set_xlabel(r'Cosmic time $m t$ (log scale)')
    axs[0, 0].set_ylabel(r'$\phi(t) / M_{\mathrm{Pl}}$')
    axs[0, 0].set_title(r'(a) Inflaton Evolution in Log-Time')
    axs[0, 0].grid(True, which='both', alpha=0.25)
    axs[0, 0].legend(loc='upper right')
    axs[0, 0].set_xlim(0.1, 500)
    axs[0, 0].set_ylim(-3.5, 16)

    # Panel (b): Effective equation of state w_phi(t) in log-time
    axs[0, 1].plot(t_log, w_log, color='#006633', lw=1.4,
                   label=r'$w_\phi(t) = P_\phi / \rho_\phi$')
    axs[0, 1].axhline(-1/3, color='#cc0000', linestyle='--',
                      label=r'$w = -1/3$ (Acceleration bound)')
    axs[0, 1].axhline(-1.0, color='gray', linestyle=':',
                      label=r'$w = -1$ (de Sitter vacuum)')
    axs[0, 1].axhline(0.0, color='#333333', linestyle='-.',
                      label=r'$\langle w_\phi \rangle = 0$ (Matter average)')
    axs[0, 1].axvline(t_end, color='#cc0000', linestyle=':', lw=1.5)
    axs[0, 1].set_xscale('log')
    axs[0, 1].set_xlabel(r'Cosmic time $m t$ (log scale)')
    axs[0, 1].set_ylabel(r'Equation of State $w_\phi$')
    axs[0, 1].set_title(r'(b) Equation of State $w_\phi(t)$ in Log-Time')
    axs[0, 1].grid(True, which='both', alpha=0.25)
    axs[0, 1].legend(loc='lower left')
    axs[0, 1].set_xlim(0.1, 500)
    axs[0, 1].set_ylim(-1.1, 1.1)

    # Panel (c): Phase portrait in (phi, dot{phi}) plane
    axs[1, 0].plot(phi, dphi / m, color='#660066', lw=1.5,
                   label='Phase Trajectory')
    axs[1, 0].plot(phi0, dphi0 / m, 'go', markersize=6,
                   label=r'Initial State ($\phi_0 = 15$)')
    axs[1, 0].plot(phi_end, dphi[end_idx] / m, 'ro', markersize=6,
                   label=r'End of Inflation ($\epsilon = 1$)')
    axs[1, 0].plot(0, 0, 'k*', markersize=8,
                   label=r'Attractor Origin $(0, 0)$')
    axs[1, 0].set_xlabel(r'$\phi / M_{\mathrm{Pl}}$')
    axs[1, 0].set_ylabel(r'$\dot{\phi} / (m M_{\mathrm{Pl}})$')
    axs[1, 0].set_title(r'(c) Phase Portrait in $(\phi, \dot{\phi})$ Plane')
    axs[1, 0].grid(True, alpha=0.3)
    axs[1, 0].legend(loc='lower right')

    # Panel (d): e-folds N(t) and energy density dilution in log-time
    color_n = '#b85d19'
    color_rho = '#006699'
    ax_d1 = axs[1, 1]
    ax_d2 = ax_d1.twinx()

```

```
    line1 = ax_d1.plot(t_log, N_log, color=color_n, lw=2.0,
                       label=r'e-folds $N(t)$')
    line2 = ax_d2.plot(
        t_log, np.log10(rho_log / rho[0]), color=color_rho,
        lw=1.8, linestyle='-', label=r'$\log_{10}(\rho_\phi / \rho_0)$'
    )

    ax_d1.axvline(t_end, color='#cc0000', linestyle=':', lw=1.5)
    ax_d1.set_xscale('log')
    ax_d2.set_xscale('log')
    ax_d1.set_xlabel(r'Cosmic time $m t$ (log scale)')
    ax_d1.set_ylabel(r'Number of e-folds $N(t)$', color=color_n)
    ax_d2.set_ylabel(r'$\log_{10}(\rho_\phi / \rho_0)$', color=color_rho)
    ax_d1.tick_params(axis='y', labelcolor=color_n)
    ax_d2.tick_params(axis='y', labelcolor=color_rho)

    lines = line1 + line2
    labels = [l.get_label() for l in lines]
    ax_d1.legend(lines, labels, loc='center left')
    ax_d1.set_title(r'(d) Expansion e-folds and Energy Density')
    ax_d1.grid(True, which='both', alpha=0.25)
    ax_d1.set_xlim(0.1, 500)

    plt.tight_layout()
    output_path = 'figures/inflaton_dynamics.pdf'
    plt.savefig(output_path, bbox_inches='tight')
    plt.close()
    print(f'Successfully generated {output_path}')

if __name__ == '__main__':
    solve_inflaton_dynamics()
```

Listing 9.1: High-precision Runge–Kutta numerical integration of the decoupled autonomous inflaton equation of motion and phase-space dynamics.

Equations in conformal time

These equations can be translated into conformal time. They then take the form

$$\mathcal{H}^2 = \frac{8\pi G_{\mathrm{N}}}{3}\left(\frac{1}{2}\phi'^2 + Va^2\right) - K, \tag{9.2.9}$$

$$\mathcal{H}' = -\frac{8\pi G_{\mathrm{N}}}{3}\left(\phi'^2 - Va^2\right), \tag{9.2.10}$$

for the Friedmann equations, and

$$\phi'' + 2\mathcal{H}\phi' + a^2 V_{,\phi} = 0, \tag{9.2.11}$$

for the Klein–Gordon equation.

Hamilton–Jacobi equations

As long as the evolution of the scalar field is monotonic, it can be useful to rewrite (9.2.5)–(9.2.7) using ϕ, instead of t or η, as a time variable. Assuming that $K = 0$, Eq. (9.2.8) takes the form

$$H_{,\phi} \equiv \frac{\mathrm{d}H(\phi)}{\mathrm{d}\phi} = -4\pi G_{\mathrm{N}}\dot{\phi}, \tag{9.2.12}$$

which defines the notation $H_{,\phi}$, so that Eq. (9.2.5) becomes

$$\left[H_{,\phi}(\phi)\right]^2 - 12\pi G_{\mathrm{N}} H^2(\phi) = -32\pi^2 G_{\mathrm{N}}^2 V(\phi). \tag{9.2.13}$$

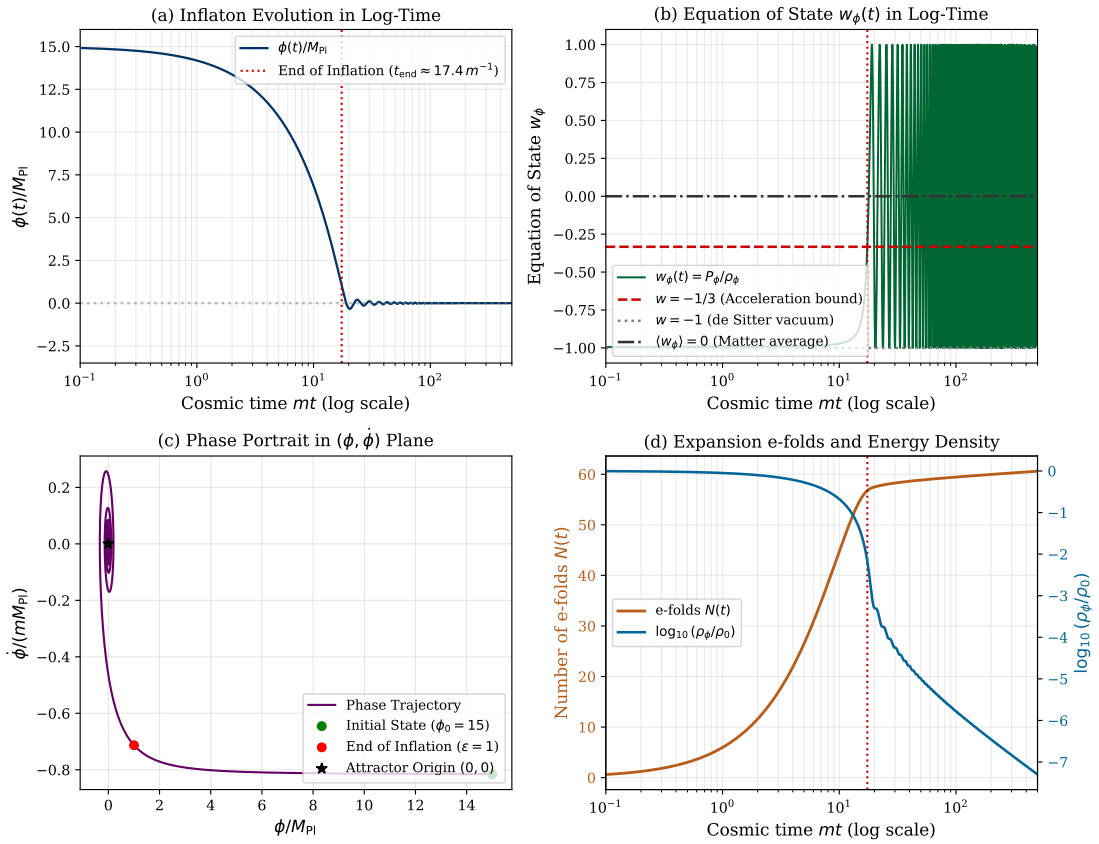


Figure 9.4: Numerical integration of the exact decoupled autonomous inflaton equation (4) for chaotic inflation $V(\phi) = \frac{1}{2}m^2\phi^2$ with $\phi_0 = 15M_{\rm Pl}$. **(a)** Inflaton field evolution $\phi(t)$ in logarithmic cosmic time $mt \in [0.1, 500]$, transitioning smoothly from slow roll ($N \approx 56.8$ e-folds) into rapid damped oscillations around the potential minimum. **(b)** Effective equation of state $w_\phi(t) = P_\phi/\rho_\phi$ in log-time, verifying $w_\phi \approx -1$ during inflation, crossing the cosmic acceleration bound $w = -1/3$ at $t_{\rm end} \approx 17.4\,m^{-1}$, and rapidly oscillating around the matter-dust cycle average $\langle w_\phi \rangle = 0$. **(c)** Phase portrait in the $(\phi, \dot\phi)$ plane illustrating attraction onto the slow-roll trajectory and spiral relaxation toward the origin limit cycle $(0, 0)$. **(d)** Cumulative expansion e-folds $N(t)$ and energy density dilution $\log_{10}(\rho_\phi/\rho_0)$ in log-time, exhibiting the inflationary slow-roll plateau followed by dust-like dilution $\rho_\phi \propto a^{-3} \propto t^{-2}$ during oscillatory reheating.

Both (9.2.12) and (9.2.13) are strictly equivalent to (9.2.5)–(9.2.8). This formalism is very efficient to obtain exact solutions of the dynamics since the choice of a function $H(\phi)$ directly determines the form of the required potential.

Consider the ADM action for a homogeneous scalar field in a flat FLRW universe $\mathrm{d}s^2 = -N^2(t)\mathrm{d}t^2 + a^2(t)\mathrm{d}\mathbf{x}^2$:

$$S = \int \mathrm{d}t \left[-\frac{3}{8\pi G_{\rm N}} \frac{a\dot{a}^2}{N} + \frac{a^3\dot{\phi}^2}{2N} - Na^3V(\phi) \right]. \tag{1}$$

In cosmic time gauge $N = 1$, the conjugate momenta are:

$$\pi_a = \frac{\partial L}{\partial \dot{a}} = -\frac{3}{4\pi G_{\rm N}} a\dot{a}, \quad \pi_\phi = \frac{\partial L}{\partial \dot{\phi}} = a^3\dot{\phi}. \tag{2}$$

The Hamiltonian constraint $\frac{\delta S}{\delta N} = 0$ yields:

$$\mathcal{H}_{\text{total}} = -\frac{2\pi G_{\text{N}}}{3}\frac{\pi_a^2}{a} + \frac{\pi_\phi^2}{2a^3} + a^3 V(\phi) = 0. \tag{3}$$

In Hamilton–Jacobi theory, we write $\pi_a = \frac{\partial S}{\partial a}$ and $\pi_\phi = \frac{\partial S}{\partial \phi}$. Using the separation ansatz $S(a,\phi) = -\frac{1}{4\pi G_{\text{N}}} a^3 H(\phi)$, we obtain:

$$\frac{\partial S}{\partial a} = -\frac{3}{4\pi G_{\text{N}}} a^2 H(\phi), \quad \frac{\partial S}{\partial \phi} = -\frac{1}{4\pi G_{\text{N}}} a^3 H_{,\phi}(\phi). \tag{4}$$

Equating $\pi_\phi = a^3\dot{\phi} = \frac{\partial S}{\partial \phi} = -\frac{1}{4\pi G_{\text{N}}} a^3 H_{,\phi}$ immediately gives:

$$\dot{\phi} = -\frac{1}{4\pi G_{\text{N}}} H_{,\phi}(\phi) \iff H_{,\phi} = -4\pi G_{\text{N}}\dot{\phi}. \tag{5}$$

Substituting $\frac{\partial S}{\partial a}$ and $\frac{\partial S}{\partial \phi}$ into the Hamiltonian constraint:

$$-\frac{2\pi G_{\text{N}}}{3a}\left(-\frac{3}{4\pi G_{\text{N}}} a^2 H\right)^2 + \frac{1}{2a^3}\left(-\frac{1}{4\pi G_{\text{N}}} a^3 H_{,\phi}\right)^2 + a^3 V(\phi) = 0, \tag{6}$$

$$a^3\left[-\frac{3}{8\pi G_{\text{N}}} H^2 + \frac{1}{32\pi^2 G_{\text{N}}^2}(H_{,\phi})^2 + V(\phi)\right] = 0, \tag{7}$$

which multiplying by $32\pi^2 G_{\text{N}}^2$ yields the exact Hamilton–Jacobi equation:

$$\boxed{[H_{,\phi}(\phi)]^2 - 12\pi G_{\text{N}} H^2(\phi) = -32\pi^2 G_{\text{N}}^2 V(\phi)}. \tag{8}$$

9.2.2 Slow-roll parameters

The slow-roll formalism is useful in reformulating most models of inflation that do not rely on first-order phase transitions (such as old inflation) in terms of the so-called slow-roll parameters. It makes it possible to compute the predictions of inflation without requiring a specification of the exact form of the inflaton's potential.

We assume that $K = 0$, since the curvature term decreases as a^{-2} and becomes negligible in comparison with the contribution from the scalar field, which has an almost constant energy density, if inflation lasts sufficiently long. This first implies that $\dot{H} < 0$ [see (9.2.8)], and thus H is always decreasing, and second that $|\dot{H}| < H^2$ if $\ddot{a} > 0$.

Principle

During inflation, the scalar field must be in a slow-roll regime. This implies that the scalar field varies little during the inflationary phase and thus satisfies

$$\dot{\phi}^2 \ll V, \quad |\ddot{\phi}| \ll |3H\dot{\phi}|. \tag{9.2.14}$$

From (9.2.4), the background inflaton can be approximated as a perfect fluid with

$$\rho_\phi \approx V(\phi), \qquad P_\phi \approx -V(\phi), \tag{1}$$

so that

$$w_\phi \equiv \frac{P_\phi}{\rho_\phi} \approx -1. \tag{2}$$

The adiabatic sound speed is therefore

$$c_s^2 \equiv \frac{\dot{P}_\phi}{\dot{\rho}_\phi} \approx \frac{-\dot{V}}{\dot{V}} = -1. \tag{3}$$

The pressure perturbation can be decomposed as

$$\delta P = c_s^2 \delta\rho + \delta P_{\rm nad}, \tag{4}$$

where $\delta P_{\rm nad}$ denotes the non-adiabatic pressure perturbation. Hence, if the inflaton perturbations are effectively adiabatic in the regime under consideration, we have

$$\delta P_{\rm nad} \approx 0, \qquad \delta P \approx -\delta\rho. \tag{5}$$

Under these conditions, the evolution equations (9.2.5) and (9.2.7) reduce to

$$H^2 \simeq \frac{8\pi G_{\rm N}}{3} V, \quad \dot{H} = -4\pi G_{\rm N}\dot{\phi}^2, \quad 3H\dot{\phi} \simeq -V_{,\phi}, \tag{9.2.15}$$

which implies (from (9.2.14))

$$\frac{|\dot{H}|}{H^2} = \frac{4\pi G_{\rm N}}{\frac{8\pi G_{\rm N}}{3}} \left(\frac{\dot{\phi}^2}{V}\right) \ll \frac{3}{2}. \tag{9.2.16}$$

Moreover, since

$$\ddot{\phi} = \frac{{\rm d}\dot{\phi}}{{\rm d}t} \simeq -\frac{\rm d}{{\rm d}t}\left(\frac{V_{,\phi}}{3H}\right) \simeq -\frac{V_{,\phi\phi}\dot{\phi}}{3H} \simeq \frac{V_{,\phi}V_{,\phi\phi}}{9H^2} \simeq \frac{1}{24\pi G_{\rm N}}\frac{V_{,\phi}V_{,\phi\phi}}{V}, \tag{9.2.17}$$

we infer that the two conditions (9.2.14) imply

$$\frac{\dot{\phi}^2}{V} = \frac{V_{,\phi}^2}{9H^2V} = \frac{V_{,\phi}^2}{9\frac{8\pi G_{\rm N}}{3}V^2} \implies \left(\frac{V_{,\phi}}{V}\right)^2 \ll 24\pi G_{\rm N}, \quad \frac{\ddot{\phi}}{3H\dot{\phi}} = \frac{\frac{V_{,\phi}V_{,\phi\phi}}{24\pi G_{\rm N}V}}{-V_{,\phi}} \implies \frac{|V_{,\phi\phi}|}{V} \ll 24\pi G_{\rm N}. \tag{9.2.18}$$

So, in order for inflation to occur, the potential should be very flat. In this slow-roll regime, we can expand (9.2.5) and (9.2.7) in terms of the derivatives of the expansion rate, H, or of the potential, V.

For slow-roll inflation, we neglect the acceleration term $\ddot{\phi}$ in the inflaton equation of motion. In contrast, during ultra-slow-roll inflation, we neglect the potential-gradient term $\partial_\phi V$, so that

$$\ddot{\phi} + 3H\dot{\phi} = 0. \tag{1}$$

Assuming H is approximately constant, this equation gives

$$\dot{\phi} \propto e^{-3Ht}, \qquad \ln\dot{\phi} = -3Ht + \text{constant}. \tag{2}$$

Using

$$\epsilon = \frac{4\pi G}{H^2}\dot{\phi}^2, \tag{3}$$

we therefore find

$$\epsilon \propto e^{-6Ht}. \tag{4}$$

Thus, while H remains nearly constant, the slow-roll parameter ϵ decreases exponentially during ultra-slow roll. In terms of the number of e-folds N, defined by $dN = Hdt$, this can be written as

$$\epsilon \propto e^{-6N}. \tag{5}$$

This rapid decrease of ϵ is the characteristic feature of the ultra-slow-roll regime.

Definitions

To formalize the previous argument, let us introduce the slow-roll parameters via the relations

$$\epsilon \equiv -\frac{\dot{H}}{H^2}, \tag{9.2.19}$$

$$\delta \equiv \epsilon - \frac{\dot{\epsilon}}{2H\epsilon} = \epsilon - \frac{1}{2}\frac{\mathrm{d}\ln\epsilon}{\mathrm{d}N}, \tag{9.2.20}$$

$$\xi \equiv \frac{\dot{\epsilon}-\dot{\delta}}{H}. \tag{9.2.21}$$

The parameter δ is sometimes called η in the literature, but we shall not use this notation, to avoid any confusion with the conformal time. The definitions (9.2.19)–(9.2.21) have the advantage of depending only on the space-time geometry, i.e. they are not restricted to the single scalar field case. Yet, in practice we shall almost always consider the case of a single scalar field. Using the equations of motion we find that the slow-roll parameters can be rewritten as

$$\epsilon = \frac{3}{2}\dot{\phi}^2\left[\frac{1}{2}\dot{\phi}^2 + V(\phi)\right]^{-1}, \tag{9.2.22}$$

$$\delta = -\frac{\ddot{\phi}}{H\dot{\phi}}. \tag{9.2.23}$$

From (9.2.22) and (9.2.4),

$$\epsilon = \frac{3}{2}\frac{\dot{\phi}^2}{\rho_\phi}, \qquad \rho_\phi = \frac{1}{2}\dot{\phi}^2 + V. \tag{1}$$

Taking a time derivative and using the Friedmann equation $\dot{H} = -4\pi G\dot{\phi}^2$, we have

$$\begin{aligned}\dot{\epsilon} &= 3\frac{\dot{\phi}\ddot{\phi}}{\rho_\phi} - \frac{3}{2}\frac{\dot{\phi}^2\dot{\rho}_\phi}{\rho_\phi^2} = -3H\delta\frac{\dot{\phi}^2}{\rho_\phi} + \frac{9H}{2}\frac{\dot{\phi}^4}{\rho_\phi^2} \\ &= -3H\delta\left(\frac{2}{3}\epsilon\right) + \frac{9H}{2}\left(\frac{2}{3}\epsilon\right)^2 = -2H\epsilon\delta + 2H\epsilon^2 \qquad (2)\\ &= 2H\epsilon(\epsilon-\delta). \qquad (3)\end{aligned}$$

Therefore,

$$\boxed{\delta = \epsilon - \frac{\dot{\epsilon}}{2H\epsilon}}. \tag{4}$$

The Friedmann equations imply that the parameter ϵ can be expressed as

$$\epsilon = \frac{1}{4\pi G_{\rm N}} \left[\frac{H'(\phi)}{H(\phi)}\right]^2, \tag{9.2.24}$$

with the notation $H' \equiv \mathrm{d}H/\mathrm{d}\phi$. This indicates that ϵ is related to the small parameter that appeared in the condition (9.2.16). It can be used to rewrite the two Friedmann equations as

$$H^2\left(1 - \frac{1}{3}\epsilon\right) = \frac{8\pi G_{\rm N}}{3}V, \quad \frac{\ddot{a}}{a} = H^2(1-\epsilon), \tag{9.2.25}$$

and the effective equation of state of the inflaton as

$$w_\phi = -1 + \frac{2}{3}\epsilon. \tag{9.2.26}$$

From (9.2.4)

$$\rho_\phi + P_\phi = \dot{\phi}^2. \tag{1}$$

Using $P_\phi = w_\phi \rho_\phi$, we obtain

$$\dot{\phi}^2 = (1 + w_\phi)\rho_\phi. \tag{2}$$

On the other hand, the first slow-roll parameter is

$$\epsilon = -\frac{\dot{H}}{H^2} = \frac{3}{2}\frac{\rho_\phi + P_\phi}{\rho_\phi} = \frac{3}{2}\frac{\dot{\phi}^2}{\rho_\phi}. \tag{3}$$

Hence,

$$\epsilon = \frac{3}{2}(1 + w_\phi), \tag{4}$$

which gives

$$w_\phi = -1 + \frac{2}{3}\epsilon. \tag{5}$$

The condition for inflation therefore reduces to

$$\ddot{a} > 0 \iff w < -\frac{1}{3} \iff \epsilon < 1. \tag{9.2.27}$$

The three parameters (9.2.19)–(9.2.21) are related via the equations of motion as

$$\frac{\dot{\epsilon}}{H} = 2\epsilon(\epsilon - \delta), \quad \frac{\dot{\delta}}{H} = 2\epsilon(\epsilon - \delta) - \xi. \tag{9.2.28}$$

The slow-roll conditions are thus satisfied as long as ϵ and δ are small compared to 1 and $\xi = \mathcal{O}(\epsilon^2, \delta^2, \epsilon\delta)$. The parameter δ also takes the form

$$\delta = \frac{1}{4\pi G_{\rm N}}\frac{H_{,\phi\phi}}{H}. \tag{9.2.29}$$

It is also useful to express these parameters in terms of conformal time as

$$\epsilon = 1 - \frac{\mathcal{H}'}{\mathcal{H}^2}, \quad \delta = 1 - \frac{\phi''}{\mathcal{H}\phi'} = \epsilon - \frac{\epsilon'}{2\mathcal{H}\epsilon}. \tag{9.2.30}$$

Expansion of the potential and of the expansion rate

The slow-roll approximation can be seen as a Taylor expansion of the potential V or of the expansion rate H in terms of their derivatives with respect to ϕ. Two sets of slow-roll parameters can be introduced, based either on the potential V and its derivatives,

$$\epsilon_V \equiv \frac{1}{16\pi G_\mathrm{N}}\left(\frac{V_{,\phi}}{V}\right)^2, \quad \eta_V \equiv \frac{1}{8\pi G_\mathrm{N}}\frac{V_{,\phi\phi}}{V}, \quad \xi_V^2 \equiv \frac{1}{8\pi G_\mathrm{N}}\frac{V_{,\phi}V_{,\phi\phi\phi}}{V^2}, \tag{9.2.31}$$

or on the expansion rate H,

$$\epsilon_H \equiv \frac{1}{4\pi G_\mathrm{N}}\left(\frac{H_{,\phi}}{H}\right)^2, \quad \delta_H \equiv \frac{1}{4\pi G_\mathrm{N}}\left(\frac{H_{,\phi\phi}}{H}\right), \quad \xi_H^2 \equiv \frac{1}{4\pi G_\mathrm{N}}\frac{H_{,\phi}H_{,\phi\phi\phi}}{H^2}. \tag{9.2.32}$$

Both these sets of parameters appear in the literature. They are not independent and one can show, using the Hamilton–Jacobi equation (9.2.13), that

$$\epsilon = \epsilon_V = \epsilon_H, \quad \delta = \delta_H = \eta_V - \epsilon_V, \quad \xi^2 = \xi_H^2 - 3\epsilon\delta + 2\epsilon^2, \quad \xi_H^2 = \xi_V^2 - 3\epsilon\eta_V + 3\epsilon^2. \tag{9.2.33}$$

From the exact Hamilton–Jacobi equation:

$$H_{,\phi}^2 - 12\pi G_\mathrm{N}H^2 = -32\pi^2 G_\mathrm{N}^2 V(\phi) \implies V(\phi) = \frac{3}{8\pi G_\mathrm{N}}H^2\left(1 - \frac{1}{3}\epsilon_H\right), \tag{1}$$

where $\epsilon_H \equiv \frac{1}{4\pi G_\mathrm{N}}\left(\frac{H_{,\phi}}{H}\right)^2$. Differentiating with respect to ϕ:

$$V_{,\phi} = \frac{3}{8\pi G_\mathrm{N}}\left[2HH_{,\phi}\left(1 - \frac{1}{3}\epsilon_H\right) - \frac{1}{3}H^2\epsilon_{H,\phi}\right]. \tag{2}$$

Since $\epsilon_{H,\phi} = \frac{1}{4\pi G_\mathrm{N}}\frac{\mathrm{d}}{\mathrm{d}\phi}\left(\frac{H_{,\phi}^2}{H^2}\right) = \frac{2}{4\pi G_\mathrm{N}}\frac{H_{,\phi}}{H}\left(\frac{H_{,\phi\phi}}{H} - \frac{H_{,\phi}^2}{H^2}\right) = 2\sqrt{4\pi G_\mathrm{N}\epsilon_H}\,(\delta_H - \epsilon_H)$, we have:

$$\frac{V_{,\phi}}{V} = \frac{2\frac{H_{,\phi}}{H}\left(1 - \frac{1}{3}\epsilon_H\right) - \frac{2}{3}\frac{H_{,\phi}}{H}(\delta_H - \epsilon_H)}{1 - \frac{1}{3}\epsilon_H} = \frac{H_{,\phi}}{H}\left[\frac{2 - \frac{2}{3}\delta_H}{1 - \frac{1}{3}\epsilon_H}\right] = 2\frac{H_{,\phi}}{H}\left[\frac{1 - \frac{1}{3}\delta_H}{1 - \frac{1}{3}\epsilon_H}\right]. \tag{3}$$

Squaring and multiplying by $\frac{1}{16\pi G_\mathrm{N}}$:

$$\epsilon_V = \frac{1}{16\pi G_\mathrm{N}}\left(\frac{V_{,\phi}}{V}\right)^2 = \frac{1}{4\pi G_\mathrm{N}}\left(\frac{H_{,\phi}}{H}\right)^2\left[\frac{1 - \frac{1}{3}\delta_H}{1 - \frac{1}{3}\epsilon_H}\right]^2 = \epsilon_H\left[\frac{1 - \frac{1}{3}\delta_H}{1 - \frac{1}{3}\epsilon_H}\right]^2 \simeq \epsilon_H + \mathcal{O}(\text{second order}). \tag{4}$$

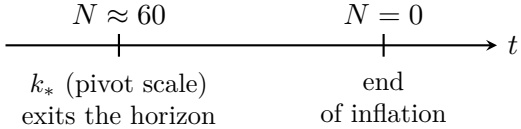
Similarly, taking the second derivative $V_{,\phi\phi}$:

$$\eta_V \equiv \frac{1}{8\pi G_\mathrm{N}}\frac{V_{,\phi\phi}}{V} = \epsilon_H + \delta_H + \mathcal{O}(\epsilon_H^2, \delta_H^2) \implies \delta_H = \eta_V - \epsilon_V. \tag{5}$$

Thus, to leading order in the slow-roll expansion, the potential and Hubble slow-roll parameters coincide according to (9.2.33).

Duration of inflation and number of e-folds

The expansion between an arbitrary time t and the end of inflation can be related to



$$N(t, t_f) = \int_t^{t_f} H\,\mathrm{d}t. \tag{9.2.34}$$

since, after integration,

$$a(t) = a_f e^{-N}, \tag{9.2.35}$$

where N, the number of e-folds, characterizes the amount of inflation between a given time and the end of the phase of accelerated expansion [$N(t_i, t_f)$ corresponds to the quantity (9.1.8)]. One can express N in terms of the inflaton as

$$N(\phi) \equiv N(\phi, \phi_f) = \int_\phi^{\phi_f} \frac{H}{\dot\phi} \mathrm{d}\phi = -4\pi G_\mathrm{N} \int_\phi^{\phi_f} \frac{H}{H'} \mathrm{d}\phi = -\sqrt{4\pi G_\mathrm{N}} \int_\phi^{\phi_f} \frac{\mathrm{d}\phi}{\sqrt{\epsilon}}, \tag{9.2.36}$$

where the explicit reference to the end of inflation ϕ_f in N is dropped when there is no ambiguity.

During inflation, a mode k becomes super-Hubble at the time t_k defined by

$$ck = a(t_k)H(t_k) = a_\mathrm{f} e^{-N} H(t_k). \tag{9.2.37}$$

It is therefore useful to relate k to the value of the inflaton at t_k by

$$k(\phi) = a_f H(\phi) \exp[-N(\phi)], \tag{9.2.38}$$

so that the speed at which scales become super-Hubble is characterized by

$$\frac{\mathrm{d}\ln k}{\mathrm{d}\phi} = \sqrt{\frac{4\pi G_\mathrm{N}}{\epsilon}}(\epsilon - 1). \tag{9.2.39}$$

Similarly, one can easily show that the relation $k = a_k H_k$ implies

$$\frac{\mathrm{d}\ln H_k}{\mathrm{d}\ln k} = -\epsilon. \tag{9.2.40}$$

The first expression points out that if $H' = 0$ (and thus $\epsilon = 0$), i.e. for a pure de Sitter phase, $\mathrm{d}\ln k/\mathrm{d}\phi = 0$ and the process of Hubble radius crossing is scale invariant.

Sensitivity to initial conditions

The Klein–Gordon equation, being a second-order differential equation, has two independent solutions. In the slow-roll regime, however, it becomes a first-order differential equation, with only one solution. One of the solutions has thus been neglected. In the Hamilton–Jacobi formalism, the solution of (9.2.13) depends on the choice of the initial condition for the scalar field. In order for the results we obtain to be robust, and for inflation to be predictive, the late times solution $H(\phi)$ should be independent of the choice of initial conditions [14].

To see this, let us assume that $H_0(\phi)$ is the slow-roll solution and let us consider a small perturbation with respect to this solution, $H(\phi) = H_0 + \delta H$. By linearizing (9.2.13), we find that the perturbation δH is a solution of $H_0' \delta H' \simeq 3H_0 \delta H / 12\pi G_\mathrm{N}$ and takes the general form

$$\delta H = \delta H(\phi_i) \exp\left\{12\pi G_\mathrm{N} \int_{\phi_i}^{\phi} \frac{H_0(\phi)}{H_0'(\phi)} \mathrm{d}\phi\right\}, \tag{9.2.41}$$

where $\delta H(\phi_i)$ represents a different choice of initial condition. The previous integral can be expressed in terms of the number of e-folds (9.2.36) as

$$\delta H = \delta H(\phi_i) \exp\left\{-3\left[N(\phi_i) - N(\phi)\right]\right\}. \tag{9.2.42}$$

The effect of the choice of initial conditions is therefore exponentially erased and has no observable effect as soon as a few e-folds of inflation have occurred before the scales of cosmological interest become super-Hubble.

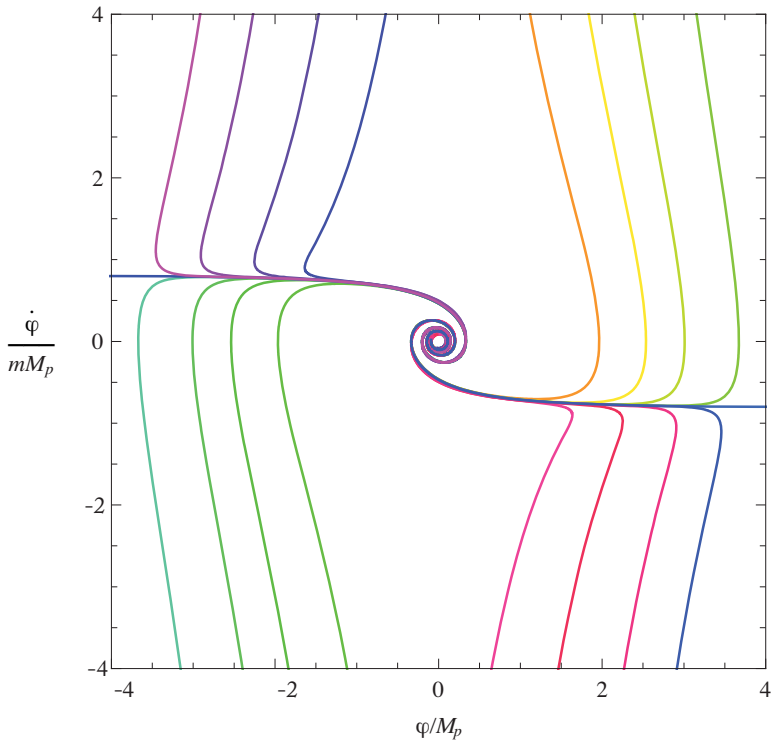


Figure 9.5: Phase portrait in the $(\phi, \dot{\phi})$-plane of the dynamics of a scalar field with a potential $V = \frac{1}{2}m^2\phi^2$, assuming $m = 10^{-6}M_{\rm Pl}$. This illustrates the mechanism of attraction toward the slow-roll solution.

Another way to illustrate that the slow-rolling solution is an attractor of the dynamics is to depict it as a phase portrait in the $(\phi, \dot{\phi})$-plane, starting from arbitrary initial conditions. Fig. 9.5 gives an example of such a phase portrait for a potential $V = \frac{1}{2}m^2\phi^2$. The spiralling around the origin describes the dynamics at the end of the inflationary period and is relevant for the reheating that will be discussed later.

We also see in the picture that when the inflationary period is short, there is a possibility that the actual solution has not converged toward the inflationary attractor at the time the modes of observational interest are becoming super-Hubble. In such a case the predictions of inflation will be discussed in terms of trajectories.

9.2.3 End of inflation

The end of inflation can be identified with the end of the slow-roll regime. We can therefore define the value of the field at the end of inflation, ϕ_f, by

$$\max(\epsilon, |\delta|) \simeq 1. \tag{9.2.43}$$

At the end of inflation, all classical inhomogeneities have been exponentially washed out, and one can consider them as non-existent at this stage. The Universe has become very flat so that curvature terms can be neglected and one can set $K = 0$ during all the primordial phase (see Section 9.8 for a discussion of the effect of spatial curvature). Moreover, all the entropy has been diluted. If the inflaton potential has a minimum, the scalar field will oscillate around this minimum right after the end of inflation (see Fig. 9.5).

Due to the Hubble expansion, these oscillations are damped and the scalar field decays into a large number of particles. During this phase, the inflationary Universe (of low entropy and dominated by the coherent oscillations of the inflaton) becomes a hot Universe (of high entropy

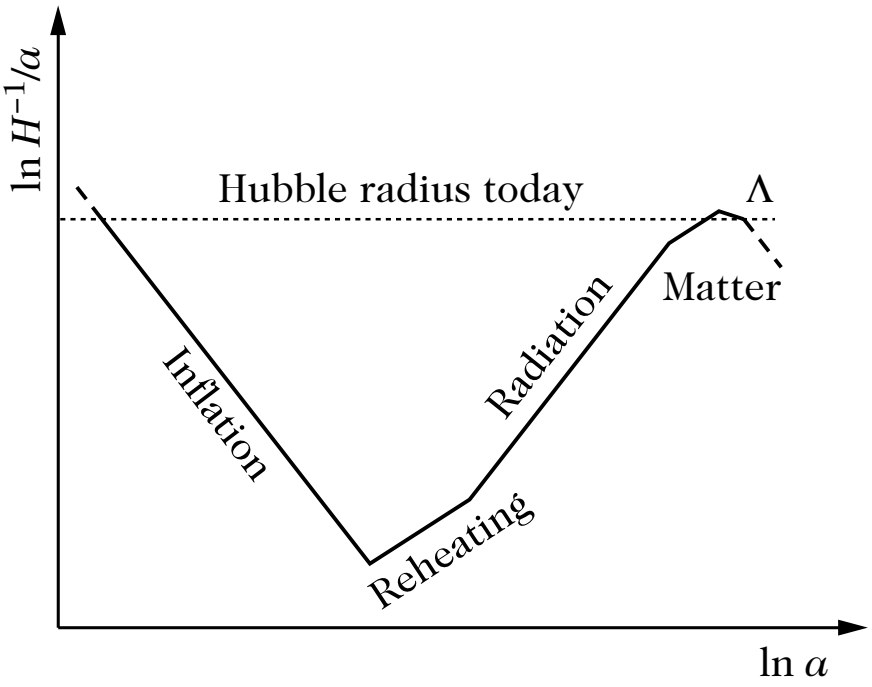


Figure 9.6: The function $\ln(aH)$ as a function of $\ln a$ illustrates the different epochs to take into account to compute the number of e-folds.

and dominated by radiation). This reheating phase, described in Section 9.5, connects inflation with the hot Big-Bang scenario.

To analyse the transition from the inflating Universe to the observable Universe, it will be useful to relate the instant at which a mode k becomes super-Hubble during inflation to that where it becomes sub-Hubble again. In order to do so, one needs a complete model of the evolution of the Universe. In the simplest case, the sequence of events is: inflation – matter era during the reheating phase (denoted with the index 'reh') – radiation era – matter era (see Fig. 9.6). Assuming that the transitions between these various regimes are instantaneous, one has

$$\frac{k}{a_0H_0}=\frac{a_kH_k}{a_0H_0}=\frac{a_k}{a_f}\frac{a_f}{a_{\rm reh}}\frac{a_{\rm reh}}{a_{\rm eq}}\frac{a_{\rm eq}}{a_0}\frac{H_k}{H_0}. \tag{9.2.44}$$

Using the typical values of these epochs, we obtain

$$N(k)\simeq 62-\ln\left(\frac{k}{a_0H_0}\right)-\ln\left(\frac{10^{16}\ {\rm GeV}}{V_k^{1/4}}\right)+\frac{1}{4}\ln\left(\frac{V_k}{V_f}\right)-\frac{1}{3}\ln\left(\frac{V_f^{1/4}}{\rho_{\rm reh}^{1/4}}\right)-\ln h. \tag{9.2.45}$$

To relate the physical scale of a perturbation mode k exiting the Hubble radius during inflation ($k=a_kH_k$) to its observable size today, we decompose the total expansion factor into a product of sequential cosmological eras:

$$\frac{k}{a_0H_0}=\frac{a_kH_k}{a_0H_0}=\left(\frac{a_k}{a_f}\right)\left(\frac{a_f}{a_{\rm reh}}\right)\left(\frac{a_{\rm reh}}{a_{\rm eq}}\right)\left(\frac{a_{\rm eq}}{a_0}\right)\left(\frac{H_k}{H_0}\right). \tag{1}$$

Taking the logarithm and noting that $N(k)\equiv\ln(a_f/a_k)$ is the number of e-folds remaining until the end of inflation:

$$N(k)=-\ln\left(\frac{k}{a_0H_0}\right)+\ln\left(\frac{H_k}{H_0}\right)+\ln\left(\frac{a_f}{a_{\rm reh}}\right)+\ln\left(\frac{a_{\rm reh}}{a_0}\right). \tag{2}$$

Evaluating each term using standard cosmological scaling laws:

1. **Reheating era ($a_f \to a_{\rm reh}$):** The oscillating inflaton behaves as dust with $\rho \propto a^{-3}$, giving:
$$\frac{a_f}{a_{\rm reh}} = \left(\frac{\rho_{\rm reh}}{\rho_f}\right)^{1/3} = \left(\frac{\rho_{\rm reh}}{V_f}\right)^{1/3}. \tag{3}$$

2. **Post-reheating thermal era ($a_{\rm reh} \to a_0$):** Comoving entropy conservation $S = g_{*S} a^3 T^3 = \text{const}$ with $\rho_{\rm reh} = \frac{\pi^2}{30} g_* T_{\rm reh}^4$ gives:
$$\frac{a_{\rm reh}}{a_0} = \left(\frac{g_{*S,0}}{g_{*S,\rm reh}}\right)^{1/3} \frac{T_0}{T_{\rm reh}} \simeq \left(\frac{43/11}{g_{*S,\rm reh}}\right)^{1/3} \left(\frac{\pi^2 g_*}{30}\right)^{1/4} \frac{T_0}{\rho_{\rm reh}^{1/4}}. \tag{4}$$

3. **Hubble parameter ratio:** Slow-roll gives $H_k = \sqrt{V_k/(3M_{\rm Pl}^2)}$. Normalizing by the GUT scale $V_k \sim (10^{16}\ \text{GeV})^4$ and current Hubble constant $H_0 = 100\,h\ \text{km s}^{-1}\text{Mpc}^{-1}$:
$$\ln\left(\frac{H_k}{H_0}\right) = \ln\left(\frac{10^{16}\ \text{GeV}}{\sqrt{3} M_{\rm Pl} H_0}\right) + \frac{1}{2}\ln\left(\frac{V_k}{(10^{16}\ \text{GeV})^4}\right) - \ln h. \tag{5}$$

Combining all terms with $T_0 = 2.7255\ \text{K} \simeq 2.35 \times 10^{-13}\ \text{GeV}$ and $g_{*S,\rm reh} \approx 106.75$ yields the master relation (9.2.45):

$$N(k) \simeq 62 - \ln\left(\frac{k}{a_0 H_0}\right) - \ln\left(\frac{10^{16}\ \text{GeV}}{V_k^{1/4}}\right) + \frac{1}{4}\ln\left(\frac{V_k}{V_f}\right) - \frac{1}{3}\ln\left(\frac{V_f^{1/4}}{\rho_{\rm reh}^{1/4}}\right) - \ln h. \tag{6}$$

For our present Hubble scale $k = a_0 H_0$, this establishes that cosmological scales crossed the horizon $N \approx 50$–60 e-folds before the end of inflation.

This relation can be modified if inflation occurs at a lower-energy scale. Moreover, the mode of wavelength equal to the Hubble radius today, $k_0 = a_0 H_0$, becomes super-Hubble at around 60 e-folds before the end of inflation. A scale of the order of 1 Mpc, typically of the order of the smallest observable scales today, becomes super-Hubble at around 9 e-folds after k_0. Our observations are thus mainly sensitive to the dynamics of inflation during about ten e-folds.

Finally, let us note that $\Omega_K(k) = -K/(a_k^2 H_k^2) = -K/(a_0^2 H_0^2) = \Omega_{K0}$ for the largest mode observable today, i.e. k_0. Constraints obtained in Chapter 4 show that it is completely legitimate to neglect the spatial curvature during inflation for the epochs during which the modes observable today become super-Hubble.

Observable scale and Largest Mode

1. **Largest Observable Mode (k_0):** Setting $a_0 = 1$ today, the mode currently crossing the Hubble radius satisfies:
$$k_0 = a_0 H_0 \approx \frac{H_0}{c} \tag{1}$$
Using $H_0 \approx 2.18 \times 10^{-18}\ \text{s}^{-1}$ in SI units and $1\ \text{Mpc} \approx 3.086 \times 10^{22}\ \text{m}$:
$$k_0 \approx \frac{2.18 \times 10^{-18}\ \text{s}^{-1}}{3.0 \times 10^{8}\ \text{m s}^{-1}} \times (3.086 \times 10^{22}\ \text{m Mpc}^{-1}) \approx 2.24 \times 10^{-4}\ \text{Mpc}^{-1} \tag{2}$$

2. **End of Inflation Scale (a_fH):** Since scale factor evolution is parameterized backwards from the end of inflation ($N = 0$) as $a(t) = a_f e^{-N}$, horizon exit occurs when $k = aH = a_f H e^{-N}$. Evaluating for k_0 at $N \approx 60$:

$$a_f H = k_0 e^{60} \approx (2.24 \times 10^{-4}\ \text{Mpc}^{-1}) \times e^{60} \approx 2.74 \times 10^{22}\ \text{Mpc}^{-1} \tag{3}$$

Hence, the final mode exiting the horizon at $N = 0$ is $k_{\text{end}} \sim 10^{22}\ \text{Mpc}^{-1}$.

3. **Smallest Observable Mode ($k \sim 1\ \text{Mpc}^{-1}$):** The smallest observable cosmological scales before non-linear structure dominates correspond to $k \sim 1\ \text{Mpc}^{-1}$. Its horizon exit e-fold count N satisfies:

$$1\ \text{Mpc}^{-1} = a_f H e^{-N} \implies N = \ln(a_f H) \approx \ln\left(2.74 \times 10^{22}\right) \approx 51.6 \approx 51 \tag{4}$$

4. **Observational Window:** Cosmological observations (CMB anisotropy and large-scale galaxy surveys) probe only:

$$\Delta N = N_{k_0} - N_{k=1} \approx 60 - 51 \approx 9\text{–}10\ \text{e-folds} \tag{5}$$

9.2.4 Some examples

Chaotic inflation

The simplest model Chaotic inflation is based on considering a free massive scalar field so that

$$V(\phi) = \frac{1}{2} m^2 \phi^2. \tag{9.2.46}$$

The Klein–Gordon equation (9.2.7) reduces to that of a damped harmonic oscillator, $\ddot{\phi} + 3H\dot{\phi} + m^2\phi = 0$. If ϕ is initially large, then the Friedmann equation (9.2.5) implies that H is also very large. The friction term becomes important and dominates the dynamics so that the field must be in the slow-roll regime. The evolution equations then reduce to the following two first-order equations,

$$3H\dot{\phi} + m^2\phi = 0, \quad H^2 = \frac{4\pi}{3}\left(\frac{m}{M_{\text{Pl}}}\right)^2 \phi^2. \tag{9.2.47}$$

They can be integrated easily to give

$$\phi(t) = \phi_i - \frac{m M_{\text{Pl}}}{\sqrt{12\pi}} t, \tag{9.2.48}$$

$$a(t) = a_i \exp\left\{\frac{2\pi}{M_{\text{Pl}}^2}\left[\phi_i^2 - \phi^2(t)\right]\right\}, \tag{9.2.49}$$

where ϕ_i and a_i are the values of the field and the scale factor at $t_i = 0$. The slow-roll parameters (9.2.19)–(9.2.21) being given by

$$\epsilon = \frac{M_{\text{Pl}}^2}{4\pi\phi^2}, \quad \delta = 0, \quad \xi_V = 0, \tag{9.2.50}$$

the slow-roll regime lasts until the field ϕ reaches $\phi_f = M_{\text{Pl}}/\sqrt{4\pi}$. We infer that the total number of e-folds is

$$N(\phi_i) = 2\pi\left(\frac{\phi_i}{M_{\text{Pl}}}\right)^2 - \frac{1}{2}. \tag{9.2.51}$$

In order to have $N \gtrsim 70$, we need $\phi_i \gtrsim 3M_{\text{Pl}}$. If ϕ takes the largest possible value compatible with the classical description adopted here, i.e. fixed by $V(\phi_i) \leq M_{\text{Pl}}^4$, we find that $\phi_i \sim M_{\text{Pl}}^2/m$.

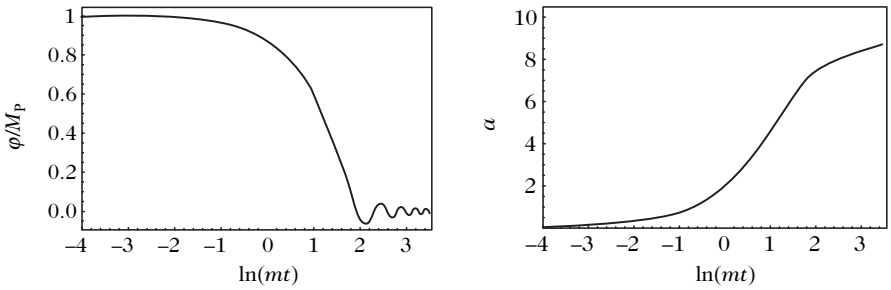


Figure 9.7: Evolution of the inflaton and scale factor during chaotic inflation for the model (9.2.46). The scalar field is initially in a slow-roll regime and the expansion of the Universe is accelerated. At the end of this regime, it starts oscillating at the minimum of its potential and it is equivalent to a pressureless fluid.

In this case, the maximal accessible number of e-folds would be $N_{\rm max} \sim 2\pi M_{\rm Pl}^2/m^2$. As will be seen later, the theory of cosmological perturbations and observations of the cosmic microwave background impose that the mass is of the order of $m \sim 10^{-6} M_{\rm Pl}$, so that $N_{\rm max} \sim 10^{13}$. The maximal number of e-folds is thus very large compared to the minimum required for solving the cosmological problems.

Consequently, if the Universe is initially composed of regions where the values of the scalar field are randomly distributed, or if we consider different Universes with different field values, then the domains where the initial value of ϕ is too small, never inflate or only for a small number of e-folds. The main contribution to the total physical volume of the Universe at the end of inflation thus comes from regions that have inflated for a long time and that had an initially large value of ϕ. These domains produce extremely flat and homogeneous zones at the end of inflation with a very large size compared to that of the observable Universe. At larger scales, the Universe has a very inhomogeneous structure. We will come back to this chaotic initial condition idea in Section 9.6.

Extensions This model can be generalized to any polynomial function of the scalar field. As an example, let us consider the family of potentials

$$V(\phi) = \frac{3M_{\rm Pl}^4}{8\pi}\lambda_n \left(\frac{\phi}{M_{\rm Pl}}\right)^n, \tag{9.2.52}$$

where λ_n is a dimensionless constant. The slow-roll parameters are given by

$$\epsilon = \frac{M_{\rm Pl}^2}{16\pi}\frac{n^2}{\phi^2}, \quad \delta = \frac{M_{\rm Pl}^2}{16\pi}\frac{n(n-2)}{\phi^2}, \tag{9.2.53}$$

and the slow-roll regime lasts as long as $\phi > \phi_f = nM_{\rm Pl}/(4\sqrt{\pi})$. We infer that the total number of e-folds is

$$N(\phi_i) = \frac{4\pi}{n}\left(\frac{\phi_i}{M_{\rm Pl}}\right)^2 - \frac{n}{4}. \tag{9.2.54}$$

So in order for $N > N_{\rm min}$, we should initially have

$$\phi_i > M_{\rm Pl}\sqrt{\frac{n}{4\pi}\left(N_{\rm min} + \frac{n}{4}\right)}. \tag{9.2.55}$$

Yet, chaotic initial conditions imply that $V(\phi_i) \sim M_{\rm Pl}^4$, imposing

$$\phi_i \sim M_{\rm Pl}\left(\frac{8\pi}{3\lambda_n}\right)^{1/n}, \tag{9.2.56}$$

a value in general very large compared to $M_{\rm Pl}$. Solving the Klein–Gordon and Friedmann equations in the slow-roll regime approximation gives

$$\phi(t) = \phi_i \left[1 + n(n-4)\frac{\sqrt{\lambda_n}}{16\pi}\left(\frac{\phi_i}{M_{\rm Pl}}\right)^{(n-4)/2} M_{\rm Pl} t\right]^{2/(4-n)}, \tag{9.2.57}$$

$$a(t) = a_i \exp\left\{\frac{4\pi}{nM_{\rm Pl}^2}\left[\phi_i^2 - \phi^2(t)\right]\right\}. \tag{9.2.58}$$

The expression in square brackets for the solution of the scalar field is positive definite as long as the slow-roll hypothesis is valid.

Notice that although expression (9.2.58) is well defined for all n, the scalar field on the other hand is singular for $n = 4$. The direct computation of this specific case shows that (9.2.58) remains valid, with $\phi(t) = \phi_i \exp\left(-\frac{1}{2\pi}\sqrt{\lambda_4} M_{\rm Pl} t\right)$.

We can reexpress these quantities in terms of the number of e-folds to obtain

$$\phi(N) = M_{\rm Pl}\sqrt{\left(\frac{\phi_i}{M_{\rm Pl}}\right)^2 - \frac{n}{4\pi}N}, \tag{9.2.59}$$

$$H(N) = M_{\rm Pl}\sqrt{\lambda_n}\left[\left(\frac{\phi_i}{M_{\rm Pl}}\right)^2 - \frac{n}{4\pi}N\right]^{n/4},$$

valid until $N(\phi_i)$. These expressions show that the Hubble parameter can vary significantly during the slow-roll phase. In particular, and as will be seen later, the amplitude of the observable modes generated roughly $50 - 70$ e-folds before the end of inflation implies that $H/M_{\rm Pl} \sim 10^{-5}$. However, this does not imply that the Hubble parameter could not have been much larger before that. Cosmological observations can thus give some inputs on the last 70 e-folds or so of the inflation era but, a priori, put no constraints on the primordial phase of inflation.

Power-law inflation: an exact solution

For most models, the slow-roll approximation is so good that there is no need to go further, even if the end of inflation is not very well taken into account. It is, however, useful to have some exact solutions.

Many models exist in the literature but the most attractive one is that of power-law inflation [15]. The advantage of this model is that the perturbation equations can also be solved analytically.

These models are constructed such that the solution of the equations of evolution are

$$a = a_0|\eta|^{1+\beta} \iff a = a_0 t^p, \tag{9.2.60}$$

and

$$1 + \beta = \frac{p}{1-p}, \quad \gamma = \frac{\beta+2}{\beta+1} = \frac{1}{p}, \tag{9.2.61}$$

$$\frac{\phi - \phi_i}{M_{\rm Pl}} = \frac{1}{2}\sqrt{\frac{\gamma}{\pi}}(1+\beta)\ln|\eta|.$$

The parameter p varies between 1 and ∞ and thus $\beta < -2$. The evolution equations imply that the potential is of the form

$$V(\phi) = V_i \exp\left[4\sqrt{\frac{\pi}{p}}\frac{(\phi - \phi_i)}{M_{\rm Pl}}\right], \tag{9.2.62}$$

so that the slow-roll parameters turn out to be constant,

$$\epsilon = \delta = \xi_H = \frac{1}{p}, \quad \xi = 0. \tag{9.2.63}$$

Thus, with such a potential, the slow-roll parameters' inflation never ends. The limit $p \to +\infty$ (i.e. $\beta \to -2$) corresponds to a de Sitter Universe.

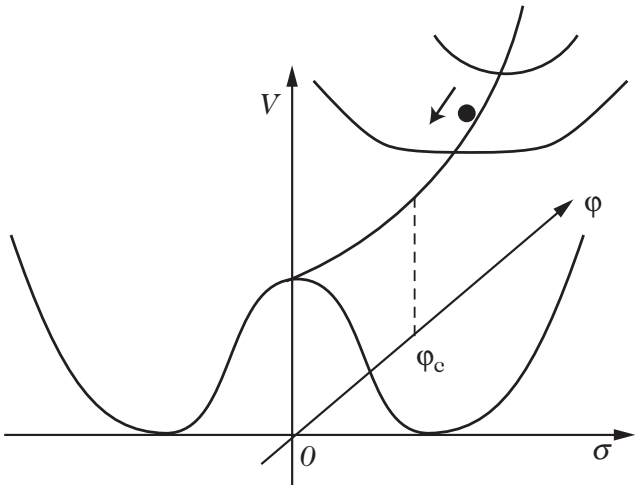


Figure 9.8: Potential for hybrid inflation with two scalar fields. Inflation occurs when $\sigma = 0$ and stops abruptly when ϕ reaches the value ϕ_c of the bifurcation point.

Hybrid inflation

Models of hybrid inflation introduce two coupled scalar fields, ϕ and σ, with the potential

$$V(\phi,\sigma) = \frac{\lambda}{4}\left(\sigma^2 - v^2\right)^2 + \frac{1}{2}g^2\phi^2\sigma^2 + V(\phi), \tag{9.2.64}$$

where v is the vacuum expectation value (the VEV, as defined in Chapter 2) of σ and $V(\phi)$ is the potential of the field ϕ (see Fig. 9.8). The effective mass of the field σ depends on the value of the second field:

$$m^2(\sigma) = g^2\left(\phi^2 - \phi_c^2\right), \quad \phi_c^2 = \frac{\lambda v^2}{g^2}. \tag{9.2.65}$$

The point $\phi = \phi_c$ is the bifurcation point where the effective mass of the field σ changes sign to become negative. As long as $\phi > \phi_c$, the potential has a valley along $\sigma = 0$; everything therefore goes as if there was only one field rolling towards the minimum of the potential $V(\phi) + \frac{\lambda}{4}v^4$. The global minimum is, however, at $\phi = 0$, $|\sigma| = v$. When the inflaton reaches the value ϕ_c, the mass of σ changes sign, which results in a spinodal instability (see Chapter 11) triggering the end of inflation. The total number of e-folds is then

$$N = 2\pi\frac{\lambda v^4}{m^2 M_{\rm Pl}^2}\ln\left(\frac{\phi_i}{\phi_c}\right), \tag{9.2.66}$$

if the potential for the inflaton is of the form $V(\phi) = \frac{1}{2}m^2\phi^2$.

9.2.5 Classification of inflationary models

For single-field inflationary models, the potential can always be characterized by two mass scales, Λ and μ,

$$V(\phi) = \Lambda^4 f\left(\frac{\phi}{\mu}\right). \tag{9.2.67}$$

These two scales typically determine the height and width of the potential.

In order to discuss the phenomenology of these models, it is useful to classify them as a function of the value of their slow-roll parameters. We distinguish between the four following classes.

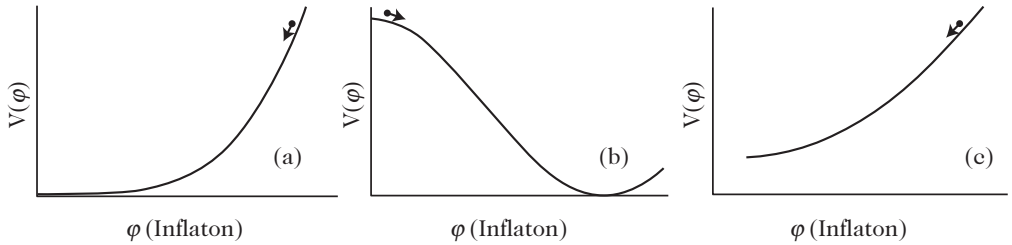


Figure 9.9: From left to right, three examples of inflationary potentials with large field values (class A), small field values (class B) and hybrid (class C).

A. Inflation with large values of the inflaton: $-\epsilon < \delta \leq \epsilon$. The archetypes of this class are the potentials of chaotic inflation, e.g. (9.2.46) and (9.2.52) or exponential potentials (9.2.62). In these models, the field is initially away from its minimum towards which it slowly relaxes. These potentials satisfy $V'' > 0$, i.e. $\eta_V > 0$. In the framework of chaotic inflation, the condition $V \sim M_{\rm Pl}^4$ implies that $\phi \gg M_{\rm Pl}$, but a value $\phi \sim M_{\rm Pl}$ is, in general, sufficient to obtain the required number of e-folds.

B. Inflation with small values of the inflaton: $\delta < -\epsilon$. These potentials are those that appear naturally in the mechanisms of spontaneous symmetry breaking and were at the origin of the so-called models of new inflation. The field is initially close to an unstable equilibrium position at $\phi_i = 0$. Inflation occurs when the field is close to the origin and the asymptotic form of the potential is not significant. We therefore have $V'' < 0$, i.e. $\eta_V < 0$ and $\epsilon > 0$ and very small (very flat potential). We infer that $\delta < -\epsilon$. An example of such a potential is

$$V(\phi) = \Lambda^4 \left[1 - \left(\frac{\phi}{\mu}\right)^p\right], \tag{9.2.68}$$

that we can consider as a Taylor expansion of the 'true' potential around the origin. In the case of a spontaneous symmetry-breaking potential of the Higgs kind, $p = 2$; μ is then the scale of the symmetry breaking. Another example is the Coleman–Weinberg potential,

$$V = \frac{\lambda}{4}\phi^4 + \frac{\lambda^2}{44\pi}\phi^4 \left[\ln\left(\frac{\phi}{\mu}\right) - \frac{25}{6}\right]. \tag{9.2.69}$$

C. Hybrid models: $0 < \delta < \epsilon$. These potentials often appear in the framework of supersymmetry. The field evolves towards its non-vanishing minimum [cf. the example given in (9.2.64)]. Inflation usually ends abruptly due to an instability of the auxiliary field. Another example is given by the potential $V(\phi) = \Lambda^4 [1 + (\phi/\mu)^p]$.

D. Linear models: $\delta = -\epsilon$. These models, characterized by

$$V(\phi) \propto \phi, \tag{9.2.70}$$

represent a limit between the models with large and small values of the field.

Of course, such a classification is not exhaustive as it does not incorporate potentials of the type $V(\phi) \propto \ln \phi$, which often appear in the framework of supersymmetry, or of inverse power laws, $V(\phi) \propto \phi^{-n}$ as in intermediate inflation. However, these models require the introduction of auxiliary fields to stop inflation.

9.3 Quantum fluctuations during inflation

One of the great successes of inflation has been to provide a way to relate the origin of the large-scale structure of the Universe to the quantum fluctuations of the inflaton that are amplified

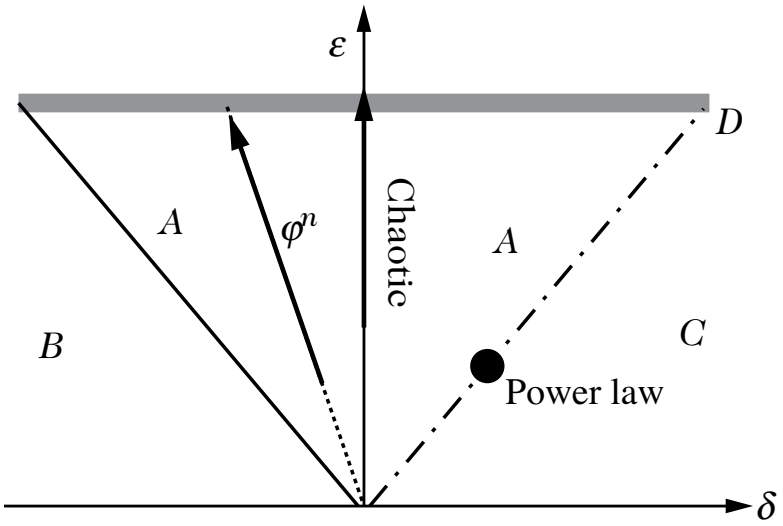


Figure 9.10: The four classes of models of single-field inflation in the space of the two first slow-roll parameters. We have indicated the trajectory of the models of chaotic inflation (9.2.46) and (9.2.52) and the point represents a model of power-law inflation (9.2.62).

during inflation. To address this central point of primordial cosmology, we start by a warm up and only consider test fields.

This study will illustrate the fact that during inflation, any light field (in the sense $m < H$) develops super-Hubble fluctuations with mean amplitude $H/2\pi$.

9.3.1 Massless test scalar field in a de Sitter space-time

The structure of the vacuum in an exponentially expanding space-time is much more complex than that in a Minkowski space-time (Chapter 2). In particular, the wavelength of any fluctuation grows exponentially. When it becomes larger than the Hubble radius, H^{-1}, this fluctuation is said to freeze as it no longer oscillates due to the friction term in the Klein–Gordon equation, and it rapidly converges toward a solution of constant amplitude.

Quantization

To quantize a scalar field in an expanding space-time, we proceed in an analogous way to what is usually done in Minkowski space-time and promote the scalar field, χ say, to the status of a quantum operator,

$$\hat{\chi}(\mathbf{x},t) = \int \frac{\mathrm{d}^3k}{(2\pi)^{3/2}} \left[\chi_k(t) e^{i\mathbf{k}\cdot\mathbf{x}} \hat{a}_{\mathbf{k}} + \chi_k^*(t) e^{-i\mathbf{k}\cdot\mathbf{x}} \hat{a}_{\mathbf{k}}^\dagger\right], \tag{9.3.1}$$

where the creation and annihilation operators satisfy the canonical commutation relations

$$[\hat{a}_{\mathbf{k}}, \hat{a}_{\mathbf{k}'}^\dagger] = \delta^{(3)}(\mathbf{k}-\mathbf{k}'), \qquad [\hat{a}_{\mathbf{k}}, \hat{a}_{\mathbf{k}'}] = [\hat{a}_{\mathbf{k}}^\dagger, \hat{a}_{\mathbf{k}'}^\dagger] = 0. \tag{9.3.2}$$

and $\mathbf{k}\cdot\mathbf{x} = \sum_{i=1}^3 k^i x^i$. The mode function $\chi_k(t)$ satisfies the classical Klein–Gordon equation,

$$\ddot{\chi}_k + 3H\dot{\chi}_k + \frac{k^2}{a^2}\chi_k = 0. \tag{9.3.3}$$

Qualitatively, this equation possesses two regimes. When the mode is sub-Hubble, $k \gg aH$, the friction term is negligible and the equation reduces to that of a harmonic oscillator analogous to that obtained in Minkowski space-time, with the only difference being that the wave number is

redshifted by the expansion. When the mode is super-Hubble, $k \ll aH$, and the gradient term is negligible. The equation reduces to $\ddot{\chi}_k + 3H\dot{\chi}_k = 0$ where the dominant solution is a constant. In conclusion, an initially sub-Hubble mode oscillates as its wavelength grows and then freezes when it becomes super-Hubble.

General solution

To solve (9.3.3) it is convenient to work in conformal time and to introduce the auxiliary field

$$v = a\chi. \tag{9.3.4}$$

The Klein–Gordon equation then takes the form

$$v_k'' + \left(k^2 - \frac{a''}{a}\right) v_k = 0, \tag{9.3.5}$$

which introduces an effective time-dependent mass term, $m^2 = -a''/a$. For a de Sitter space, the scale factor is given by (9.1.20) so that v satisfies

$$v_k'' + \left(k^2 - \frac{2}{\eta^2}\right) v_k = 0. \tag{9.3.6}$$

It has the general solution

$$v_k(\eta) = \sqrt{-\eta}\left[A(k)H^{(1)}_{3/2}(-k\eta) + B(k)H^{(2)}_{3/2}(-k\eta)\right], \tag{9.3.7}$$

where $H^{(1,2)}_{3/2}$ are the Hankel functions of first and second kind (Appendix B). Using

$$H^{(1)}_{3/2}(z) = \left[H^{(2)}_{3/2}(z)\right]^* = -\sqrt{\frac{2}{\pi z}}e^{iz}\left(1 + \frac{i}{z}\right), \quad H^{(2)}_{3/2}(z) = -\sqrt{\frac{2}{\pi z}}e^{-iz}\left(1 - \frac{i}{z}\right), \tag{9.3.8}$$

we infer the general form

$$v_k(\eta) = A(k)e^{-ik\eta}\left(1 + \frac{1}{ik\eta}\right) + B(k)e^{ik\eta}\left(1 - \frac{1}{ik\eta}\right). \tag{9.3.9}$$

At this stage, we need to postulate a choice of initial conditions to determine the two arbitrary functions $A(k)$ and $B(k)$.

Initial condition

To perform a canonical quantization, we impose the commutation rules on the operators $\hat{v}$ and $\hat{\pi}$, namely $[\hat{v}(\mathbf{x},\eta), \hat{v}(\mathbf{x}',\eta)] = [\hat{\pi}(\mathbf{x},\eta), \hat{\pi}(\mathbf{x}',\eta)] = 0$ and $[\hat{v}(\mathbf{x},\eta), \hat{\pi}(\mathbf{x}',\eta)] = i\delta^{(3)}(\mathbf{x} - \mathbf{x}')$ on constant time hypersurfaces, $\hat{\pi}$ being the conjugate momentum of $\hat{v}$. From the commutation rules of the annihilation and creation operators, this implies that

$$v_k v_k^{*\prime} - v_k^* v_k' = i, \tag{9.3.10}$$

which determines the normalization of the Wronskian. The choice of a specific mode function $v_k(\eta)$ corresponds to the choice of a prescription for the physical vacuum $|0\rangle$, defined by $\hat{a}_{\mathbf{k}}|0\rangle = 0$.

To choose the vacuum, let us note that for very high frequency modes, $k\eta \gg 1$, the curvature term is negligible so that the field can be quantized as if it were in a Minkowski space-time (see Chapter 2). We thus pick up the solution that corresponds adiabatically to the usual Minkowski vacuum, keeping only the positive frequencies. We therefore impose

$$v_k \to \frac{e^{-ik\eta}}{\sqrt{2k}}, \quad k\eta \to -\infty. \tag{9.3.11}$$

Note that the particle-like interpretation of these states is valid only on subhorizon scales. Once the modes become super-Hubble, this interpretation breaks down, and it is therefore not meaningful to regard their statistics as describing the statistics of "physical" particles, such as the distribution of underdensities and overdensities.

This choice is referred to as the Bunch–Davies vacuum.[2] We infer that the solution satisfying this initial condition is

$$\chi_k = \frac{H\eta}{\sqrt{2k}}\left(1+\frac{1}{ik\eta}\right)e^{-ik\eta}. \tag{9.3.12}$$

Let us emphasize that, at this stage, we have not justified why we have to apply the condition (9.3.11) to v_k rather than to χ_k. Such a justification will be made later.

To determine the arbitrary integration coefficients $A(k)$ and $B(k)$ in the general de Sitter mode solution (9.3.9), we proceed in two systematic steps:

Step 1: Positive-Frequency Eigenstate Condition on χ_k In the asymptotic past $\eta \to -\infty$, physical wavelengths are deep inside the Hubble horizon ($k \gg aH$). We define the physical vacuum by requiring that the test scalar field mode $\chi_k(\eta)$ is an instantaneous positive-frequency eigenstate of the timelike conformal derivative operator $i\partial_\eta$ at an initial conformal time η_i:

$$i\partial_\eta \chi_k(\eta_i) = k\chi_k(\eta_i) \iff \chi_k'(\eta_i) = -ik\chi_k(\eta_i). \tag{1}$$

To translate this condition into a boundary condition on the canonical variable $v_k(\eta) = a(\eta)\chi_k(\eta)$, we express $\chi_k = v_k/a$ and compute its conformal time derivative:

$$\chi_k' = \left(\frac{v_k}{a}\right)' = \frac{v_k' - \mathcal{H}v_k}{a}, \qquad \text{where} \quad \mathcal{H} \equiv \frac{a'}{a}. \tag{2}$$

In de Sitter spacetime, the scale factor is $a(\eta) = -1/(H\eta)$, which gives the conformal Hubble parameter $\mathcal{H}(\eta) = -1/\eta$. Substituting χ_k' and χ_k into the eigenstate condition (1):

$$\frac{v_k'(\eta_i) - \mathcal{H}(\eta_i)v_k(\eta_i)}{a(\eta_i)} = -ik\frac{v_k(\eta_i)}{a(\eta_i)} \implies v_k'(\eta_i) - \mathcal{H}(\eta_i)v_k(\eta_i) = -ikv_k(\eta_i), \tag{3}$$

which yields the exact boundary condition on v_k:

$$\boxed{v_k'(\eta_i) + \left(\frac{1}{\eta_i} + ik\right)v_k(\eta_i) = 0}. \tag{4}$$

Step 2: Linear Relation Between $A(k)$ and $B(k)$ We now substitute the general solution (9.3.9) for $v_k(\eta)$ into the boundary condition (4). Decomposing $v_k(\eta) = A(k)u_1(\eta) + B(k)u_2(\eta)$ with:

$$u_1(\eta) \equiv e^{-ik\eta}\left(1+\frac{1}{ik\eta}\right) = e^{-ik\eta}\left(1-\frac{i}{k\eta}\right), \qquad u_2(\eta) \equiv e^{ik\eta}\left(1-\frac{1}{ik\eta}\right) = e^{ik\eta}\left(1+\frac{i}{k\eta}\right), \tag{5}$$

[2] Note that the quantum states have a clear physical interpretation only once v_k has been chosen. The normalization condition (9.3.10) does not fix v_k completely and any function of the form $u_k = \alpha_k v_k + \beta_k v_k^*$, with α_k and β_k two complex numbers satisfying $|\alpha_k|^2 - |\beta_k|^2 = 1$ will also be a solution normalized according to (9.3.10). The asymptotic condition (9.3.11) completely defines the quantum states.

their conformal time derivatives are:

$$u_1'(\eta) = -ike^{-ik\eta}\left(1 - \frac{i}{k\eta} - \frac{1}{k^2\eta^2}\right), \tag{6}$$

$$u_2'(\eta) = ike^{ik\eta}\left(1 + \frac{i}{k\eta} - \frac{1}{k^2\eta^2}\right). \tag{7}$$

Evaluating the combination $u' + (\frac{1}{\eta} + ik)u$ for each component:

$$u_1'(\eta) + \left(\frac{1}{\eta} + ik\right)u_1(\eta) = \frac{e^{-ik\eta}}{\eta}, \tag{8}$$

$$u_2'(\eta) + \left(\frac{1}{\eta} + ik\right)u_2(\eta) = \frac{e^{ik\eta}}{\eta}(2ik\eta - 1). \tag{9}$$

Applying the boundary condition (4) at $\eta = \eta_i$:

$$A(k)\left[\frac{e^{-ik\eta_i}}{\eta_i}\right] + B(k)\left[\frac{e^{ik\eta_i}}{\eta_i}(2ik\eta_i - 1)\right] = 0. \tag{10}$$

Multiplying by $\eta_i e^{ik\eta_i}$ establishes the exact relation between $A(k)$ and $B(k)$:

$$A(k) = (1 - 2ik\eta_i)\, e^{2ik\eta_i} B(k) \iff B(k) = \frac{e^{-2ik\eta_i}}{1 - 2ik\eta_i} A(k). \tag{11}$$

In the asymptotic past $\eta_i \to -\infty$ (or for any mode well within the Hubble radius, $-k\eta_i \gg 1$), the ratio vanishes:

$$\left|\frac{B(k)}{A(k)}\right| = \frac{1}{|1 - 2ik\eta_i|} = \frac{1}{\sqrt{1 + 4k^2\eta_i^2}} \xrightarrow{\eta_i \to -\infty} 0 \implies \boxed{B(k) = 0}. \tag{12}$$

Step 3: Normalization Condition and Final Mode Function To preserve the canonical commutation relations $[\hat{v}(\mathbf{x}, \eta), \hat{\pi}(\mathbf{x}', \eta)] = i\delta^{(3)}(\mathbf{x} - \mathbf{x}')$, the mode function $v_k(\eta)$ must satisfy the Wronskian normalization condition:

$$W[v_k, v_k^*] \equiv v_k v_k^{*\prime} - v_k^* v_k' = i. \tag{13}$$

Evaluating this Wronskian explicitly using the general solution $v_k = A(k)u_1(\eta) + B(k)u_2(\eta)$:

$$v_k v_k^{*\prime} - v_k^* v_k' = 2ik\left(|A(k)|^2 - |B(k)|^2\right) = i \implies \boxed{|A(k)|^2 - |B(k)|^2 = \frac{1}{2k}}. \tag{14}$$

Substituting $B(k) = 0$ into this normalization condition and choosing the standard phase convention fixes:

$$|A(k)|^2 = \frac{1}{2k} \implies \boxed{A(k) = \frac{1}{\sqrt{2k}}}. \tag{15}$$

Substituting $A(k) = 1/\sqrt{2k}$ and $B(k) = 0$ back into (9.3.9) uniquely yields the **Bunch–Davies vacuum** mode functions:

$$\boxed{v_k(\eta) = \frac{1}{\sqrt{2k}}\left(1 - \frac{i}{k\eta}\right)e^{-ik\eta} \iff \chi_k(\eta) = \frac{H\eta}{\sqrt{2k}}\left(1 + \frac{1}{ik\eta}\right)e^{-ik\eta}}. \tag{16}$$

Properties on super-Hubble scales

On super-Hubble scales, the field χ acquires a constant amplitude

$$|\chi_k|(k\eta \ll 1) = \frac{H}{\sqrt{2k^3}}, \tag{9.3.13}$$

where we have used $a(\eta) = -1/(H\eta)$ for a pure de Sitter inflationary phase, H being a constant in this case.

Let us emphasize that, in this limit, the field actually behaves as

$$\hat{\chi} \to \int \frac{\mathrm{d}^3k}{(2\pi)^{3/2}} \hat{\chi}_k e^{i\mathbf{k}\cdot\mathbf{x}} = \int \frac{\mathrm{d}^3k}{(2\pi)^{3/2}} \frac{H}{\sqrt{2k^3}} (\hat{a}_{\mathbf{k}} + \hat{a}^{\dagger}_{-\mathbf{k}}) e^{i\mathbf{k}\cdot\mathbf{x}}. \tag{9.3.14}$$

At this point, all the modes being proportional to $(\hat{a}_{\mathbf{k}} + \hat{a}^{\dagger}_{-\mathbf{k}})$, the quantum operators $\hat{\chi}_k$ commute. Therefore, $\hat{\chi}$ has actually the same statistical properties as a Gaussian classical stochastic field. Effectively, on super-Hubble scales, the quantum operator $\hat{\chi}$ can be replaced by a stochastic field with Gaussian statistics. Introducing a unit Gaussian random variable, $e_v(\mathbf{k})$, which satisfies

$$\langle e_v(\mathbf{k}) \rangle = 0, \quad \langle e_v(\mathbf{k}) e_v^*(\mathbf{k}') \rangle = \delta^{(3)}(\mathbf{k} - \mathbf{k}'), \tag{9.3.15}$$

the mode operators are replaced by stochastic variables through

$$\hat{v}_k \to v_k = v_k(\eta) e_v(\mathbf{k}) \tag{9.3.16}$$

and we can switch from the vacuum quantum average $\langle 0| \dots |0\rangle$ to the classical ensemble average $\langle \dots \rangle$.

statistics of quantum field → statistics of initial condition

This explains why the super-Hubble modes have Gaussian statistics. A description of the quantum to classical transition can be found in Ref. [16].

Power spectrum

The correlation function of v is defined as

$$\xi_v \equiv \langle 0|\hat{v}(\mathbf{x},\eta)\hat{v}(\mathbf{x}',\eta)|0\rangle, \tag{9.3.17}$$

and takes the simple form

$$\xi_v = \int \frac{\mathrm{d}^3k}{(2\pi)^3} |v_k|^2 e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')} = \int \frac{\mathrm{d}k}{k} \frac{k^3}{2\pi^2} |v_k|^2 \frac{\sin kr}{kr}, \tag{9.3.18}$$

where the isotropy of the background space-time has been used to integrate over angles. Because of the symmetries of the background, it is a function of $r = |\mathbf{x} - \mathbf{x}'|$ only. We thus define the power spectra of v, using the definitions of Section B.5 in Appendix B for the normalization of the power spectrum,

$$P_v(k) = |v_k|^2, \quad \mathcal{P}_v(k) = \frac{k^3}{2\pi^2} |v_k|^2. \tag{9.3.19}$$

In the stochastic picture, the correlator of v_k is simply given by

$$\langle v_k v_{k'}^* \rangle = P_v(k) \delta^{(3)}(\mathbf{k} - \mathbf{k}'), \tag{9.3.20}$$

from which one easily deduces that the power spectrum of χ is

$$P_\chi(k) = \frac{2\pi^2}{k^3} \mathcal{P}_\chi(k) = |\chi_k|^2 = \frac{|v_k|^2}{a^2}. \tag{9.3.21}$$

It is thus a scale invariant power spectrum,

$$\mathcal{P}_\chi(k) = \left(\frac{H}{2\pi}\right)^2. \tag{9.3.22}$$

WKB approximation

In more general cases, where e.g. the function $a(\eta)$ is not known analytically, we may not have exact solutions for the mode functions. One can generalize the previous construction by relying on a WKB approach in which one introduces the WKB mode functions

$$v_k^{\rm WKB}(\eta) = \frac{1}{\sqrt{2\omega(k,\eta)}} \exp\left[\pm i \int^\eta \omega(k,\eta')\mathrm{d}\eta'\right], \tag{9.3.23}$$

which, as is easily checked, are solutions of

$$v_k^{\rm WKB\prime\prime} + \left[\omega^2(k,\eta) - Q_{\rm WKB}\right] v_k^{\rm WKB} = 0, \tag{9.3.24}$$

where the potential is

$$Q_{\rm WKB} = \frac{3}{4}\left(\frac{\omega'}{\omega}\right)^2 - \frac{1}{2}\frac{\omega''}{\omega}. \tag{9.3.25}$$

The WKB solution is thus a good approximation as long as the WKB condition

$$\left|\frac{Q_{\rm WKB}}{\omega^2}\right| \ll 1 \tag{9.3.26}$$

holds. For a function satisfying an equation such as (9.3.6), $\omega^2(k,\eta) = k^2 - 2/\eta^2$, the WKB condition reduces to $|k\eta| \gg 1$, so that on sub-Hubble scales, the mode function is actually close to its WKB approximation.

One might thus reinterpret the quantization procedure by noting the fact that there exists an adiabatic (WKB) solution on sub-Hubble scales.

Interpretation and infra-red divergences

Expression (9.3.12) can be used to express the variance of the field, $\langle\chi^2\rangle$, as a function of the physical momentum, $p = a^{-1}(t)k$,

$$\langle\chi^2\rangle = \int \frac{\mathrm{d}^3p}{(2\pi)^3 p}\left(\frac{1}{2} + \frac{H^2}{2p^2}\right). \tag{9.3.27}$$

The first term is simply the zero-point energy of a harmonic oscillator in Minkowski space-time (Chapter 2). This term leads to an ultraviolet divergence that can be eliminated by renormalization. The second term only appears in a curved space-time and is characteristic of de Sitter space-time. It can be interpreted as the existence of

$$n_p = \frac{H^2}{2p^2} \tag{9.3.28}$$

particles of physical momentum p. This term diverges in the infra-red. There are actually two divergences, one when $p \to 0$ and the second one when $\eta \to 0$.

The infra-red divergence at $p = 0$ appears because the expansion of the field in creation and annihilation modes do not include the zero mode. For a massless field in Minkowski space in a box of volume V, a complete description of the solutions requires the introduction of the position and momentum operators, $\hat{\phi}_0$ and $\hat{\pi}_{\phi 0}$, that satisfy the relation $[\hat{\phi}_0, \hat{\pi}_{\phi 0}] = i$. We can then define the vacuum by imposing $\hat{\pi}_{\phi 0}|0\rangle = 0$ and $\hat{a}_{\mathbf{k}}|0\rangle = 0$, which amounts to describing the state in a space that is the product of a Fock space and a Hilbert space. In a Minkowski space of dimension larger than 2, the continuum limit ($V \to \infty$) exists because the phase space volume factor compensates the divergence. However, it persists in a de Sitter space-time.

If this divergence is regularised by imposing a frequency cutoff $p > p_0 \sim H$ then

$$\langle\chi^2\rangle \sim \frac{H^2}{4\pi^2}\int_{He^{-Ht}} \frac{\mathrm{d}k}{k} \sim \frac{H^3}{4\pi^2}t. \tag{9.3.29}$$

This result is general and the resolution of the infra-red divergence at $k = 0$ does not resolve this second problem. The physical origin of this divergence is found in the eternal expansion that brings modes towards $p = 0$ exponentially.

However, these divergences only appear for massless particles in a de Sitter space that characterizes a phase of eternal inflation and naturally disappear in quasi-de Sitter space-times.

9.3.2 Massive test field in de Sitter

We can extend the previous case to that of a test field of mass m_χ. Equation (9.3.5) then takes the form

$$\ddot{\chi}_k + 3H\dot{\chi}_k + \left(\frac{k^2}{a^2} + m_\chi^2\right)\chi_k = 0, \tag{9.3.30}$$

when considering a potential $\frac{1}{2}m_\chi^2\chi^2$. In conformal time, introducing $v_k = a\chi_k$, it can be rewritten as

$$v_k'' + \left[k^2 - \frac{1}{\eta^2}\left(\nu_\chi^2 - \frac{1}{4}\right)\right]v_k = 0, \tag{9.3.31}$$

with

$$\nu_\chi^2 \equiv \frac{9}{4} - \frac{m_\chi^2}{H^2}. \tag{9.3.32}$$

If $\nu_\chi^2 \geq 0$, the general solution that satisfies the initial condition (9.3.11) when $k\eta \to -\infty$ is

$$v_k = \frac{\sqrt{-\pi\eta}}{2} i^{(\nu_\chi+1/2)} H_{\nu_\chi}^{(1)}(-k\eta) = \frac{\sqrt{-\pi\eta}}{2} i^{(\nu_\chi+1/2)} H_{\nu_\chi}^{(2)}(k\eta), \quad \nu_\chi^2 \geq 0. \tag{9.3.33}$$

On super-Hubble scales, the field freezes to a constant amplitude,

$$|\chi_k|(k\eta \ll 1) = \frac{H}{\sqrt{2k^3}} 2^{\nu_\chi - 3/2} \frac{\Gamma(\nu_\chi)}{\Gamma(3/2)} \left(\frac{k}{aH}\right)^{3/2-\nu_\chi}, \tag{9.3.34}$$

and has a power spectrum

$$\mathcal{P}_\chi(k) = 2^{2\nu_\chi - 3} \left[\frac{\Gamma(\nu_\chi)}{\Gamma(3/2)}\right]^2 \left(\frac{H}{2\pi}\right)^2 \left(\frac{k}{aH}\right)^{3-2\nu_\chi}, \tag{9.3.35}$$

using the definitions from Section B.5 of Appendix B for the normalization of the power spectrum.

If the field has a mass larger than $3H/2$ then $\nu_\chi^2 < 0$. The solution can be obtained by exchanging ν_χ with $i|\nu_\chi|$ in (9.3.33). The sub-Hubble solution is not affected and the initial conditions are therefore identical. Nevertheless, the mode function in the super-Hubble limit is also oscillating. After some algebra, we find that the power spectrum is now given by

$$\mathcal{P}_\chi(k) \propto \left(\frac{H}{2\pi}\right)^2 \frac{H}{m_\chi} \left(\frac{k}{aH}\right)^3. \tag{9.3.36}$$

The amplitude of the spectrum is thus reduced by a factor $H/m_\chi > 3/2$ and the spectrum decreases rapidly for long wavelengths.

This result is one of the generic predictions of inflation, practically independent of the model at hand:

> *During inflation, any light test field, $m < H$, develops super-Hubble fluctuations of characteristic amplitude $H/2\pi$, with an almost scale invariant power spectrum.*

9.3.3 Massive test field during slow-roll inflation

In the two previous examples, inflation was described as a pure de Sitter phase so that H was constant. Of course, this is only an approximation, and in a more realistic model, H decreases slowly in time during inflation.

This time dependence will have an influence on the spectrum of any light field. For a massless scalar field, expression (9.3.22) means that the spectrum of the super-Hubble modes is given by the value of H at the time where the mode becomes super-Hubble, $\mathcal{P}_\chi(k) \propto H_k^2$. In a de Sitter space-time, H is constant so that H_k does not depend on k and thus the power spectrum is scale invariant. Relation (9.2.40) indicates that in the slow-roll regime, $H_k \propto k^{-\epsilon}$, so that we expect that the spectrum of a massless test scalar field behaves as $\mathcal{P}_\chi \propto k^{-2\epsilon}$. Its spectrum will therefore be red-tilted (since $\epsilon \geq 0$) if inflation is realized in the slow-roll regime.

Evolution of the scale factor

The expansion of the Universe is now quasi-de Sitter. Integrating by parts, the conformal time is given by

$$\eta = \int \frac{\mathrm{d}a}{a^2 H} = -\frac{1}{aH} + \int \frac{\epsilon}{a^2 H}\mathrm{d}a. \tag{9.3.37}$$

Assuming ϵ is constant, we find that the scale factor evolves as

$$a(\eta) = -\frac{1}{H\eta}\frac{1}{1-\epsilon}. \tag{9.3.38}$$

From (9.2.28), this hypothesis amounts to neglecting all terms of 2nd order in the slow-roll parameters since $\dot{\epsilon} = \mathcal{O}(\epsilon^2, \delta\epsilon)$. This first-order slow-roll approximation thus amounts to considering only the two parameters ϵ and δ and to assume that they are constant.

Solution of the Klein–Gordon equation

To solve (9.3.30), one should first evaluate

$$\frac{a''}{a} = a^2\left(H^2 + \frac{\dot{a}}{a}\right) = a^2(2H^2 + \dot{H}) = a^2H^2(2-\epsilon) = \frac{2-\epsilon}{(1-\epsilon)^2\eta^2}. \tag{9.3.39}$$

We infer that, to first order in ϵ,

$$\frac{a''}{a} \simeq \frac{2+3\epsilon}{\eta^2}. \tag{9.3.40}$$

The Klein–Gordon equation therefore takes the form (9.3.31) with

$$\nu_\chi^2 = \frac{9}{4} + 3\epsilon - \frac{m_\chi^2}{H^2}. \tag{9.3.41}$$

For a light field, the power spectrum then takes the form (9.3.35), which corresponds to a spectral index

$$n_\chi - 1 \equiv \frac{\mathrm{d}\ln\mathcal{P}_\chi}{\mathrm{d}\ln k} = 3 - 2\nu_\chi \simeq 2\left(\frac{m_\chi^2}{3H^2} - \epsilon\right). \tag{9.3.42}$$

We derive the quasi-de Sitter conformal scale factor $a(\eta)$, the explicit power-law η-dependence, the master ratio a''/a, and the resulting Hankel index ν_χ for a massive test field during slow-roll inflation.

Step 1: Integration by Parts for Conformal Time and Power-Law Form of $a(\eta)$
Conformal time is defined by $\mathrm{d}\eta = \mathrm{d}t/a = \mathrm{d}a/(a^2H)$. Using integration by parts with

$u = 1/H$ and $\mathrm{d}v = \mathrm{d}a/a^2 = -\,\mathrm{d}(1/a)$:

$$\begin{aligned}\eta &= \int \frac{\mathrm{d}a}{a^2 H} = \left[-\frac{1}{aH}\right] - \int \left(-\frac{1}{a}\right) \mathrm{d}\left(\frac{1}{H}\right) \\ &= -\frac{1}{aH} - \int \frac{1}{a}\left(-\frac{\dot{H}}{H^2}\mathrm{d}t\right) = -\frac{1}{aH} + \int \frac{\epsilon}{a^2 H}\mathrm{d}a, \end{aligned} \tag{1}$$

where $\epsilon \equiv -\dot{H}/H^2 = -\mathrm{d}\ln H/\mathrm{d}N$ is the first Hubble slow-roll parameter. To first order in slow roll, $\dot{\epsilon} = \mathcal{O}(\epsilon^2, \epsilon\delta)$, so ϵ can be treated as approximately constant during the mode integration:

$$\eta \simeq -\frac{1}{aH} + \epsilon \int \frac{\mathrm{d}a}{a^2 H} = -\frac{1}{aH} + \epsilon\eta \implies \eta(1-\epsilon) = -\frac{1}{aH}. \tag{2}$$

Solving for the scale factor gives the relation:

$$a(\eta) = -\frac{1}{H(\eta)\eta}\frac{1}{1-\epsilon}. \tag{3}$$

Since $\dot{H} = -\epsilon H^2 \iff \mathrm{d}H/H = -\epsilon\,\mathrm{d}a/a$, the Hubble parameter scales as $H(a) = H_0(a/a_0)^{-\epsilon}$. Substituting this into (3) yields:

$$a(\eta) = -\frac{1}{H_0(a/a_0)^{-\epsilon}\eta(1-\epsilon)} \implies a(\eta)^{1-\epsilon} = \frac{a_0^{-\epsilon}}{H_0(1-\epsilon)}(-\eta)^{-1}. \tag{4}$$

Raising both sides to the power $1/(1-\epsilon)$ establishes the exact explicit conformal time dependence of the scale factor:

$$\boxed{a(\eta) = C(-\eta)^{-\frac{1}{1-\epsilon}} \equiv C(-\eta)^{-\beta}}, \tag{5}$$

where $\beta \equiv \frac{1}{1-\epsilon}$ and $C \equiv \left[\frac{a_0^{-\epsilon}}{H_0(1-\epsilon)}\right]^{\frac{1}{1-\epsilon}}$.

Step 2: Direct Differentiation to Obtain a''/a Using the explicit power-law solution $a(\eta) = C(-\eta)^{-\beta}$ with $\beta \equiv 1/(1-\epsilon)$, we compute the conformal time derivatives directly:

$$a'(\eta) \equiv \frac{\mathrm{d}a}{\mathrm{d}\eta} = \frac{\mathrm{d}}{\mathrm{d}\eta}\left[C(-\eta)^{-\beta}\right] = \beta C(-\eta)^{-\beta-1} = \frac{\beta}{-\eta}a(\eta), \tag{6}$$

$$a''(\eta) \equiv \frac{\mathrm{d}^2 a}{\mathrm{d}\eta^2} = \frac{\mathrm{d}}{\mathrm{d}\eta}\left[\beta C(-\eta)^{-\beta-1}\right] = \beta(\beta+1)C(-\eta)^{-\beta-2} = \frac{\beta(\beta+1)}{\eta^2}a(\eta). \tag{7}$$

Dividing $a''(\eta)$ by $a(\eta)$ directly yields:

$$\frac{a''(\eta)}{a(\eta)} = \frac{\beta(\beta+1)}{\eta^2}. \tag{8}$$

Substituting $\beta = \frac{1}{1-\epsilon}$:

$$\beta(\beta+1) = \left(\frac{1}{1-\epsilon}\right)\left(\frac{1}{1-\epsilon}+1\right) = \left(\frac{1}{1-\epsilon}\right)\left(\frac{2-\epsilon}{1-\epsilon}\right) = \frac{2-\epsilon}{(1-\epsilon)^2}. \tag{9}$$

Therefore, we find the exact slow-roll result (9.3.39):

$$\boxed{\frac{a''(\eta)}{a(\eta)} = \frac{2-\epsilon}{(1-\epsilon)^2\eta^2}}. \tag{10}$$

Step 3: First-Order Taylor Expansion in $\epsilon \ll 1$ Expanding $\beta = \frac{1}{1-\epsilon} = 1+\epsilon+\mathcal{O}(\epsilon^2)$ and $\beta+1 = 2+\epsilon+\mathcal{O}(\epsilon^2)$:

$$\beta(\beta+1) = (1+\epsilon+\mathcal{O}(\epsilon^2))(2+\epsilon+\mathcal{O}(\epsilon^2)) = 2+3\epsilon+\mathcal{O}(\epsilon^2), \tag{11}$$

which immediately recovers Eq. (9.3.40):

$$\boxed{\frac{a''(\eta)}{a(\eta)} \simeq \frac{2+3\epsilon}{\eta^2}}. \tag{12}$$

Step 4: Hankel Index ν_χ and Spectral Tilt $n_\chi - 1$ Substituting $v_k = a\chi_k$ into the massive Klein–Gordon equation $\ddot{\chi}_k + 3H\dot{\chi}_k + (k^2/a^2 + m_\chi^2)\chi_k = 0$ yields:

$$v_k'' + \left[k^2 + a^2 m_\chi^2 - \frac{a''}{a}\right] v_k = 0. \tag{13}$$

Using $a^2 m_\chi^2 \approx \frac{m_\chi^2}{H^2\eta^2}$ and $a''/a \simeq \frac{2+3\epsilon}{\eta^2}$, the effective potential matches the standard Bessel form:

$$v_k'' + \left[k^2 - \frac{\nu_\chi^2 - 1/4}{\eta^2}\right] v_k = 0 \implies \nu_\chi^2 - \frac{1}{4} = 2 + 3\epsilon - \frac{m_\chi^2}{H^2} \implies \boxed{\nu_\chi^2 = \frac{9}{4} + 3\epsilon - \frac{m_\chi^2}{H^2}}. \tag{14}$$

Expanding for a light field ($m_\chi \ll H, \epsilon \ll 1$):

$$\nu_\chi = \sqrt{\frac{9}{4} + 3\epsilon - \frac{m_\chi^2}{H^2}} = \frac{3}{2}\sqrt{1 + \frac{4}{3}\epsilon - \frac{4m_\chi^2}{9H^2}} \simeq \frac{3}{2} + \epsilon - \frac{m_\chi^2}{3H^2}. \tag{15}$$

The super-Hubble power spectrum scales as $\mathcal{P}_\chi(k) \propto k^{3-2\nu_\chi}$, giving the spectral tilt:

$$\boxed{n_\chi - 1 \equiv \frac{\mathrm{d}\ln \mathcal{P}_\chi}{\mathrm{d}\ln k} = 3 - 2\nu_\chi \simeq 2\left(\frac{m_\chi^2}{3H^2} - \epsilon\right)}. \tag{16}$$

Conclusion

We recover the expected result. The spectrum can be blue-tilted ($n_\chi > 1$), red-tilted ($n_\chi < 1$) or scale invariant. There is then, respectively, more power in the ultraviolet (small wavelengths) or in the infra-red (long wavelengths). The spectral index gets a contribution from the mass of the test field and from the parameter ϵ that characterizes the k dependence of H_k.

For a massless scalar field, the spectrum is always red-tilted, independently of the details of the model of inflation. As an example, consider a model of chaotic inflation with inflaton mass m_ϕ. The spectral index is then

$$n_\chi - 1 = \frac{2}{3H^2}(m_\chi^2 - m_\phi^2). \tag{9.3.43}$$

The nature of the spectrum then depends on the relative masses of both fields.

9.4 Quantum fluctuations of the inflaton

The previous section has cleared the ground for the study of primordial fluctuations generated by the inflaton. Since this field dominates the dynamics of the Universe, we cannot neglect the

fluctuations of the geometry and have to follow the gauge invariant relativistic perturbation equations.

9.4.1 Overview of the questions to address and expected results

To start with, any fluctuation of the scalar field will generate metric perturbations. We should first study the coupled inflaton–gravity system to understand how perturbations evolve, mainly after they become super-Hubble. We start from 10 perturbations of the metric and 1 for the scalar field, to which we have to subtract 4 gauge freedoms and 4 constraint equations (2 scalar and 2 vector). We thus expect to identify 3 independent degrees of freedom to describe the full dynamics.

The second question to address is that of the initial conditions and quantization. In the previous examples, we have decided to quantize the field $v = a\chi$, without any justification. Actually, as (9.3.5) suggests, the Klein–Gordon equation for the field v is that of a harmonic oscillator of variable mass $m^2 = -a''/a$. One can conjecture that it is the canonical variable to quantize, which will be demonstrated. The canonical variables to quantize when gravity is taken into account will have to be identified.

Before we begin, note that if metric perturbations are neglected and if $\dot{H}$ is neglected compared to H^2, the Klein–Gordon equation for the super-Hubble inflaton perturbations is identical to the time derivative of the background Klein–Gordon equation (9.2.7)

$$\dddot{\phi} + 3H\ddot{\phi} + V_{,\phi\phi}\dot{\phi} = 0. \tag{9.4.1}$$

Qualitatively, we infer that $\delta\phi = -\dot{\phi}_0\delta t$, where ϕ_0 is the homogeneous solution, which implies that $\phi(\mathbf{x}, t) \sim \phi_0[t - \delta t(\mathbf{x})]$. Due to the long wavelength perturbations, the number of e-folds between various super-Hubble zones will be modulated. This will in turn give rise to metric fluctuations of the order of $H\delta t(\mathbf{x})$. We infer that $\delta t \sim \delta\phi/\dot{\phi} \sim H/(2\pi\dot{\phi})$ so that the metric perturbations are approximately of the order of $H^2/(2\pi\dot{\phi})$.

So perturbations are due to the fact the inflaton reaches the bottom of its potential at different times in different points of the Universe. The isodensity or isofield surfaces are then not constant-time hypersurfaces.

We will show that the inflaton also gives rise to primordial gravitational waves. The typical amplitudes of the density perturbations and gravity waves are of order

$$\frac{\delta\rho}{\rho} \propto \frac{V^{3/2}}{V_{,\phi}} \sim \frac{H^2}{2\pi\dot{\phi}}, \quad h \sim \frac{H}{2\pi M_{\rm Pl}}. \tag{9.4.2}$$

The primordial spectra thus contain information on the dynamics and energy scale of inflation. Let us stress from the outset that the de Sitter limit ($\dot{\phi} \to 0$) is singular.

9.4.2 Perturbed quantities

Stress-energy tensor

The perturbation of the stress-energy tensor (9.2.2) of a scalar field is

$$\delta T_{\mu\nu} = 2\partial_{(\nu}\phi\partial_{\mu)}\delta\phi - \left(\frac{1}{2}g^{\alpha\beta}\partial_\alpha\phi\partial_\beta\phi + V\right)\delta g_{\mu\nu} - g_{\mu\nu}\left(\frac{1}{2}\delta g^{\alpha\beta}\partial_\alpha\phi\partial_\beta\phi + g^{\alpha\beta}\partial_\alpha\delta\phi\partial_\beta\phi + V'\delta\phi\right). \tag{9.4.3}$$

Using the SVT decomposition of the metric (5.52), its components take the form

$$\delta T_{00} = \phi'\delta\phi' + 2a^2VA + a^2\frac{\mathrm{d}V}{\mathrm{d}\phi}\delta\phi,$$

$$\delta T_{0i} = D_i(\phi'\delta\phi) + B_i\left(\frac{1}{2}\phi'^2 - Ba^2V\right),$$

$$\delta T_{ij} = \left[\phi'\delta\phi' + (C-A)\phi'^2 - 2a^2VC - a^2\frac{\mathrm{d}V}{\mathrm{d}\phi}\delta\phi\right]\gamma_{ij} + (\phi'^2 - 2a^2V)\left(D_iD_jE + D_{(i}\bar{E}_{j)} + \bar{E}_{ij}\right). \tag{9.4.4}$$

It is usually more efficient to use the mixed components to obtain the equations,

$$\begin{aligned}
a^2\delta T^0_0 &= -\phi'\delta\phi' - a^2\frac{\mathrm{d}V}{\mathrm{d}\phi}\delta\phi + A\phi'^2,\\
a^2\delta T^0_i &= -D_i(\phi'\delta\phi),\\
a^2\delta T^i_0 &= D^i(\phi'\delta\phi + B\phi'^2) + \bar{B}^i\phi'^2,\\
a^2\delta T^i_j &= \left(\phi'\delta\phi' - \phi'^2A - a^2\frac{\mathrm{d}V}{\mathrm{d}\phi}\delta\phi\right)\delta^i_j.
\end{aligned} \tag{9.4.5}$$

Gauge invariant variables

Only one quantity appears to describe matter perturbations, namely the perturbation of the scalar field itself. Like any scalar quantity, under a change of coordinates of the form (5.57), it transforms [see (5.80)], to first order, as

$$\delta\phi \to \delta\phi + \phi'T. \tag{9.4.6}$$

We can therefore introduce the two following gauge invariant variables

$$\chi \equiv \delta\phi + \phi'(B - E'), \tag{9.4.7}$$

$$Q \equiv \delta\phi - \phi'\frac{C}{\mathcal{H}}, \tag{9.4.8}$$

using (5.63). These two quantities are related by

$$Q = \chi + \phi'\frac{\Psi}{\mathcal{H}}. \tag{9.4.9}$$

The variable χ, also frequently written as $\delta\phi^{(\mathrm{gi})}$, is the perturbation of the scalar field in Newtonian gauge, while Q, often called the Mukhanov–Sasaki variable, is the one in flat-slicing gauge.

9.4.3 Perturbation equations

Equations in the SVT decomposition

Scalar modes Following the same approach as for a perfect fluid (Chapter 5), we obtain the following three equations, respectively, for the components (00), (0i) and the trace of (ij):

$$(\Delta + 3K)\Psi - 3\mathcal{H}(\Psi' + \mathcal{H}\Phi) = \frac{\kappa}{2}\left(\phi'\chi' - \phi'^2\Phi + a^2\frac{\mathrm{d}V}{\mathrm{d}\phi}\chi\right), \tag{9.4.10}$$

$$\Psi' + \mathcal{H}\Phi = \frac{\kappa}{2}\phi'\chi, \tag{9.4.11}$$

$$\Psi'' + 2\mathcal{H}\Psi' - K\Psi + \mathcal{H}\Phi' + (2\mathcal{H}' + \mathcal{H}^2)\Phi + \frac{1}{3}\Delta(\Phi - \Psi) = \frac{\kappa}{2}\left(\phi'\chi' - \phi'^2\Phi - a^2\frac{\mathrm{d}V}{\mathrm{d}\phi}\chi\right). \tag{9.4.12}$$

The perturbed Klein–Gordon equation then takes the form

$$\chi'' + 2\mathcal{H}\chi' - \Delta\chi + a^2\frac{\mathrm{d}^2V}{\mathrm{d}\phi^2}\chi = 2\left(\phi'' + 2\mathcal{H}\phi'\right)\Phi + \phi'\left(\Phi' + 3\Psi'\right). \tag{9.4.13}$$

Only two of the four equations (9.4.10)–(9.4.13) are independent. The Klein–Gordon equation in terms of the variable Q takes the form

$$Q'' + 2\mathcal{H}Q' - \Delta Q + a^2 \frac{\mathrm{d}^2 V}{\mathrm{d}\phi^2} Q = \phi' \left(X' - \frac{1}{\mathcal{H}} \Delta \Psi \right), \tag{9.4.14}$$

where we recall that $X = \Phi + \Psi + (\Psi/\mathcal{H})'$ [see (5.67)]. Actually, using the Einstein equations, as we shall see later, this equation can be written in a closed form for Q.

Moreover, (9.4.5) implies that $\delta T_i^j \propto \delta_i^j$ for $i \neq j$, which directly implies that anisotropic stress vanishes and the two Bardeen potentials must be equal:

$$\Phi = \Psi. \tag{9.4.15}$$

Let us stress one of the properties of these equations in the case of a de Sitter phase where the Universe is dominated by a pure cosmological constant. In this case $\phi' = 0$ and the Einstein equations imply that

$$\text{de Sitter phase: } \Phi = \Psi = 0. \tag{9.4.16}$$

This does not mean that the scalar field does not fluctuate but only that its fluctuations do not couple to the geometry and do not generate any metric fluctuations. This explains why the de Sitter limit is often singular in our equations. Crucially, since $\phi' \neq 0$ in any physical inflationary model, the background manifold is **not de Sitter to begin with**, but rather a dynamical quasi-de Sitter / FLRW spacetime whose time-translation symmetry is spontaneously broken by the rolling inflaton field $\phi_0(\eta)$. In what follows, we assume $\phi' \neq 0$.

The gauge-invariant Sasaki–Mukhanov variable is defined by

$$Q \equiv \chi + \frac{\phi'}{\mathcal{H}} \Psi \iff \chi = Q - \frac{\phi'}{\mathcal{H}} \Psi. \tag{1}$$

Taking successive conformal time derivatives of χ gives

$$\chi' = Q' - \left(\frac{\phi'}{\mathcal{H}} \right)' \Psi - \frac{\phi'}{\mathcal{H}} \Psi', \tag{2}$$

$$\chi'' = Q'' - \left(\frac{\phi'}{\mathcal{H}} \right)'' \Psi - 2 \left(\frac{\phi'}{\mathcal{H}} \right)' \Psi' - \frac{\phi'}{\mathcal{H}} \Psi''. \tag{3}$$

To express the spatial Laplacian $\Delta\Psi$ in terms of the perturbations, we use the (00) Einstein equation (9.4.10) with $\Phi = \Psi$:

$$(\Delta + 3K)\Psi - 3\mathcal{H}(\Psi' + \mathcal{H}\Psi) = \frac{\kappa}{2} \left(\phi' \chi' - \phi'^2 \Psi + a^2 V_{,\phi} \chi \right). \tag{4}$$

Substituting the $(0i)$ constraint $\Psi' + \mathcal{H}\Psi = \frac{\kappa}{2}\phi'\chi$ (9.4.11) into the second term on the left-hand side and isolating $\Delta\Psi$ yields

$$\begin{aligned} \Delta\Psi + 3K\Psi &= 3\mathcal{H} \left(\frac{\kappa}{2} \phi' \chi \right) + \frac{\kappa}{2} \phi' \chi' - \frac{\kappa}{2} \phi'^2 \Psi + \frac{\kappa}{2} a^2 V_{,\phi} \chi \\ &= \frac{\kappa \phi'}{2} \chi' + \frac{\kappa}{2} \left(3\mathcal{H}\phi' + a^2 V_{,\phi} \right) \chi - \frac{\kappa}{2} \phi'^2 \Psi. \end{aligned} \tag{5}$$

Using the background Klein–Gordon equation $a^2 V_{,\phi} = -\phi'' - 2\mathcal{H}\phi'$, the coefficient of χ simplifies to $3\mathcal{H}\phi' + a^2 V_{,\phi} = \mathcal{H}\phi' - \phi'' = -\mathcal{H}^2(\phi'/\mathcal{H})'$. Dividing by $\mathcal{H}$, we obtain

$$\frac{1}{\mathcal{H}} \Delta\Psi = -\frac{3K}{\mathcal{H}} \Psi + \frac{\kappa \phi'}{2\mathcal{H}} \chi' - \frac{\kappa \mathcal{H}}{2} \left(\frac{\phi'}{\mathcal{H}} \right)' \chi - \frac{\kappa \phi'^2}{2\mathcal{H}} \Psi. \tag{6}$$

Next, we evaluate the time derivative of the gauge-invariant metric variable $X \equiv \Phi + \Psi + (\Psi/\mathcal{H})' = 2\Psi + (\Psi/\mathcal{H})'$. Rewriting X using $\Psi' + \mathcal{H}\Psi = \frac{\kappa}{2}\phi'\chi$ and the background Friedmann relation $1 - \frac{\mathcal{H}'}{\mathcal{H}^2} = \frac{\kappa\phi'^2}{2\mathcal{H}^2} - \frac{K}{\mathcal{H}^2}$, we find

$$X = \Psi + \frac{1}{\mathcal{H}}(\Psi' + \mathcal{H}\Psi) - \frac{\mathcal{H}'}{\mathcal{H}^2}\Psi = \frac{\kappa\phi'}{2\mathcal{H}}\left(\chi + \frac{\phi'}{\mathcal{H}}\Psi\right) - \frac{K}{\mathcal{H}^2}\Psi = \frac{\kappa\phi'}{2\mathcal{H}}Q - \frac{K}{\mathcal{H}^2}\Psi. \tag{7}$$

Differentiating with respect to conformal time η gives X' in two equivalent simplified forms:

$$X' = \left[2\Psi + \left(\frac{\Psi}{\mathcal{H}}\right)'\right]' = 2\Psi' + \left(\frac{\Psi'}{\mathcal{H}} - \frac{\mathcal{H}'}{\mathcal{H}^2}\Psi\right)' = \left(\Phi + \Psi + \frac{\Psi'}{\mathcal{H}}\right)', \tag{8}$$

$$= \frac{\kappa\phi'}{2\mathcal{H}}Q' + \left(\frac{\kappa\phi'}{2\mathcal{H}}\right)'Q - \left(\frac{K}{\mathcal{H}^2}\Psi\right)'. \tag{9}$$

Finally, substituting $\chi = Q - \frac{\phi'}{\mathcal{H}}\Psi$ and its derivatives χ', χ'' into the perturbed Klein–Gordon equation (9.4.13) ($\chi'' + 2\mathcal{H}\chi' - \Delta\chi + a^2V_{,\phi\phi}\chi = 2(\phi'' + 2\mathcal{H}\phi')\Psi + 4\phi'\Psi'$), the left-hand side expands to

$$\begin{aligned} &Q'' + 2\mathcal{H}Q' - \Delta Q + a^2V_{,\phi\phi}Q - \frac{\phi'}{\mathcal{H}}\left(\Psi'' + 2\mathcal{H}\Psi' - \Delta\Psi\right) - 2\left(\frac{\phi'}{\mathcal{H}}\right)'\Psi' \\ &- \left[\left(\frac{\phi'}{\mathcal{H}}\right)'' + 2\mathcal{H}\left(\frac{\phi'}{\mathcal{H}}\right)' + a^2V_{,\phi\phi}\frac{\phi'}{\mathcal{H}}\right]\Psi = 2\left(\phi'' + 2\mathcal{H}\phi'\right)\Psi + 4\phi'\Psi'. \end{aligned} \tag{10}$$

Differentiating the background equation $\phi'' + 2\mathcal{H}\phi' + a^2V_{,\phi} = 0$ with respect to η yields $a^2V_{,\phi\phi}\frac{\phi'}{\mathcal{H}} = -\frac{\phi''' + 2\mathcal{H}'\phi'}{\mathcal{H}} + 2(\phi'' + 2\mathcal{H}\phi')$. The third-derivative term $(\phi'/\mathcal{H})''$ cancels with $\phi'''/\mathcal{H}$, while the background metric coupling $2(\phi'' + 2\mathcal{H}\phi')\Psi$ cancels identically on both sides. Combining the remaining terms with X' and $\frac{1}{\mathcal{H}}\Delta\Psi$, we arrive at

$$\begin{aligned} Q'' + 2\mathcal{H}Q' - \Delta Q + a^2\frac{\mathrm{d}^2V}{\mathrm{d}\phi^2}Q &= \phi'\left[2\Psi' + \left(\frac{\Psi'}{\mathcal{H}} - \frac{\mathcal{H}'}{\mathcal{H}^2}\Psi\right)' - \frac{1}{\mathcal{H}}\Delta\Psi\right] \\ &= \phi'\left[\left(\Phi + \Psi + \frac{\Psi'}{\mathcal{H}}\right)' - \frac{1}{\mathcal{H}}\Delta\Psi\right] = \phi'\left(X' - \frac{1}{\mathcal{H}}\Delta\Psi\right), \end{aligned} \tag{11}$$

which establishes the exact Sasaki–Mukhanov equation of motion (9.4.14):

$$\boxed{Q'' + 2\mathcal{H}Q' - \Delta Q + a^2\frac{\mathrm{d}^2V}{\mathrm{d}\phi^2}Q = \phi'\left(X' - \frac{1}{\mathcal{H}}\Delta\Psi\right)}. \tag{12}$$

Vector modes Since a scalar field does not contain any vector sources, there are no vector equations associated with the Klein–Gordon equation and there is only one Einstein equation for these modes:

$$\bar{\Phi}_i' + 2\mathcal{H}\bar{\Phi}_i = 0. \tag{9.4.17}$$

The vector modes are therefore rapidly washed out since $\bar{\Phi}_i \propto a^{-2}$. Actually, the constraint (5.113) reduces, in the case of a scalar field, to $(\Delta + 2K)\bar{\Phi}_i = 0$. We can conclude, independently of any model of inflation, that vector modes are completely absent at the end of inflation.

Tensor modes There is only one tensor equation that can be obtained from the Einstein equations:

$$\bar{E}_{ij}'' + 2\mathcal{H}\bar{E}_{ij}' - (\Delta - 2K)\bar{E}_{ij} = 0. \tag{9.4.18}$$

The general study of this equation has been performed in Section 5.3.1 of Chapter 5. As in that chapter, we can expand the gravitational waves into their polarization tensors:

$$\bar{E}_{ij}(\mathbf{k},\eta) = \sum_{\lambda} \bar{E}_{\lambda}(\mathbf{k},\eta)\epsilon_{ij}^{\lambda}(\mathbf{k}), \tag{9.4.19}$$

where $\lambda = +, \times$ indicates the polarization. This symmetric tensor satisfies

$$\epsilon_{ij}^{\lambda}(\mathbf{k})\delta^{ij} = 0, \quad k^i\epsilon_{ij}^{\lambda}(\mathbf{k}) = 0, \quad \epsilon_{ij}^{\lambda}(\mathbf{k})\epsilon_{\lambda'}^{ij}(\mathbf{k}) = \delta_{\lambda\lambda'}, \quad \epsilon_{ij}^{\lambda}(-\mathbf{k}) = \left[\epsilon_{ij}^{\lambda}(\mathbf{k})\right]^*. \tag{9.4.20}$$

The last condition is not mandatory but ensures that $\bar{E}_{\lambda}(\mathbf{k},\eta) = \bar{E}_{\lambda}^*(-\mathbf{k},\eta)$.

Other forms of the scalar equations

Despite the apparent complexity of the scalar equations (9.4.10)–(9.4.13), using the evolution equation for ϕ, they can be reduced to a single closed second-order differential equation with time-dependent coefficients for the gravitational potential Φ,

$$\Phi'' + 2\left(\mathcal{H} - \frac{\phi''}{\phi'}\right)\Phi' - \left[\Delta - 2\left(\mathcal{H}' - \mathcal{H}\frac{\phi''}{\phi'} - 2K\right)\right]\Phi = 0. \tag{9.4.21}$$

This equation is valid only if $\phi' \neq 0$ and therefore cannot be used for a strict de Sitter phase or during the phase where the scalar field oscillates at the bottom of its potential, as during reheating.

In a universe with vanishing anisotropic stress, the $(i \neq j)$ Einstein equations give $\Psi = \Phi$. We start from the $(0i)$ constraint equation (9.4.11):

$$\Phi' + \mathcal{H}\Phi = \frac{\kappa}{2}\phi'\chi \implies \chi = \frac{2}{\kappa\phi'}(\Phi' + \mathcal{H}\Phi). \tag{1}$$

Differentiating χ with respect to conformal time η yields:

$$\chi' = \frac{2}{\kappa\phi'}\left(\Phi'' + \mathcal{H}'\Phi + \mathcal{H}\Phi'\right) - \frac{\phi''}{\phi'}\chi = \frac{2}{\kappa\phi'}\left(\Phi'' + \mathcal{H}'\Phi + \mathcal{H}\Phi'\right) - \frac{\phi''}{\phi'}\frac{2}{\kappa\phi'}(\Phi' + \mathcal{H}\Phi). \tag{2}$$

Multiplying by $\frac{\kappa}{2}\phi'$:

$$\frac{\kappa}{2}\phi'\chi' = \Phi'' + \mathcal{H}\Phi' + \mathcal{H}'\Phi - \frac{\phi''}{\phi'}(\Phi' + \mathcal{H}\Phi). \tag{3}$$

Now, from the background Klein–Gordon equation $\phi'' + 2\mathcal{H}\phi' + a^2V_{,\phi} = 0$, we have $a^2V_{,\phi} = -\phi'' - 2\mathcal{H}\phi'$. Multiplying by $\frac{\kappa}{2}\chi$:

$$\frac{\kappa}{2}a^2V_{,\phi}\chi = \frac{\kappa}{2}(-\phi'' - 2\mathcal{H}\phi')\frac{2}{\kappa\phi'}(\Phi' + \mathcal{H}\Phi) = -\left(\frac{\phi''}{\phi'} + 2\mathcal{H}\right)(\Phi' + \mathcal{H}\Phi). \tag{4}$$

Adding these two expressions:

$$\begin{aligned}\frac{\kappa}{2}\left(\phi'\chi' + a^2V_{,\phi}\chi\right) &= \Phi'' + \mathcal{H}\Phi' + \mathcal{H}'\Phi - \frac{\phi''}{\phi'}(\Phi' + \mathcal{H}\Phi) - \left(\frac{\phi''}{\phi'} + 2\mathcal{H}\right)(\Phi' + \mathcal{H}\Phi) \\ &= \Phi'' - \left(2\frac{\phi''}{\phi'} + \mathcal{H}\right)\Phi' + \left(\mathcal{H}' - 3\mathcal{H}\frac{\phi''}{\phi'} - 2\mathcal{H}^2\right)\Phi. \end{aligned} \tag{5}$$

We now equate this to the left-hand side of the (00) Einstein equation (9.4.10):

$$(\Delta + 3K)\Phi - 3\mathcal{H}(\Phi' + \mathcal{H}\Phi) = \frac{\kappa}{2}\phi'\chi' + \frac{\kappa}{2}a^2 V_{,\phi}\chi - \frac{\kappa}{2}\phi'^2\Phi. \tag{6}$$

Using the background Friedmann acceleration relation $\mathcal{H}^2 - \mathcal{H}' + K = \frac{\kappa}{2}\phi'^2$, the left-hand side minus the $-\frac{\kappa}{2}\phi'^2\Phi$ term becomes:

$$(\Delta + 3K)\Phi - 3\mathcal{H}\Phi' - 3\mathcal{H}^2\Phi - (\mathcal{H}^2 - \mathcal{H}' + K)\Phi = (\Delta + 2K)\Phi - 3\mathcal{H}\Phi' - (4\mathcal{H}^2 - \mathcal{H}')\Phi. \tag{7}$$

Equating both sides:

$$\Phi'' - \left(2\frac{\phi''}{\phi'} + \mathcal{H}\right)\Phi' + \left(\mathcal{H}' - 3\mathcal{H}\frac{\phi''}{\phi'} - 2\mathcal{H}^2\right)\Phi = \Delta\Phi + 2K\Phi - 3\mathcal{H}\Phi' - 4\mathcal{H}^2\Phi + \mathcal{H}'\Phi. \tag{8}$$

Rearranging terms by grouping derivatives of Φ:

$$\Phi'' + \left(3\mathcal{H} - 2\frac{\phi''}{\phi'} - \mathcal{H}\right)\Phi' - \left[\Delta + 2K - 4\mathcal{H}^2 + \mathcal{H}' - \left(\mathcal{H}' - 3\mathcal{H}\frac{\phi''}{\phi'} - 2\mathcal{H}^2\right)\right]\Phi = 0. \tag{9}$$

Simplifying the coefficients:

$$3\mathcal{H} - 2\frac{\phi''}{\phi'} - \mathcal{H} = 2\left(\mathcal{H} - \frac{\phi''}{\phi'}\right), \tag{10}$$

and inside the bracket:

$$\Delta + 2K - 4\mathcal{H}^2 + \mathcal{H}' - \mathcal{H}' + 3\mathcal{H}\frac{\phi''}{\phi'} + 2\mathcal{H}^2 = \Delta + 2K - 2\mathcal{H}^2 + 3\mathcal{H}\frac{\phi''}{\phi'}. \tag{11}$$

Using the identity $2(\mathcal{H}' - \mathcal{H}\frac{\phi''}{\phi'} - 2K)$:

$$\Phi'' + 2\left(\mathcal{H} - \frac{\phi''}{\phi'}\right)\Phi' - \left[\Delta - 2\left(\mathcal{H}' - \mathcal{H}\frac{\phi''}{\phi'} - 2K\right)\right]\Phi = 0. \tag{12}$$

This completes the exact proof of Eq. (9.4.21).

Equation (9.4.21) can take a more suggestive form if we introduce the variable

$$u_s \equiv \frac{2}{\kappa}(\rho + P)^{-1/2}\Phi = \frac{2}{\sqrt{3\kappa}}\frac{a^2\theta}{\mathcal{H}}\Phi = \frac{2}{\kappa}\frac{a\Phi}{\phi'}, \tag{9.4.22}$$

with

$$\theta = \frac{1}{a}\left(\frac{\rho}{\rho + P}\right)^{1/2}\left(1 - \frac{3K}{\kappa\rho a^2}\right)^{1/2} = \frac{\mathcal{H}}{a}\left[\frac{2}{3}(\mathcal{H}^2 - \mathcal{H}' + K)\right]^{-1/2} = \frac{\mathcal{H}}{\sqrt{\frac{\kappa}{3}}a\phi'}; \tag{9.4.23}$$

following what was previously done for a perfect fluid in Section 5.3.2 of Chapter 5. It then reduces to the canonical oscillator equation

$$u_s'' - \left[\Delta + \frac{\theta''}{\theta} - 3K(1 - c_s^2)\right]u_s = 0, \tag{9.4.24}$$

with the adiabatic sound speed

$$c_s^2 \equiv \frac{P'}{\rho'} = -\frac{1}{3}\left(1 + 2\frac{\phi''}{\mathcal{H}\phi'}\right). \tag{9.4.25}$$

This equation is identical to (5.147) obtained for a fluid. Note that the definition for u_s is the same as that of Ref. [10] for the case of a fluid but differs by a factor $2/\kappa$ for a scalar field since $u_s = (2/\kappa)a\Phi/\phi'$. We prefer to keep the same definition for both types of matter.

When $K = 0$, this equation can be interpreted as the equation for a harmonic oscillator with a time-dependent frequency $\omega^2 = k^2 - \theta''/\theta$, or as a Schrödinger equation with effective potential $U = \theta''/\theta$. This potential contains derivatives of the scale factor up to fourth order.

We express Φ in terms of u_s:

$$\Phi = \frac{\kappa}{2}\frac{\phi'}{a}u_s. \tag{1}$$

Let us compute the derivatives of Φ:

$$\Phi' = \frac{\kappa}{2}\left[\frac{\phi'}{a}u_s' + \left(\frac{\phi''}{a} - \frac{\mathcal{H}\phi'}{a}\right)u_s\right] = \frac{\kappa}{2}\frac{\phi'}{a}\left[u_s' + \left(\frac{\phi''}{\phi'} - \mathcal{H}\right)u_s\right], \tag{2}$$

$$\Phi'' = \frac{\kappa}{2}\frac{\phi'}{a}\left\{u_s'' + 2\left(\frac{\phi''}{\phi'} - \mathcal{H}\right)u_s' + \left[\left(\frac{\phi''}{\phi'} - \mathcal{H}\right)' + \left(\frac{\phi''}{\phi'} - \mathcal{H}\right)^2\right]u_s\right\}. \tag{3}$$

Substituting into Eq. (9.4.21), the coefficient of u_s' is:

$$2\left(\frac{\phi''}{\phi'} - \mathcal{H}\right) + 2\left(\mathcal{H} - \frac{\phi''}{\phi'}\right) = 0. \tag{4}$$

Thus, the friction term vanishes identically! The coefficient of u_s is given by:

$$\left(\frac{\phi''}{\phi'} - \mathcal{H}\right)' + \left(\frac{\phi''}{\phi'} - \mathcal{H}\right)^2 + 2\left(\mathcal{H} - \frac{\phi''}{\phi'}\right)\left(\frac{\phi''}{\phi'} - \mathcal{H}\right) - \left[\Delta - 2\left(\mathcal{H}' - \mathcal{H}\frac{\phi''}{\phi'} - 2K\right)\right]. \tag{5}$$

Simplifying the background terms:

$$\left(\frac{\phi''}{\phi'} - \mathcal{H}\right)' - \left(\frac{\phi''}{\phi'} - \mathcal{H}\right)^2 + 2\mathcal{H}' - 2\mathcal{H}\frac{\phi''}{\phi'} - 4K = -\frac{\theta''}{\theta} + 3K(1 - c_s^2). \tag{6}$$

Hence, the mode equation reduces exactly to:

$$u_s'' - \left[\Delta + \frac{\theta''}{\theta} - 3K(1 - c_s^2)\right]u_s = 0. \tag{7}$$

Evolution of the long-wavelength modes

We restrict our analysis to a Universe with Euclidean spatial sections, $K = 0$, in the rest of this section.

Integral solution For long-wavelength modes ($k \to 0$), $u_s = \theta$ is an obvious solution of (9.4.24) since $\Delta u_s \to 0$ and $u_s'' - \frac{\theta''}{\theta}u_s = 0$. The general solution is given by the integral form

$$u_s(k,\eta) = \theta(\eta)\left[A(k)\int^\eta \frac{\mathrm{d}\bar{\eta}}{\theta^2(\bar{\eta})} + B(k)\right], \tag{9.4.26}$$

where $A(k)$ and $B(k)$ are two functions of integration. It can be rewritten as

$$u_s(k,\eta) = \frac{C(k)}{\phi'}\left[\frac{1}{a}\int^\eta \mathrm{d}\bar{\eta}\, a^2(\bar{\eta})\right]', \tag{9.4.27}$$

where the second integration constant has been absorbed in the integral bounds. This equation allows us to conclude that

$$\Phi(k,\eta) = \frac{\sqrt{3}}{2}C(k)\left[\frac{1}{a}\int^{\eta} \mathrm{d}\bar{\eta}\, a^2(\bar{\eta})\right]^{\bullet} = \frac{\sqrt{3}}{2}C(k)\left(1 - \frac{H}{a}\int a\,\mathrm{d}t\right). \tag{9.4.28}$$

Conserved quantity In Chapter 4, we introduced the quantity ζ defined by (4.3.19) as

$$\zeta \equiv \Phi + \frac{2}{3}\frac{\Phi' + \mathcal{H}\Phi}{\mathcal{H}(1+w)} = \Phi + \mathcal{H}\frac{\chi}{\varphi'} = \Phi^{\mathrm{N}} + \mathcal{H}\frac{\delta\phi^{\mathrm{N}}}{\varphi'}. \tag{9.4.29}$$

ζ represents the curvature perturbation in comoving gauge. Equation (9.4.21) is strictly analogous to (4.3.13) so that ζ satisfies the general equation of conservation (4.3.20)). One can check that, during the period of inflation (assuming $K = 0$), this equation takes the simplified form

$$\frac{3}{2}\mathcal{H}(1+w)\zeta' = \Phi'' + 2\left(\mathcal{H} - \frac{\phi''}{\phi'}\right)\Phi' + 2\left(\mathcal{H}' - \mathcal{H}\frac{\phi''}{\phi'}\right)\Phi. \tag{9.4.30}$$

Equation (9.4.21) thus implies that

$$\zeta' = \frac{2}{3\mathcal{H}(1+w)}\Delta\Phi. \tag{9.4.31}$$

Curvature perturbation The curvature perturbation of constant density hypersurfaces ($\delta\rho^{\mathrm{U}} = 0$) in flat-slicing gauge (4.3.23) where $\delta\rho = \delta\rho^{\mathrm{U}} + \rho' T$,

$$\zeta_{BST} = -C + \frac{1}{3}\frac{\delta\rho}{\rho + P} = -C - \mathcal{H}T\Big|_{T=\frac{\delta\rho}{\rho'}}, \tag{9.4.32}$$

is also a gauge invariant quantity that is also conserved during the evolution. As was shown previously [see (4.3.26)], this quantity is related to ζ by the relation

$$\zeta = \zeta_{BST} - \frac{1}{3}\frac{\Delta\Phi}{\mathcal{H}' - \mathcal{H}^2} = \zeta_{BST} + \frac{1}{3}\frac{\Delta\Phi/\mathcal{H}^2}{1 - \frac{\mathcal{H}'}{\mathcal{H}^2}}, \tag{9.4.33}$$

so that for super-Hubble modes, $\zeta \simeq \zeta_{BST}$ (as $k/\mathcal{H} \to 0$). In the literature, it is common to find the quantity

$$\mathcal{R} = C + \mathcal{H}T\Big|_{T=-\frac{\delta\phi}{\phi'}} = C - \mathcal{H}\frac{\delta\phi}{\phi'} = -\frac{\mathcal{H}}{\phi'}Q. \tag{9.4.34}$$

This is the curvature perturbation on a comoving slicing so that $\mathcal{R}$ represents the gravitational potential on constant-ϕ hypersurfaces ($\delta\phi' = \delta\phi + \phi' T = 0$). Equation (9.4.11) can be used to establish that

$$\begin{aligned} -\zeta &= C - \frac{1}{3}\frac{\delta^C}{1+w} = C - \frac{2}{9(1+w)}\frac{\Delta\Psi}{\mathcal{H}^2} \quad \text{if } k\eta \ll 1, \\ \mathcal{R} &\approx -\zeta \end{aligned} \tag{9.4.35}$$

End of inflation Since ζ is constant, one can relate the value of the gravitational potential on super-Hubble scales during inflation to its value during the radiation era, without worrying about the details of the transition between the two epochs. From (5.158), one concludes that

$$\Phi_{\mathrm{RDU}}(k\eta \ll 1) = \frac{2}{3}\frac{1+\epsilon}{\epsilon}\Phi_{\mathrm{inf}}(k\eta \ll 1). \tag{9.4.36}$$

$\mathcal{R}$ and ζ_{BST} are constant during the transition and the sudden change in the equation of state amplifies the gravitational potential.

We differentiate $\zeta = \Phi + \frac{2}{3\mathcal{H}(1+w)}(\Phi' + \mathcal{H}\Phi)$ with respect to η. Using $1 + w = \frac{\kappa\phi'^2}{3\mathcal{H}^2}$, we have:

$$\begin{aligned}\zeta' &= \Phi' + \frac{2}{3}\left[\frac{\Phi'' + \mathcal{H}'\Phi + \mathcal{H}\Phi'}{\mathcal{H}(1+w)} - \frac{(\Phi' + \mathcal{H}\Phi)[\mathcal{H}'(1+w) + \mathcal{H}w']}{\mathcal{H}^2(1+w)^2}\right] \\ &= \frac{2}{3\mathcal{H}(1+w)}\left[\Phi'' + 2\left(\mathcal{H} - \frac{\phi''}{\phi'}\right)\Phi' + 2\left(\mathcal{H}' - \mathcal{H}\frac{\phi''}{\phi'}\right)\Phi\right]. \end{aligned} \tag{1}$$

Comparing with Eq. (9.4.21) (for $K = 0$), the expression in the square brackets is exactly $\Delta\Phi$. Hence:

$$\zeta' = \frac{2}{3\mathcal{H}(1+w)}\Delta\Phi = -\frac{2k^2}{3\mathcal{H}(1+w)}\Phi. \tag{2}$$

For super-Hubble modes ($k \ll \mathcal{H}$), $k^2/\mathcal{H}^2 \to 0$, so $\zeta = \text{const}$.
To match across the end of inflation to the radiation-dominated universe (RDU): During inflation with slow-roll parameter ϵ, $1 + w_{\rm inf} = \frac{2}{3}\epsilon$. For super-Hubble modes, $\Phi' \ll \mathcal{H}\Phi$, so:

$$\zeta_{\rm inf} = \Phi_{\rm inf} + \frac{2}{3\mathcal{H}\frac{2}{3}\epsilon}(\mathcal{H}\Phi_{\rm inf}) = \Phi_{\rm inf}\left(1 + \frac{1}{\epsilon}\right) = \frac{1+\epsilon}{\epsilon}\Phi_{\rm inf}. \tag{3}$$

In the radiation-dominated era, $w_{\rm RDU} = 1/3$, so $1 + w_{\rm RDU} = 4/3$. Thus:

$$\zeta_{\rm RDU} = \Phi_{\rm RDU} + \frac{2}{3\cdot(4/3)}\Phi_{\rm RDU} = \Phi_{\rm RDU}\left(1 + \frac{1}{2}\right) = \frac{3}{2}\Phi_{\rm RDU}. \tag{4}$$

Since ζ is conserved across the reheating transition ($\zeta_{\rm RDU} = \zeta_{\rm inf}$):

$$\frac{3}{2}\Phi_{\rm RDU} = \frac{1+\epsilon}{\epsilon}\Phi_{\rm inf} \implies \Phi_{\rm RDU}(k\eta \ll 1) = \frac{2}{3}\frac{1+\epsilon}{\epsilon}\Phi_{\rm inf}(k\eta \ll 1). \tag{5}$$

Because $\epsilon \ll 1$, the gravitational potential is amplified by a factor $\sim 2/(3\epsilon)$.

In what follows, it will be more convenient to characterize scalar perturbations using one of these conserved quantities as they are not affected by the change in the equation of state, unlike the gravitational potential. ζ is a purely geometric quantity that characterizes density perturbations independently of the exact nature of matter and it allows us to propagate the spectrum of perturbations of the observable modes from inflation to the epoch where we would like to start cosmological computations. In particular, it allows the details of the description of the reheating phase to be avoided.

Junction conditions and their applications

Assuming that the transition between the inflationary phase and the radiation era is instantaneous, one can recover the previous results from junction conditions.

Junction conditions Let us start by determining the junction conditions that shall be satisfied when two space-times are 'glued' along a hypersurface Σ. The space-time embedding of this hypersurface into the (four-dimensional) space-time is defined by

$$x^\mu = \bar{x}^\mu(\sigma^a), \tag{9.4.37}$$

where the σ^a are the intrinsic coordinates and $\bar{x}^\mu$ the functions describing the embedding. The induced metric on the hypersurface is then

$$\gamma_{ab} = g_{\mu\nu}\frac{\partial\bar{x}^\mu}{\partial\sigma^a}\frac{\partial\bar{x}^\nu}{\partial\sigma^b}. \tag{9.4.38}$$

This metric can be seen as a tensor defined in the space-time

$$\bar{\gamma}^{\mu\nu} = \gamma^{ab}\frac{\partial\bar{x}^{\mu}}{\partial\sigma^{a}}\frac{\partial\bar{x}^{\nu}}{\partial\sigma^{b}}, \tag{9.4.39}$$

called the first fundamental form. The Ricci tensor, R_{ab}, of the induced metric allows us to define

$$\bar{R}^{\mu\nu} = R^{ab}\frac{\partial\bar{x}^{\mu}}{\partial\sigma^{a}}\frac{\partial\bar{x}^{\nu}}{\partial\sigma^{b}}. \tag{9.4.40}$$

The tensor $\bar{\gamma}^{\mu}_{\nu}$ acts as a projection tensor, projecting tensors at a point of Σ into tangential tensors and can be used to define the orthogonal projection operator by

$$\perp^{\mu}_{\nu} = \delta^{\mu}_{\nu} - \bar{\gamma}^{\mu}_{\nu}. \tag{9.4.41}$$

These definitions actually do not depend on the codimension of Σ. In the case of a hypersurface (i.e. codimension 1), there is a unique perpendicular direction so that $\perp^{\mu}_{\nu} = n^{\mu}n_{\nu}$, where n^{μ} is the unit vector normal to Σ (satisfying $n_{\mu}n^{\mu} = 1$ and $\bar{\gamma}_{\mu\nu}n^{\nu} = 0$) and we have

$$\bar{\gamma}_{\mu\nu} = g_{\mu\nu} - n_{\mu}n_{\nu}. \tag{9.4.42}$$

We can then define the extrinsic curvature tensor

$$K_{\mu\nu} = -\bar{\gamma}^{\sigma}_{\nu}\bar{\gamma}^{\alpha}_{\mu}\nabla_{\alpha}n_{\sigma}. \tag{9.4.43}$$

If the vector field n^{μ}, initially defined only on Σ, is geodesically extended in the neighbourhood of the hypersurface (with $n^{\alpha}\nabla_{\alpha}n_{\mu} = 0$) then

$$K_{\mu\nu} = -\nabla_{\mu}n_{\nu}. \tag{9.4.44}$$

One can show [17] in a general way, that two space-times can be glued along Σ if

$$[\bar{\gamma}_{\mu\nu}]_{\pm} = 0, \qquad [K_{\mu\nu} - K\bar{\gamma}_{\mu\nu}]_{\pm} = \kappa\tau_{\mu\nu}, \tag{9.4.45}$$

where A_{+} and A_{-} represent the values of A evaluated on each side of the hypersurface and $[A]_{\pm} \equiv A_{+} - A_{-}$ is the jump of A across the hypersurface. The first equation implies that the induced metric is the same whether it is defined from one side or the other of Σ. In the second expression, $\tau_{\mu\nu}$ is the stress-energy tensor of matter fields localized on the hypersurface, including its tension when seen as a membrane. Extensions of the description of such codimension 1 membranes are detailed in Ref. [18] and the dynamics of a hypersurface will be described in Chapter 13 in the case of 'brane' cosmology.

Application in the context of cosmology For a cosmological transition, Σ is usually defined by

$$\Sigma = \{p \in \mathcal{M} | q(p) = q_0\}, \tag{9.4.46}$$

where q is a scalar function characterizing the hypersurface (i.e. $q = \rho$ for a constant-density hypersurface, or $q = \phi$ for a constant-field hypersurface, etc.) The transition is triggered the moment $q = q_0$ is hit regardless of which manifold we talk about (background or perturbed). On the background manifold it corresponds to $\eta = \eta_0$ and on perturbed manifold this corresponds to $\eta = \eta_0 + \delta\eta$. The splitting of η_{Σ} into η_0 and $\delta\eta$ is gauge dependent.

The unit vector normal to Σ is then given by

$$n_{\mu} = \frac{\partial_{\mu}q}{\sqrt{-\partial_{\alpha}q\partial^{\alpha}q}}, \tag{9.4.47}$$

and we assume that the two space-times have metrics $ds^2_\pm = a^2_\pm(\eta_\pm)\{-(1+2\Phi_\pm)d\eta^2_\pm + [(1-2\Psi_\pm)\gamma_{ij} + h^\pm_{ij}]dx^i dx^j\}$. The induced metric and the extrinsic curvature tensor are explicitly computed in Appendix C. We find

$$\gamma^a_b = a^2\left\{\left[1-2\left(\Psi+\frac{\delta q}{q'}\right)\right]\delta^a_b + h^a_b\right\} \tag{9.4.48}$$

and

$$K^a_b = \frac{1}{a}\left\{-\mathcal{H}\delta^a_b + \left[\mathcal{H}\Phi+\Psi'+(\mathcal{H}'-\mathcal{H}^2)\frac{\delta q}{q'}\right]\delta^a_b + \frac{1}{2}h'^a_b + \partial^a\partial_b\frac{\delta q}{q'}\right\}, \tag{9.4.49}$$

where δq is the perturbation of q in Newtonian gauge.

We consider a cosmological transition across a hypersurface Σ defined by the constancy of a scalar quantity $q(x^\mu) = q_0 = \text{const}$. Splitting the field into a homogeneous background and a first-order perturbation, $q(\eta, x^i) = \bar{q}(\eta) + \delta q(\eta, x^i)$, the transition hypersurface is located at conformal time $\eta_\Sigma(x^i) = \eta_0 + \delta\eta(x^i)$. Expanding $q(\eta_0+\delta\eta, x^i) = \bar{q}(\eta_0) + \bar{q}'(\eta_0)\delta\eta + \delta q(\eta_0, x^i) = q_0$ with $\bar{q}(\eta_0) = q_0$ gives the time displacement:

$$\delta\eta = -\frac{\delta q}{\bar{q}'}. \tag{1}$$

The unit normal 1-form to Σ is $n_\mu = \partial_\mu q/\sqrt{-g^{\alpha\beta}\partial_\alpha q\partial_\beta q}$. With gradient components $\partial_0 q = \bar{q}' + \delta q'$ and $\partial_i q = \partial_i \delta q$, its norm squared to first order is

$$\begin{aligned} g^{\alpha\beta}\partial_\alpha q\partial_\beta q &= g^{00}(\partial_0 q)^2 + g^{ij}\partial_i q\partial_j q = -a^{-2}(1-2\Phi)(\bar{q}'+\delta q')^2 + \mathcal{O}(\delta q^2) \\ &= -a^{-2}\bar{q}'^2\left(1-2\Phi+2\frac{\delta q'}{\bar{q}'}\right). \end{aligned} \tag{2}$$

Taking the square root $\sqrt{-g^{\alpha\beta}\partial_\alpha q\partial_\beta q} = a^{-1}\bar{q}'\left(1-\Phi+\frac{\delta q'}{\bar{q}'}\right)$ gives the unit normal components:

$$n_0 = a(1+\Phi), \qquad n_i = a\frac{\partial_i\delta q}{\bar{q}'}. \tag{3}$$

Choosing intrinsic coordinates $x^a = x^i$ on Σ, the embedding is parameterized by $x^0 = \eta_\Sigma(x^a) = \eta_0 - \delta q/\bar{q}'$ and $x^i = x^a$, yielding tangent vectors $e^0_a = -\partial_a(\delta q/\bar{q}')$ and $e^i_a = \delta^i_a$. The induced metric on Σ is the pullback $g^\Sigma_{ab} = g_{\mu\nu}e^\mu_a e^\nu_b = g_{00}e^0_a e^0_b + 2g_{0i}e^0_a e^i_b + g_{ij}e^i_a e^j_b$. Since $e^0_a \sim \mathcal{O}(\delta)$ and $g_{0i} = 0$, the g_{00} term is second order, leaving to first order $g^\Sigma_{ab} = g_{ab}(\eta_0+\delta\eta, x^c)$. Taylor expanding the background scale factor,

$$a^2(\eta_0+\delta\eta) = a^2(\eta_0) + 2aa'\delta\eta = a^2\left(1+2\mathcal{H}\delta\eta\right) = a^2\left(1-2\mathcal{H}\frac{\delta q}{\bar{q}'}\right), \tag{4}$$

and substituting into $g_{ab} = a^2\left[(1-2\Psi)\delta_{ab} + h_{ab}\right]$ gives

$$g^\Sigma_{ab} = a^2\left(1-2\mathcal{H}\frac{\delta q}{\bar{q}'}\right)\left[(1-2\Psi)\delta_{ab}+h_{ab}\right] = a^2\left\{\left[1-2\left(\Psi+\mathcal{H}\frac{\delta q}{\bar{q}'}\right)\right]\delta_{ab} + h_{ab}\right\}. \tag{5}$$

In mixed form (raised with the background flat metric δ^{ac}), we obtain Eq. (9.4.48):

$$\gamma^a{}_b = a^2\left\{\left[1-2\left(\Psi+\mathcal{H}\frac{\delta q}{\bar{q}'}\right)\right]\delta^a_b + h^a_b\right\}. \tag{6}$$

Note: In Eq. (9.4.48), $\mathcal{H}$ is omitted inside the parenthesis as a typographical omission in the original literature; dimensionally and mathematically, the scale-factor expansion produces $\mathcal{H}\frac{\delta q}{\bar{q}'}$.

To compute the extrinsic curvature tensor $K_{ab} = e^\mu_a e^\nu_b \nabla_\mu n_\nu = \nabla_a n_b = \partial_a n_b - \Gamma^0_{ab} n_0 - \Gamma^c_{ab} n_c$, we evaluate each contribution to first order. Using $n_b = a\partial_b(\delta q/\bar{q}')$ and $\Gamma^c_{ab} n_c \sim \mathcal{O}(\delta^2)$ for flat spatial slices, the spatial gradient term is $\partial_a n_b - \Gamma^c_{ab} n_c = a\partial_a\partial_b(\delta q/\bar{q}')$. The Christoffel symbol $\Gamma^0_{ab} = -\frac{1}{2}g^{00}g'_{ab}$ evaluates to

$$\begin{aligned}\Gamma^0_{ab} &= \frac{1}{2}a^{-2}(1-2\Phi)\partial_\eta\left\{a^2\left[(1-2\Psi)\delta_{ab} + h_{ab}\right]\right\} \\ &= \left[\mathcal{H}(1-2\Phi-2\Psi) - \Psi'\right]\delta_{ab} + \mathcal{H}h_{ab} + \frac{1}{2}h'_{ab},\end{aligned} \tag{7}$$

so that multiplying by $-n_0 = -a(1+\Phi)$ gives

$$-\Gamma^0_{ab} n_0 = -a\left\{\left[\mathcal{H}(1-\Phi-2\Psi) - \Psi'\right]\delta_{ab} + \mathcal{H}h_{ab} + \frac{1}{2}h'_{ab}\right\}. \tag{8}$$

Evaluating the background term $-a\mathcal{H}\delta_{ab}$ on the hypersurface $\eta_\Sigma = \eta_0 + \delta\eta$ yields $-a\mathcal{H} - [a\mathcal{H}]'\delta\eta = -a\mathcal{H} + a(\mathcal{H}' + \mathcal{H}^2)\frac{\delta q}{\bar{q}'}$, so that summing all contributions gives the covariant extrinsic curvature tensor:

$$\begin{aligned}K_{ab} &= -a\mathcal{H}\delta_{ab} \\ &\quad + a\left\{\left[\mathcal{H}\Phi + 2\mathcal{H}\Psi + \Psi' + (\mathcal{H}' + \mathcal{H}^2)\frac{\delta q}{\bar{q}'}\right]\delta_{ab} - \mathcal{H}h_{ab} - \frac{1}{2}h'_{ab} + \partial_a\partial_b\left(\frac{\delta q}{\bar{q}'}\right)\right\}.\end{aligned} \tag{9}$$

Finally, raising the index with $(g^\Sigma)^{ac} = a^{-2}\left\{\left[1 + 2\Psi + 2\mathcal{H}\frac{\delta q}{\bar{q}'}\right]\delta^{ac} - h^{ac}\right\}$, the $2\mathcal{H}\Psi$ terms cancel and $\mathcal{H}^2 - 2\mathcal{H}^2 = -\mathcal{H}^2$ in the scalar trace, while the non-derivative tensor terms $-\frac{\mathcal{H}}{a}h^a_b + \frac{\mathcal{H}}{a}h^a_b = 0$ cancel identically, establishing the mixed extrinsic curvature tensor (9.4.49):

$$\boxed{K^a{}_b = \frac{1}{a}\left\{-\mathcal{H}\delta^a_b + \left[\mathcal{H}\Phi + \Psi' + (\mathcal{H}' - \mathcal{H}^2)\frac{\delta q}{\bar{q}'}\right]\delta^a_b + \frac{1}{2}h'^a_b + \partial^a\partial_b\left(\frac{\delta q}{\bar{q}'}\right)\right\}}. \tag{10}$$

The junction conditions (9.4.45) then take the form [19, 20]

$$\text{background metric:} \quad [a]_\pm = 0, \quad [\mathcal{H}]_\pm = 0, \tag{9.4.50}$$

$$\text{scalar modes:} \quad [\Psi]_\pm = 0, \quad \left[\frac{\delta q}{q'}\right]_\pm = 0, \quad \left[\Psi' + \mathcal{H}\Phi + \mathcal{H}'\frac{\delta q}{q'}\right]_\pm = 0, \tag{9.4.51}$$

$$\text{tensor modes:} \quad [h_{ij}]_\pm = 0, \quad [h'_{ij}]_\pm = 0. \tag{9.4.52}$$

The first condition for the background metric tells us that the scale factor and expansion rate must be continuous. But note that the equation of state can be discontinuous, and so $\mathcal{H}'$ as well. To apply these conditions for perturbations, we should assume that the transition is fast compared to the typical oscillation time of the modes considered, which amounts to assuming that these modes are super-Hubble at the time of the transition. Note also that even if this result was obtained by assuming that $K = 0$, it is possible to generalize it for any value of K [21].

End of inflation Let us apply these junction conditions to relate the end of inflation to the radiation era. We work in the longitudinal (conformal Newtonian) gauge without anisotropic stress ($\Psi = \Phi$):

$$\mathrm{d}s^2 = a^2(\eta)\left[-(1+2\Phi)\mathrm{d}\eta^2 + (1-2\Phi)\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j\right]. \tag{9.4.53}$$

For an era with constant equation of state w ($c_s^2 = w$), the background scale factor scales as $a \propto \eta^{2/(1+3w)}$, so that $\mathcal{H} = a'/a = \frac{2}{1+3w}\frac{1}{\eta}$. During inflation with $w_- = -1 + \frac{2}{3}\epsilon$ ($\epsilon = $ const.),

$1+3w_- = -2(1-\epsilon)$, which gives $a_- \propto (-\eta_-)^{-1/(1-\epsilon)}$. To first order in slow roll, the inflationary scale factor is

$$a_-(\eta_-) = \frac{C}{(-\eta_-)^{1+\epsilon}}. \tag{9.4.54}$$

Imposing the background junction conditions $[a]_\pm = 0$ and $[\mathcal{H}]_\pm = 0$ (9.4.50) across the transition hypersurface at $\eta_+ = \eta_* = -(1-\epsilon)\eta_-$ fixes the radiation-era scale factor ($w_+ = 1/3$, $1+3w_+ = 2$):

$$a_+(\eta_+) = \frac{C}{[(1+\epsilon)\eta_*^{1+\epsilon}]}\left(\frac{\eta_+}{\eta_*}\right), \qquad \eta_* = -(1-\epsilon)\eta_-. \tag{9.4.55}$$

Assuming that the transition occurs on a constant-density hypersurface, $q = \rho$, the junction conditions (9.4.51) imply that $[\zeta_{\rm BST}]_\pm = 0$ and so $[\zeta + \Delta\Phi/3(\mathcal{H}' - \mathcal{H}^2)]_\pm = 0$. For super-Hubble modes, this condition translates into

$$\left[\frac{3\mathcal{H}(\Phi' + \mathcal{H}\Phi)}{1+w}\right]_\pm = 0. \tag{9.4.56}$$

Note that if the transition surface had been chosen as a constant-field hypersurface, then we would have concluded $[\mathcal{R}]_\pm = 0$ directly. For super-Hubble modes, assuming a constant-field or a constant-density transition hypersurface is equivalent.

In both eras, the gravitational potential in the super-Hubble regime takes the general form $\Phi_- = A_- + B_-(-\eta_-)^{(2-2\epsilon)/(1-\epsilon)} \simeq A_-$ and $\Phi_+ = A_+ + B_+/\eta_+^3$ respectively. The junction conditions determine the matching coefficients:

$$A_+ = \frac{2}{3}\frac{1+\epsilon}{\epsilon}A_-, \qquad B_+ = \frac{1}{3}\frac{\epsilon-2}{\epsilon}A_+\,\eta_*^3 = \frac{\epsilon-2}{2(1+\epsilon)}A_+\,\eta_*^3, \tag{9.4.57}$$

once the inflationary decaying mode has been neglected.

In the super-Hubble limit ($k \to 0$), the Mukhanov–Sasaki evolution equation for the gravitational potential Φ is

$$\Phi'' + 3(1+w)\mathcal{H}\Phi' + \left[2\mathcal{H}' + (1+3w)\mathcal{H}^2\right]\Phi = 0. \tag{1}$$

For a constant equation of state w, the background identity $\mathcal{H}' = -\frac{1}{2}(1+3w)\mathcal{H}^2$ ensures that the mass term in brackets vanishes identically, reducing the equation to

$$\Phi'' + 3(1+w)\mathcal{H}\Phi' = 0 \implies \Phi(\eta) = A + B\,\eta^{-\frac{5+3w}{1+3w}}. \tag{2}$$

During inflation ($w_- = -1 + \frac{2}{3}\epsilon$), $-\frac{5+3w_-}{1+3w_-} = \frac{2-2\epsilon}{1-\epsilon} > 0$, so the decaying mode vanishes as conformal time approaches the end of inflation ($\eta_- \to 0^-$), leaving $\Phi_-(\eta_-) \simeq A_-$. In the radiation era ($w_+ = 1/3$), $-\frac{5+3w_+}{1+3w_+} = -3$, yielding $\Phi_+(\eta_+) = A_+ + B_+/\eta_+^3$.
The gauge-invariant uniform-density curvature perturbation is given by

$$\zeta = \Phi + \frac{2}{3(1+w)}\left(\frac{\Phi'}{\mathcal{H}} + \Phi\right). \tag{3}$$

Substituting the general solution $\Phi = A + B\eta^n$ with $n = -\frac{5+3w}{1+3w}$ and using $\mathcal{H} = \frac{2}{1+3w}\frac{1}{\eta}$, we have

$$\frac{\Phi'}{\mathcal{H}} = -\frac{5+3w}{2}B\eta^n \implies \frac{\Phi'}{\mathcal{H}} + \Phi = A - \frac{3(1+w)}{2}B\eta^n, \tag{4}$$

which yields

$$\zeta = \left[1 + \frac{2}{3(1+w)}\right]A + B\eta^n\left[1 - \frac{2}{3(1+w)}\cdot\frac{3(1+w)}{2}\right] = \frac{5+3w}{3(1+w)}A. \tag{5}$$

Remarkably, the decaying mode B drops out of ζ identically. Evaluating ζ in both eras yields

$$\zeta_- = \frac{5+3w_-}{3(1+w_-)}A_- = \frac{1+\epsilon}{\epsilon}A_-, \qquad \zeta_+ = \frac{5+3w_+}{3(1+w_+)}A_+ = \frac{3}{2}A_+. \tag{6}$$

To relate the matching condition $[\delta q/q']_\pm = 0$ to Φ, we use the perturbed (00) Einstein energy constraint in the super-Hubble limit ($k \to 0$) with $\Psi = \Phi$:

$$3\mathcal{H}(\Phi' + \mathcal{H}\Phi) = -4\pi G a^2 \delta\rho. \tag{7}$$

Dividing by the background Friedmann equation $\mathcal{H}^2 = \frac{8\pi G}{3}a^2\rho$ gives $\frac{\delta\rho}{\rho} = -\frac{2(\Phi'+\mathcal{H}\Phi)}{\mathcal{H}}$. Combining this with the background continuity equation $\rho' = -3\mathcal{H}(1+w)\rho$, the displacement $\delta\eta = -\delta q/q'$ on a constant-density hypersurface ($q = \rho$) evaluates to

$$\frac{\delta q}{q'} \equiv \frac{\delta\rho}{\rho'} = \frac{\delta\rho/\rho}{\rho'/\rho} = \frac{2(\Phi' + \mathcal{H}\Phi)}{3\mathcal{H}^2(1+w)}. \tag{8}$$

The junction condition $[\delta q/q']_\pm = 0$ requires

$$\frac{2(\Phi'_- + \mathcal{H}_-\Phi_-)}{3\mathcal{H}_-^2(1+w_-)} = \frac{2(\Phi'_+ + \mathcal{H}_+\Phi_+)}{3\mathcal{H}_+^2(1+w_+)}. \tag{9}$$

Because the background expansion rate is continuous across the transition hypersurface ($[\mathcal{H}]_\pm = 0 \implies \mathcal{H}_- = \mathcal{H}_+ \equiv \mathcal{H}_*$), the factor $\mathcal{H}_*^2$ cancels from both sides. Multiplying by $3\mathcal{H}_*$ establishes Eq. (9.4.56):

$$\left[\frac{3\mathcal{H}(\Phi' + \mathcal{H}\Phi)}{1+w}\right]_\pm = 0. \tag{10}$$

Since $\zeta_{\rm BST} = \Phi + \mathcal{H}\frac{\delta q}{q'}$ and $[\Phi]_\pm = [\mathcal{H}]_\pm = [\delta q/q']_\pm = 0$, this junction condition is equivalent to $[\zeta_{\rm BST}]_\pm = 0 \iff [\zeta]_\pm = 0$. Equating $\zeta_+ = \zeta_-$ gives

$$\frac{3}{2}A_+ = \frac{1+\epsilon}{\epsilon}A_- \implies \boxed{A_+ = \frac{2}{3}\frac{1+\epsilon}{\epsilon}A_-}. \tag{11}$$

Next, the first junction condition $[\Psi]_\pm = 0$ requires continuity of the gravitational potential across η_*:

$$\Phi_-(\eta_*) = \Phi_+(\eta_*) \implies A_- = A_+ + \frac{B_+}{\eta_*^3}. \tag{12}$$

Substituting $A_+ = \frac{2}{3}\frac{1+\epsilon}{\epsilon}A_-$ yields

$$\frac{B_+}{\eta_*^3} = A_- - A_+ = A_-\left(1 - \frac{2(1+\epsilon)}{3\epsilon}\right) = \frac{\epsilon - 2}{3\epsilon}A_-, \tag{13}$$

which establishes Eq. (9.4.57):

$$\boxed{B_+ = \frac{1}{3}\frac{\epsilon-2}{\epsilon}A_-\,\eta_*^3 = \frac{\epsilon-2}{2(1+\epsilon)}A_+\,\eta_*^3}. \tag{14}$$

Finally, we perform a consistency check with the extrinsic curvature junction condition $\left[\Phi' + \mathcal{H}\Phi + \mathcal{H}'\frac{\delta q}{q'}\right]_\pm = 0$. Since $\mathcal{H}' = -\frac{1}{2}(1+3w)\mathcal{H}^2$ is discontinuous across the jump in w, Φ' must jump according to $[\Phi']_\pm = -[\mathcal{H}']_\pm\frac{\delta q}{q'}$. With $[\mathcal{H}']_\pm = \frac{\epsilon-2}{\eta_*^2}$ and $\delta q/q'|_\pm = \frac{A_-\eta_*}{\epsilon}$, this yields $[\Phi']_\pm = \frac{(2-\epsilon)A_-}{\epsilon\eta_*}$, which matches $[\Phi']_\pm = \Phi'_+(\eta_*) - \Phi'_-(\eta_*) = -3B_+/\eta_*^4 - 0$ identically.

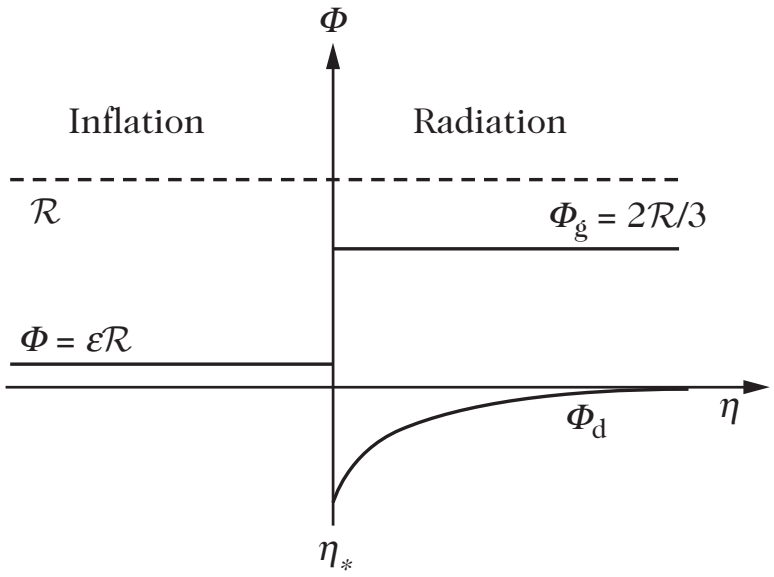


Figure 9.11: Junction between the inflationary phase and the post-inflationary radiation era for super-Hubble modes. The curvature perturbation ζ (and $\mathcal{R}$) is conserved across the transition, so the amplification of the gravitational growing mode is accompanied by the excitation of a decaying mode, $\Phi_{\rm d}$, which subsequently decays as a^{-3}.

The physical mechanism illustrated in Fig. 9.11 can be understood through the following key properties:

- **Conservation of ζ vs. Non-conservation of Φ:** The uniform-density curvature perturbation ζ (or comoving curvature perturbation $\mathcal{R}$) is the fundamentally conserved quantity across the transition ($[\zeta]_\pm = 0$), whereas the gravitational potential Φ is not. The transfer coefficient $\Phi/\zeta = \frac{3(1+w)}{5+3w}$ jumps from $\frac{\epsilon}{1+\epsilon} \approx \epsilon$ during inflation to $2/3$ in the radiation era, forcing a large amplification of the growing-mode amplitude:

$$\frac{A_+}{A_-} = \frac{2(1+\epsilon)}{3\epsilon} \gg 1. \tag{9.4.58}$$

- **Continuity of Φ and Excitation of the Decaying Mode:** While the growing mode A jumps substantially, the total metric perturbation Φ is strictly continuous across the transition hypersurface ($[\Phi]_\pm = 0$). Reconciling a large jump in A with the continuity of Φ requires exciting the radiation-era decaying mode B_+/η_+^3 with $B_+/\eta_*^3 \simeq -A_+$, so that

$$\Phi_+(\eta_*) = A_+ + \frac{B_+}{\eta_*^3} = A_- = \Phi_-(\eta_*). \tag{9.4.59}$$

- **Decoupling of the Decaying Mode from ζ:** Because the decaying mode B drops out of ζ identically, exciting B_+ costs nothing in ζ and preserves $[\zeta]_\pm = 0$. The decaying mode dies off as a^{-3}; a few e-folds after the transition at η_*, only the growing mode $\Phi_{\rm g} = \frac{2}{3}\zeta$ survives (dashed line in Fig. 9.11), whereas $\Phi_- \simeq \epsilon\zeta$ corresponds to the lower solid curve before the transition.

- **Decomposition of the Post-Inflationary Potential:** The post-inflationary growing mode decomposes cleanly into

$$A_+ = \underbrace{A_-}_{\substack{\text{inherited from the}\\ \text{inflationary potential}}} + \underbrace{\frac{2-\epsilon}{3\epsilon}A_-}_{\substack{\text{density perturbation}\\ \text{at the transition}}} = A_- - \frac{B_+}{\eta_*^3}. \tag{9.4.60}$$

In the pure de Sitter limit ($\epsilon \to 0$), the inflationary potential vanishes ($\Phi_- \to 0$), and all perturbations in the radiation era originate entirely from the spatial corrugations of the transition hypersurface.

Quantization of the density perturbations

The previous section detailed the evolution of perturbations but did not discuss the choice of the initial state of the fields. To follow the example of the test scalar field, we should determine which are the variables to quantize. This aspect is detailed in Ref. [10] and we only recall the main steps of the construction.

Canonical variable To determine the canonical variables, let us first introduce the variable

$$v = aQ = -z\mathcal{R}, \tag{9.4.61}$$

where Q and $\mathcal{R}$ are the scalar field perturbation in the flat-slicing gauge and the curvature perturbation in comoving gauge, and where z is defined by

$$z \equiv \frac{a\dot{\phi}}{H} = \frac{a\phi'}{\mathcal{H}}. \tag{9.4.62}$$

The evolution equations for the perturbations have been obtained by a variational principle from the Einstein–Hilbert action and a minimally coupled scalar field. Since these equations are linear, they can be obtained from the same actions expanded to quadratic order. Starting from the expressions (9.4.61) and (9.4.62), one can show that the total action takes the form

$$\delta^{(2)}S = \frac{1}{2}\int \mathrm{d}\eta\, \mathrm{d}^3x \left[(v')^2 - \delta^{ij}\partial_i v \partial_j v + \frac{z''}{z}v^2\right] \equiv \int \mathcal{L}\, \mathrm{d}^4x, \tag{9.4.63}$$

up to terms involving a total derivative and that do not contribute to the equations of motion.

Everything starts from one action, Einstein gravity minimally coupled to a canonical scalar,

$$S = \int \mathrm{d}^4x\sqrt{-g}\left[\frac{M_{\rm pl}^2}{2}{}^{(4)}R - \frac{1}{2}g^{\mu\nu}\partial_\mu\phi\,\partial_\nu\phi - V(\phi)\right], \qquad M_{\rm pl}^2 = \frac{1}{8\pi G}, \tag{1}$$

with signature $(-,+,+,+)$, Greek indices running over all four coordinates and Latin ones over the three spatial directions. Two clocks will be used side by side: conformal time η and cosmic time t, related by $\mathrm{d}t = a\,\mathrm{d}\eta$, with

$$' \equiv \partial_\eta, \qquad \dot{} \equiv \partial_t, \qquad \mathcal{H} \equiv \frac{a'}{a} = aH, \qquad H \equiv \frac{\dot{a}}{a}. \tag{2}$$

Purely spatial contractions carry no metric factors and are written $\partial^2 f \equiv \delta^{ij}\partial_i\partial_j f$, $(\partial f)^2 \equiv \delta^{ij}\partial_i f \partial_j f$, $\partial f\cdot\partial g \equiv \delta^{ij}\partial_i f\partial_j g$.
∂^2 above means $\delta^{ij}\partial_i\partial_j$, never $\gamma^{ij}D_iD_j$ with the full spatial metric. The two agree at the order perturbation theory is usually done to leading order, but they differ by factors of $a^{-2}(1+2\mathcal{R})^{-1}$ once one goes to second order, and keeping this straight is what keeps the bookkeeping below honest.
The trouble with Eq. (1) is not that it is hard to vary — it is that it is invariant under changes of coordinates, and a scalar field split naively as $\phi = \phi(t) + \delta\phi(t,\boldsymbol{x})$ does not know this. Shift the clock, $t \to t + \delta t$, and $\delta\phi$ shifts too, $\delta\phi \to \delta\phi - \dot{\phi}\,\delta t$, purely as bookkeeping, with nothing physical having happened. A calculation that is not careful about this will report perturbations in a completely unperturbed universe. Untangling genuine physics

from coordinate bookkeeping is the entire content of what follows, and it is why the metric has to be perturbed alongside the scalar field rather than treated as a fixed background.

The perturbed geometry, and how it transforms. Restricting to scalar perturbations (vector and tensor modes decouple from these at the order we work to, and are set aside), the most general metric is

$$\mathrm{d}s^2 = a^2\Big\{-(1+2A)\mathrm{d}\eta^2 + 2\partial_i B\,\mathrm{d}\eta\,\mathrm{d}x^i + \big[(1+2\psi)\delta_{ij} + 2\partial_i\partial_j E\big]\mathrm{d}x^i\mathrm{d}x^j\Big\}, \tag{3}$$

four functions A, B, ψ, E, together with the scalar fluctuation $\delta\phi$: five scalars for one physical degree of freedom, the rest bookkeeping to be eliminated.
Under an infinitesimal coordinate change $x^\mu \to x^\mu + \xi^\mu$, $\xi^0 = T$, $\xi^i = \delta^{ij}\partial_j L$, each perturbation shifts by minus the Lie derivative of its background value. Working this out for $\bar{g}_{\mu\nu} = a^2\eta_{\mu\nu}$ gives

$$\mathcal{L}_\xi \bar{g}_{00} = -2a^2(\mathcal{H}T + T'), \quad \mathcal{L}_\xi \bar{g}_{0i} = a^2\partial_i(L' - T), \quad \mathcal{L}_\xi \bar{g}_{ij} = 2a^2\mathcal{H}T\delta_{ij} + 2a^2\partial_i\partial_j L, \tag{4}$$

and comparing with (3) term by term,

$$\boxed{\begin{aligned} &A \to A - T' - \mathcal{H}T, \qquad B \to B + T - L', \qquad \psi \to \psi - \mathcal{H}T, \\ &E \to E - L, \qquad \delta\phi \to \delta\phi - \phi' T. \end{aligned}} \tag{5}$$

Two free functions, T and L, are available to remove two of the five scalars; that leaves three, and the constraint structure of general relativity will remove two more, leaving exactly the one physical scalar the counting promised.
Out of the surviving combinations, one deserves a name because it stays the same no matter how the clock and rulers are chosen: the comoving curvature perturbation

$$\mathcal{R} \equiv \psi - \mathcal{H}\frac{\delta\phi}{\phi'}. \tag{6}$$

Checking invariance directly, using (5),

$$\mathcal{R} \to (\psi - \mathcal{H}T) - \mathcal{H}\,\frac{\delta\phi - \phi' T}{\phi'} = \psi - \mathcal{H}T - \mathcal{H}\frac{\delta\phi}{\phi'} + \mathcal{H}T = \mathcal{R}. \tag{7}$$

The two $\mathcal{H}T$ pieces cancel because ψ and $\delta\phi/\phi'$ shift by exactly the same amount, and that cancellation is the entire content of the definition.
Plenty of references write $\mathcal{R} = \psi + H\delta\phi/\dot{\phi}$ instead, which comes from choosing the opposite sign for the time shift ξ^0, or parametrising the spatial metric as $(1-2\psi)$. Nothing physical hinges on it. The one fact used below, in either convention, is that in the gauge where $\delta\phi$ is set to zero, $\mathcal{R}$ reduces to ψ.

Choosing the gauge. With two functions to spend, the obvious — and, as it turns out, extremely convenient — choice is $T = \delta\phi/\phi'$, $L = E$, which sets

$$\delta\phi = 0, \qquad E = 0. \tag{8}$$

This is the comoving, or unitary, gauge: surfaces of constant time coincide with surfaces of constant ϕ, the inflaton fluctuation has been entirely absorbed into the metric, and (as long as $\phi' \neq 0$) no residual freedom is left over. Relabelling $\psi \to \mathcal{R}$, since by (6) that is exactly what it has become, the line element left standing is

$$\boxed{\mathrm{d}s^2 = a^2(\eta)\Big\{-(1+2A)\mathrm{d}\eta^2 + 2\partial_i B\,\mathrm{d}\eta\,\mathrm{d}x^i + (1+2\mathcal{R})\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j\Big\}.} \tag{9}$$

Three functions remain, $A, B, \mathcal{R}$. What is coming next will show that A and B carry no dynamics of their own — they are the lapse and the shift of the spacetime foliation, and their job is to enforce constraints, not to propagate. $\mathcal{R}$ is left standing alone as the single physical scalar.
The condition $\phi' \neq 0$ used to fix the gauge is not a technicality to wave away: in exact de Sitter, $\phi' = 0$ and the whole construction collapses, because $\mathcal{R}$ is then pure gauge. The final result below will be visibly proportional to a slow-roll parameter, and this is the reason why.

The background hums along. Before touching perturbations of the metric it is worth extracting the equations that govern the homogeneous background, because they will be needed repeatedly, and the cleanest way to obtain them is to treat the background itself as a mild perturbation problem: expand the action to first order in homogeneous fluctuations of the lapse, the spatial scale factor, and the scalar field, and read off one field equation per fluctuation.
Take the ansatz

$$\mathrm{d}s^2 = a^2(\eta)\Big[-(1+2A(\eta))\mathrm{d}\eta^2 + (1+2\psi(\eta))\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j\Big], \qquad \phi_{\rm full} = \phi(\eta) + \delta\phi(\eta), \tag{10}$$

so that the lapse is $N = 1 + A$ to first order. (A common slip is to write the unperturbed metric with an extra factor of a^2 multiplying the spatial part inside the lapse term, i.e. $a^2[-N^2\mathrm{d}\eta^2 + a^2\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j]$; this double-counts a factor of a^2 and, once carried through, spoils the coefficients below. The consistent statement is $\mathrm{d}s^2 = a^2[-N^2\mathrm{d}\eta^2 + \delta_{ij}\mathrm{d}x^i\mathrm{d}x^j]$, exactly as in (10).)
Expanding $\sqrt{-g}$, the Ricci scalar, and the matter Lagrangian to first order in $A, \psi, \delta\phi$,

$$\sqrt{-g} = a^4[1 + A + 3\psi] + \mathcal{O}(2), \tag{11}$$

$$a^{2\,(4)}R = 6(\mathcal{H}^2 + \mathcal{H}') - 12(\mathcal{H}^2 + \mathcal{H}')A - 6\mathcal{H}A' + 6\psi'' + 18\mathcal{H}\psi' + \mathcal{O}(2), \tag{12}$$

$$-\tfrac{1}{2}g^{00}\phi'^2_{\rm full} = \frac{\phi'^2}{2a^2} + \frac{1}{a^2}(\phi'\delta\phi' - A\phi'^2) + \mathcal{O}(2), \qquad -V(\phi+\delta\phi) = -V - V_{,\phi}\delta\phi + \mathcal{O}(2), \tag{13}$$

and assembling $L = \sqrt{-g}[\frac{M_{\rm pl}^2}{2}{}^{(4)}R - \frac{1}{2}g^{00}\phi'^2_{\rm full} - V]$, the linear piece comes out to

$$\begin{aligned} L^{(1)} = &-\frac{1}{2}a^2\Big(6M_{\rm pl}^2\mathcal{H}^2 + 6M_{\rm pl}^2\mathcal{H}' + \phi'^2 + 2a^2V\Big)A - 3M_{\rm pl}^2a^2\mathcal{H}A' \\ &+ \frac{3}{2}a^2\Big(6M_{\rm pl}^2\mathcal{H}^2 + 6M_{\rm pl}^2\mathcal{H}' + \phi'^2 - 2a^2V\Big)\psi + 3M_{\rm pl}^2a^2(\psi'' + 3\mathcal{H}\psi') \\ &+ a^2\phi'\,\delta\phi' - a^4V_{,\phi}\,\delta\phi, \end{aligned} \tag{14}$$

with, for later reference, the zeroth-order piece

$$L^{(0)} = 3M_{\rm pl}^2a^2(\mathcal{H}' + \mathcal{H}^2) + \tfrac{1}{2}a^2\phi'^2 - a^4V. \tag{15}$$

Equation (14) still has A', ψ', ψ'', $\delta\phi'$ sitting in it, and none of these can be varied against directly. Peel them off one at a time, using $(a^2)' = 2a^2\mathcal{H}$ and $(a^2)'' = a^2(4\mathcal{H}^2 + 2\mathcal{H}')$:

$$-3M_{\rm pl}^2a^2\mathcal{H}A' \simeq +3M_{\rm pl}^2(a^2\mathcal{H})'A = a^2(6M_{\rm pl}^2\mathcal{H}^2 + 3M_{\rm pl}^2\mathcal{H}')A, \tag{16}$$

$$\begin{aligned} 3M_{\rm pl}^2a^2\psi'' + 9M_{\rm pl}^2a^2\mathcal{H}\psi' &\simeq 3M_{\rm pl}^2a^2(4\mathcal{H}^2 + 2\mathcal{H}')\psi - 9M_{\rm pl}^2a^2(2\mathcal{H}^2 + \mathcal{H}')\psi \\ &= -3M_{\rm pl}^2a^2(2\mathcal{H}^2 + \mathcal{H}')\psi, \end{aligned} \tag{17}$$

$$a^2\phi'\delta\phi' \simeq -a^2(\phi'' + 2\mathcal{H}\phi')\delta\phi. \tag{18}$$

Folding these back into (14) leaves every probe field undifferentiated,

$$L^{(1)} = a^2\Big(3M_{\rm pl}^2\mathcal{H}^2 - \tfrac{1}{2}\phi'^2 - a^2V\Big)A + \frac{3}{2}a^2\Big[2M_{\rm pl}^2(2\mathcal{H}' + \mathcal{H}^2) + \phi'^2 - 2a^2V\Big]\psi \\ - a^2\Big(\phi'' + 2\mathcal{H}\phi' + a^2V_{,\phi}\Big)\delta\phi. \tag{19}$$

Stationarity of $S^{(1)} = \int \mathrm{d}\eta\, \mathrm{d}^3x\, L^{(1)}$ then just means each bracket vanishes, and out come the three background equations in one stroke: varying A gives the Friedmann constraint,

$$\boxed{3M_{\rm pl}^2\mathcal{H}^2 = \tfrac{1}{2}\phi'^2 + a^2V}, \tag{20}$$

varying ψ, and using (20) to remove a^2V from the result, gives

$$\boxed{2M_{\rm pl}^2(\mathcal{H}' - \mathcal{H}^2) = -\phi'^2}, \tag{21}$$

and varying $\delta\phi$ gives the Klein–Gordon equation,

$$\boxed{\phi'' + 2\mathcal{H}\phi' + a^2V_{,\phi} = 0}. \tag{22}$$

Solved for the two matter quantities, Eqs. (20)–(21) give two identities that will be leaned on constantly below, since they are the only place the background equations of motion ever enter the perturbative calculation:

$$\boxed{\phi'^2 = 2M_{\rm pl}^2(\mathcal{H}^2 - \mathcal{H}'), \qquad a^2V = M_{\rm pl}^2(2\mathcal{H}^2 + \mathcal{H}').} \tag{23}$$

It is also convenient to name the first Hubble-flow parameter,

$$\varepsilon \equiv 1 - \frac{\mathcal{H}'}{\mathcal{H}^2} = -\frac{\dot H}{H^2} = \frac{\phi'^2}{2M_{\rm pl}^2\mathcal{H}^2} = \frac{\dot\phi^2}{2M_{\rm pl}^2H^2}, \tag{24}$$

so that the first identity in (23) reads simply $\phi'^2 = 2M_{\rm pl}^2\mathcal{H}^2\varepsilon$.
In conformal time, $\varepsilon = 1 - \mathcal{H}'/\mathcal{H}^2$, not $-\mathcal{H}'/\mathcal{H}^2$; the 1 comes from $\mathcal{H} = aH$ and dropping it introduces an error that looks disturbingly like a legitimate slow-roll correction and is easy to miss downstream.

Splitting spacetime into space and time. The perturbed action could now be expanded directly in (9), and this can indeed be done — it is carried out in full in Appendix 171, as a cross-check — but the direct route buries the calculation in second time derivatives of the metric that then have to be removed one by one, after the fact, by integration by parts, at the point where the expression is already the longest. A cleaner path exists: rearrange the gravitational action *before* perturbing, so that time derivatives never appear beyond first order to begin with. This is the Arnowitt–Deser–Misner, or ADM, formulation of gravity, and setting it up is worth doing carefully because everything downstream depends on getting the bookkeeping right once.

Foliating spacetime. Slice the four-dimensional spacetime into a stack of spacelike hypersurfaces Σ_t labelled by a time coordinate t, and let n^μ be the future-pointing unit normal to each slice, $n_\mu n^\mu = -1$. The metric induced on a slice, and the projector onto it, are

$$\gamma_{\mu\nu} = g_{\mu\nu} + n_\mu n_\nu, \qquad \gamma^\mu{}_\nu = \delta^\mu{}_\nu + n^\mu n_\nu. \tag{25}$$

The vector $t^\mu = \partial x^\mu/\partial t$ that flows from one slice to the next need not be orthogonal to the slices; splitting it into a piece along the normal and a piece tangent to Σ_t, $t^\mu = Nn^\mu + N^\mu$,

defines the lapse N (how much proper time elapses between neighbouring slices) and the shift N^i (how the spatial coordinates get dragged sideways in the process). In these variables the line element becomes

$$\mathrm{d}s^2 = -N^2\mathrm{d}t^2 + \gamma_{ij}\big(\mathrm{d}x^i + N^i\mathrm{d}t\big)\big(\mathrm{d}x^j + N^j\mathrm{d}t\big), \tag{26}$$

from which the metric components, their inverses, and the volume element follow directly:

$$g_{00} = \gamma_{ij}N^iN^j - N^2, \qquad g_{0i} = \gamma_{ij}N^j \equiv N_i, \qquad g_{ij} = \gamma_{ij}, \tag{27}$$

$$g^{00} = -N^{-2}, \qquad g^{0i} = N^{-2}N^i, \qquad g^{ij} = \gamma^{ij} - N^{-2}N^iN^j, \tag{28}$$

$$\sqrt{-g} = N\sqrt{\gamma}, \qquad n_\mu = (-N, 0, 0, 0), \qquad n^\mu = \Big(\frac{1}{N}, -\frac{N^i}{N}\Big). \tag{29}$$

The practical upshot, for reading ADM data off any metric handed to us in the usual $g_{\mu\nu}$ form, is the short dictionary

$$\boxed{\gamma_{ij} = g_{ij}, \qquad N_i = g_{0i}, \qquad N^i = \gamma^{ij}N_j, \qquad N^2 = -g_{00} + N_iN^i.} \tag{30}$$

How a slice bends: extrinsic curvature. The induced metric γ_{ij} captures the geometry *within* each slice, but nothing about how the slices sit inside the ambient spacetime. That is the job of the extrinsic curvature,

$$K_{\mu\nu} \equiv \gamma^\rho{}_\mu \nabla_\rho n_\nu, \tag{31}$$

which turns out to be symmetric, purely spatial ($K_{\mu\nu}n^\mu = 0$), and equal to half the Lie derivative of $\gamma_{\mu\nu}$ along n. In terms of the lapse and shift,

$$\boxed{K_{ij} = \frac{1}{2N}\Big(\dot\gamma_{ij} - D_iN_j - D_jN_i\Big) \equiv \frac{E_{ij}}{N}, \qquad K \equiv \gamma^{ij}K_{ij} = \frac{E}{N},} \tag{32}$$

where D_i is the covariant derivative built from γ_{ij}, and $E_{ij} \equiv NK_{ij}$, $E \equiv \gamma^{ij}E_{ij}$ are introduced because that combination, not K_{ij} itself, is what turns up in the action without stray inverse powers of N. One more geometric object is needed: the acceleration of the flow of normal vectors,

$$a^\mu \equiv n^\nu\nabla_\nu n^\mu, \qquad a^0 = 0, \qquad a^i = \gamma^{ij}\partial_j \ln N, \tag{33}$$

the vanishing of a^0 following from $a^\mu n_\mu = 0$ together with $n_\mu = (-N, \mathbf{0})$.

Trading the four-dimensional curvature for the three-dimensional one. The whole point of the exercise is to rewrite ${}^{(4)}R$ — which, expanded directly, contains second time derivatives of γ_{ij} — in terms of ${}^{(3)}R$, K_{ij}, and only *first* time derivatives. Start from the exact split

$${}^{(4)}R = g^{\mu\nu}g^{\rho\sigma}R_{\mu\rho\nu\sigma} = (\gamma^{\mu\nu} - n^\mu n^\nu)(\gamma^{\rho\sigma} - n^\rho n^\sigma)R_{\mu\rho\nu\sigma} = \gamma^{\mu\nu}\gamma^{\rho\sigma}R_{\mu\rho\nu\sigma} - 2R_{\mu\nu}n^\mu n^\nu, \tag{34}$$

where the two cross terms coincide by the symmetries of the Riemann tensor and the term with four factors of n vanishes by antisymmetry. The purely spatial projection of the Riemann tensor is the Gauss equation,

$$\gamma^{\mu\nu}\gamma^{\rho\sigma}R_{\mu\rho\nu\sigma} = {}^{(3)}R + K^2 - K_{\mu\nu}K^{\mu\nu}, \tag{35}$$

which leaves ${}^{(4)}R = {}^{(3)}R + K^2 - K_{\mu\nu}K^{\mu\nu} - 2R_{\mu\nu}n^\mu n^\nu$ still to be dealt with in the last term. That term is where the total derivative announced above is born, so it is worth

deriving in full rather than quoting. Starting from the Ricci identity on the normal vector, $[\nabla_\mu, \nabla_\nu] n^\mu = R_{\sigma\nu} n^\sigma$, and contracting with n^ν,

$$R_{\mu\nu} n^\mu n^\nu = n^\nu \nabla_\mu \nabla_\nu n^\mu - n^\nu \nabla_\nu \nabla_\mu n^\mu . \tag{36}$$

Move the surviving n^ν inside each derivative in turn:

$$n^\nu \nabla_\mu \nabla_\nu n^\mu = \nabla_\mu (n^\nu \nabla_\nu n^\mu) - (\nabla_\mu n^\nu)(\nabla_\nu n^\mu) = \nabla_\mu a^\mu - (\nabla_\mu n^\nu)(\nabla_\nu n^\mu), \tag{37}$$

$$n^\nu \nabla_\nu \nabla_\mu n^\mu = \nabla_\nu (n^\nu K) - K^2, \tag{38}$$

using $K = \nabla_\mu n^\mu$ in the second line. For the remaining quadratic piece, decompose $\nabla_\mu n_\nu = K_{\mu\nu} - n_\mu a_\nu$ (this is Eq. (32) run backwards, with the acceleration restoring the piece along the normal that the purely spatial $K_{\mu\nu}$ misses), so that

$$(\nabla_\mu n^\nu)(\nabla_\nu n^\mu) = (K_\mu{}^\nu - n_\mu a^\nu)(K_\nu{}^\mu - n_\nu a^\mu) = K_{\mu\nu} K^{\mu\nu}, \tag{39}$$

every cross term dying against $K_{\mu\nu} n^\nu = 0$ or $a_\nu n^\nu = 0$, and the pure aa term dying against $n_\mu a^\mu = 0$. Collecting everything,

$$\boxed{R_{\mu\nu} n^\mu n^\nu = K^2 - K_{\mu\nu} K^{\mu\nu} + \nabla_\mu v^\mu, \qquad v^\mu \equiv a^\mu - n^\mu K,} \tag{40}$$

and substituting back,

$$\boxed{{}^{(4)}R = {}^{(3)}R + K_{ij} K^{ij} - K^2 - 2\nabla_\mu v^\mu,} \tag{41}$$

where $K_{\mu\nu} K^{\mu\nu} = K_{ij} K^{ij}$ has been used, since $K_{\mu\nu}$ is purely spatial and its indices are raised with γ^{ij}.
With $N = 1$, $N^i = 0$, $\gamma_{ij} = a^2 \delta_{ij}$ in cosmic time, ${}^{(3)}R = 0$ and $K_{ij} = \frac{1}{2}\dot\gamma_{ij} = a\dot a\, \delta_{ij}$, so $K = 3H$, $K_{ij} K^{ij} = 3H^2$; the acceleration vanishes, so $v^\mu = -3Hn^\mu$ and $\nabla_\mu v^\mu = -a^{-3}\partial_t(3Ha^3) = -3\dot H - 9H^2$. Then (41) gives ${}^{(4)}R = 3H^2 - 9H^2 + 6\dot H + 18H^2 = 6(\dot H + 2H^2)$, the textbook FLRW value.

Naming what gets thrown away. Substituting (41) into the gravitational part of the action, and using $\sqrt{-g} = N\sqrt{\gamma}$,

$$\frac{M_{\rm pl}^2}{2}\int \mathrm{d}^4x\, \sqrt{-g}\, {}^{(4)}R = \frac{M_{\rm pl}^2}{2}\int \mathrm{d}^4x\, N\sqrt{\gamma}\left[{}^{(3)}R + K_{ij}K^{ij} - K^2\right] \; \underbrace{- M_{\rm pl}^2 \int \mathrm{d}^4x\, \partial_\mu(\sqrt{-g}\, v^\mu)}_{\equiv\ \Delta S_{\rm bdy}}, \tag{42}$$

using $\sqrt{-g}\nabla_\mu v^\mu = \partial_\mu(\sqrt{-g}\, v^\mu)$. That last term is the whole of what a casual "$\simeq$" would sweep aside, and there is no need to leave it unexamined: from (29), (33) and $K = E/N$,

$$\sqrt{-g}\, v^0 = N\sqrt{\gamma}\Big(0 - \frac{K}{N}\Big) = -\sqrt{\gamma}\, K, \tag{43}$$

$$\sqrt{-g}\, v^i = N\sqrt{\gamma}\Big(\gamma^{ij}\partial_j \ln N + \frac{K N^i}{N}\Big) = \sqrt{\gamma}\Big(\gamma^{ij}\partial_j N + K N^i\Big), \tag{44}$$

so that

$$\boxed{\Delta S_{\rm bdy} = M_{\rm pl}^2 \int \mathrm{d}^4x \left\{-\partial_t\Big(\sqrt{\gamma}\, K\Big) + \partial_i\Big[\sqrt{\gamma}(\gamma^{ij}\partial_j N + K N^i)\Big]\right\}.} \tag{45}$$

Two remarks make it clear why this term never bothers anything that follows. The time piece, $-M_{\rm pl}^2\big[\oint \mathrm{d}^3x \sqrt{\gamma} K\big]_{t_i}^{t_f}$, is precisely minus the Gibbons–Hawking–York term; had that

boundary term been carried along with the action from the start, as is required for a well-posed variational problem when γ_{ij} is held fixed on the boundary, this piece would be cancelled outright rather than merely dropped. The spatial piece is a three-divergence, and integrates to a surface term at spatial infinity that vanishes for any field configuration decaying at large distance — equivalently, it never contributes to any Fourier mode with $\boldsymbol{k} \neq 0$. Neither piece can move an equation of motion or a correlation function; the only thing it changes is the off-shell value of the Lagrangian density, a fact that becomes concrete below when the zeroth-order piece of the two computations are compared directly. The identity $\sqrt{-g}\,{}^{(4)}R = \sqrt{\gamma}[N^{(3)}R + N^{-1}(E_{ij}E^{ij} - E^2)] - 2\partial_\mu(\sqrt{-g}\,v^\mu)$, with $\sqrt{-g}\,v^\mu$ exactly as above, has been checked with computer algebra order by order up to second order in the perturbed metric used below; the residual vanishes identically at $\mathcal{O}(\epsilon^0)$, $\mathcal{O}(\epsilon^1)$, and $\mathcal{O}(\epsilon^2)$ alike (Appendix 171).

The ADM action. The matter sector transforms just as easily. With $g^{00} = -N^{-2}$, $g^{0i} = N^{-2}N^i$, $g^{ij} = \gamma^{ij} - N^{-2}N^iN^j$,

$$-\tfrac{1}{2}g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi = \frac{1}{2N^2}\big(\dot\phi - N^i\partial_i\phi\big)^2 - \tfrac{1}{2}\gamma^{ij}\partial_i\phi\partial_j\phi, \tag{46}$$

which, once the comoving gauge sets $\partial_i\phi = 0$, collapses to $\dot\phi^2/(2N^2)$. Putting gravity and matter together,

$$\boxed{S = \int \mathrm{d}t\,\mathrm{d}^3x\,\sqrt{\gamma}\left\{\frac{M_{\rm pl}^2}{2}\left[N\,{}^{(3)}R + \frac{1}{N}\Big(E_{ij}E^{ij} - E^2\Big)\right] + \frac{\dot\phi^2}{2N} - N\,V\right\}.} \tag{47}$$

Three things about this form matter for everything that follows. First, only first time derivatives appear, of γ_{ij} and ϕ alone; the variational problem is well posed as written. Second, N and N^i appear with no time derivatives whatsoever, marking them out as Lagrange multipliers: their variations give the Hamiltonian and momentum constraints, and because they never carry dynamics, those constraints can be solved and substituted straight back into the action without disturbing the equations obeyed by whatever is left. Third, $E_{ij}E^{ij} - E^2$ plays the role of a kinetic term for γ_{ij}, moving in a potential ${}^{(3)}R$.
If a field λ enters $S[q,\lambda]$ with no derivatives, and $\lambda_\star[q]$ solves $\delta S/\delta\lambda = 0$, then $\delta S_{\rm red}/\delta q = \delta S/\delta q\big|_{\lambda_\star} + (\delta S/\delta\lambda)\big|_{\lambda_\star}\,\delta\lambda_\star/\delta q$, and the second term vanishes by construction, since $\lambda_\star$ was chosen to make it vanish. So $S_{\rm red}[q] \equiv S[q,\lambda_\star[q]]$ gives the correct equations for q. This is precisely why point two above matters, and it would fail the moment λ carried its own time derivatives. It is also why only the *first-order* constraint solutions are ever needed to get the action right to *second* order: one order of sloppiness in $\lambda_\star$ costs two orders in the action, not one.

Two roads to the quadratic action. With the machinery in place, the second-order action for $\mathcal{R}$ can now be built. It will be built twice, by two routes that share nothing beyond the master formula (47): once in conformal time with the metric written as $\gamma_{ij} = a^2(1+2\mathcal{R})\delta_{ij}$, and once in cosmic time with the metric written as $\gamma_{ij} = a^2e^{2\zeta}\delta_{ij}$. The two parametrisations expand completely differently, the intermediate expressions look nothing alike, and the constraint solutions take different forms — so finding the same answer at the end is a genuine check on everything that came before, not a restatement of it.

First road: conformal time, linear parametrisation. Feed the comoving-gauge metric (9) into the dictionary (30): this is exact, no expansion performed yet,

$$\gamma_{ij} = a^2(1+2\mathcal{R})\delta_{ij}, \qquad \gamma^{ij} = \frac{\delta^{ij}}{a^2(1+2\mathcal{R})}, \qquad \sqrt{\gamma} = a^3(1+2\mathcal{R})^{3/2}, \tag{48}$$

$$N_i = a^2\partial_i B, \qquad N^i = \frac{\partial_i B}{1+2\mathcal{R}}, \qquad N^2 = a^2(1+2A) + \frac{a^2(\partial B)^2}{1+2\mathcal{R}}. \tag{49}$$

Attach a bookkeeping parameter ϵ to every perturbation and expand to $\mathcal{O}(\epsilon^2)$. It helps to strip the background factors first, defining

$$\mathcal{S} \equiv \frac{\sqrt{\gamma}}{a^3}, \qquad n \equiv \frac{N}{a}, \qquad m \equiv \frac{a}{N} = n^{-1}, \qquad r \equiv a^2\,{}^{(3)}\!R, \qquad K \equiv E_{ij}E^{ij} - E^2. \tag{50}$$

The volume factor is the easiest: $(1+2\mathcal{R})^{3/2} = 1 + 3\mathcal{R} + \frac{3}{2}\mathcal{R}^2 + \dots$, so

$$\mathcal{S} = 1 + 3\mathcal{R} + \tfrac{3}{2}\mathcal{R}^2. \tag{51}$$

The lapse follows the same way, noting that $(\partial B)^2$ is already second order so the $(1+2\mathcal{R})^{-1}$ multiplying it can be replaced by 1:

$$n = \sqrt{1 + 2A + (\partial B)^2} = 1 + A + \tfrac{1}{2}(\partial B)^2 - \tfrac{1}{2}A^2, \qquad m = 1 - A - \tfrac{1}{2}(\partial B)^2 + \tfrac{3}{2}A^2. \tag{52}$$

The three-curvature takes a little more care. For any conformally flat spatial metric $\gamma_{ij} = e^{2\omega}\delta_{ij}$ one has the standard formula ${}^{(3)}\!R = -2e^{-2\omega}[2\partial^2\omega + (\partial\omega)^2]$; here $e^{2\omega} = a^2(1+2\mathcal{R})$, so $\partial_i\omega = \partial_i\mathcal{R}/(1+2\mathcal{R})$ and $\partial^2\omega = \partial^2\mathcal{R}/(1+2\mathcal{R}) - 2(\partial\mathcal{R})^2/(1+2\mathcal{R})^2$. Expanding these to second order, $2\partial^2\omega + (\partial\omega)^2 = 2\partial^2\mathcal{R} - 4\mathcal{R}\partial^2\mathcal{R} - 3(\partial\mathcal{R})^2$ and $e^{-2\omega} = a^{-2}(1 - 2\mathcal{R} + 4\mathcal{R}^2)$, so that

$$r = -4\partial^2\mathcal{R} + 16\mathcal{R}\partial^2\mathcal{R} + 6(\partial\mathcal{R})^2. \tag{53}$$

Last comes the extrinsic curvature. Using $\partial_\eta\gamma_{ij} = 2a^2[\mathcal{H}(1+2\mathcal{R}) + \mathcal{R}']\delta_{ij}$, and the Christoffel symbols of the conformally flat spatial metric, $\Gamma^k{}_{ij}[\gamma] = \delta^k{}_i\partial_j\omega + \delta^k{}_j\partial_i\omega - \delta_{ij}\partial_k\omega$, a direct computation of $D_iN_j = \partial_iN_j - \Gamma^k{}_{ij}N_k$ gives the trace

$$E = 3\mathcal{H} + 3\mathcal{R}' - \partial^2 B + \Big[-6\mathcal{R}\mathcal{R}' + 2\mathcal{R}\partial^2 B - \partial B\cdot\partial\mathcal{R}\Big] \tag{54}$$

and the quadratic invariant

$$\begin{aligned} K = E_{ij}E^{ij} - E^2 = &-6\mathcal{H}^2 + \Big[4\mathcal{H}\partial^2 B - 12\mathcal{H}\mathcal{R}'\Big] \\ &+ \Big[-6\mathcal{R}'^2 + 4\mathcal{R}'\partial^2 B + 24\mathcal{H}\mathcal{R}\mathcal{R}' - 8\mathcal{H}\mathcal{R}\partial^2 B + 4\mathcal{H}\,\partial B\cdot\partial\mathcal{R} \\ &\quad - ((\partial^2 B)^2 - \partial_i\partial_j B\,\partial_i\partial_j B)\Big], \end{aligned} \tag{55}$$

the three brackets running through orders $\epsilon^0, \epsilon^1, \epsilon^2$ in turn.
Substituting $\sqrt{\gamma} = a^3\mathcal{S}$, $N = an$, $N^{-1} = m/a$, ${}^{(3)}\!R = r/a^2$ into (47) and pulling an a^2 out front collapses the whole Lagrangian into one line,

$$\boxed{L = a^2\,\mathcal{S}\,W, \qquad W \equiv \frac{M_{\rm pl}^2}{2}(n\,r + m\,K) + \frac{\phi'^2}{2}m - a^2 V\,n.} \tag{56}$$

Every subsequent step is bookkeeping starting from this one line. Writing $\mathcal{S} = \sum_k \mathcal{S}_k$, $W = \sum_k W_k$ order by order,

$$L^{(0)} = a^2 W_0, \qquad L^{(1)} = a^2(W_1 + 3\mathcal{R}\,W_0), \qquad L^{(2)} = a^2(W_2 + 3\mathcal{R}\,W_1 + \tfrac{3}{2}\mathcal{R}^2 W_0), \tag{57}$$

using $\mathcal{S}_1 = 3\mathcal{R}$, $\mathcal{S}_2 = \frac{3}{2}\mathcal{R}^2$ from (51). The needed pieces of n, m, r, K, read off (52)–(55), are $n_1 = A$, $n_2 = \frac{1}{2}[(\partial B)^2 - A^2]$, $m_1 = -A$, $m_2 = \frac{3}{2}A^2 - \frac{1}{2}(\partial B)^2$, $r_0 = 0$, $r_1 = -4\partial^2\mathcal{R}$,

$r_2 = 16\mathcal{R}\partial^2\mathcal{R} + 6(\partial\mathcal{R})^2$, $K_0 = -6\mathcal{H}^2$, $K_1 = 4\mathcal{H}\partial^2 B - 12\mathcal{H}\mathcal{R}'$, and K_2 the third bracket of (55).
At zeroth order,

$$W_0 = -3M_{\rm pl}^2\mathcal{H}^2 + \frac{\phi'^2}{2} - a^2V, \qquad L^{(0)} = a^2 W_0. \tag{58}$$

Comparing with the covariant $L^{(0)}$ of (15), the two differ by

$$L^{(0)}_{\rm cov} - L^{(0)}_{\rm ADM} = 3M_{\rm pl}^2 a^2\mathcal{H}' + 6M_{\rm pl}^2 a^2\mathcal{H}^2 = 3M_{\rm pl}^2(a^2\mathcal{H})', \tag{59}$$

which is exactly $-\Delta S_{\rm bdy}$ of (45) evaluated on the unperturbed background, where $\sqrt{\gamma} = a^3$, $N = a$, $E = 3\mathcal{H}$, $K = E/N = 3\mathcal{H}/a$, so $\sqrt{\gamma}K = 3a^2\mathcal{H}$ and the temporal term reads $-M_{\rm pl}^2\partial_\eta(3a^2\mathcal{H})$ while the spatial one vanishes identically. The two Lagrangians differ by nothing but the discarded divergence, exactly as advertised.
At first order,

$$\begin{aligned} W_1 &= \frac{M_{\rm pl}^2}{2}(r_1 + K_1 - AK_0) - \frac{\phi'^2}{2}A - a^2VA \\ &= M_{\rm pl}^2\Big(-2\partial^2\mathcal{R} + 2\mathcal{H}\partial^2 B - 6\mathcal{H}\mathcal{R}' + 3\mathcal{H}^2 A\Big) - \frac{\phi'^2}{2}A - a^2VA, \end{aligned} \tag{60}$$

so that

$$L^{(1)} = a^2\Big[M_{\rm pl}^2(-2\partial^2\mathcal{R} + 2\mathcal{H}\partial^2 B - 6\mathcal{H}\mathcal{R}') + \underbrace{\Big(3M_{\rm pl}^2\mathcal{H}^2 - \tfrac{1}{2}\phi'^2 - a^2V\Big)}_{=0 \text{ by } (20)}A + 3\mathcal{R}W_0\Big]. \tag{61}$$

The coefficient of A vanishes on the Friedmann constraint, as it must — the lapse is a Lagrange multiplier for the Hamiltonian constraint, and this is that statement showing up in the perturbation theory. What remains reduces to a total derivative once the background equations hold: the $\partial^2\mathcal{R}$ and $\partial^2 B$ pieces are spatial divergences outright, and $-6M_{\rm pl}^2 a^2\mathcal{H}\mathcal{R}' \simeq 6M_{\rm pl}^2(a^2\mathcal{H})'\mathcal{R} = 6a^2\mathcal{R}\,a^2V$ (using (23)) cancels $3\mathcal{R}W_0 = -6a^2V\mathcal{R}$ (using $W_0 = -2a^2V$ on shell, again from (20)). So $S^{(1)} = 0$, confirming that the background genuinely extremises the action.
At second order, assembling

$$W_2 = \frac{M_{\rm pl}^2}{2}\Big(r_2 + A\,r_1 + K_2 - A\,K_1 + m_2K_0\Big) + \frac{\phi'^2}{2}m_2 - a^2V\,n_2 \tag{62}$$

and adding $3\mathcal{R}W_1 + \frac{3}{2}\mathcal{R}^2W_0$, and collecting like structures, gives

$$\begin{aligned} \frac{L^{(2)}}{a^2} = {} & -3M_{\rm pl}^2\mathcal{R}'^2 + 3M_{\rm pl}^2(\partial\mathcal{R})^2 + 2M_{\rm pl}^2\mathcal{R}\,\partial^2\mathcal{R} - 2M_{\rm pl}^2 A\,\partial^2\mathcal{R} \\ & + 2M_{\rm pl}^2\mathcal{R}'\partial^2 B - 2M_{\rm pl}^2\mathcal{H}A\,\partial^2 B + 6M_{\rm pl}^2\mathcal{H}A\mathcal{R}' \\ & + 2M_{\rm pl}^2\mathcal{H}\mathcal{R}\,\partial^2 B + 2M_{\rm pl}^2\mathcal{H}\,\partial B\cdot\partial\mathcal{R} - 6M_{\rm pl}^2\mathcal{H}\mathcal{R}\mathcal{R}' \\ & - \frac{M_{\rm pl}^2}{2}\Big[(\partial^2 B)^2 - \partial_i\partial_j B\,\partial_i\partial_j B\Big] \\ & + \Big(-\tfrac{9}{2}M_{\rm pl}^2\mathcal{H}^2 + \tfrac{3}{4}\phi'^2 + \tfrac{1}{2}a^2V\Big)A^2 + \Big(9M_{\rm pl}^2\mathcal{H}^2 - \tfrac{3}{2}\phi'^2 - 3a^2V\Big)A\mathcal{R} \\ & + \Big(-\tfrac{9}{2}M_{\rm pl}^2\mathcal{H}^2 + \tfrac{3}{4}\phi'^2 - \tfrac{3}{2}a^2V\Big)\mathcal{R}^2 + \Big(\tfrac{3}{2}M_{\rm pl}^2\mathcal{H}^2 - \tfrac{1}{4}\phi'^2 - \tfrac{1}{2}a^2V\Big)(\partial B)^2. \end{aligned} \tag{63}$$

As a check on the bookkeeping: the coefficient of $\mathcal{R}\partial^2\mathcal{R}$ gets $\frac{M_{\rm pl}^2}{2}\times 16$ from r_2 and $3\mathcal{R}\times(-2M_{\rm pl}^2\partial^2\mathcal{R})$ from $3\mathcal{R}W_1$, giving $8M_{\rm pl}^2-6M_{\rm pl}^2=2M_{\rm pl}^2$ as claimed; the coefficient of $\mathcal{H}\mathcal{R}\mathcal{R}'$ gets $\frac{M_{\rm pl}^2}{2}\times 24$ from K_2 and $3\mathcal{R}\times(-6M_{\rm pl}^2\mathcal{H}\mathcal{R}')$, giving $12M_{\rm pl}^2-18M_{\rm pl}^2=-6M_{\rm pl}^2$, likewise as claimed.
Now feed in the substitution rules (23). Three of the four purely algebraic coefficients collapse outright:

$$A\mathcal{R}: \quad 9M_{\rm pl}^2\mathcal{H}^2-3M_{\rm pl}^2(\mathcal{H}^2-\mathcal{H}')-3M_{\rm pl}^2(2\mathcal{H}^2+\mathcal{H}')=0, \tag{64}$$

$$(\partial B)^2: \quad \tfrac{3}{2}M_{\rm pl}^2\mathcal{H}^2-\tfrac{1}{2}M_{\rm pl}^2(\mathcal{H}^2-\mathcal{H}')-\tfrac{1}{2}M_{\rm pl}^2(2\mathcal{H}^2+\mathcal{H}')=0, \tag{65}$$

$$A^2: \quad -\tfrac{9}{2}M_{\rm pl}^2\mathcal{H}^2+\tfrac{3}{4}\cdot 2M_{\rm pl}^2(\mathcal{H}^2-\mathcal{H}')+\tfrac{1}{2}M_{\rm pl}^2(2\mathcal{H}^2+\mathcal{H}')$$
$$=-M_{\rm pl}^2(2\mathcal{H}^2+\mathcal{H}')=-\tfrac{1}{2}(6M_{\rm pl}^2\mathcal{H}^2-\phi'^2), \tag{66}$$

$$\mathcal{R}^2: \quad -\tfrac{9}{2}M_{\rm pl}^2\mathcal{H}^2+\tfrac{3}{2}M_{\rm pl}^2(\mathcal{H}^2-\mathcal{H}')-\tfrac{3}{2}M_{\rm pl}^2(2\mathcal{H}^2+\mathcal{H}')=-6M_{\rm pl}^2\mathcal{H}^2-3M_{\rm pl}^2\mathcal{H}'. \tag{67}$$

The cross term between A and $\mathcal{R}$ vanishes, and so does the gradient term for the shift; the $\mathcal{R}^2$ coefficient will shortly be cancelled by an integration by parts, and the A^2 coefficient is the mass term that will survive into the final unreduced action.
Four total derivatives finish the job, and it is worth writing every one of them out rather than asserting that "after some integration by parts" things simplify. First, the shift's quadratic piece: since

$$\begin{aligned}\partial_i\Big(\partial_j B\,\partial_i\partial_j B-\partial_i B\,\partial^2 B\Big)&=\partial_i\partial_j B\,\partial_i\partial_j B+\partial_j B\,\partial_j(\partial^2 B)\\&\quad-(\partial^2 B)^2-\partial_j B\,\partial_j(\partial^2 B)=\partial_i\partial_j B\,\partial_i\partial_j B-(\partial^2 B)^2,\end{aligned} \tag{68}$$

and the prefactor does not depend on position,

$$-\frac{M_{\rm pl}^2a^2}{2}\Big[(\partial^2 B)^2-\partial_i\partial_j B\,\partial_i\partial_j B\Big]=\frac{M_{\rm pl}^2a^2}{2}\partial_i\Big(\partial_j B\,\partial_i\partial_j B-\partial_i B\,\partial^2 B\Big)\simeq 0. \tag{69}$$

Second, the $\mathcal{H}\mathcal{R}B$ pair collapses into a pure spatial divergence,

$$2M_{\rm pl}^2a^2\mathcal{H}\Big[\mathcal{R}\partial^2 B+\partial B\cdot\partial\mathcal{R}\Big]=2M_{\rm pl}^2a^2\mathcal{H}\,\partial_i(\mathcal{R}\partial_i B)\simeq 0. \tag{70}$$

Third, $\mathcal{R}\partial^2\mathcal{R}$ integrates by parts to $-(\partial\mathcal{R})^2$,

$$2M_{\rm pl}^2a^2\mathcal{R}\partial^2\mathcal{R}\simeq -2M_{\rm pl}^2a^2(\partial\mathcal{R})^2, \tag{71}$$

which, added to the explicit $+3M_{\rm pl}^2a^2(\partial\mathcal{R})^2$ already sitting in (63), leaves $+M_{\rm pl}^2a^2(\partial\mathcal{R})^2$ standing. Fourth and last, the $\mathcal{H}\mathcal{R}\mathcal{R}'$ term is a total time derivative in disguise,

$$-6M_{\rm pl}^2a^2\mathcal{H}\mathcal{R}\mathcal{R}'=-3M_{\rm pl}^2a^2\mathcal{H}(\mathcal{R}^2)'\simeq+3M_{\rm pl}^2(a^2\mathcal{H})'\mathcal{R}^2=3M_{\rm pl}^2a^2(2\mathcal{H}^2+\mathcal{H}')\mathcal{R}^2, \tag{72}$$

which cancels (67) exactly. No mass term for $\mathcal{R}$ survives the dust settling — the first hint that $\mathcal{R}$ will end up conserved once a mode leaves the horizon.
Collecting everything that is left standing,

$$\boxed{\begin{aligned}S^{(2)}=\int\mathrm{d}\eta\,\mathrm{d}^3x\;a^2\Big[&-3M_{\rm pl}^2\mathcal{R}'^2+2M_{\rm pl}^2\mathcal{R}'\partial^2 B-2M_{\rm pl}^2\mathcal{H}A\partial^2 B+M_{\rm pl}^2(\partial\mathcal{R})^2\\&-2M_{\rm pl}^2A\partial^2\mathcal{R}+6M_{\rm pl}^2\mathcal{H}A\mathcal{R}'-\tfrac{1}{2}(6\mathcal{H}^2M_{\rm pl}^2-\phi'^2)A^2\Big].\end{aligned}} \tag{73}$$

It carries the wrong-looking sign only because A has not yet been eliminated: the $A\partial^2\mathcal{R}$ term still to be dealt with will hand back a gradient contribution of the opposite sign

once the constraint is solved. Reading off the speed of sound before the auxiliary fields are removed is a reliable way to scare oneself for no reason.
Neither A' nor B' ever appears in (73), confirming both are auxiliary as promised. Varying B, whose only appearances are through $\partial^2 B$, gives the momentum constraint

$$\partial^2\Big[2M_{\rm pl}^2 a^2\mathcal{R}' - 2M_{\rm pl}^2 a^2\mathcal{H}A\Big] = 0 \implies \boxed{\mathcal{H}A = \mathcal{R}'}, \tag{74}$$

dropping the homogeneous solution of $\partial^2(\cdot) = 0$ as an unwanted non-decaying mode. Varying A gives the Hamiltonian constraint,

$$(\phi'^2 - 6M_{\rm pl}^2\mathcal{H}^2)A + 2M_{\rm pl}^2(3\mathcal{H}\mathcal{R}' - \partial^2\mathcal{R}) = 2M_{\rm pl}^2\mathcal{H}\,\partial^2 B, \tag{75}$$

which, using $\phi'^2 - 6M_{\rm pl}^2\mathcal{H}^2 = -2M_{\rm pl}^2(2\mathcal{H}^2 + \mathcal{H}')$ and $A = \mathcal{R}'/\mathcal{H}$ from (74), solves to

$$\partial^2 B = -\frac{\partial^2\mathcal{R}}{\mathcal{H}} + \frac{\phi'^2}{2M_{\rm pl}^2\mathcal{H}^2}\,\mathcal{R}' = -\frac{\partial^2\mathcal{R}}{\mathcal{H}} + \varepsilon\,\mathcal{R}'. \tag{76}$$

This explicit solution for B is, in fact, never needed, for a reason worth noting: putting $A = \mathcal{R}'/\mathcal{H}$ into (73), the two terms carrying B combine as $2M_{\rm pl}^2 a^2\partial^2 B(\mathcal{R}' - \mathcal{H}A) = 0$ identically. This is a general feature of these constrained systems — the shift always multiplies the momentum constraint, and the momentum constraint has just been set to zero by its own variation.
What survives, with $A = \mathcal{R}'/\mathcal{H}$ substituted, is

$$\frac{L^{(2)}_{\rm red}}{a^2} = -3M_{\rm pl}^2\mathcal{R}'^2 + M_{\rm pl}^2(\partial\mathcal{R})^2 - \frac{2M_{\rm pl}^2}{\mathcal{H}}\mathcal{R}'\partial^2\mathcal{R} + 6M_{\rm pl}^2\mathcal{R}'^2 - \frac{1}{2}(6\mathcal{H}^2M_{\rm pl}^2 - \phi'^2)\frac{\mathcal{R}'^2}{\mathcal{H}^2}. \tag{77}$$

The kinetic coefficients add up to

$$-3M_{\rm pl}^2 + 6M_{\rm pl}^2 - 3M_{\rm pl}^2 + \frac{\phi'^2}{2\mathcal{H}^2} = \frac{\phi'^2}{2\mathcal{H}^2} \equiv Q_s, \tag{78}$$

so the three pure-gravity contributions cancel completely and the entire kinetic term for $\mathcal{R}$ turns out to come from the matter sector alone, which is exactly why Q_s is proportional to ϕ'^2, and hence to ε, and why the whole edifice degenerates the moment $\phi' \to 0$. The mixed term is dispatched by integrating by parts twice, first in space and then in time,

$$-\frac{2M_{\rm pl}^2 a^2}{\mathcal{H}}\mathcal{R}'\partial^2\mathcal{R} \simeq \frac{2M_{\rm pl}^2 a^2}{\mathcal{H}}\partial\mathcal{R}'\cdot\partial\mathcal{R} = \frac{M_{\rm pl}^2 a^2}{\mathcal{H}}\partial_\eta\big[(\partial\mathcal{R})^2\big] \simeq -M_{\rm pl}^2\left(\frac{a^2}{\mathcal{H}}\right)'(\partial\mathcal{R})^2, \tag{79}$$

and since $(a^2/\mathcal{H})' = 2a^2 - a^2\mathcal{H}'/\mathcal{H}^2 = a^2(1+\varepsilon)$, the total gradient coefficient is

$$M_{\rm pl}^2 a^2 - M_{\rm pl}^2 a^2(1+\varepsilon) = -M_{\rm pl}^2 a^2\varepsilon = -a^2 Q_s. \tag{80}$$

Everything assembles into

$$\boxed{S^{(2)} = \int \mathrm{d}\eta\,\mathrm{d}^3x\; a^2 Q_s\Big[\mathcal{R}'^2 - c_s^2(\partial\mathcal{R})^2\Big], \qquad Q_s = \frac{\phi'^2}{2\mathcal{H}^2} = M_{\rm pl}^2\varepsilon, \qquad c_s^2 = 1.} \tag{81}$$

$Q_s > 0$ is automatic for a canonical scalar with $\phi' \neq 0$, and is the statement that $\mathcal{R}$ is not a ghost; $c_s^2 = 1$ reflects the fact that the kinetic term of ϕ was canonical to begin with — had the matter Lagrangian been a general function $P(X,\phi)$ instead, the same manipulations would have produced $c_s^2 = P_{,X}/(P_{,X} + 2XP_{,XX}) \neq 1$.

Second road: cosmic time, exponential parametrisation. To make sure the last result was not an artefact of a lucky choice of variables, the whole calculation can be redone from scratch in cosmic time, writing the spatial metric with an exponential rather than a linear factor,

$$\gamma_{ij} = a^2(t) e^{2\zeta(t,\boldsymbol{x})}\delta_{ij}, \qquad N = 1+\alpha, \qquad N_i = \partial_i\beta, \qquad \phi = \phi(t). \tag{82}$$

Background relations in cosmic time read $\dot\phi^2 = -2M_{\rm pl}^2\dot H$, $V = M_{\rm pl}^2(3H^2+\dot H)$, $\varepsilon = -\dot H/H^2$, translating (23) and (24). Because the exponential form is used, several quantities are exact rather than merely expanded to some order:

$$\dot\gamma_{ij} = 2a^2(H+\dot\zeta)e^{2\zeta}\delta_{ij}, \qquad \gamma^{ij} = a^{-2}e^{-2\zeta}\delta^{ij}, \qquad \dot\gamma^{ij} = -2a^{-2}(H+\dot\zeta)e^{-2\zeta}\delta^{ij}, \tag{83}$$

$$\Gamma^k_{\ ij}[\gamma] = \delta^k_{\ i}\partial_j\zeta + \delta^k_{\ j}\partial_i\zeta - \delta_{ij}\partial_k\zeta, \quad {}^{(3)}\!R = -\frac{2e^{-2\zeta}}{a^2}\Big[2\partial^2\zeta + (\partial\zeta)^2\Big]. \tag{84}$$

Differentiating $\gamma^{ik}\gamma_{kj} = \delta^i_{\ j}$ gives $\dot\gamma^{ik}\gamma_{kj} = -\gamma^{ik}\dot\gamma_{kj}$: the inverse metric's time derivative picks up a minus sign relative to the metric's own. Missing this is, in practice, the single most common slip in this entire calculation.

Expanding (47) to $\mathcal{O}(\epsilon^2)$ in α,β,ζ and using the background relations to simplify coefficients as they appear,

$$\begin{aligned}\frac{L^{(2)}}{a^3} = {}& \Big(-3M_{\rm pl}^2H^2 + \tfrac12\dot\phi^2\Big)\alpha^2 + 6M_{\rm pl}^2H\alpha\dot\zeta - \Big(3M_{\rm pl}^2\dot H + \tfrac32\dot\phi^2\Big)\alpha\zeta - \frac{2M_{\rm pl}^2}{a^2}\alpha\,\partial^2\zeta - \frac{2M_{\rm pl}^2H}{a^2}\alpha\,\partial^2\beta \\ & - \Big(27M_{\rm pl}^2H^2 + \tfrac92 M_{\rm pl}^2\dot H - \tfrac94\dot\phi^2\Big)\zeta^2 - 18M_{\rm pl}^2H\zeta\dot\zeta + \frac{2M_{\rm pl}^2H}{a^2}\Big(\zeta\partial^2\beta + \partial\beta\cdot\partial\zeta\Big) \\ & - \frac{2M_{\rm pl}^2}{a^2}\zeta\partial^2\zeta - \frac{M_{\rm pl}^2}{a^2}(\partial\zeta)^2 - 3M_{\rm pl}^2\dot\zeta^2 + \frac{2M_{\rm pl}^2}{a^2}\dot\zeta\partial^2\beta - \frac{M_{\rm pl}^2}{a^4}\Big[(\partial^2\beta)^2 - \partial_i\partial_j\beta\,\partial_i\partial_j\beta\Big].\end{aligned} \tag{85}$$

On the background, the $\alpha\zeta$ coefficient already vanishes, $3M_{\rm pl}^2\dot H + \frac32\dot\phi^2 = 3M_{\rm pl}^2\dot H - 3M_{\rm pl}^2\dot H = 0$, and the α^2 coefficient is $-M_{\rm pl}^2(3H^2+\dot H)$.

Vary β. Three sets of terms involve it: the last term of (85), which is the three-dimensional divergence already dealt with above; the pair $\frac{2M_{\rm pl}^2H}{a^2}(\zeta\partial^2\beta + \partial\beta\cdot\partial\zeta) = \frac{2M_{\rm pl}^2H}{a^2}\partial_i(\zeta\partial_i\beta)$, likewise a pure divergence; and the genuine pair $2M_{\rm pl}^2a(\dot\zeta - H\alpha)\partial^2\beta$. Only the last one survives variation, giving

$$\partial^2(\dot\zeta - H\alpha) = 0 \implies \boxed{\alpha = \frac{\dot\zeta}{H}}, \tag{86}$$

the cosmic-time twin of (74). Varying α,

$$2(-3M_{\rm pl}^2H^2 - M_{\rm pl}^2\dot H)\alpha + 6M_{\rm pl}^2H\dot\zeta - \frac{2M_{\rm pl}^2}{a^2}\partial^2\zeta - \frac{2M_{\rm pl}^2H}{a^2}\partial^2\beta = 0, \tag{87}$$

and substituting $\alpha = \dot\zeta/H$ makes most of the $\dot\zeta$ terms cancel, leaving

$$\frac{\partial^2\beta}{a^2} = -\frac{\dot H}{H^2}\dot\zeta - \frac{\partial^2\zeta}{Ha^2} = \varepsilon\,\dot\zeta - \frac{\partial^2\zeta}{Ha^2}, \tag{88}$$

solved by

$$\boxed{\beta = -\frac{\zeta}{H} + a^2\varepsilon\,\partial^{-2}\dot\zeta,} \tag{89}$$

where ∂^{-2} denotes the inverse Laplacian. (As a check: $\partial^2(-\zeta/H)/a^2 = -\partial^2\zeta/(Ha^2)$ and $\partial^2(a^2\varepsilon\partial^{-2}\dot\zeta)/a^2 = \varepsilon\dot\zeta$, matching the equation just solved.)
Substituting $\alpha = \dot\zeta/H$ into (85), and dropping the two divergences already identified — which is what lets β disappear without ever invoking (89) explicitly —

$$\frac{L^{(2)}_{\rm red}}{a^3} = -M_{\rm pl}^2\frac{\dot H}{H^2}\dot\zeta^2 - \Big(27M_{\rm pl}^2H^2 + 9M_{\rm pl}^2\dot H\Big)\zeta^2 - 18M_{\rm pl}^2H\zeta\dot\zeta - \frac{2M_{\rm pl}^2}{a^2}\zeta\partial^2\zeta - \frac{M_{\rm pl}^2}{a^2}(\partial\zeta)^2 - \frac{2M_{\rm pl}^2}{Ha^2}\dot\zeta\partial^2\zeta, \tag{90}$$

where $\dot\phi^2 = -2M_{\rm pl}^2\dot H$ has already turned $-\frac{9}{2}M_{\rm pl}^2\dot H + \frac{9}{4}\dot\phi^2$ into $-9M_{\rm pl}^2\dot H$. Three cleanups finish it. The kinetic term is immediately $-M_{\rm pl}^2(\dot H/H^2)\dot\zeta^2 = M_{\rm pl}^2\varepsilon\dot\zeta^2$. The mass terms cancel much as before:

$$-18M_{\rm pl}^2a^3H\zeta\dot\zeta = -9M_{\rm pl}^2a^3H(\zeta^2)^{\cdot} \simeq +9M_{\rm pl}^2(a^3H)^{\cdot}\zeta^2 = 9M_{\rm pl}^2a^3(3H^2+\dot H)\zeta^2, \tag{91}$$

exactly cancelling $-(27M_{\rm pl}^2H^2 + 9M_{\rm pl}^2\dot H)a^3\zeta^2$. And the gradient terms combine as, first, $-2M_{\rm pl}^2a\,\zeta\partial^2\zeta \simeq +2M_{\rm pl}^2a(\partial\zeta)^2$, leaving $+M_{\rm pl}^2a(\partial\zeta)^2$ once added to the explicit $-M_{\rm pl}^2a(\partial\zeta)^2$, and then

$$-\frac{2M_{\rm pl}^2a}{H}\dot\zeta\partial^2\zeta \simeq \frac{2M_{\rm pl}^2a}{H}\partial\dot\zeta\cdot\partial\zeta = \frac{M_{\rm pl}^2a}{H}\partial_t\big[(\partial\zeta)^2\big] \simeq -M_{\rm pl}^2\Big(\frac{a}{H}\Big)^{\cdot}(\partial\zeta)^2 = -M_{\rm pl}^2a(1+\varepsilon)(\partial\zeta)^2, \tag{92}$$

using $(a/H)^{\cdot} = a(1+\varepsilon)$, so the net gradient coefficient is $M_{\rm pl}^2a - M_{\rm pl}^2a(1+\varepsilon) = -M_{\rm pl}^2a\varepsilon$. Altogether,

$$\boxed{S^{(2)} = \int \mathrm{d}t\,\mathrm{d}^3x\;a^3\,M_{\rm pl}^2\varepsilon\left[\dot\zeta^2 - \frac{(\partial\zeta)^2}{a^2}\right] = \frac{1}{2}\int \mathrm{d}t\,\mathrm{d}^3x\;a^3\frac{\dot\phi^2}{H^2}\left[\dot\zeta^2 - \frac{(\partial\zeta)^2}{a^2}\right].} \tag{93}$$

Translating back to conformal time with $\mathrm{d}t = a\mathrm{d}\eta$, $\dot\zeta = \mathcal{R}'/a$, $\zeta = \mathcal{R}$,

$$S^{(2)} = \int a\,\mathrm{d}\eta\,\mathrm{d}^3x\;a^3M_{\rm pl}^2\varepsilon\frac{\mathcal{R}'^2 - (\partial\mathcal{R})^2}{a^2} = \int \mathrm{d}\eta\,\mathrm{d}^3x\;a^2M_{\rm pl}^2\varepsilon\Big[\mathcal{R}'^2 - (\partial\mathcal{R})^2\Big], \tag{94}$$

which, since $M_{\rm pl}^2\varepsilon = \phi'^2/(2\mathcal{H}^2) = Q_s$, is exactly (81) with $c_s^2 = 1$. The two roads — one built on a linear conformal-factor expansion in conformal time, the other on an exponential one in cosmic time, with entirely different intermediate expressions and constraint solutions — arrive at the same destination.

What the action says. Varying (81) directly,

$$\frac{\partial L}{\partial\mathcal{R}'} = 2a^2Q_s\mathcal{R}', \qquad \frac{\partial L}{\partial(\partial_i\mathcal{R})} = -2a^2Q_sc_s^2\partial_i\mathcal{R}, \qquad \frac{\partial L}{\partial\mathcal{R}} = 0, \tag{95}$$

gives $\big(a^2Q_s\mathcal{R}'\big)' - a^2Q_sc_s^2\partial^2\mathcal{R} = 0$, or

$$\boxed{\mathcal{R}'' + \frac{(a^2Q_s)'}{a^2Q_s}\mathcal{R}' - c_s^2\partial^2\mathcal{R} = 0,} \tag{96}$$

a friction term $2\mathcal{H} + Q_s'/Q_s$ that is Hubble friction plus a slow-roll correction. Passing to Fourier space, $\mathcal{R}(\eta,\boldsymbol{x}) = (2\pi)^{-3}\int \mathrm{d}^3\boldsymbol{k}\,\mathcal{R}_k(\eta)e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$ and $\partial^2 \to -k^2$ mode by mode,

$$\boxed{\mathcal{R}_k'' + \frac{(a^2Q_s)'}{a^2Q_s}\mathcal{R}_k' + c_s^2k^2\mathcal{R}_k = 0.} \tag{97}$$

This equation has no mass term, and that is the reason $\mathcal{R}$ freezes once a mode leaves the horizon: the two independent solutions on super-horizon scales are $\mathcal{R} = \text{const}$ and $\mathcal{R} \propto \int \mathrm{d}\eta/(a^2 Q_s)$, the second decaying during inflation, leaving only the constant mode. That the mass term cancelled — in Route A at step (D4), in Route B in the cancellation of ζ^2 terms — is exactly this statement in embryo, worked out well before any horizon-crossing argument is invoked.

Defining the Mukhanov–Sasaki variable

$$z \equiv a\sqrt{2Q_s} = a\frac{|\phi'|}{\mathcal{H}} = aM_{\rm pl}\sqrt{2\varepsilon}, \qquad v \equiv z\mathcal{R}, \tag{98}$$

so that $a^2 Q_s = \frac{1}{2}z^2$, the action becomes canonically normalised,

$$S^{(2)} = \frac{1}{2}\int \mathrm{d}\eta\, \mathrm{d}^3x \Big[v'^2 - c_s^2(\partial v)^2 + \frac{z''}{z}v^2\Big]. \tag{99}$$

Writing $\mathcal{R} = v/z$ and using $z^{-2}(z^2\mathcal{R}')' = \mathcal{R}'' + 2(z'/z)\mathcal{R}' = z^{-1}[v'' - (z''/z)v]$, Eq. (97) becomes

$$\boxed{v_k'' + \Big(c_s^2 k^2 - \frac{z''}{z}\Big)v_k = 0,} \tag{100}$$

formally the equation of a scalar field in flat space with a time-dependent mass $m^2(\eta) = -z''/z$: every trace of gravity's interaction with the inflaton has been repackaged into that one function of time. This is exactly the form needed to quantise the field with Bunch–Davies initial conditions, $v_k \to e^{-ic_s k\eta}/\sqrt{2c_s k}$ as $c_s k\eta \to -\infty$, though following that road further belongs elsewhere.

In brief, then: the Friedmann and Klein–Gordon equations came from expanding the action to first order in a homogeneous fluctuation; the geometry of a perturbed universe was organised so that one combination, $\mathcal{R}$, survived every change of coordinates; casting gravity in ADM form made the lapse and shift into Lagrange multipliers that could be solved for and substituted away exactly; doing so twice, by unrelated routes, both times produced

$$S^{(2)} = \int \mathrm{d}\eta\, \mathrm{d}^3x\; a^2 Q_s\Big[\mathcal{R}'^2 - (\partial\mathcal{R})^2\Big], \qquad Q_s = M_{\rm pl}^2\varepsilon, \tag{101}$$

a single canonically normalisable scalar degree of freedom, massless on super-horizon scales, with a kinetic normalisation controlled entirely by how fast the inflaton is rolling.

The direct route, without ADM.

As a cross-check, the same second-order action can be reached by expanding $\sqrt{-g}^{(4)}R$ directly in the comoving-gauge metric (9), never passing through the extrinsic curvature of Section 3 at all. This route is not shortcut anywhere below: every tensor between the metric and the Ricci scalar is written down and contracted explicitly, first order completely by hand, second order by the identical method carried one step further, with only the longest mechanical substitution handed to a computer algebra system rather than to arithmetic patience.

Write the metric as $g_{\mu\nu} = \bar{g}_{\mu\nu} + h_{\mu\nu}$, with $\bar{g}_{00} = -a^2$, $\bar{g}_{ij} = a^2\delta_{ij}$ the background and

$$h_{00} = -2a^2 A, \qquad h_{0i} = a^2\partial_i B, \qquad h_{ij} = 2a^2\mathcal{R}\delta_{ij} \tag{102}$$

the perturbation — exact, not truncated, since $A, B, \mathcal{R}$ enter the metric linearly by construction. Two things are not linear in them: the inverse metric, and everything built from it.

The determinant needs no perturbative expansion at all, only a standard fact about block matrices. Writing $g_{\mu\nu}$ in $(0, i)$ block form with $g_{00} = -N^2 + \gamma_{ij}N^iN^j$, $g_{0i} = \gamma_{ij}N^j$, $g_{ij} = \gamma_{ij}$

(purely a relabelling of the metric components, not an appeal to any extrinsic curvature or foliation physics), row-reducing the 0-row against the spatial rows removes the off-diagonal blocks and leaves $\det g = -N^2 \det\gamma$, i.e. $\sqrt{-g} = N\sqrt{\gamma}$. Here $\gamma_{ij} = a^2(1+2\mathcal{R})\delta_{ij}$, $\sqrt{\gamma} = a^3(1+2\mathcal{R})^{3/2}$, and $N^2 = -g_{00} + \gamma^{ij}g_{0i}g_{0j} = a^2(1+2A) + a^2(\partial B)^2/(1+2\mathcal{R})$, so that

$$\begin{aligned}\sqrt{-g} &= a^4(1+2\mathcal{R})^{3/2}\sqrt{1+2A+\frac{(\partial B)^2}{1+2\mathcal{R}}} \\ &= a^4\Big[1 + (A+3\mathcal{R}) + \Big(3\mathcal{R}^2 + 3A\mathcal{R} - \tfrac{1}{2}A^2 + \tfrac{1}{2}(\partial B)^2\Big)\Big] + \mathcal{O}(3),\end{aligned} \tag{103}$$

expanding the square root and the $3/2$ power to second order exactly as in Section 5. The inverse metric is genuinely non-linear in $h_{\mu\nu}$, and its expansion is the matrix Neumann series $g^{-1} = (\bar g + h)^{-1} = \bar g^{-1} - \bar g^{-1}h\bar g^{-1} + \bar g^{-1}h\bar g^{-1}h\bar g^{-1} - \dots$, truncated at two powers of h. It is convenient to raise both indices of h with the background metric, $H^{\mu\nu} \equiv \bar g^{\mu\alpha}\bar g^{\nu\beta}h_{\alpha\beta}$ — itself exactly first order — so that $g^{\mu\nu} = \bar g^{\mu\nu} - H^{\mu\nu} + H^{\mu}{}_{\rho}H^{\rho\nu} + \mathcal{O}(3)$. Componentwise, with $\bar g^{00} = -a^{-2}$, $\bar g^{ij} = a^{-2}\delta^{ij}$,

$$H^{00} = -\frac{2A}{a^2}, \qquad H^{0i} = \frac{\partial^i B}{a^2}, \qquad H^{ij} = \frac{2\mathcal{R}}{a^2}\delta^{ij}, \tag{104}$$

and carrying the series to second order gives

$$\begin{aligned} g^{00} &= -\frac{1 - 2A + 4A^2 - (\partial B)^2}{a^2}, \qquad g^{0i} = \frac{\partial^i B\,(1 - 2A - 2\mathcal{R})}{a^2} + \mathcal{O}(3), \\ g^{ij} &= \frac{\delta^{ij}(1 - 2\mathcal{R} + 4\mathcal{R}^2) - \partial^i B\partial^j B}{a^2} + \mathcal{O}(3). \end{aligned} \tag{105}$$

With $g_{\mu\nu}$ and $g^{\mu\nu}$ both in hand, the Christoffel symbols follow from their definition, $\Gamma^{\lambda}_{\mu\nu} = \frac{1}{2}g^{\lambda\kappa}(\partial_\mu g_{\kappa\nu} + \partial_\nu g_{\kappa\mu} - \partial_\kappa g_{\mu\nu})$, order by order in ϵ. At first order only $\bar g^{\lambda\kappa}$ multiplies ∂h, plus one cross term where $H^{\lambda\kappa}$ multiplies the background piece $\partial\bar g$; the latter is exactly $-H^{\lambda}{}_{\sigma}\bar\Gamma^{\sigma}_{\mu\nu}$ (with $H^{\lambda}{}_{\sigma} \equiv \bar g^{\lambda\kappa}h_{\kappa\sigma}$), so

$$\delta^{(1)}\Gamma^{\lambda}{}_{\mu\nu} = \tfrac{1}{2}\bar g^{\lambda\kappa}(\partial_\mu h_{\kappa\nu} + \partial_\nu h_{\kappa\mu} - \partial_\kappa h_{\mu\nu}) - H^{\lambda}{}_{\sigma}\bar\Gamma^{\sigma}_{\mu\nu}. \tag{106}$$

Two components worked out explicitly show the pattern. For $\Gamma^0{}_{00}$, only $\bar g^{00} = -a^{-2}$ enters and $\bar\Gamma^0_{00} = \mathcal{H}$ needs no correction since $H^0{}_0\bar\Gamma^0_{00}$ would require $\bar\Gamma^0_{00}$ evaluated at background order times a first-order H, which is indeed present:

$$\begin{aligned}\delta^{(1)}\Gamma^0{}_{00} &= \tfrac{1}{2}(-a^{-2})(2\partial_\eta h_{00} - \partial_\eta h_{00}) - H^0{}_0\mathcal{H} = \tfrac{1}{2}(-a^{-2})\partial_\eta(-2a^2A) - 2A\mathcal{H} \\ &= \partial_\eta(a^2A)/a^2 - 2A\mathcal{H} = A' + 2A\mathcal{H} - 2A\mathcal{H} = A'.\end{aligned} \tag{107}$$

For $\Gamma^i{}_{00}$, using $\bar g^{ij} = a^{-2}\delta^{ij}$ and $\bar\Gamma^i_{00} = 0$ (no correction term needed),

$$\delta^{(1)}\Gamma^i{}_{00} = \tfrac{1}{2}a^{-2}\delta^{ij}(2\partial_\eta h_{j0} - \partial_j h_{00}) = \tfrac{1}{2}a^{-2}(2\partial_\eta(a^2\partial^i B) + 2a^2\partial^i A) = \partial^i A + (\partial^i B)' + \mathcal{H}\partial^i B. \tag{108}$$

The remaining four independent components follow the same substitution into (106) and are simple enough to record directly:

$$\delta^{(1)}\Gamma^0{}_{0i} = \partial_i A + \mathcal{H}\partial_i B, \qquad \delta^{(1)}\Gamma^0{}_{ij} = [\mathcal{R}' + 2\mathcal{H}(\mathcal{R} - A)]\delta_{ij} - \partial_i\partial_j B, \tag{109}$$

$$\delta^{(1)}\Gamma^i{}_{0j} = \mathcal{R}'\delta^i{}_j, \qquad \delta^{(1)}\Gamma^i{}_{jk} = \delta^i{}_j\partial_k\mathcal{R} + \delta^i{}_k\partial_j\mathcal{R} - \delta_{jk}\partial^i\mathcal{R} - \mathcal{H}\delta_{jk}\partial^i B. \tag{110}$$

The Ricci tensor is at most *quadratic* in the connection — there are no higher powers in $R_{\mu\nu} = \partial_\lambda \Gamma^\lambda{}_{\mu\nu} - \partial_\nu \Gamma^\lambda{}_{\lambda\mu} + \Gamma^\lambda{}_{\lambda\sigma}\Gamma^\sigma{}_{\mu\nu} - \Gamma^\lambda{}_{\nu\sigma}\Gamma^\sigma{}_{\lambda\mu}$ — and since $\delta\Gamma \equiv \Gamma - \bar{\Gamma}$ is a genuine tensor, substituting $\Gamma = \bar{\Gamma} + \delta\Gamma$ into this definition and using $\bar{\nabla}$ (the background covariant derivative) to package the partial derivatives gives an identity that holds to *all* orders at once, not just to leading order:

$$R_{\mu\nu} = \bar{R}_{\mu\nu} + \Big[\bar{\nabla}_\lambda \delta\Gamma^\lambda{}_{\mu\nu} - \bar{\nabla}_\nu \delta\Gamma^\lambda{}_{\lambda\mu}\Big] + \Big[\delta\Gamma^\lambda{}_{\lambda\sigma}\delta\Gamma^\sigma{}_{\mu\nu} - \delta\Gamma^\lambda{}_{\nu\sigma}\delta\Gamma^\sigma{}_{\lambda\mu}\Big]. \tag{111}$$

Expanding $\delta\Gamma = \delta^{(1)}\Gamma + \delta^{(2)}\Gamma + \dots$ and collecting powers of ϵ, the first-order piece needs only the square bracket linear in $\delta\Gamma$,

$$\delta^{(1)} R_{\mu\nu} = \bar{\nabla}_\lambda \delta^{(1)}\Gamma^\lambda{}_{\mu\nu} - \bar{\nabla}_\nu \delta^{(1)}\Gamma^\lambda{}_{\lambda\mu}. \tag{112}$$

Working this out (background covariant derivatives contracted against (109)–(110), using the background Christoffels $\bar{\Gamma}^0_{00} = \mathcal{H}$, $\bar{\Gamma}^0_{ij} = \mathcal{H}\delta_{ij}$, $\bar{\Gamma}^i_{0j} = \mathcal{H}\delta^i{}_j$) gives the three independent components,

$$\delta^{(1)} R_{00} = \partial^2 A - 3\mathcal{R}'' + \big(\partial^2 B\big)' + 3\mathcal{H}A' + \mathcal{H}\partial^2 B - 3\mathcal{H}\mathcal{R}', \tag{113}$$

$$\delta^{(1)} R_{0i} = -2\partial_i \mathcal{R}' + 2\mathcal{H}\partial_i A + \big(\mathcal{H}' + \mathcal{H}^2\big)\partial_i B, \tag{114}$$

$$\begin{aligned}\delta^{(1)} R_{ij} = {}& \delta_{ij}\Big[2(\mathcal{R} - A)(\mathcal{H}' + 2\mathcal{H}^2) + \mathcal{R}'' - \partial^2\mathcal{R} - \mathcal{H}A' - \mathcal{H}\partial^2 B + 5\mathcal{H}\mathcal{R}'\Big] \\ & - \partial_i\partial_j\big[A + \mathcal{R} + B' + 2\mathcal{H}B\big].\end{aligned} \tag{115}$$

The off-diagonal check is immediate: for $i \neq j$ the bracket in (115) drops out and $\delta^{(1)}R_{xy} = -\partial_x\partial_y[A + \mathcal{R} + B' + 2\mathcal{H}B]$, which is exactly what a direct contraction of (109)–(110) gives for that single component — a useful cross-check that costs one line.

Two background quantities are needed to turn $\delta^{(1)}R_{\mu\nu}$ into $\delta_1{}^{(4)}R$: the background Ricci tensor itself, obtained from the same defining formula evaluated on $\bar{\Gamma}^0_{00} = \mathcal{H}$, $\bar{\Gamma}^0_{ij} = \mathcal{H}\delta_{ij}$, $\bar{\Gamma}^i_{0j} = \mathcal{H}\delta^i{}_j$ alone,

$$\bar{R}_{00} = -3\mathcal{H}', \qquad \bar{R}_{ij} = (\mathcal{H}' + 2\mathcal{H}^2)\delta_{ij}, \tag{116}$$

which one checks satisfies $a^2{}^{(4)}\bar{R} = -\bar{R}_{00} + \delta^{ij}\bar{R}_{ij}\cdot\frac{a^2}{a^2} = 3\mathcal{H}' + 3(\mathcal{H}' + 2\mathcal{H}^2) = 6(\mathcal{H}' + \mathcal{H}^2)$, reproducing the background trace used throughout Section 3. Since ${}^{(4)}R = g^{\mu\nu}R_{\mu\nu}$, the first-order piece of this product is

$$\delta_1{}^{(4)}R = \delta^{(1)}g^{\mu\nu}\bar{R}_{\mu\nu} + \bar{g}^{\mu\nu}\delta^{(1)}R_{\mu\nu}, \tag{117}$$

an ordinary product rule with nothing hidden in it. Reading $\delta^{(1)}g^{\mu\nu}$ off (105), the first term contracts to $a^{-2}\big[2A(-3\mathcal{H}') + (-2\mathcal{R})\cdot 3(\mathcal{H}' + 2\mathcal{H}^2)\big] = -a^{-2}\big[6A\mathcal{H}' + 6\mathcal{R}(\mathcal{H}' + 2\mathcal{H}^2)\big]$, while the second term, using $\bar{g}^{00} = -a^{-2}$ and $\bar{g}^{ij} = a^{-2}\delta^{ij}$ on (113)–(115), contracts to $a^{-2}\big[-\delta^{(1)}R_{00} + 3\big(2(\mathcal{R} - A)(\mathcal{H}' + 2\mathcal{H}^2) + \mathcal{R}'' - \partial^2\mathcal{R} - \mathcal{H}A' - \mathcal{H}\partial^2 B + 5\mathcal{H}\mathcal{R}'\big) - \partial^2(A + \mathcal{R} + B' + 2\mathcal{H}B)\big]$. Adding the two and cancelling the $\mathcal{R}(\mathcal{H}' + 2\mathcal{H}^2)$ and $A(\mathcal{H}' + 2\mathcal{H}^2)$ pieces that appear with opposite sign in each (six such terms in all, the same cancellation pattern already seen among the algebraic coefficients of Section 5) leaves, after multiplying through by a^2,

$$a^2\delta_1{}^{(4)}R = -2\partial^2 A - 12(\mathcal{H}' + \mathcal{H}^2)A - 6\mathcal{H}A' + 6\mathcal{R}'' + 18\mathcal{H}\mathcal{R}' - 4\partial^2\mathcal{R} - 2\big(\partial^2 B\big)' - 6\mathcal{H}\partial^2 B, \tag{118}$$

exactly the first-order Ricci scalar used in Section 3, now obtained without appeal to $\sqrt{-g}$ or the extrinsic curvature at all.

The second order proceeds by the identical logic, and one structural fact keeps it from doubling in length: the second-order Christoffel correction is not a new object. Carrying the Neumann series for $g^{\lambda\kappa}$ one power further and substituting into the definition of Γ, the two pieces that arise — one from $H^{\lambda}_{\ \rho}H^{\rho\kappa}$ multiplying the background bracket, one from $-H^{\lambda\kappa}$ multiplying the first-order bracket — combine so that every term built purely from H and $\bar{\Gamma}$ cancels, leaving

$$\boxed{\delta^{(2)}\Gamma^{\lambda}_{\ \mu\nu} = -H^{\lambda}_{\ \kappa}\,\delta^{(1)}\Gamma^{\kappa}_{\ \mu\nu}\,,} \tag{119}$$

with $H^{\lambda}_{\ \kappa} \equiv \bar{g}^{\lambda\rho}h_{\rho\kappa}$ exactly as before. This identity was checked component by component (all 64 combinations of λ,μ,ν) directly against the non-linear definition of Γ and holds without exception. It says that nothing qualitatively new happens at second order in the connection: the correction is just the first-order correction, reprocessed once more through the metric perturbation itself.

The same is true one level up. Collecting the $\mathcal{O}(\epsilon^2)$ piece of the exact identity (111), and using (119) to eliminate $\delta^{(2)}\Gamma$ in favour of $\delta^{(1)}\Gamma$,

$$\begin{aligned}\delta^{(2)}R_{\mu\nu} = &-\bar{\nabla}_{\lambda}(H^{\lambda}_{\ \kappa}\delta^{(1)}\Gamma^{\kappa}_{\ \mu\nu}) + \bar{\nabla}_{\nu}(H^{\lambda}_{\ \kappa}\delta^{(1)}\Gamma^{\kappa}_{\ \lambda\mu})\\ &+\delta^{(1)}\Gamma^{\lambda}_{\ \lambda\sigma}\delta^{(1)}\Gamma^{\sigma}_{\ \mu\nu} - \delta^{(1)}\Gamma^{\lambda}_{\ \nu\sigma}\delta^{(1)}\Gamma^{\sigma}_{\ \lambda\mu},\end{aligned} \tag{120}$$

built entirely out of $H^{\lambda}_{\ \kappa}$ and $\delta^{(1)}\Gamma$, both already tabulated in (109)–(110). No new kind of object appears anywhere in the calculation from here to the end; what remains is exclusively the work of expanding the background covariant derivatives and multiplying out roughly a dozen contractions of already-known tensors. That step was carried out symbolically rather than by hand, precisely because it is long without being illuminating — every one of the resulting terms is a mechanical product of quantities already written down above — and the result was cross-checked two independent ways: once by evaluating (120) term by term, and once by expanding Γ to second order directly from its non-linear definition with no appeal to (111) at all. The two agree termwise, and contracting either against $g^{\mu\nu} = \bar{g}^{\mu\nu} + \delta^{(1)}g^{\mu\nu} + \delta^{(2)}g^{\mu\nu}$ at second order gives

$$\begin{aligned}a^2\delta_2{}^{(4)}R = {}& 24(\mathcal{H}^2+\mathcal{H}')A^2 + 24\mathcal{H}AA' + 4A\partial^2 A + 2(\partial A)^2\\ &- 12A\mathcal{R}'' - 36\mathcal{H}A\mathcal{R}' - 6A'\mathcal{R}' + 12\mathcal{H}A\partial^2 B + 4A(\partial^2 B)' + 2A'\partial^2 B\\ &+ 4\mathcal{R}\partial^2 A - 2\partial A\cdot\partial\mathcal{R} + 6\mathcal{H}\partial A\cdot\partial B\\ &- 12\mathcal{R}\mathcal{R}'' - 36\mathcal{H}\mathcal{R}\mathcal{R}' + 16\mathcal{R}\partial^2\mathcal{R} + 6(\partial\mathcal{R})^2 + 12\mathcal{H}\mathcal{R}\partial^2 B + 4\mathcal{R}(\partial^2 B)'\\ &- 6\mathcal{H}\partial B\cdot\partial\mathcal{R} - 8\partial B\cdot\partial\mathcal{R}' - 4\mathcal{R}'\partial^2 B - 2\partial\mathcal{R}\cdot\partial B'\\ &- 6(\mathcal{H}^2+\mathcal{H}')(\partial B)^2 - 6\mathcal{H}\partial B\cdot\partial B' + (\partial^2 B)^2 - \partial_i\partial_j B\,\partial_i\partial_j B.\end{aligned} \tag{121}$$

As a partial hand check on (121): the piece of it coming purely from $\delta^{(2)}g^{\mu\nu}\bar{R}_{\mu\nu}$ — the easiest of the three contractions in the second-order analogue of (117) — is, using $\delta^{(2)}g^{00} = a^{-2}[-4A^2 + (\partial B)^2]$ and the trace of $\delta^{(2)}g^{ij}$ from (105) together with (116), exactly $12A^2\mathcal{H}' + 12\mathcal{R}^2\mathcal{H}' + 24\mathcal{R}^2\mathcal{H}^2 - 4\mathcal{H}'(\partial B)^2 - 2\mathcal{H}^2(\partial B)^2$, a clean six-term result that already accounts for part of the A^2, $\mathcal{R}^2$ and $(\partial B)^2$ coefficients appearing above; the remaining two contractions, $\delta^{(1)}g^{\mu\nu}\delta^{(1)}R_{\mu\nu}$ and $\bar{g}^{\mu\nu}\delta^{(2)}R_{\mu\nu}$, are individually about as long as (121) itself and supply the rest.

The second time derivatives $\mathcal{R}''$ and $A'\mathcal{R}'$ in (121), and the mixed $(\partial^2 B)'$ structures, are exactly what the extrinsic-curvature rewriting of Section 3 removes before any expansion is performed — this appendix simply shows where they come from and that they cancel correctly once the background equations of motion, (23), are imposed. Assembling

$L = \sqrt{-g}[\frac{M_{\rm pl}^2}{2}{}^{(4)}R - \frac{1}{2}g^{00}\phi'^2 - V]$, isolating the $\mathcal{O}(2)$ piece, and imposing (23), gives a Lagrangian that differs from the integrand of (73) only by a total four-divergence, checked directly by computer algebra in the closing appendix below.

Where the computer algebra was used.
Every non-trivial algebraic step above was checked with SymPy. A bookkeeping parameter ϵ multiplies each perturbation, and `trunc` discards $\mathcal{O}(\epsilon^3)$ and higher. The first four blocks below are excerpts from one running conformal-time session and share its setup (the metric, its inverse, the Christoffel symbols, and the ADM quantities N, γ_{ij}, E_{ij} built from them) rather than standing alone; the fifth, in cosmic time, is a separate session and is written out in full for exactly that reason.

The Gauss–Codazzi rewriting, boundary vector included.

```
    LHS  = trunc(sp.expand(sqrtmg*R4))                    # sqrt(-g) * {}^(4)R
    K    = trE/N                                          # K = E/N
    bulk = sg*( N*R3 + (EE - trE**2)/N )
    J0   = -sg*K
    Ji   = [ sg*( sum(gi[i,j]*sp.diff(N,Y[j]) for j in range(3)) + K*Nu[i] ) for i in
range(3)]
    div  = sp.diff(J0,eta) + sum(sp.diff(Ji[i],Y[i]) for i in range(3))
    RHS  = ser(sp.expand(bulk - 2*div))
    print([sp.simplify(sp.Poly(trunc(sp.expand(LHS-RHS)),eps).coeff_monomial(eps**k))
    for k in range(3)])
    # -> [0, 0, 0]
```

Covariant $L^{(2)\cdot}$ against Eq. (73)

```
    from sympy.calculus.euler import euler_equations
    H  = sp.diff(a,eta)/a
    lap = lambda f: sp.diff(f,x,2)+sp.diff(f,y,2)+sp.diff(f,z,2)
    gr2 = lambda f: sp.diff(f,x)**2+sp.diff(f,y)**2+sp.diff(f,z)**2
    Lq = a**2*(-3*Mp**2*sp.diff(R,eta)**2 + 2*Mp**2*sp.diff(R,eta)*lap(B)
    - 2*Mp**2*H*A*lap(B) + Mp**2*gr2(R) - 2*Mp**2*A*lap(R)
    + 6*Mp**2*H*A*sp.diff(R,eta)
    - sp.Rational(1,2)*(6*H**2*Mp**2 - pp**2)*A**2)
    D = sp.expand(L2 - Lq).subs(V, Mp**2*(2*H**2+sp.diff(H,eta))/a**2)
    D = sp.expand(D.subs(pp, sp.sqrt(2)*Mp*sp.sqrt(H**2 - sp.diff(H,eta))))
    print([sp.simplify(e.lhs-e.rhs) for e in euler_equations(D,[A,B,R],[eta,x,y,z])])
    # -> [0, 0, 0]
```

Reduction to the canonical form, first route.

```
    Rp = sp.diff(R,eta)
    Lc = a**2*(-3*Mp**2*Rp**2 + Mp**2*gr2(R) - 2*Mp**2*(Rp/H)*lap(R)
    + 6*Mp**2*Rp**2 - sp.Rational(1,2)*(6*H**2*Mp**2-pp**2)*(Rp/H)**2)
    Lt = a**2*(pp**2/(2*H**2))*(Rp**2 - gr2(R))
    D  = sp.expand(Lc-Lt).subs(pp, sp.sqrt(2)*Mp*sp.sqrt(H**2-sp.diff(H,eta)))
    print([sp.simplify(e.lhs-e.rhs) for e in euler_equations(D,[R],[eta,x,y,z])])   #
-> [0]
```

Second route. This block is an independent Python session in cosmic time — its `a`, `x`, `y`, `z` are not the conformal-time symbols used above, and none of `sg`, `R3`, `EE`, `trE` carry over from the earlier blocks: reusing those would silently evaluate the old (Route A) geometry instead of this one, which is the one failure mode worth flagging explicitly. Every geometric quantity is therefore rebuilt from scratch for the metric $\gamma_{ij} = a^2 e^{2\epsilon\zeta}\delta_{ij}$.

```
    t,x,y,z = sp.symbols('t x y z', real=True); Y = [x,y,z]
```

```
    a  = sp.Function('a')(t)
    ze = sp.Function('zeta')(t,x,y,z); al = sp.Function('alpha')(t,x,y,z); be = sp.
Function('beta')(t,x,y,z)
    pd = sp.Function('phidot')(t)

    gam = sp.Matrix(3,3, lambda i,j: a**2*sp.exp(2*eps*ze) if i==j else 0)
    gi  = sp.Matrix(3,3, lambda i,j: sp.exp(-2*eps*ze)/a**2 if i==j else 0)
    sg  = a**3*sp.exp(3*eps*ze)
    N   = 1 + eps*al
    Nl  = [eps*sp.diff(be,Y[i]) for i in range(3)]

    G = [[[0]*3 for _ in range(3)] for _ in range(3)]
    for l in range(3):
    for m in range(3):
    for n in range(m,3):
    s = sum(gi[l,k]*(sp.diff(gam[k,m],Y[n])+sp.diff(gam[k,n],Y[m])-sp.diff(gam[m,n],Y
[k]))
    for k in range(3))
    s = sp.simplify(s/2); G[l][m][n] = s; G[l][n][m] = s
    Ric3 = sp.zeros(3,3)
    for m in range(3):
    for n in range(m,3):
    s = 0
    for l in range(3):
    s += sp.diff(G[l][m][n],Y[l]) - sp.diff(G[l][m][l],Y[n])
    for k in range(3):
    s += G[l][l][k]*G[k][m][n] - G[l][n][k]*G[k][m][l]
    Ric3[m,n] = sp.expand(s); Ric3[n,m] = Ric3[m,n]
    R3  = sp.expand(sum(gi[m,n]*Ric3[m,n] for m in range(3) for n in range(3)))

    DN  = sp.Matrix(3,3, lambda i,j: sp.diff(Nl[j],Y[i]) - sum(G[k][i][j]*Nl[k] for k
 in range(3)))
    E   = sp.Matrix(3,3, lambda i,j: sp.Rational(1,2)*(sp.diff(gam[i,j],t) - DN[i,j]
- DN[j,i]))
    Eup = gi*E*gi
    trE = sp.expand(sum(gi[i,j]*E[i,j] for i in range(3) for j in range(3)))
    EE  = sp.expand(sum(E[i,j]*Eup[i,j] for i in range(3) for j in range(3)))

    Vb  = Mp**2*(3*(sp.diff(a,t)/a)**2 + sp.diff(sp.diff(a,t)/a,t))
    L   = sg*( Mp**2/2*(N*R3 + (EE-trE**2)/N) + pd**2/(2*N) - N*Vb )
    L   = trunc(sp.expand(sp.series(sp.expand(L), eps, 0, 3).removeO()))
    L2  = sp.Poly(L,eps).coeff_monomial(eps**2)
    H   = sp.diff(a,t)/a
    L2c = sp.expand(L2.subs(al, sp.diff(ze,t)/H).doit())
    tgt = a**3*(pd**2/(2*H**2))*(sp.diff(ze,t)**2 - sum(sp.diff(ze,v)**2 for v in Y)/
a**2)
    D   = sp.expand(L2c-tgt).subs(pd, sp.sqrt(2)*Mp*sp.sqrt(-sp.diff(H,t)))
    print([sp.simplify(e.lhs-e.rhs) for e in euler_equations(D,[ze,be],[t,x,y,z])])
# -> [0, 0]
```

This rather technical calculation is performed explicitly in Refs. [10, 22] (see also Ref. [84] for a generalization to Bianchi space-times). Deriving the evolution equations to first order is strictly equivalent to varying the action expanded to second order, however, the second approach gives us more information on the structure of the theory. This action is that of a scalar field with a time-dependent mass, $m^2 = -z''/z$ in a Minkowski space-time. It is interesting to stress that the natural variable is neither Φ nor ζ but v. Note that if we had started from the evolution equations [see, e.g., (9.4.24)], we could have wrongly concluded that u_s was the canonical variable. This equivalence implies that we can quantize this theory as we would quantize a scalar field evolving in a time-dependent exterior field [23], where here the time dependence would find its

origin in the space-time dynamics [24]. Note also the similarity with the case of a test scalar field in an expanding space-time, as considered in Section 8.3. It is therefore the field v that should be promoted to the status of quantum operator in second quantization (see Chapter 2),

$$\hat{v}(\mathbf{x},\eta) = \int \frac{\mathrm{d}^3k}{(2\pi)^{3/2}} \left[v_k(\eta) e^{i\mathbf{k}\cdot\mathbf{x}} \hat{a}_{\mathbf{k}} + v_k^*(\eta) e^{-i\mathbf{k}\cdot\mathbf{x}} \hat{a}_{\mathbf{k}}^\dagger \right]. \tag{9.4.64}$$

In what follows we will work in the Heisenberg representation, in which the time-dependence is carried by the operators.

The first step in performing a canonical quantization is to introduce the conjugate momentum of v,

$$\pi = \frac{\delta \mathcal{L}}{\delta v'} = v', \tag{9.4.65}$$

which is also promoted to the status of operator, $\hat{\pi}$. We can then get the Hamiltonian

$$H = \int (v'\pi - \mathcal{L})\, \mathrm{d}^4x = \frac{1}{2} \int \left[\pi^2 + \delta^{ij} \partial_i v \partial_j v - \frac{z''}{z} v^2 \right] \mathrm{d}^4x. \tag{9.4.66}$$

The operators $\hat{v}$ and $\hat{\pi}$ have to satisfy canonical commutation relations on constant time hypersurfaces

$$[\hat{v}(\mathbf{x},\eta), \hat{v}(\mathbf{y},\eta)] = [\hat{\pi}(\mathbf{x},\eta), \hat{\pi}(\mathbf{y},\eta)] = 0, \qquad [\hat{v}(\mathbf{x},\eta), \hat{\pi}(\mathbf{y},\eta)] = i\delta^{(3)}(\mathbf{x}-\mathbf{y}). \tag{9.4.67}$$

The equation of motion for v obtained from the Lagrangian (9.4.63) provides the field equation for the operator $\hat{v}$, which is similar to the Klein–Gordon equation in flat space-time,

$$\hat{v}'' - \left(\Delta + \frac{z''}{z} \right) \hat{v} = 0. \tag{9.4.68}$$

Indeed, this is also the classical Klein–Gordon equation that can be derived from (9.4.14) after performing the change of variables (9.4.61) from Q to v. This equation is actually equivalent to the Heisenberg equations

$$\hat{v}' = i[\hat{H}, \hat{v}], \qquad \hat{\pi}' = i[\hat{H}, \hat{\pi}], \tag{9.4.69}$$

where $\hat{H}$ is simply the Hamiltonian (9.4.66) in terms of $\hat{v}$ and $\hat{\pi}$.

The classical equation of motion for the Mukhanov–Sasaki canonical variable v (or field operator $\hat{v}$) is given by

$$v'' - \left(\Delta + \frac{z''}{z} \right) v = 0, \tag{1}$$

where $z \equiv \frac{a\phi'}{\mathcal{H}} = a\sqrt{2\epsilon}$. If the background quantity $z(\eta)$ experiences an abrupt transition or step discontinuity at conformal time $\eta = \eta_0$, it can be modeled as

$$z(\eta) = z_- + (z_+ - z_-)\theta(\eta - \eta_0), \tag{2}$$

where θ is the Heaviside step function. Consequently, the successive conformal time derivatives z' and z'' behave as singular distributions across the transition:

$$z'(\eta) = (z_+ - z_-)\delta(\eta - \eta_0), \qquad z''(\eta) = (z_+ - z_-)\delta'(\eta - \eta_0) + \ldots \tag{3}$$

A direct integration of the mode equation across an infinitesimal interval $[\eta_0 - \epsilon_0, \eta_0 + \epsilon_0]$ (with $\epsilon_0 > 0$),

$$\int_{\eta_0-\epsilon_0}^{\eta_0+\epsilon_0} \frac{z''}{z} v \, \mathrm{d}\eta, \tag{4}$$

is mathematically ill-defined and divergent because it involves products of discontinuous functions and derivatives of distributions (δ'/z). To deduce unambiguous, physically meaningful matching conditions across the singular transition, the equation of motion must be recast into a non-singular, total-derivative form before integration.
This regularization can be carried out equivalently in two complementary ways. First, expressing the dynamics in terms of the comoving curvature perturbation $\mathcal{R} \equiv v/z$ (where $v = z\mathcal{R}$), the time derivatives expand to

$$v' = z'\mathcal{R} + z\mathcal{R}', \qquad v'' = z''\mathcal{R} + 2z'\mathcal{R}' + z\mathcal{R}''. \tag{5}$$

Substituting these expressions into the Mukhanov–Sasaki equation $v'' - \Delta v - \frac{z''}{z}v = 0$ yields

$$(z''\mathcal{R} + 2z'\mathcal{R}' + z\mathcal{R}'') - z\Delta\mathcal{R} - z''\mathcal{R} = 0, \tag{6}$$

in which the dangerously singular $z''\mathcal{R}$ terms cancel identically. Multiplying the resulting equation $z\mathcal{R}'' + 2z'\mathcal{R}' - z\Delta\mathcal{R} = 0$ by z groups the first two terms into a total conformal time derivative:

$$(z^2\mathcal{R}')' - z^2\Delta\mathcal{R} = 0. \tag{7}$$

Integrating this regularized equation across the transition layer $[\eta_0 - \epsilon_0, \eta_0 + \epsilon_0]$ in the limit $\epsilon_0 \to 0$,

$$\lim_{\epsilon_0\to 0}\int_{\eta_0-\epsilon_0}^{\eta_0+\epsilon_0}(z^2\mathcal{R}')'\,\mathrm{d}\eta - \lim_{\epsilon_0\to 0}\int_{\eta_0-\epsilon_0}^{\eta_0+\epsilon_0} z^2\Delta\mathcal{R}\,\mathrm{d}\eta = 0. \tag{8}$$

Because the spatial Laplacian $z^2\Delta\mathcal{R}$ contains no delta-function singularities, the second integral vanishes identically over the infinitesimal support $2\epsilon_0$, leaving

$$\lim_{\epsilon_0\to 0}\left[z^2\mathcal{R}'\right]_{\eta_0-\epsilon_0}^{\eta_0+\epsilon_0} = 0 \implies \boxed{\left[z^2\mathcal{R}'\right]_{\pm} = 0}. \tag{9}$$

Since $z^2\mathcal{R}'$ is bounded and continuous across η_0, the derivative $\mathcal{R}'$ has at most finite step discontinuities, which in turn guarantees that $\mathcal{R}$ is continuous across the hypersurface:

$$\boxed{[\mathcal{R}]_{\pm} = 0}. \tag{10}$$

Alternatively, the same junction conditions emerge directly from the Wronskian identity form of the differential operator:

$$\frac{1}{z}\left(zv' - z'v\right)' = \frac{1}{z}\left(z'v' + zv'' - z''v - z'v'\right) = v'' - \frac{z''}{z}v. \tag{11}$$

Substituting this identity into $v'' - \frac{z''}{z}v - \Delta v = 0$ and multiplying by z yields

$$\frac{\mathrm{d}}{\mathrm{d}\eta}(zv' - z'v) - z\Delta v = 0. \tag{12}$$

Integrating across the singular surface η_0 gives

$$\lim_{\epsilon_0\to 0}\int_{\eta_0-\epsilon_0}^{\eta_0+\epsilon_0}\frac{\mathrm{d}}{\mathrm{d}\eta}(zv' - z'v)\,\mathrm{d}\eta = \lim_{\epsilon_0\to 0}\int_{\eta_0-\epsilon_0}^{\eta_0+\epsilon_0} z\Delta v\,\mathrm{d}\eta = 0, \tag{13}$$

which establishes the conservation of the generalized momentum quantity across the jump:

$$\boxed{[zv' - z'v]_{\pm} = 0}. \tag{14}$$

Relating this expression back to the field ratio via the identity $\frac{\mathrm{d}}{\mathrm{d}\eta}(v/z) = \frac{zv'-z'v}{z^2}$, integration across $[\eta_0-\epsilon_0, \eta_0+\epsilon_0]$ immediately enforces $\left[\frac{v}{z}\right]_\pm = 0$, confirming $[\mathcal{R}]_\pm = 0$. Although the quadratic actions for v and $\mathcal{R}$ differ only by a total time derivative, $S_v - S_\mathcal{R} = \int \mathrm{d}\eta\, \mathrm{d}^3x\, \frac{\mathrm{d}}{\mathrm{d}\eta}(\dots)$, which does not affect the bulk equations of motion away from transitions, casting the perturbation equation in the total-derivative form is essential to obtain mathematically well-defined junction conditions free of distributional products. Across any sharp transition in the background parameter $z(\eta)$, the unique matching conditions are therefore given by the conserved pair $[\mathcal{R}]_\pm = 0$ and $[z^2\mathcal{R}']_\pm = 0$ (or equivalently $[zv' - z'v]_\pm = 0$).

As for quantization in Minkowski space-time (see Chapter 2), the creation and annihilation operators appearing in the decomposition (9.4.64) satisfy the standard commutation rules

$$[\hat{a}_\mathbf{k}, \hat{a}_\mathbf{p}] = [\hat{a}^\dagger_\mathbf{k}, \hat{a}^\dagger_\mathbf{p}] = 0, \qquad [\hat{a}_\mathbf{k}, \hat{a}^\dagger_\mathbf{p}] = \delta^{(3)}(\mathbf{k}-\mathbf{p}). \tag{9.4.70}$$

These commutation rules are consistent with the commutation rules (9.4.67) only if the field v_k is normalized according to

$$W(k) \equiv v_k v_k'^* - v_k^* v_k' = i, \tag{9.4.71}$$

since

$$[\hat{v}(\mathbf{x},\eta), \hat{\pi}(\mathbf{y},\eta)] = \int \frac{\mathrm{d}^3k}{(2\pi)^3} e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{y})} W(k).$$

This determines the normalization of their Wronskian, W. Then, one needs to construct the Fock representation of the Hilbert space on which the operators $\hat{v}$ and $\hat{\pi}$ act. In Minkowski space-time, the vacuum state is uniquely defined by the condition $\hat{a}_\mathbf{k}|0\rangle = 0$ for all $\mathbf{k}$ and the other states are obtained by repeated action of the creation operator $\hat{a}^\dagger_\mathbf{k}$. This amounts to keeping only negative-frequency modes. The vacuum state $|0\rangle$ is thus defined by the condition that this state is annihilated by all the operators $\hat{a}_\mathbf{k}$,

$$\forall \mathbf{k}, \quad \hat{a}_\mathbf{k}|0\rangle = 0. \tag{9.4.72}$$

The sub-Hubble modes, i.e. the high-frequency modes compared to the expansion of the Universe, must behave as in a flat space-time. So the asymptotic condition to impose for a state in the vacuum is

$$v_k(\eta) \to \frac{1}{\sqrt{2k}} e^{-ik\eta}, \qquad k\eta \to -\infty. \tag{9.4.73}$$

Let us note that in an expanding space-time, the notion of time is fixed by the background evolution that provides a preferred direction. The notion of positive and negative frequency is not time invariant, which implies that during the evolution time of positive frequency will be generated. Note that this quantization procedure amounts to treating gravity in a quantum way to linear order since the gravitational potential is promoted to the status of operator.

Particle creation If we pick up an arbitrary initial time η_i, then

$$v_k(\eta_i) = \frac{1}{\sqrt{\omega(\eta_i,k)}}, \qquad v_k'(\eta_i) = i\sqrt{\omega(\eta_i,k)}, \tag{9.4.74}$$

defines the positive frequency solutions at η_i, indeed for oscillating modes. We can thus define a vacuum state, $|0_i\rangle$ following the previous procedure. At a later time $\eta_1 > \eta_i$, the modes defined by (9.4.74) will no longer be positive frequency and an observer at η_1 will define another vacuum state $|0_1\rangle$ satisfying $\hat{c}_\mathbf{k}|0_1\rangle = 0$ for all $\mathbf{k}$, $\hat{c}_\mathbf{k}$ being the annihilation operators for the positive-frequency modes at η_1. Since the mode equation is linear, it is clear that the positive-frequency modes at η_1, w_k, are related to mode functions v_k by a linear transformation

$$w_k = \alpha_k v_k + \beta_k v_k^*, \tag{9.4.75}$$

with the two complex coefficients satisfying $|\alpha_k|^2 - |\beta_k|^2 = 1$. The number of particles of mode k at η_1 will thus be defined by $N_k^{(1)} = \hat{c}_{\mathbf{k}}^\dagger \hat{c}_{\mathbf{k}}$. Since the two sets of creation and annihilation operators are related by $\hat{a}_{\mathbf{k}} = \alpha_k \hat{c}_{\mathbf{k}} + \beta_k^* \hat{c}_{\mathbf{k}}^\dagger$ implying that $\hat{c}_{\mathbf{k}} = \alpha_k^* \hat{a}_{\mathbf{k}} - \beta_k^* \hat{a}_{\mathbf{k}}^\dagger$, we deduce that the state $|0_i\rangle$ is seen as a state containing $\langle 0_i | N_k^{(1)} | 0_i \rangle = |\beta_k|^2$ particles by the observer at η_1. This illustrates the time dependence of the notion of positive frequency and vacuum state. Fortunately the ambiguity of the definition of the vacuum does not make important differences for small wavelength modes for which the WKB approximation holds. This mechanism of quantum particle creation is discussed in detail in Section 10.2 of Chapter 10.

Relation with the perturbation variables The previous definitions allow us to conclude that

$$z = \sqrt{\frac{3}{\kappa}} \theta^{-1}, \qquad \Delta u_s = -z \left(\frac{v}{z} \right)'. \tag{9.4.76}$$

Einstein's equations imply that the gravitational potential is given by

$$\Delta\Phi = \frac{\kappa}{2} \frac{\phi'^2}{\mathcal{H}} \left(\frac{v}{z} \right)' \qquad \text{and} \qquad \left(\frac{a^2 \Phi}{\mathcal{H}} \right)' = \frac{\kappa}{2} z v, \tag{9.4.77}$$

and we also have the useful relation

$$2\mathcal{H} X = \frac{\kappa}{a} \phi' v, \tag{9.4.78}$$

where X has been defined in (5.67).

Following the same procedure as for a test scalar field considered in Section 8.3, (9.4.61) implies that the power spectrum of the metric perturbations is

$$\mathcal{P}_{\mathcal{R}} = \mathcal{P}_\zeta = \frac{k^3}{2\pi^2} \left| \frac{v_k}{z} \right|^2, \tag{9.4.79}$$

using the definitions of Section B.5 of Appendix B for the normalization of the power spectrum. The solution of (9.4.68) with initial conditions (9.4.73) allows us to determine the gravitational potential and the power spectrum of $\mathcal{R}$ and ζ. The analysis of Section 8.4.5 then allows us to determine the curvature perturbation and the gravitational potential at the beginning of the post-inflationary phase.

Gravitational waves

Quantization The same procedure can be followed for gravitational waves. Restricting ourselves to tensor degrees of freedom, the action at second order in perturbations takes the form

$$\delta^{(2)} S = \frac{M_{\rm Pl}^2}{64\pi} \int a^2(\eta) \left[(h_{ij}')^2 - (D_k h_{ij})^2 - 2K (h_{ij})^2 \right] \sqrt{\gamma}\, \mathrm{d}^3 x\, \mathrm{d}\eta. \tag{9.4.80}$$

Unsurprisingly, the variation of this action gives the equation of propagation (9.4.18). Using the decomposition (9.4.19), and defining

$$\mu_\lambda(k,\eta) = \sqrt{\frac{M_{\rm Pl}^2}{8\pi}} a(\eta) \bar{E}_\lambda(k,\eta) = \sqrt{\frac{M_{\rm Pl}^2}{32\pi}} a(\eta) h_\lambda(k,\eta), \tag{9.4.81}$$

the action at second order takes the form

$$\delta^{(2)} S = \frac{1}{2} \sum_\lambda \int \left[(\mu_\lambda')^2 - \gamma^{ij} \partial_i \mu_\lambda \partial_j \mu_\lambda - 2K \mu_\lambda^2 + \frac{a''}{a} \mu_\lambda^2 \right] \sqrt{\gamma}\, \mathrm{d}^3 x\, \mathrm{d}\eta. \tag{9.4.82}$$

In this form, the action for the gravitational waves reduces to the action for two independent scalar fields, μ_λ, with time-dependent mass $m^2 = -a''/a$ evolving in a static space-time. We can

therefore proceed as for the scalar perturbations and quantize these two degrees of freedom. In the Heisenberg representation, we promote μ_λ to the status of quantum operator

$$\hat{\mu}_\lambda(\mathbf{x},\eta) = \int \frac{\mathrm{d}^3k}{(2\pi)^{3/2}} \left[\mu_{k,\lambda}(\eta) e^{i\mathbf{k}\cdot\mathbf{x}} \hat{a}_{\mathbf{k},\lambda} + \mu^*_{k,\lambda}(\eta) e^{-i\mathbf{k}\cdot\mathbf{x}} \hat{a}^\dagger_{\mathbf{k},\lambda}\right]. \tag{9.4.83}$$

If the modes $\mu_{k,\lambda}$ satisfy the normalization (9.4.71), then the creation and annihilation operators satisfy the commutation relations

$$[\hat{a}_{\mathbf{k},\lambda}, \hat{a}_{\mathbf{p},\lambda'}] = [\hat{a}^\dagger_{\mathbf{k},\lambda}, \hat{a}^\dagger_{\mathbf{p},\lambda'}] = 0, \qquad [\hat{a}_{\mathbf{k},\lambda}, \hat{a}^\dagger_{\mathbf{p},\lambda'}] = \delta_{\lambda\lambda'}\delta^{(3)}(\mathbf{k}-\mathbf{p}), \tag{9.4.84}$$

and for each polarization, the vacuum is defined by

$$\forall \mathbf{k}, \quad \hat{a}_{\mathbf{k},\lambda}|0\rangle = 0, \tag{9.4.85}$$

and we set the initial conditions in the same way as for scalar perturbations by imposing

$$\mu_{k,\lambda}(\eta) \to \frac{1}{\sqrt{2k}} e^{-ik\eta}, \qquad k\eta \to -\infty. \tag{9.4.86}$$

The evolution equation for $\mu_{k,\lambda}$, that we will write simply as μ_k when there is no ambiguity, is

$$\mu_k'' + \left(k^2 + 2K - \frac{a''}{a}\right)\mu_k = 0. \tag{9.4.87}$$

As previously, this equation can be solved by using the WKB approximation (9.3.23) with $\omega^2_\lambda(k,\eta) = k^2 + 2K - a''/a$.

Power spectrum The power spectrum of gravitational waves is obtained by solving the evolution equation (9.4.87) with initial conditions (9.4.73). This allows us to define the amplitude of super-Hubble tensor modes, which remains constant. The power spectrum, $\mathcal{P}_T$, of the tensor modes is defined from the variable h_{ij} and is related to the power spectrum $\mathcal{P}_{\bar{E}}$ of one polarization E_λ by

$$\langle \bar{E}_\lambda(\mathbf{k},\eta_f)\bar{E}^*_{\lambda'}(\mathbf{k}',\eta_f)\rangle = \frac{2\pi^2}{k^3}\mathcal{P}_T(k)\delta^{(3)}(\mathbf{k}-\mathbf{k}')\delta_{\lambda\lambda'}, \tag{9.4.88}$$

with

$$\mathcal{P}_{\bar{E}} = \frac{k^3}{2\pi^2}\frac{8\pi}{M_{\rm Pl}^2}\left|\frac{\mu_k}{a}\right|^2 \tag{9.4.89}$$

for each polarization. Since $\mathcal{P}_T = 8\mathcal{P}_{\bar{E}}$, where the factor 8 arises from the sum over polarizations and tensor normalization, we conclude that

$$\mathcal{P}_T = \frac{k^3}{2\pi^2}\frac{64\pi}{M_{\rm Pl}^2}\left|\frac{\mu_k}{a}\right|^2, \tag{9.4.90}$$

where each polarization contributes in an identical way.

Gravitational wave density The variation of the action (9.4.80) with respect to the metric allows us to define the stress-energy tensor for the gravitational waves

$$t_{\mu\nu} = \frac{1}{16\pi G_{\rm N}}\sum_\lambda \left(\partial_\mu\bar{E}_\lambda\partial_\nu\bar{E}_\lambda - \bar{g}_{\mu\nu}\partial_\alpha\bar{E}_\lambda\partial^\alpha\bar{E}_\lambda\right). \tag{9.4.91}$$

The gravitational wave amplitude at any time after the inflationary phase can be decomposed as

$$\bar{E}_\lambda(k,\eta) = T(k,\eta)\bar{E}_\lambda(k,\eta_f), \tag{9.4.92}$$

where $T(k,\eta)$ is a transfer function obtained by solving (5.109). The gravitational wave density can then be defined as $\rho_{\rm GW} = -\langle t_0^0(\mathbf{x},\eta)\rangle$. As soon as the mode becomes sub-Hubble, it oscillates and one can average over a few periods to extract its amplitude. This procedure is well defined as long as the amplitude varies slowly with respect to the wave period. In practice, this amounts to replacing $T(k,\eta)$ by $\bar{T}(k,\eta)$, which is the average value of T over a few periods. We then obtain the energy density per decade

$$\frac{\mathrm{d}\rho_{\rm GW}}{\mathrm{d}\ln k} = \frac{1}{8\pi G_{\rm N}}\left(\frac{k}{a}\right)^2 \mathcal{P}_T(k)\bar{T}(k,\eta), \tag{9.4.93}$$

or equivalently in terms of the density parameter,

$$\frac{\mathrm{d}\Omega_{\rm GW}}{\mathrm{d}\ln k} = \frac{1}{3}\left(\frac{k}{aH}\right)^2 \mathcal{P}_T(k)\bar{T}(k,\eta). \tag{9.4.94}$$

As long as the mode is super-Hubble, its amplitude is constant (see Chapter 5) and then it enters a damped oscillation regime on sub-Hubble scales. The damping between the time where the mode becomes super-Hubble and today is typically of order $\bar{T}(k,\eta) \simeq (a_k/a_0)^2$ where a_k is the value of the scale factor when $k = aH$. We infer that the gravitational wave spectrum today is

$$\frac{\mathrm{d}\Omega_{\rm GW}}{\mathrm{d}\ln k} \propto k^2 a_k^2 \mathcal{P}_T(k). \tag{9.4.95}$$

If the mode becomes sub-Hubble during the matter-dominated era, $a_k \propto k^{-2}$, whereas during the radiation era $a_k \propto k^{-1}$. We infer that

$$\frac{\mathrm{d}\Omega_{\rm GW}}{\mathrm{d}\ln k} \propto \mathcal{P}_T(k) \times \begin{cases} k^{-2} & k < k_{\rm eq} \\ k^0 & k > k_{\rm eq} \end{cases} \tag{9.4.96}$$

This contribution is represented in Fig. 11.15 for a primordial scale invariant spectrum, $\mathcal{P}_T \propto k^0$.

9.4.4 Perturbations in the slow-roll regime

We will now apply the previous formalism to inflation in the slow-roll regime. In this approximation, one only needs to determine the background functions $z(\eta)$ and $a(\eta)$ that then allow for the integration of the mode equations (9.4.68) and (9.4.87). It is useful to note already at this early stage that

$$z = a\frac{M_{\rm Pl}}{\sqrt{4\pi}}\sqrt{\epsilon}. \tag{9.4.97}$$

An exact solution: power-law inflation

Before studying the general case, it is instructive to consider the case of power-law inflation (9.2.62) for which the slow-roll parameters remain strictly constant:

$$\epsilon = \delta = \frac{1}{p}, \qquad \xi = 0. \tag{9.4.98}$$

In this case, the expression (9.3.38) is exact, so that

$$aH = \mathcal{H} = -\frac{1}{1-\epsilon}\frac{1}{\eta}, \tag{9.4.99}$$

and, as for a de Sitter space-time, conformal time η varies between $-\infty$ (the distant past) and 0 (the infinite future).

Scalar modes The conformal time evolution of the scale factor (9.2.62) and scalar field (9.2.64) are determined explicitly: $a(\eta) \propto (-\eta)^{1+\beta}$ with $\beta \leq -2$. The variable z, defined by (9.4.62), is then given by $z = \sqrt{p\pi}\,\eta^{1+\beta}$, so that

$$\frac{z''}{z} = \frac{\beta(\beta+1)}{\eta^2}. \tag{9.4.100}$$

The Mukhanov–Sasaki equation (9.4.68) then takes the form of a Bessel equation

$$v_k'' + \left(k^2 - \frac{\nu^2 - 1/4}{\eta^2}\right) v_k = 0, \qquad \text{with} \quad \nu = -\beta - \frac{1}{2} = \frac{3}{2} + \frac{1}{p-1}. \tag{9.4.101}$$

The solution of this equation that satisfies the Bunch–Davies initial condition (9.4.73) can be expressed in terms of a Hankel function of the first kind as

$$v_k(\eta) = \frac{\sqrt{\pi}}{2} e^{i(\nu+1/2)\pi/2} \sqrt{-\eta}\, H_\nu^{(1)}(-k\eta). \tag{9.4.102}$$

For super-Hubble modes ($-k\eta \to 0$, i.e. $k/\mathcal{H} \ll 1$), this solution asymptotically behaves as

$$v_k(\eta) \to 2^{\nu-3/2} \frac{\Gamma(\nu)}{\Gamma(3/2)} e^{i(\nu-1/2)\pi/2} \frac{1}{\sqrt{2k}} (-k\eta)^{-\nu+1/2}. \tag{9.4.103}$$

The expression (9.4.79) allows us to find that the power spectrum of comoving curvature perturbations is given by

$$\mathcal{P}_\zeta(k) = \frac{H^2}{\pi M_{\rm Pl}^2 \epsilon} \left[2^{\nu-3/2} \frac{\Gamma(\nu)}{\Gamma(3/2)}\right]^2 \left(\nu - \frac{1}{2}\right)^{-2\nu+1} \left(\frac{k}{aH}\right)^{-2\nu+3}. \tag{9.4.104}$$

The coefficient $(\nu - 1/2)^{-2\nu+1}$ arises from the contribution $(1-\epsilon)^{2\nu-1}$ that appears when replacing $k\eta$ using (9.4.99). If we evaluate this spectrum at the horizon exit epoch $aH = k$, it can be rewritten as

$$\mathcal{P}_\zeta(k) = \frac{1}{\pi M_{\rm Pl}^2} \left[2^{\nu-3/2} \frac{\Gamma(\nu)}{\Gamma(3/2)}\right]^2 \left(\nu - \frac{1}{2}\right)^{-2\nu+1} \left(\frac{H^2}{\epsilon}\right)_{k=aH}, \tag{9.4.105}$$

which encapsulates the wavenumber dependence into the factor (H^2/ϵ) evaluated at the moment the mode k exits the Hubble horizon.

Gravitational waves The study of gravitational waves in power-law inflation is formally identical to that of the scalar modes since in this particular model $z \propto a$. Equation (9.4.87) is thus identical to (9.4.68) and the solution for μ_k is given by (9.4.102). We infer from (9.4.90) that

$$\mathcal{P}_T = \frac{64\pi}{M_{\rm Pl}^2} \left(\frac{z}{a}\right)^2 \mathcal{P}_\zeta. \tag{9.4.106}$$

Using $z/a = M_{\rm Pl}\sqrt{\epsilon/(4\pi)}$ from (9.4.97), the gravitational wave power spectrum is thus given by

$$\mathcal{P}_T = 16\epsilon \mathcal{P}_\zeta. \tag{9.4.107}$$

Spectral indices Note that the amplitude of gravitational waves is determined solely by the inflationary energy scale H during the slow-roll regime, while that of the scalar modes depends on both H and the slow-roll parameter ϵ. The spectral indices, defined by

$$n_s - 1 \equiv \frac{\mathrm{d}\ln \mathcal{P}_\zeta}{\mathrm{d}\ln k}, \qquad n_T \equiv \frac{\mathrm{d}\ln \mathcal{P}_T}{\mathrm{d}\ln k}, \tag{9.4.108}$$

are, respectively, given by

$$n_s - 1 = -2\nu + 3 = 2\beta + 4 = -\frac{2}{p-1}, \qquad n_T = -2\nu + 3 = 2\beta + 4 = -\frac{2}{p-1}. \tag{9.4.109}$$

Note that when $\epsilon \ll 1$, i.e. when $p \gg 1$, then $n_T = -2/(p-1) \simeq -2/p = -2\epsilon$ to first order in the slow-roll parameters. We deduce the fundamental consistency relation to lowest order in the slow-roll parameters:

$$\frac{\mathcal{P}_T}{\mathcal{P}_\zeta} = -8n_T. \tag{9.4.110}$$

For power-law inflation $a(t) = a_0(t/t_0)^p$, conformal time is $\eta = \int \frac{\mathrm{d}t}{a(t)} = \frac{t_0}{(1-p)a_0}(t/t_0)^{1-p} \equiv \eta_0(-\eta)^{1+\beta}$, where $\beta = \frac{1}{p-1} - 2 \leq -2$. The Hubble parameter in conformal time is $\mathcal{H} = \frac{a'}{a} = \frac{1+\beta}{\eta} = -\frac{p}{(p-1)\eta}$. The slow-roll parameters are constant:

$$\epsilon = 1 - \frac{\mathcal{H}'}{\mathcal{H}^2} = 1 - \frac{-(1+\beta)/\eta^2}{(1+\beta)^2/\eta^2} = 1 + \frac{1}{1+\beta} = \frac{2+\beta}{1+\beta} = \frac{1}{p}. \tag{1}$$

The background function $z \equiv a\phi'/\mathcal{H} = aM_{\rm Pl}\sqrt{\epsilon/(4\pi)}$ scales proportionally to $a(\eta) \propto (-\eta)^{1+\beta}$. Computing its second derivative:

$$z' = (1+\beta)\frac{z}{\eta}, \qquad z'' = \beta(1+\beta)\frac{z}{\eta^2} \implies \frac{z''}{z} = \frac{\beta(\beta+1)}{\eta^2}. \tag{2}$$

Matching to the standard Bessel parameter form $\frac{z''}{z} = \frac{\nu^2 - 1/4}{\eta^2}$:

$$\nu^2 - \frac{1}{4} = \beta^2 + \beta \implies \nu^2 = \left(\beta + \frac{1}{2}\right)^2 \implies \nu = -\beta - \frac{1}{2} = \frac{3}{2} + \frac{1}{p-1}. \tag{3}$$

The Mukhanov–Sasaki equation becomes the canonical Bessel ODE:

$$v_k'' + \left(k^2 - \frac{\nu^2 - 1/4}{\eta^2}\right) v_k = 0. \tag{4}$$

The general solution is a linear combination of Hankel functions $\sqrt{-\eta}H_\nu^{(1)}(-k\eta)$ and $\sqrt{-\eta}H_\nu^{(2)}(-k\eta)$. In the asymptotic past $-k\eta \to \infty$:

$$H_\nu^{(1)}(x) \to \sqrt{\frac{2}{\pi x}} \exp\left[i\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right)\right]. \tag{5}$$

Imposing the Bunch–Davies vacuum condition $v_k(\eta) \to \frac{1}{\sqrt{2k}}e^{-ik\eta}$ selects exclusively $H_\nu^{(1)}$:

$$v_k(\eta) = \frac{\sqrt{\pi}}{2} e^{i(\nu+1/2)\pi/2}\sqrt{-\eta}\, H_\nu^{(1)}(-k\eta). \tag{6}$$

In the super-Hubble limit $-k\eta \ll 1$, using $H_\nu^{(1)}(x) \simeq -\frac{i}{\pi}\Gamma(\nu)(x/2)^{-\nu}$:

$$|v_k(\eta)| \simeq \frac{1}{\sqrt{2k}} 2^{\nu-3/2}\frac{\Gamma(\nu)}{\Gamma(3/2)}(-k\eta)^{-\nu+1/2}. \tag{7}$$

The comoving curvature power spectrum is:

$$\mathcal{P}_\zeta(k) = \frac{k^3}{2\pi^2}\left|\frac{v_k}{z}\right|^2 = \frac{k^3}{2\pi^2}\frac{1}{2k}\left[2^{\nu-3/2}\frac{\Gamma(\nu)}{\Gamma(3/2)}\right]^2 (-k\eta)^{-2\nu+1}\frac{1}{z^2}. \tag{8}$$

Using $z = aM_{\rm Pl}\sqrt{\epsilon/(4\pi)}$ and $-k\eta = \frac{k}{aH(1-\epsilon)} = \frac{k}{aH}(\nu - 1/2)^{-1}$, we obtain:

$$\mathcal{P}_\zeta(k) = \frac{H^2}{\pi M_{\rm Pl}^2 \epsilon}\left[2^{\nu-3/2}\frac{\Gamma(\nu)}{\Gamma(3/2)}\right]^2\left(\nu-\frac{1}{2}\right)^{-2\nu+1}\left(\frac{k}{aH}\right)^{-2\nu+3}. \tag{9}$$

This confirms the exact solution (9.4.104) and the consistency relation $\mathcal{P}_T/\mathcal{P}_\zeta = 16\epsilon = -8n_T$.

General case

In the general case, the slow-roll parameters vary with time, and the evolution equations cannot be integrated in closed analytical form. We must therefore construct a systematic perturbation expansion scheme in terms of the slow-roll parameters.

Expansion scheme The case of power-law inflation corresponds to constant slow-roll parameters. In an arbitrary model, these parameters vary according to the background flow equations (9.2.28), and the mode equations (9.4.68) and (9.4.87) must be solved perturbatively.

To establish the expansion, we relate the conformal time η to the Hubble scale by performing successive integrations by parts:

$$\eta = \int\frac{\mathrm{d}a}{a\mathcal{H}} = -\frac{1}{\mathcal{H}} + \int\frac{\epsilon\,\mathrm{d}a}{a\mathcal{H}} = -\frac{1}{\mathcal{H}} + \epsilon\int\frac{\mathrm{d}a}{a\mathcal{H}} - \int\frac{\epsilon'\,\mathrm{d}a}{a\mathcal{H}^2}, \tag{9.4.111}$$

which yields

$$\eta = -\frac{1}{\mathcal{H}}\left[\frac{1}{1-\epsilon} - 2\epsilon(\epsilon-\delta)\right] + \ldots. \tag{9.4.112}$$

Assuming ϵ and δ are small parameters ($\epsilon, |\delta| \ll 1$), to first order in the slow-roll parameters we have

$$\eta = -\frac{1}{\mathcal{H}(1+\epsilon)} + \mathcal{O}(2). \tag{9.4.113}$$

We distinguish the leading-order expansion (containing only lowest powers of slow-roll parameters) from the next-to-leading order expansion (containing first-order corrections in ϵ, δ, ξ).

Expansion of the functions z''/z and a''/a The two functions z''/z and a''/a govern the mode equations (9.4.68) and (9.4.87). Starting from (9.4.62):

$$\frac{z'}{z} = \mathcal{H} + \frac{\phi''}{\phi'} - \frac{\mathcal{H}'}{\mathcal{H}} = \mathcal{H}(1+\epsilon-\delta), \tag{9.4.114}$$

where the second equality follows from relations (9.2.30). Differentiating this relation with respect to η and using (9.2.28) and (9.2.30) gives the exact result:

$$\begin{aligned}\frac{z''}{z} &= \mathcal{H}^2\left[2+2\epsilon-3\delta+\delta^2-\epsilon\delta+\xi\right]\\ &= \mathcal{H}^2\left[2+2\epsilon-3\delta+2\epsilon^2+\delta^2-4\epsilon\delta+\xi_H^2\right].\end{aligned} \tag{9.4.115}$$

In an identical way, since $a''/a = \mathcal{H}' + \mathcal{H}^2$, we obtain

$$\frac{a''}{a} = (2-\epsilon)\mathcal{H}^2. \tag{9.4.116}$$

Substituting the conformal time relation (9.4.113) yields the first-order asymptotic expansions:

$$\frac{z''}{z} = \frac{2+6\epsilon-3\delta}{\eta^2} + \mathcal{O}(2), \qquad \frac{a''}{a} = \frac{2+3\epsilon}{\eta^2} + \mathcal{O}(2). \tag{9.4.117}$$

Scalar modes As in the power-law case, (9.4.68) reduces to a Bessel equation of the form (9.4.101) with effective index

$$\nu \simeq \frac{3}{2} + 2\epsilon - \delta. \tag{9.4.118}$$

The solution satisfying the Bunch–Davies vacuum condition (9.4.73) is given by the Hankel function (9.4.102). Expanding this solution for super-Hubble modes and using $\Gamma(a+h)/\Gamma(a) \sim 1 + h\psi(a)$ with $\psi(3/2) = 2 - \gamma_E - 2\ln 2$, Stewart and Lyth [25] obtained:

$$\mathcal{P}_\zeta(k) = \frac{1}{\pi}\frac{H^2}{M_{\rm Pl}^2\epsilon}\left[1 - 2(2C+1)\epsilon + 2C\delta\right]\left(\frac{k}{aH}\right)^{2\delta-4\epsilon}, \tag{9.4.119}$$

where $C \equiv \gamma_E + \ln 2 - 2 \approx -0.7296$ (with $\gamma_E \approx 0.5772$ being Euler's constant). The spectrum in the slow-roll regime takes the two equivalent forms:

$$\mathcal{P}_\zeta(k) = \frac{1}{\pi}\frac{H^2}{M_{\rm Pl}^2\epsilon}\left[1 - 2(2C+1)\epsilon + 2C\delta + (2\delta - 4\epsilon)\ln\left(\frac{k}{aH}\right)\right] \tag{9.4.120}$$

$$= \frac{1}{\pi}\left[1 - 2(2C+1)\epsilon + 2C\delta\right]\left(\frac{H^2}{M_{\rm Pl}^2\epsilon}\right)_{k=aH}. \tag{9.4.121}$$

Gravitational waves For tensor perturbations, the mode equation (9.4.87) reduces to a Bessel equation with index

$$\nu \simeq \frac{3}{2} + \epsilon. \tag{9.4.122}$$

Noticing that $\mathcal{P}_T = 16\epsilon\mathcal{P}_\zeta^{[\delta=\epsilon]}$ from (9.4.121), expanding to next-to-leading order yields:

$$\mathcal{P}_T(k) = \frac{16}{\pi}\frac{H^2}{M_{\rm Pl}^2}\left[1 - 2(C+1)\epsilon\right]\left(\frac{k}{aH}\right)^{-2\epsilon}. \tag{9.4.123}$$

The tensor spectrum in the slow-roll regime takes the two equivalent forms:

$$\mathcal{P}_T(k) = \frac{16}{\pi}\frac{H^2}{M_{\rm Pl}^2}\left[1 - 2(C+1)\epsilon - 2\epsilon\ln\left(\frac{k}{aH}\right)\right] \tag{9.4.124}$$

$$= \frac{16}{\pi}\left[1 - 2(C+1)\epsilon\right]\left(\frac{H^2}{M_{\rm Pl}^2}\right)_{k=aH}. \tag{9.4.125}$$

Consistency relation at leading order Cosmological predictions are conventionally parameterized in terms of the amplitudes

$$A_S^2 \equiv \frac{4}{25}\mathcal{P}_\zeta(k), \qquad A_T^2 \equiv \frac{1}{100}\mathcal{P}_T(k). \tag{9.4.126}$$

The scalar spectral index from (9.4.120) is:

$$n_s - 1 \equiv \frac{\mathrm{d}\ln\mathcal{P}_\zeta}{\mathrm{d}\ln k} = \frac{\mathrm{d}\ln A_S^2}{\mathrm{d}\ln k} = 2\delta - 4\epsilon. \tag{9.4.127}$$

Evaluating $n_s - 1$ directly from (9.4.121) using $\mathrm{d}\ln k = \mathrm{d}\ln a_k = H\mathrm{d}t$:

$$n_s - 1 = 2\frac{\mathrm{d}\ln H_k}{\mathrm{d}\ln k} - \frac{\mathrm{d}\ln\epsilon_k}{\mathrm{d}\ln k} = -2\epsilon - \frac{\dot\epsilon}{\epsilon H} = 2\delta - 4\epsilon. \tag{9.4.128}$$

The tensor spectral index from (9.4.124) is:

$$n_T \equiv \frac{\mathrm{d}\ln\mathcal{P}_T}{\mathrm{d}\ln k} = \frac{\mathrm{d}\ln A_T^2}{\mathrm{d}\ln k} = -2\epsilon. \tag{9.4.129}$$

Alternatively, from (9.4.125):

$$n_T = 2\frac{\mathrm{d}\ln H_k}{\mathrm{d}\ln k} = -2\epsilon. \tag{9.4.130}$$

The ratio of tensor to scalar power spectrum amplitudes is:

$$R \equiv \frac{A_T^2}{A_S^2} = \epsilon. \tag{9.4.131}$$

Combining (9.4.129) and (9.4.131) yields the celebrated **consistency relation**:

$$R = -\frac{n_T}{2}. \tag{9.4.132}$$

This relation is an incontrovertible, model-independent prediction of all single-field slow-roll inflation models.

A fundamental question in inflationary model building is whether the field excursion $\Delta\phi \equiv |\phi_* - \phi_{\rm end}|$ during the observable e-folds is sub-Planckian ($\Delta\phi < M_{\rm Pl}$) or super-Planckian ($\Delta\phi > M_{\rm Pl}$).
From the definition of the number of e-folds (9.2.36) during slow roll:

$$\mathrm{d}N = H\,\mathrm{d}t = \frac{H}{\dot\phi}\,\mathrm{d}\phi = -\sqrt{\frac{4\pi G_{\rm N}}{\epsilon}}\,\mathrm{d}\phi = -\frac{1}{\sqrt{2\epsilon}M_{\rm Pl}}\,\mathrm{d}\phi. \tag{1}$$

The total field excursion between the horizon exit of the CMB scale ($N_* \sim 60$) and the end of inflation ($N = 0$) is:

$$\frac{\Delta\phi}{M_{\rm Pl}} = \int_0^{N_*} \sqrt{2\epsilon(N)}\,\mathrm{d}N. \tag{2}$$

Using the tensor-to-scalar ratio consistency relation $r = 16\epsilon$ (or $r = 8R = 8\epsilon$), we can express the integrand directly in terms of the observable tensor ratio $r(N)$:

$$\frac{\Delta\phi}{M_{\rm Pl}} = \int_0^{N_*} \sqrt{\frac{r(N)}{8}}\,\mathrm{d}N. \tag{3}$$

If $r(N)$ is monotonically decreasing or approximately constant over the $\Delta N_{\rm CMB} \approx 30$–$60$ e-folds of observable inflation, we obtain the celebrated **Lyth bound** [15]:

$$\boxed{\frac{\Delta\phi}{M_{\rm Pl}} \gtrsim \sqrt{\frac{r}{8}}\,\Delta N_{\rm CMB} \simeq \left(\frac{r}{0.01}\right)^{1/2}\left(\frac{\Delta N_{\rm CMB}}{30}\right) \times 1.06\,M_{\rm Pl}}. \tag{4}$$

This derivation establishes a profound physical dichotomy:

- **Large-Field Models ($r \gtrsim 0.01$):** Require super-Planckian field excursions $\Delta\phi > M_{\rm Pl}$, making the theory sensitive to an infinite tower of Planck-suppressed higher-dimension operators $\mathcal{O}_{d>4} \sim (\phi/M_{\rm Pl})^n V(\phi)$ unless protected by an exact shift symmetry (e.g. natural inflation, axion monodromy).
- **Small-Field Models ($r \ll 0.001$):** Feature sub-Planckian field excursions $\Delta\phi \ll M_{\rm Pl}$ and produce undetectable primordial gravitational wave amplitudes.

Consistency relation at next-to-leading order At next order, taking into account the scale dependence of the slow-roll parameters via (9.2.39) and (9.2.40), the logarithmic derivatives of (9.4.121) and (9.4.125) give:

$$n_s - 1 = 2\delta - 4\epsilon - \left[8(C+1)\epsilon^2 - (6+10C)\epsilon\delta + 2C\xi_H^2\right], \tag{9.4.133}$$

$$n_T = -2\epsilon\left[1 + (3 + 2C)\epsilon - 2(1 + C)\delta\right]. \tag{9.4.134}$$

The ratio between A_S and A_T at next order becomes:

$$\epsilon = R\left[1 - 2C(\epsilon - \delta)\right]. \tag{9.4.135}$$

Moreover, the ratio between (9.4.121) and (9.4.125) gives:

$$n_T = -2R(1 + 3\epsilon - 2\delta). \tag{9.4.136}$$

Replacing ϵ and δ inside the brackets by their leading-order values yields the next-to-leading order consistency relation:

$$n_T = -2R\left[1 - \frac{R}{2} + (1 - n_s)\right]. \tag{9.4.137}$$

Remarkably, this relation does not involve any derivatives of the spectral indices.

The running of the spectral indices is defined by:

$$\alpha_s \equiv \frac{\mathrm{d}n_s}{\mathrm{d}\ln k}, \qquad \alpha_T \equiv \frac{\mathrm{d}n_T}{\mathrm{d}\ln k}. \tag{9.4.138}$$

Using the flow equations (9.2.28), $\frac{\mathrm{d}\epsilon}{\mathrm{d}\ln k} = 2\epsilon(\delta - \epsilon)$ and $\frac{\mathrm{d}\delta}{\mathrm{d}\ln k} = 2\epsilon(\epsilon - \delta) - \xi$, at lowest order we find:

$$\alpha_s = -4\epsilon(\epsilon - \delta) - 2\xi, \qquad \alpha_T = -4\epsilon(\epsilon - \delta). \tag{9.4.139}$$

This implies the existence of a second consistency relation [27]:

$$\alpha_T = 2R\left[2R + (n_s - 1)\right]. \tag{9.4.140}$$

Expanded around a chosen pivot scale k_* (typically $k_* \sim 0.01\,h\,\mathrm{Mpc}^{-1}$ or $0.05\,\mathrm{Mpc}^{-1}$), the primordial power spectra take the standard logarithmic Taylor form:

$$\ln A_S^2(k) = \ln A_S^2(k_*) + \left[n_s(k_*) - 1\right]\ln\left(\frac{k}{k_*}\right) + \frac{1}{2}\alpha_s \ln^2\left(\frac{k}{k_*}\right) + \ldots, \tag{9.4.141}$$

and similarly for $A_T^2(k)$.

Table 9.1: Summary of the scalar and tensor perturbation spectra in the slow-roll regime.

Quantity	**Leading Order**	**Next-to-Leading Order**
$R \equiv A_T^2/A_S^2$	ϵ	$R\left[1 - 2C(\epsilon - \delta)\right]$
$n_s - 1$	$2\delta - 4\epsilon$	$2\delta - 4\epsilon - \left[8(C+1)\epsilon^2 - (6 + 10C)\epsilon\delta + 2C\xi_H^2\right]$
n_T	-2ϵ	$-2\epsilon\left[1 + (3 + 2C)\epsilon - 2(1 + C)\delta\right]$
Consistency 1	$n_T = -2R$	$n_T = -2R\left[1 - \frac{R}{2} + (1 - n_s)\right]$
α_s	0	$-4\epsilon(\epsilon - \delta) - 2\xi$
α_T	0	$-4\epsilon(\epsilon - \delta)$
Consistency 2	—	$\alpha_T = 2R\left[2R + (n_s - 1)\right]$

We expand the Hankel function $H_\nu^{(1)}(x)$ for small argument $x \to 0$ up to next order in ϵ, δ:

$$H_\nu^{(1)}(x) \approx -\frac{i}{\pi}\Gamma(\nu)\left(\frac{x}{2}\right)^{-\nu}\left[1 - \frac{x^2}{4(1-\nu)} + \ldots\right]. \tag{1}$$

The index is $\nu = \frac{3}{2} + h$, where $h = 2\epsilon - \delta \ll 1$. Expanding the Gamma function:

$$\Gamma\left(\frac{3}{2}+h\right) = \Gamma\left(\frac{3}{2}\right)\left[1 + h\psi\left(\frac{3}{2}\right) + \mathcal{O}(h^2)\right], \tag{2}$$

where $\psi(3/2) = 2 - \gamma_E - 2\ln 2 = -C - \ln 2$, with $C \equiv \gamma_E + \ln 2 - 2$. Expanding $(x/2)^{-\nu} = (x/2)^{-3/2}(x/2)^{-h} = (x/2)^{-3/2}[1 - h\ln(x/2)]$:

$$\left|\frac{\Gamma(\nu)}{\Gamma(3/2)}\left(\frac{x}{2}\right)^{-\nu}\right|^2 = \left(\frac{x}{2}\right)^{-3}\left[1 + 2h\psi(3/2) - 2h\ln\left(\frac{x}{2}\right)\right] = \left(\frac{x}{2}\right)^{-3}\left[1 - 2h(C + \ln x)\right]. \tag{3}$$

Substituting $x = -k\eta = \frac{k}{aH(1+\epsilon)} \approx \frac{k}{aH}(1-\epsilon)$:

$$\ln x = \ln\left(\frac{k}{aH}\right) - \epsilon. \tag{4}$$

Substituting $h = 2\epsilon - \delta$ and grouping terms linear in ϵ and δ:

$$\mathcal{P}_\zeta(k) = \frac{1}{\pi}\frac{H^2}{M_{\rm Pl}^2\epsilon}\left[1 - 2(2C+1)\epsilon + 2C\delta\right]\left(\frac{k}{aH}\right)^{2\delta - 4\epsilon}. \tag{5}$$

Differentiating $\ln \mathcal{P}_\zeta$ with respect to $\ln k$ yields the exact next-to-leading order spectral index (9.4.133).

9.4.5 Relation to observations

As long as we work in the slow-roll regime, perturbations are fully characterized by the two primordial spectra A_T and A_S. For super-Hubble modes, ζ is constant during cosmic evolution. Computations of the matter power spectrum and CMB temperature anisotropies require the gravitational potential Φ:

$$\Phi(k\eta \ll 1, \eta < \eta_{\rm eq}) = \frac{2}{3}\zeta, \qquad \Phi(k\eta \ll 1, \eta > \eta_{\rm eq}) = \frac{3}{5}\zeta. \tag{9.4.142}$$

Cosmic microwave background

Restricting ourselves to large angular scales (small multipoles ℓ), the CMB temperature angular power spectrum for scalar modes from the Sachs–Wolfe effect (Chapter 6) is:

$$C_\ell^S \simeq \pi A_S^2(k_0)\left[\frac{\pi}{2}\frac{\Gamma(3-n_s)}{2^{3-n_s}}\frac{\Gamma\left(\ell + \frac{n_s-1}{2}\right)}{\Gamma^2\left(2 - \frac{n_s}{2}\right)\Gamma\left(\ell + \frac{5-n_s}{2}\right)}\right], \tag{9.4.143}$$

with pivot $k_* = k_0 = 1/\eta_0$. For a scale-invariant spectrum $n_s \sim 1$, $\ell(\ell+1)C_\ell^S = \frac{\pi}{2}A_S^2(k_0)$. Observations of CMB anisotropies by WMAP [28] and Planck constrain the quadrupole:

The connection between a scale-invariant primordial spectrum and the angular power spectrum C_ℓ^S stems directly from the line-of-sight Sachs–Wolfe integral. On large angular scales ($\ell \lesssim 30$, where causal microphysics, acoustic oscillations, and Silk damping are inactive), the Sachs–Wolfe effect dominates the temperature anisotropy:

$$\Theta_{\rm SW}(\boldsymbol{e}) = \frac{1}{3}\Phi(\boldsymbol{x}_{\rm LSS}, \eta_{\rm LSS}). \tag{1}$$

Expanding into spherical harmonics and projecting Fourier modes k onto the last scattering sphere at comoving distance $r_{\rm LSS} = \eta_0 - \eta_{\rm LSS} \simeq \eta_0$ introduces the spherical Bessel functions $j_\ell(kr_{\rm LSS})$:

$$C_\ell^S = \frac{2}{\pi}\int_0^\infty \mathrm{d}k\, k^2\, j_\ell^2(k\eta_0)\left|\frac{1}{3}\Phi_k\right|^2 = \frac{2}{\pi}\int_0^\infty \frac{\mathrm{d}k}{k}\, j_\ell^2(k\eta_0)\,\frac{1}{9}\mathcal{P}_\Phi(k), \tag{2}$$

where $\mathcal{P}_\Phi(k) \equiv \frac{k^3}{2\pi^2}|\Phi_k|^2$ is the dimensionless power spectrum of the gravitational potential. In a matter-dominated background, super-Hubble modes relate to the comoving curvature perturbation ζ via $\Phi = \frac{3}{5}\zeta$, giving $\mathcal{P}_\Phi(k) = \frac{9}{25}\mathcal{P}_\zeta(k) \equiv \frac{9}{4}A_S^2(k)$. Thus:

$$C_\ell^S = \pi\int_0^\infty \frac{\mathrm{d}k}{k}\, j_\ell^2(k\eta_0)\, A_S^2(k). \tag{3}$$

For a general power-law primordial spectrum parameterized with spectral index n_s:

$$\mathcal{P}_\zeta(k) \propto k^{n_s-1} \implies A_S^2(k) = A_S^2(k_0)\left(\frac{k}{k_0}\right)^{n_s-1}. \tag{4}$$

Substituting $x \equiv k\eta_0$ and using the standard Weber–Schafheitlin integral for spherical Bessel functions:

$$\int_0^\infty \mathrm{d}x\, x^{n_s-2}\, j_\ell^2(x) = \frac{\pi}{2^{4-n_s}}\frac{\Gamma(3-n_s)\,\Gamma\left(\ell+\frac{n_s-1}{2}\right)}{\Gamma^2\left(\frac{4-n_s}{2}\right)\Gamma\left(\ell+\frac{5-n_s}{2}\right)}, \tag{5}$$

which yields the exact expression (9.4.143).
To understand what scale invariance implies for the observable C_ℓ^S:

1. **Exact scale invariance ($n_s = 1$):** When $n_s = 1$ (the Harrison–Zel'dovich spectrum), the dimensionless primordial variance per logarithmic wavenumber interval is strictly constant:

   $$\mathcal{P}_\zeta(k) = \text{constant} \iff \Delta_\Phi^2(k) \equiv \mathcal{P}_\Phi(k) = \text{constant}. \tag{6}$$

   The Bessel integral then reduces via $\int_0^\infty \frac{\mathrm{d}x}{x} j_\ell^2(x) = \frac{1}{2\ell(\ell+1)}$ directly to:

   $$C_\ell^S = \frac{\pi}{2}\frac{A_S^2(k_0)}{\ell(\ell+1)} \iff \ell(\ell+1)C_\ell^S = \frac{\pi}{2}A_S^2(k_0) = \text{constant}. \tag{7}$$

   Therefore, scale invariance of primordial metric fluctuations maps onto the condition that the quantity $\ell(\ell+1)C_\ell$ (the Sachs–Wolfe plateau) is independent of multipole ℓ.

2. **Physical requirement across scales:** A flat plateau $\ell(\ell+1)C_\ell = \text{const}$ requires that the fluctuation amplitude entering the horizon at each epoch carries identical gravitational potential depth Φ. Because a multipole ℓ subtends an angle $\theta \sim \pi/\ell$, the angular variance per logarithmic interval in multipole space is:

   $$\frac{\mathrm{d}\langle(\Delta T/T)^2\rangle}{\mathrm{d}\ln\ell} \simeq \frac{2\ell+1}{4\pi}\ell\, C_\ell \approx \frac{\ell(\ell+1)}{2\pi}C_\ell = \text{constant}. \tag{8}$$

   Hence, scale invariance demands that the temperature fluctuation power per decade in angular scale ℓ is strictly constant, mirroring the scale independence $\mathrm{d}\langle\zeta^2\rangle/\mathrm{d}\ln k = \text{constant}$ in 3D wavenumber space.

3. **Slow-roll tilt:** In realistic slow-roll inflation, the spacetime departs slightly from pure de Sitter ($\epsilon \neq 0, \eta_V \neq 0$). The spectral index becomes slightly tilted ($n_s - 1 = 2\delta - 4\epsilon \approx -0.035$). Using the asymptotic ratio of Gamma functions $\frac{\Gamma(\ell+a)}{\Gamma(\ell+b)} \approx \ell^{a-b}$ for large ℓ:
$$\ell(\ell+1)C_\ell^S \propto \ell^{n_s-1}. \tag{9}$$
Thus, an exact constant plateau requires $n_s = 1$; the slight red tilt predicted by inflation induces a mild downward slope across the low-ℓ Sachs–Wolfe plateau.

$$Q_{\rm rms-PS} \equiv \Theta_0\sqrt{\frac{5C_2^S}{4\pi}} = 18 \times 10^{-6}\,{\rm K}. \tag{9.4.144}$$

For $\Theta_0 = 2.73\,{\rm K}$, $A_S^2(k_0) \approx 4.16 \times 10^{-10}$. Using (9.4.104) and (9.2.31), we infer the energy scale of inflation:
$$\left(\frac{V}{\epsilon}\right)^{1/4} \simeq 0.0047\,M_{\rm Pl} \simeq 1.5 \times 10^{17}\,{\rm GeV}. \tag{9.4.145}$$
This represents an upper bound since part of the low-ℓ power may originate from gravitational waves.

For gravitational waves, the ratio of quadrupole contributions is:
$$r \equiv \frac{C_2^T}{C_2^S} = f(h, \Omega_{\rm m}, \Omega_\Lambda, \dots)\epsilon, \tag{9.4.146}$$
where $f \sim 10$ for the standard ΛCDM model ($\Omega_{\Lambda 0} \approx 0.7$). WMAP and Planck observational limits impose:
$$\epsilon < 0.032, \tag{9.4.147}$$
which implies
$$\frac{H_*}{M_{\rm Pl}} \lesssim 1.4 \times 10^{-5} \iff \left(\frac{V_*}{M_{\rm Pl}^4}\right)^{1/4} \lesssim 2.2 \times 10^{-3}. \tag{9.4.148}$$

Link with the number of e-folds

Observations constrain the parameter pair $(r, n_s - 1)$, which directly relates to the potential and the number of e-folds N between horizon exit and the end of inflation.

For power-law chaotic potentials $V(\phi) \propto \phi^n$ (9.2.52), relations (9.2.53) and (9.4.127) give:
$$r = \frac{f}{2}\left(\frac{n}{n+2}\right)(1 - n_s). \tag{9.4.149}$$

In terms of the number of e-folds N:
$$r = \frac{fn}{4(N + n/4)}, \qquad 1 - n_s = \frac{n+2}{2(N+n/4)}. \tag{9.4.150}$$

Figure 9.12 displays these model predictions in the $(n_s - 1, r)$ plane against observational confidence regions.

The symmetry between horizon exit ($\eta < \eta_e$) and horizon re-entry ($\eta > \eta_e$) bounds the total observable e-folds:
$$N < 60.9 + \frac{1}{2}\ln\left(\frac{H_*}{10^{15}\,{\rm GeV}}\right) - \ln\left(\frac{k}{0.02\,{\rm Mpc}^{-1}}\right). \tag{9.4.151}$$

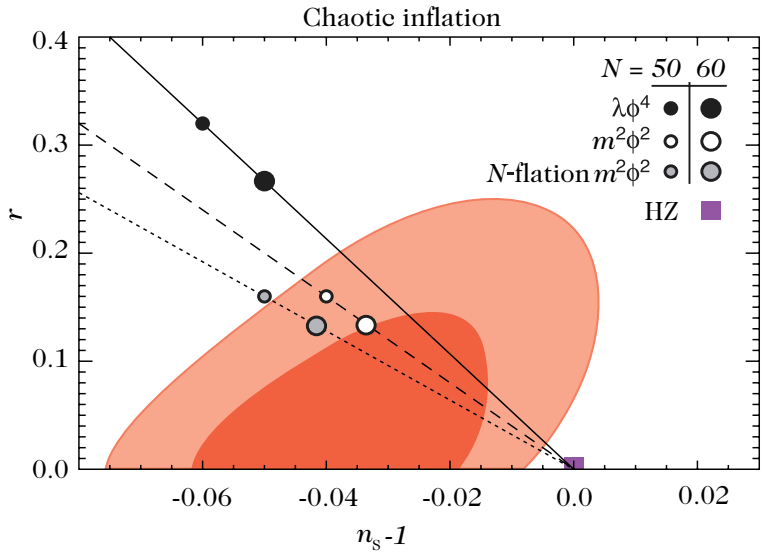


Figure 9.12: Constraints on single-field chaotic inflationary models with potentials $V(\phi) \propto \phi^n$ with $n = 2$ (dashed), $n = 4$ (solid) and N-flation (dotted) in the $(n_s - 1, r)$ plane. HZ denotes the Harrison–Zel'dovich scale-invariant spectrum ($n_s = 1, r = 0$). The predictions for $N = 50$ and $N = 60$ e-folds are shown. The ϕ^4 model is strongly excluded. Adapted from Ref. [29].

Using $H_*/M_{\rm Pl} \lesssim 2 \times 10^{-5}$ and $k = a_0 H_0$:

$$N < 62.5 + \ln\left(\frac{0.6}{h}\right). \tag{9.4.152}$$

Assuming the Sachs–Wolfe plateau is entirely generated by tensor modes yields the absolute upper bound $V_*^{1/4} < 3 \times 10^{16}\,\text{GeV}$.

Consistency relation and the Lyth bound

WMAP constraints give $C_{10}^T/C_{10}^S < 0.32$. During an interval of ΔN e-folds, the inflaton field displacement is:

$$\frac{\Delta\phi}{M_{\rm Pl}} = \sqrt{\frac{\epsilon}{4\pi}}\Delta N, \tag{9.4.153}$$

which, using (9.4.146), yields the celebrated **Lyth bound** [32]:

$$\frac{\Delta\phi}{M_{\rm Pl}} = \sqrt{\frac{16\pi r}{f}}\Delta N \sim 2.37\left(\frac{r}{0.07}\right)^{1/2}\left(\frac{f}{10}\right)^{-1/2}\left(\frac{\Delta N}{4}\right). \tag{9.4.154}$$

Hence, detecting $r \gtrsim 0.01$ implies a **super-Planckian field excursion** $\Delta\phi > M_{\rm Pl}$.

Due to cosmic variance and lensing B-mode foregrounds, the primordial gravitational wave detection floor is $r \gtrsim 6 \times 10^{-4}$ [34], corresponding to an inflation energy scale bound:

$$V_*^{1/4} > 3.2 \times 10^{15}\,\text{GeV} \iff \left(\frac{V_*}{M_{\rm Pl}^4}\right)^{1/4} \gtrsim 2.62 \times 10^{-4}. \tag{9.4.155}$$

9.4.6 Reconstruction of the potential

Principle

Measuring the cosmological parameters $(n_s, n_T, R, \alpha_s, \alpha_T)$ allows one, in principle, to invert the perturbation equations and reconstruct the inflaton potential $V(\phi)$ around the horizon-exit value

ϕ_*:

$$V(\phi) = V_* + V'_*\Delta\phi + \frac{1}{2}V''_*(\Delta\phi)^2 + \dots. \tag{9.4.156}$$

The potential derivatives in terms of the slow-roll parameters are:

$$V_* = \frac{M_{\rm Pl}^2}{8\pi}H^2(3-\epsilon), \tag{9.4.157}$$

$$V'_* = -\frac{M_{\rm Pl}}{\sqrt{4\pi}}H^2\sqrt{\epsilon}(3-\delta), \tag{9.4.158}$$

$$V''_* = H^2(3\epsilon + 3\delta - \delta^2 - \xi_H^2). \tag{9.4.159}$$

Expressed in terms of observational quantities:

$$V_* = \frac{75}{32}M_{\rm Pl}^4 A_T^2(k_*)\left[1 + \left(\frac{5}{3} + 2C\right)R_*\right]. \tag{9.4.160}$$

The higher derivatives are given by [35]:

$$V'_* = -\frac{75}{8}\sqrt{\pi}M_{\rm Pl}^3 R_*^{3/2}A_S^2(k_*)\left[1 + \left(3C + \frac{4}{3}\right)R_* - \frac{1}{2}\left(C - \frac{1}{3}\right)(1-n_{s*})\right], \tag{9.4.161}$$

$$V''_* = \frac{25}{4}\sqrt{\pi}M_{\rm Pl}^2 R_* A_S^2(k_*)\left[9R_* - \frac{3}{2}(1-n_{s*}) + (36C+2)R_*^2 - \frac{1}{4}(1-n_{s*})^2 \right. \tag{9.4.162}$$
$$\left. -(12C-6)R_*(1-n_{s*}) - \frac{3C-1}{2}\alpha_{s*}\right].$$

Table 9.2: Summary of slow-roll parameters and observables required to reconstruct the potential derivatives.

	Parameters		**Observables**	
Derivative	**Leading Order**	**Next Order**	**Leading Order**	**Next Order**
V	H	H, ϵ	A_T^2	A_T^2, R
V'	H, ϵ	H, ϵ, δ	A_T^2, R	A_T^2, R, n_s
V''	H, ϵ, δ	H, ϵ, δ, ξ	A_T^2, R, n_s	A_T^2, R, n_s, α_s
V'''	H, ϵ, δ, ξ	—	A_T^2, R, n_s, α_s	—

Flow equations

An alternative approach consists of generating inflationary trajectories stochastically using flow equations [36]. We define the infinite hierarchy of slow-roll parameters:

$$^{\ell}\lambda_H \equiv \left(\frac{M_{\rm Pl}^2}{4\pi}\right)^{\ell}\frac{(H')^{\ell-1}}{H^{\ell}}\frac{{\rm d}^{\ell+1}H}{{\rm d}\phi^{\ell+1}}, \tag{9.4.163}$$

with $\delta_H = {}^1\lambda_H$ and $\xi_H^2 = {}^2\lambda_H$. Using the e-fold number N as the independent variable ($\mathrm{d}/\mathrm{d}N = M_{\rm Pl}\sqrt{\epsilon/(4\pi)}\,\mathrm{d}/\mathrm{d}\phi$), the flow hierarchy is:

$$\frac{{\rm d}\epsilon_H}{{\rm d}N} = 2\epsilon_H(\delta_H - \epsilon_H), \qquad \frac{{\rm d}\,{}^{\ell}\lambda_H}{{\rm d}N} = \left[(\ell-1)\,{}^1\lambda_H - \ell\epsilon_H\right]{}^{\ell}\lambda_H + {}^{\ell+1}\lambda_H. \tag{9.4.164}$$

Truncating this hierarchy at order $\ell = M$, one can numerically sample the space of viable inflationary potentials. Figure 9.13 displays the distribution of 10^6 randomly generated models in the observable parameter space.

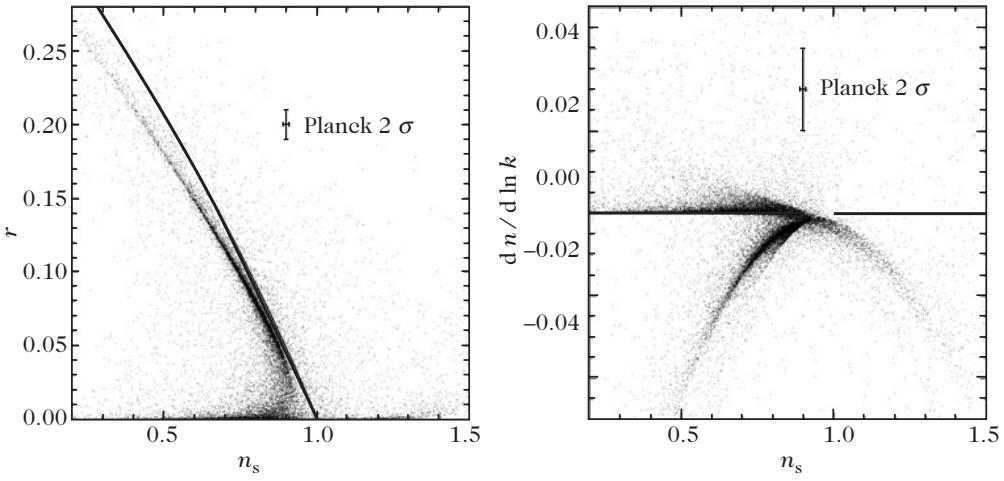


Figure 9.13: Distribution of 10^6 inflationary models with truncation order $M = 5$ and e-fold range $N = 40$–70 in the observable planes (n_s, r) and $(n_s, \mathrm{d}n_s/\mathrm{d}\ln k)$. Adapted from Ref. [36].

9.5 End of inflation and reheating

The theory of reheating after inflation is one of the most important applications of quantum field theory in curved space-time. Almost all the matter present in our Universe would have been created through a process of particle production during the inflaton decay.

9.5.1 Perturbative reheating

During inflation, all the energy is concentrated in the slowly rolling inflaton. Shortly after the end of inflation, the Universe is cold and 'frozen' in a state of low entropy where the field oscillates around the minimum of its potential (Fig. 9.5). The coherent oscillations of the inflaton can be considered as a collection of independent scalar particles. If they couple to other particles, the inflaton can then decay perturbatively to produce light particles. The interaction of the inflaton should therefore give rise to an effective decay rate, Γ_ϕ, and reheating would only occur after the expansion rate had decreased to a value $H \sim \Gamma_\phi$. This also implies that during the first $\mathcal{O}(m/\Gamma_\phi)$ oscillations of the inflaton, nothing interesting happens.

Let us illustrate this mechanism in the case of a model of chaotic inflation with the potential (9.2.52). As illustrated in Fig. 9.5 for $n = 2$, the inflaton behaves as a pressureless fluid during the oscillatory phase and the scale factor evolves as $a \propto t^{2/3}$. To see this, let us rewrite the Klein–Gordon equation as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(a^3\rho_\phi\right) = -P_\phi \frac{\mathrm{d}a^3}{\mathrm{d}t}. \tag{9.5.1}$$

During the oscillatory phase, $H < m$ and the inflaton undergoes several oscillations during a time H^{-1}. It is thus reasonable to use the mean value of the pressure and density over several oscillations. For a harmonic oscillator $\langle\dot{\phi}^2/2\rangle = (n/2)\langle V(\phi)\rangle$, so that $\langle P_\phi\rangle = \frac{n-2}{n+2}\langle\rho_\phi\rangle$. The scalar field thus behaves as a dust fluid for $n = 2$ and as a radiation fluid for $n = 4$.

In order for the inflaton to decay, it should be coupled to other fields. Let us restrict ourselves to the case $n = 2$ and, as an example, let us consider couplings described by the Lagrangian

$$-\mathcal{L}_{\mathrm{int}} = \sigma v \phi \chi^2 + h\phi\bar{\psi}\psi, \tag{9.5.2}$$

where σ and h are dimensionless coupling constants and v is a mass scale. No mass term has been introduced for the bosons χ and the fermions ψ that are assumed to be very light compared to the inflaton. This is necessary in order for the decay rate of the inflaton into these particles to be noticeable. As seen earlier, the mass scale of the inflaton should be of the order of $m \sim 10^{-6} M_{\mathrm{Pl}}$.

Evolution of the inflaton

Including quantum corrections, the Klein–Gordon equation becomes

$$\ddot{\phi} + 3H\dot{\phi} + \left[m^2 + \Pi(m)\right]\phi = 0, \tag{9.5.3}$$

where $\Pi(m)$ is the polarization operator of the inflaton. The real part of $\Pi(m)$ corresponds to the mass correction. Even though $\mathrm{Im}[\Pi(m)] \ll m^2$, Π has an imaginary part

$$\mathrm{Im}[\Pi(m)] = m\Gamma_\phi. \tag{9.5.4}$$

Since m is much larger than both Γ_ϕ and H at the end of inflation, we can solve the Klein–Gordon equation by assuming that both these quantities are constant during an oscillation. We infer that the inflaton evolves as

$$\phi = \Phi(t)\sin mt, \qquad \Phi(t) = \phi_0 \exp\left[-\frac{1}{2}\int (3H + \Gamma_\phi)\,\mathrm{d}t\right]. \tag{9.5.5}$$

As long as $3H > \Gamma_\phi$, the decrease in the inflaton energy caused by the expansion (Hubble friction) dominates over particle decay. Since $H = 2/(3t)$, we infer that

$$\Phi(t) = \phi_f \frac{t_f}{t} = \frac{M_{\rm Pl}}{m}\frac{1}{\sqrt{3\pi}\,t}, \tag{9.5.6}$$

where we have used $\phi_f = M_{\rm Pl}/\sqrt{4\pi}$, $t_f = 2/(3H_f)$ and $H_f^2 = \frac{4\pi}{3}(m/M_{\rm Pl})^2\phi_f^2$. We recover $\langle\rho_\phi\rangle = \frac{1}{2}m^2\Phi^2 \propto a^{-3}$.

Reheating temperature

During the oscillatory phase following inflation, the inflaton condensate behaves as a pressureless matter fluid with cycle-averaged energy density $\langle\rho_\phi\rangle = \frac{1}{2}m^2\Phi^2(t) \propto a^{-3}$. Taking into account the decay width Γ_ϕ arising from interactions with light fields, the oscillation amplitude (9.5.5) evolves according to

$$\Phi(t) = \frac{M_{\rm Pl}}{m}\frac{1}{\sqrt{3\pi}\,t}e^{-\Gamma_\phi t/2}. \tag{9.5.7}$$

As long as the Hubble expansion rate satisfies $3H(t) \gg \Gamma_\phi$, the redshift caused by cosmic expansion dominates over particle decay, and the energy remains almost entirely locked in the coherent inflaton oscillations. Perturbative reheating becomes efficient only when the expansion rate drops to a value comparable to the decay rate:

$$\Gamma_\phi \simeq 3H(t_{\rm reh}) \iff H(t_{\rm reh}) = \frac{\Gamma_\phi}{3}, \tag{9.5.8}$$

which defines the characteristic time of reheating $t_{\rm reh} \simeq 2/\Gamma_\phi$.

Using the flat Friedmann equation $H^2 = \frac{8\pi G}{3}\rho_\phi = \frac{8\pi}{3M_{\rm Pl}^2}\rho_\phi$, the energy density stored in the oscillating inflaton at the moment of reheating is directly determined by

$$\rho_\phi(t_{\rm reh}) = \frac{3M_{\rm Pl}^2}{8\pi}H^2(t_{\rm reh}) = \frac{3M_{\rm Pl}^2}{8\pi}\left(\frac{\Gamma_\phi}{3}\right)^2 = \frac{\Gamma_\phi^2 M_{\rm Pl}^2}{24\pi} \equiv \rho_{\rm reh}. \tag{9.5.9}$$

To determine the resulting reheating temperature $T_{\rm reh}$, we assume the idealized limit of **instantaneous reheating and thermalization**: whatever energy density was stored in the inflaton field at $t = t_{\rm reh}$ is assumed to be instantaneously and completely converted into relativistic particles (radiation), such that $\rho_r(t_{\rm reh}) = \rho_\phi(t_{\rm reh}) = \rho_{\rm reh}$. The energy density of the relativistic thermal plasma at temperature $T_{\rm reh}$ is governed by the Stefan–Boltzmann law:

$$\rho_{\rm reh} = \frac{\pi^2}{30}g_* T_{\rm reh}^4, \tag{9.5.10}$$

where $g_* = g_*(T_{\rm reh})$ denotes the effective number of relativistic degrees of freedom in thermal equilibrium ($g_* \approx 106.75$ for the full Standard Model at high energies).

Equating the radiation energy density (9.5.10) to the inflaton energy density (9.5.9) yields

$$\frac{\pi^2}{30} g_* T_{\rm reh}^4 = \frac{\Gamma_\phi^2 M_{\rm Pl}^2}{24\pi} \implies T_{\rm reh}^4 = \frac{30}{\pi^2 g_*} \left(\frac{\Gamma_\phi^2 M_{\rm Pl}^2}{24\pi} \right) = \frac{5}{4\pi^3 g_*} \Gamma_\phi^2 M_{\rm Pl}^2. \tag{9.5.11}$$

Taking the fourth root, the reheating temperature is explicitly given by

$$T_{\rm reh} = \left(\frac{5}{4\pi^3 g_*} \right)^{1/4} \sqrt{\Gamma_\phi M_{\rm Pl}} \simeq 0.14 \left(\frac{100}{g_*} \right)^{1/4} \sqrt{\Gamma_\phi M_{\rm Pl}} \ll 10^{15}\ \text{GeV}. \tag{9.5.12}$$

The theoretical upper bound $T_{\rm reh} \ll 10^{15}$ GeV follows directly from the perturbative consistency condition $\Gamma_\phi \ll m \sim 10^{-6} M_{\rm Pl}$, assuming $g_* \gtrsim 100$. This result carries significant cosmological implications:

- **Grand Unified Baryogenesis:** Since $T_{\rm reh} < E_{\rm GUT} \sim 10^{16}$ GeV (the characteristic scale of Grand Unified symmetry breaking, see Chapter 9), the Universe is never reheated to GUT temperatures after inflation, precluding standard GUT-scale thermal baryogenesis.
- **The Gravitino Constraint:** In supersymmetric models, thermal scatterings in the hot plasma can produce dangerous long-lived gravitinos whose late decay would disrupt Big-Bang nucleosynthesis (BBN). Avoiding gravitino overproduction requires a stringent upper limit on the reheating temperature, typically $T_{\rm reh} \lesssim 10^9$–$10^{10}$ GeV.

Decay rate

Γ_ϕ is the total decay rate of the inflaton and can be decomposed as $\Gamma_\phi = \Gamma(\phi \to \chi\chi) + \Gamma(\phi \to \bar{\psi}\psi)$ where the decay rates into bosons and fermions are, respectively, given by

$$\Gamma(\phi \to \chi\chi) = \frac{\sigma^2 v^2}{8\pi m}, \qquad \Gamma(\phi \to \bar{\psi}\psi) = \frac{h^2 m}{8\pi}. \tag{9.5.13}$$

These expressions are only valid in the limit where $\Gamma_\phi \ll m$, i.e. $h \ll 1$ and $\sigma v \ll m$. Assuming $h \sim 10^{-2}$, the decay rate into fermions is of the order of $\Gamma(\phi \to \bar{\psi}\psi) \sim 10^8$ GeV. The reheating temperature (9.5.12) is then $T_{\rm reh} \sim 10^{13}$ GeV and the inflaton undergoes about 10^5 oscillations before decaying. For bosons, assuming $\sigma \sim 10^{-2}$ and $v \sim 10^{11}$ GeV, we find that $\Gamma(\phi \to \chi\chi) \sim 10^3$ GeV only. The reheating temperature is of the order of $T_{\rm reh} \sim 10^{10}$ GeV and the inflaton undergoes about 10^{10} oscillations.

Note that the decay rate for an interaction of the type $\mathcal{L}_{\rm int} = -\frac{1}{2} g^2 \phi^2 \chi^2$ is

$$\Gamma(\phi\phi \to \chi\chi) = \frac{g^4 \Phi^2}{8\pi m}. \tag{9.5.14}$$

Unlike 'three-leg' interactions, this rate depends on the value of the inflaton. Since $\phi^2 \sim a^{-3} \sim t^{-2}$, this decay rate decreases quicker than the Hubble rate H. Perturbative reheating can therefore never be produced by 'four-leg' interactions.

Complete reheating can only occur if Γ_ϕ decreases slower than t^{-1}. Hence assuming that reheating is complete imposes constraints on the structure of the theory and on the couplings of the inflaton with other fields.

Phenomenological formulation

To finish, let us mention a simplified version of this mechanism in which we introduce the coupling between the scalar field and radiation. The conservation equations are then transformed into

$$\dot{\rho}_\phi + H(\rho_\phi + P_\phi) = -\Gamma_\phi \rho_\phi, \tag{9.5.15}$$

$$\dot{\rho}_r + 4H\rho_r = \Gamma_\phi \rho_\phi. \tag{9.5.16}$$

Reheating can then be described phenomenologically as a transfer between two fluids.

9.5.2 Theory of preheating

The theory of perturbative reheating is simple and intuitive in many aspects. However, the decay of the inflaton can start much earlier in a phase of preheating (parametric reheating) where particles are produced by parametric resonance.[3] The preheating process can be decomposed into three stages: (1) non-perturbative production of particles, (2) perturbative stage and (3) thermalization of the produced particles.

Parametric resonance

In order to illustrate this mechanism, let us consider a model of chaotic inflation with potential (9.2.46) and an interaction of the form

$$V_{\rm int} = \frac{1}{2} g^2 \phi^2 \chi^2. \tag{9.5.17}$$

The scalar field χ can be decomposed, in second quantization, into

$$\hat{\chi} = \int \frac{\mathrm{d}^3 k}{(2\pi)^{3/2}} \left[\chi_k(t) \hat{a}_{\mathbf{k}}(t) e^{-i\mathbf{k}\cdot\mathbf{x}} + \text{h.c.} \right]. \tag{9.5.18}$$

Considering only the interaction between this quantum operator and the classical field ϕ, then the evolution of each mode is dictated by

$$\ddot{\chi}_k + 3H\dot{\chi}_k + \left(\frac{k^2}{a^2} + g^2\phi^2 \right) \chi_k = 0. \tag{9.5.19}$$

One can rewrite this equation in terms of the variable $X_k = a^{3/2}\chi_k$ as

$$\ddot{X}_k + \omega_k^2 X_k = 0, \qquad \omega_k^2 = \frac{k^2}{a^2} + g^2\phi^2 - \frac{3}{4}\left(2\dot{H} + 3H^2\right). \tag{9.5.20}$$

Using the time variable $z = mt$ and defining the quantities

$$q \equiv \frac{g^2 \Phi^2(t)}{4m^2}, \qquad A_k \equiv 2q + \frac{k^2}{m^2 a^2}, \tag{9.5.21}$$

where Φ is defined by (9.5.5), this equation takes the form

$$\frac{\mathrm{d}^2 X_k}{\mathrm{d}z^2} + \left(A_k - 2q\cos 2z + \Delta \right) X_k = 0, \tag{9.5.22}$$

where $\Delta = \left[m_\chi^2 - \frac{3}{4}(2\dot{H} + 3H^2) \right] / m^2$. This term is, in general, negligible.

[3] Defining a parametric oscillator as an oscillator with time-dependent frequency, the associated resonance, when it exists, will be a parametric resonance.

Minkowski space-time case

For a Minkowski space-time, the coefficients A_k and q are constant (we set $a = 1$). Then this equation is a Mathieu equation. Floquet's theorem implies that its solutions are of the form

$$X_k(z) = \exp(\mu_k z)\phi_1(z) + \exp(-\mu_k z)\phi_2(z), \tag{9.5.23}$$

where ϕ_i are two periodic functions of period π and $\mu_k = \mu_k(A_k, q)$ is the Floquet exponent. When μ_k has a non-vanishing real part, the amplitude of the mode X_k grows exponentially. This corresponds to the instability zones of the parameter space (A_k, q) (Fig. 9.14). Note that the Mathieu equation is simply the Schrödinger equation for an electron in a one-dimensional periodic potential. The instability zones then correspond to the allowed energy bands for the electron.

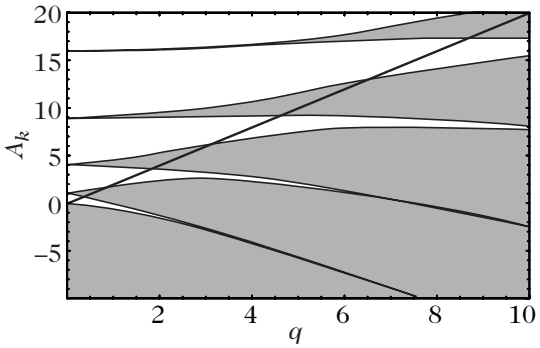


Figure 9.14: Stability diagram of the Mathieu equation. The shaded zones are instability zones where X_k grows exponentially. As $q \to 0$, we find narrow resonances for $A_k \simeq n^2$ ($n \in \mathbb{N}$) whereas for $q \gg 1$, the instability zones are large. The line $A_k = 2q$ corresponds to $k = 0$. In an expanding Universe, a given mode evolves by moving along a line of slope greater than 2 ($A_k = 2q + k^2/(m^2a^2)$) towards the origin (as $q \propto a^{-3}$ for $V = \frac{1}{2}m^2\phi^2$). Adapted from Ref. [38].

In the narrow resonance regime ($q \ll 1$), we analyze the instability of the Mathieu equation (9.5.22) in static Minkowski spacetime:

$$\frac{\mathrm{d}^2 X_k}{\mathrm{d}z^2} + (A_k - 2q\cos 2z)\, X_k = 0. \tag{1}$$

For $q = 0$, the unperturbed solutions are free oscillations $X_k^{(0)}(z) = \cos(nz)$ or $\sin(nz)$ when $A_k = n^2$ ($n = 1, 2, \dots$). The strongest resonance occurs at the first band $n = 1$ ($A_k \approx 1$).
We perturb around $A_k = 1$ by setting $A_k = 1 + qA_1 + q^2A_2 + \dots$, and seek a Floquet solution of the form:

$$X_k(z) = e^{\mu z}\phi(z), \qquad \phi(z) = \sin(z - \theta) + q\phi_1(z) + \dots, \quad \mu = q\mu_1 + q^2\mu_2 + \dots \tag{2}$$

Substituting this ansatz into the Mathieu equation and collecting terms order by order in q:

$$\text{Order } q^0: \quad \phi_0'' + \phi_0 = 0 \implies \phi_0(z) = \sin(z - \theta), \tag{3}$$

$$\text{Order } q^1: \quad \phi_1'' + \phi_1 = -2\mu_1\phi_0' - A_1\phi_0 + 2\cos(2z)\phi_0. \tag{4}$$

Using the trigonometric identity $2\cos(2z)\sin(z - \theta) = \sin(3z - \theta) - \sin(z + \theta) = \sin(3z - \theta) - \cos(2\theta)\sin(z - \theta) - \sin(2\theta)\cos(z - \theta)$:

$$\phi_1'' + \phi_1 = \sin(3z - \theta) - [A_1 + \cos(2\theta)]\sin(z - \theta) - [2\mu_1 + \sin(2\theta)]\cos(z - \theta). \tag{5}$$

To avoid secular runaway resonant terms at frequency $n = 1$, the coefficients of $\sin(z-\theta)$ and $\cos(z-\theta)$ on the right-hand side must vanish simultaneously:

$$A_1 + \cos(2\theta) = 0, \qquad 2\mu_1 + \sin(2\theta) = 0. \tag{6}$$

Squaring and adding both relations eliminates the phase θ:

$$\cos^2(2\theta) + \sin^2(2\theta) = 1 \implies A_1^2 + 4\mu_1^2 = 1 \implies \mu_1 = \frac{1}{2}\sqrt{1 - A_1^2}. \tag{7}$$

Re-expressing in terms of the physical parameters $q = qA_1 = A_k - 1$:

$$\mu_k = \frac{1}{2}\sqrt{q^2 - (A_k - 1)^2}. \tag{8}$$

Instability ($\mathrm{Re}(\mu_k) > 0$) occurs in the band $|A_k - 1| \leq q$, with maximal growth rate:

$$\mu_{k,\max} = \frac{q}{2}. \tag{9}$$

Since $q = g^2\Phi^2/(4m^2)$, the mode grows exponentially as $X_k \propto \exp\left(\frac{g^2\Phi^2}{8m}t\right)$, demonstrating explosive bosonic particle creation.

As illustrated in Fig. 9.14, we obtain frequency bands Δn_k to which we can associate an integer index n, and inside which the mode grows exponentially, $X_k \propto \exp(\mu_k z)$ that can be interpreted as a production of particles. In this regime, called preheating, the perturbative description of Γ_ϕ is not applicable.

Resonance regimes

Two regimes must be distinguished:

1. **Narrow parametric resonance ($q \ll 1$):** Instability zones are centered around $A_k \simeq n^2$ ($n \in \mathbb{N}$) and the most important band corresponds to $n = 1$. The width of this instability band is $\Delta k \sim qm$ and the Floquet parameter is $\mu_k \leq q/2$. The maximal growth corresponds to $A_k = 1$ ($k \approx m$). The field amplitude thus grows as $\chi_k \propto \exp(\mu_k m t)$ with a rate $\mu_k m \sim g^2\Phi_0^2/(8m)$, which can be very large compared to Γ_ϕ. The total number of particles in a comoving volume grows as

$$n_\chi \propto \int \mathrm{d}^3k\, |X_k(t)|^2 \propto \exp(2\mu_k m t). \tag{9.5.24}$$

2. **Broad parametric resonance ($q \gg 1$):** In this regime, the system is in an instability zone for a large range of values of k. Particle creation can be interpreted as a breakdown of the adiabatic approximation. Defining the adiabaticity condition as

$$\left|\frac{\dot{\omega}_k}{\omega_k^2}\right| \ll 1, \tag{9.5.25}$$

where $\omega_k^2 \simeq k^2 + g^2\phi^2$, we find that this condition is satisfied except when $\phi \approx 0$. At each passage of ϕ through zero, particles with $k \leq k_* \sim (gm\Phi)^{1/2}$ are created. The time during which adiabaticity is broken is $\Delta t_* \sim (gm\Phi)^{-1/2}$, which corresponds to a phase $\Delta z \sim q^{-1/4} \ll 1$. The number of particles

$$n_k = \frac{\omega_k}{2}\left(\frac{|\dot{\chi}_k|^2}{\omega_k^2} + |\chi_k|^2\right) - \frac{1}{2} \tag{9.5.26}$$

increases exponentially as $n_k \propto \exp(2\mu_k m t)$ with $\mu_k \approx \frac{1}{2\pi}\ln 3 \simeq 0.17$.

In the broad resonance regime, the instability band extends up to $k_{\max} \sim mq^{1/4} \gg m$, which implies that particles with large momentum can be produced. The estimate of the behaviour of the parametric resonance given here is to be compared with the numerical solution of Fig. 9.15. One should note that the behaviour of n_k is much more regular and physically transparent than that of χ_k.

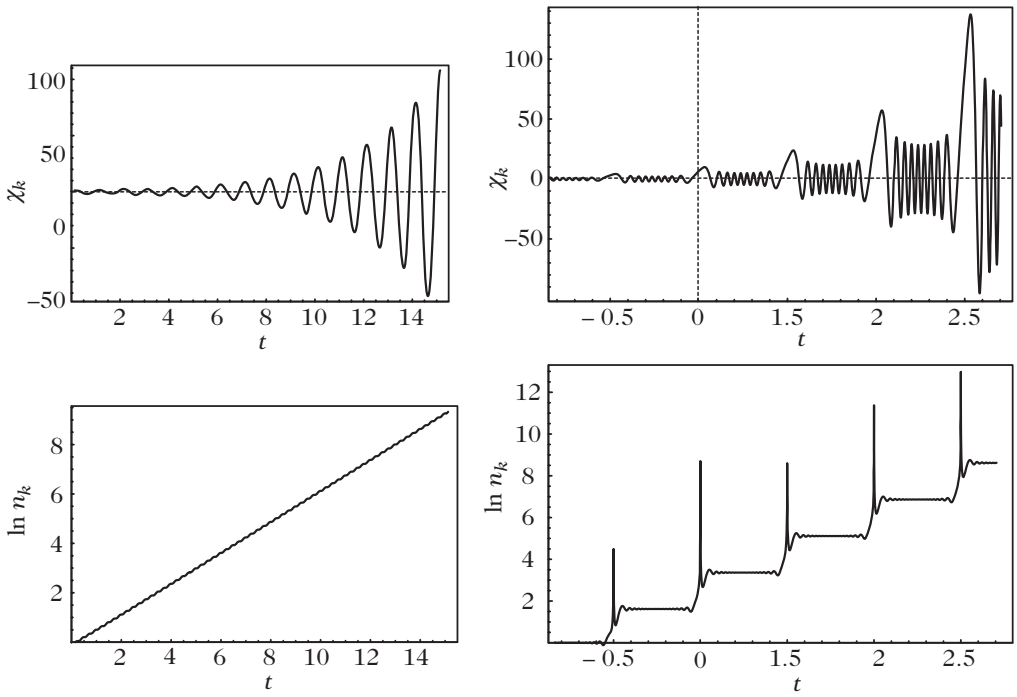


Figure 9.15: Evolution of χ_k and of the number of particles n_k as a function of time in units of $m/2\pi$ in Minkowski space-time. (left): Narrow parametric resonance with $q \sim 0.1$. The number of particles grows in an exponential way with slope $\mu_k \sim q/2 \sim 0.05$. (right): Large resonance with $q \sim 200$. At each oscillation of ϕ, χ_k oscillates several times. The peaks correspond to the moments where $\phi = 0$, where n_k is not defined. Adapted from Ref. [38].

Expanding space-time

In an expanding space-time, the parameter $q \propto a^{-3}$ and the physical momentum $k/a(t)$ vary with time. In particular, this changes the phase of χ_k at the time when $\phi = 0$ in an unpredictable way so that the number of particles can either increase or decrease. This is what is called the phase of **stochastic resonance**. One can show that n_k increases statistically three times more often than it decreases so that in the end n_k still increases exponentially.

If $q \gg 1$ initially, there is a phase of stochastic resonance associated with large resonances. As q decreases, the preheating continues through narrow resonances and then stops. This succession of events is illustrated in Fig. 9.16. The complete theory describing this scenario is developed in Ref. [38].

Backreaction

The dynamics presented above is, however, too simplified because the field χ has been treated as a test field evolving in a space-time described by $a(t)$ and $\phi(t)$ that are not themselves affected by the particle production. In fact, due to the exponential growth of χ_k, this field must backreact on the space-time dynamics and on the evolution of the field ϕ, in particular via a contribution of order $g^2\langle\chi^2\rangle$ to the effective mass of ϕ.

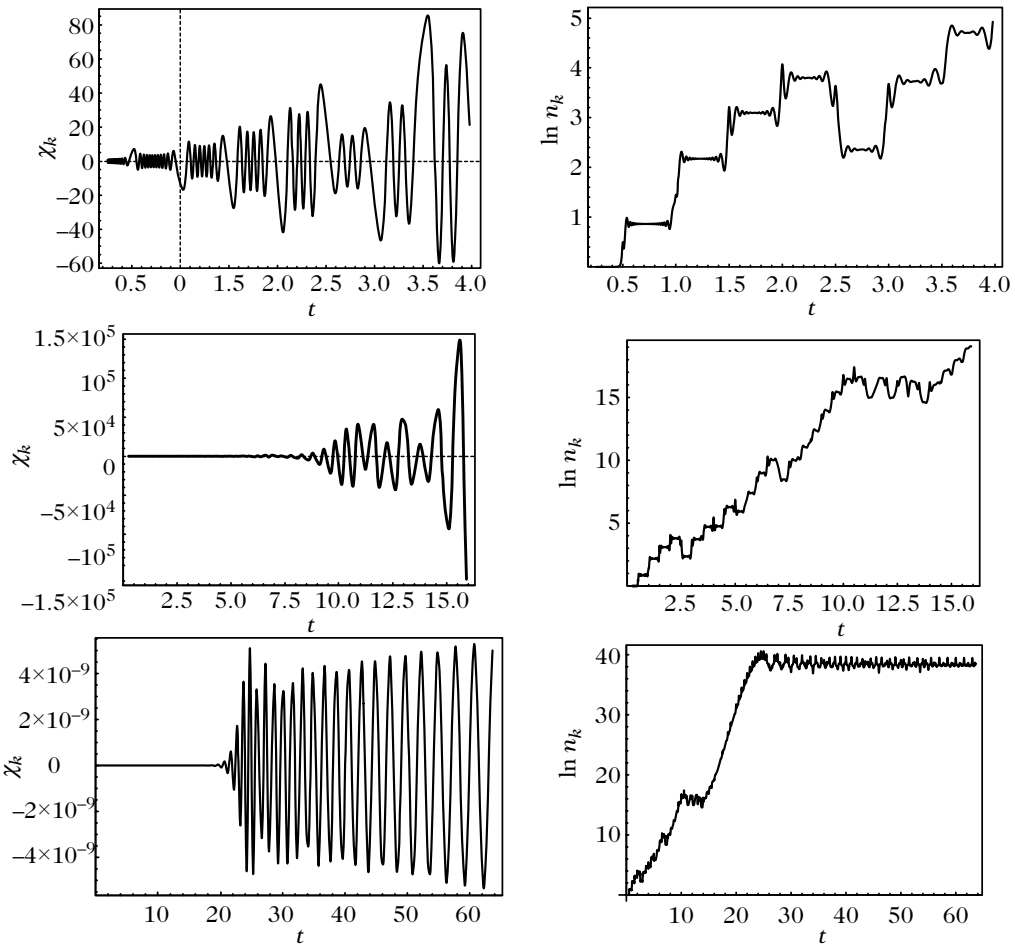


Figure 9.16: Evolution of χ_k and of the number of particles n_k as a function of time in units of $m/2\pi$ (i.e. number of oscillations) in an expanding space-time for $q = 3 \times 10^3$ initially ($g = 5\times10^{-4}, m = 10^{-6} M_{\rm Pl}$). During the first oscillations (top), n_k increases but can occasionally decrease, a purely quantum effect that would be impossible if the particles were created in a state of thermodynamic equilibrium. As t increases, $q \propto t^{-2}$ and the resonance becomes narrow (middle), then it disappears and n_k becomes constant (bottom). Adapted from Ref. [38].

A more precise description requires, first, to take into account the contribution of χ to the evolution of the Universe,

$$H^2 = \frac{8\pi G_{\rm N}}{3}\left[\frac{1}{2}\dot{\phi}^2 + \frac{1}{2}\dot{\chi}^2 + \frac{1}{2}m^2\phi^2 + \rho_\chi + \rho_\phi\right], \tag{9.5.27}$$

where ρ_χ and ρ_ϕ are the energy densities of the particles (i.e. $k \neq 0$) of χ and ϕ. We should then describe the coupled evolution of the perturbations of χ and ϕ, which introduces the polarization tensor for each field,

$$\ddot{X}_k + \left(\frac{k^2}{a^2} + g^2\Phi^2 \sin^2 mt\right) X_k = -\int X_k(t')\Pi_\chi(t,t';k)\,{\rm d}t', \tag{9.5.28}$$

$$\ddot{v}_k + \left(\frac{k^2}{a^2} + g^2\Phi^2 \sin^2 mt\right) v_k = -\int v_k(t')\Pi_\phi(t,t';k)\,{\rm d}t', \tag{9.5.29}$$

where $v = a^{3/2}\delta\phi$. To go further, we should therefore compute the polarization tensors. The description of the preheating phase then becomes more complicated and requires numerical simulations to be performed.

A simplified way to address the question is to work in the Hartree approximation for which the backreaction of χ on the evolution of the homogeneous solution is taken into account as

$$\ddot{\phi} + 3H\dot{\phi} + \left(m^2 + g^2\langle\chi^2\rangle\right)\phi = 0, \tag{9.5.30}$$

with

$$\langle\chi^2\rangle = \int \frac{k^2\,\mathrm{d}k}{2\pi^2 a^3}|X_k(t)|^2. \tag{9.5.31}$$

This term tends to make the oscillations of the inflaton incoherent and to suppress the resonance. Typically, the inflaton undergoes a few dozen oscillations and produces a number of particles of the order of $n_k \sim 10^2 g^{-2}$.

Discussion

The results of preheating are very model dependent. For instance, with the potential

$$V = \frac{\lambda}{4}\phi^4 + \frac{1}{2}g^2\phi^2\chi^2, \tag{9.5.32}$$

the resonance never becomes stochastic as all the parameters of the resonance have the same behaviour in Φ. The most general case with a term in $m^2\phi^2$ and a term in $\lambda\phi^4$ gives rise to various possibilities depending on the value of the parameters [38].

Another important case is that of **tachyonic reheating** [39] that relies on an effective potential of the form $V = \frac{\lambda}{4}(\phi^2 - v^2)^2 = \frac{\lambda}{4}\phi^4 - \frac{1}{2}m^2\phi^2 + \ldots$. The beginning of the phase transition ($\phi \sim 0$) is associated with a tachyonic instability ($V'' < 0$). Any mode $k < m$ grows exponentially and

$$\langle\delta\phi^2\rangle = \int_0^m \frac{k\,\mathrm{d}k}{4\pi^2} e^{2t\sqrt{m^2-k^2}}, \tag{9.5.33}$$

until $\langle\delta\phi^2\rangle \sim v^2/4$. This happens at around $t_* \sim (2m)^{-1}\ln(\pi^2/\lambda)$ and the occupation number grows until

$$n_k \sim \exp(2mt_*) \sim \frac{\pi^2}{\lambda} \gg 1. \tag{9.5.34}$$

In this case, the field decays before reaching the minimum of the potential and the reheating phase is quasi-instantaneous.

The dynamics of preheating is therefore very different from the picture proposed by perturbative reheating. A question that has not been addressed is that of the thermalization of the produced particles. Interesting phenomena can also occur in this out-of-equilibrium phase, such as the restoration of symmetries and the possibility of forming topological defects, or of undergoing a phase of inflation at low energy.

Let us mention that we have not addressed the effect of the large, time-dependent, field inhomogeneities on the space-time metric. In particular they act as a source of gravitational radiation [40]. This may be one of the only observational signatures of this phase. We refer to Ref. [41] for a general discussion.

A numerical code in free access [42], `LATTICEEASY`, simulates the evolution of the interacting scalar fields in an expanding space-time and can be used to study these phases of reheating.

9.6 Eternal inflation

The previous sections have detailed the observational signatures, mostly sensitive to the dynamics of inflation during the e-folds when the observable modes become super-Hubble. However, this zone of e-folds only corresponds to a small region of the dynamics of inflation.

The discovery of the self-reproduction process, that we describe here, is a major development for inflation and cosmology. This mechanism was known for the models of old and new inflation and was then extended to models of chaotic inflation [43].

9.6.1 Heuristic argument

Mechanism

In inflationary models with large values of the inflaton (class A), quantum fluctuations can locally increase the value of the inflaton. The expansion of these regions is then quicker and their own quantum fluctuations generate new inflationary domains. This process naturally leads to a self-reproducing Universe where there are always inflating regions.

To understand this mechanism, we give a heuristic argument. We recall that regions separated by distances greater than H^{-1} can be considered to be evolving independently. So any region of size H^{-1} will be considered as an independent Universe decoupled from other regions.

Let us consider such a region of size H^{-1} for which the scalar field is homogeneous enough and has a value $\phi \gg M_{\rm Pl}$ and let us consider a potential of the form (9.2.46). In a time interval $\Delta t \sim H^{-1}$, the field classically decreases by $\Delta\phi \sim \dot\phi \Delta t \sim \dot\phi/H$. The Klein–Gordon equation in the slow-roll regime then implies that $\Delta\phi \sim -M_{\rm Pl}^2/(4\pi\phi)$. This value should be compared with the typical amplitude of quantum fluctuations, $|\delta\phi| \sim H/(2\pi) \sim m\phi/(\sqrt{3\pi}M_{\rm Pl})$.

Classical and quantum fluctuations have the same amplitude for

$$|\Delta\phi| \sim |\delta\phi| \iff \phi \sim \phi_* \equiv \frac{M_{\rm Pl}}{2}\sqrt{\frac{M_{\rm Pl}}{m}}. \tag{9.6.1}$$

We can thus distinguish between three phases in the evolution of the inflaton (Fig. 9.17): a phase during which the quantum fluctuations are of the same order (or larger) as the classical field variation, a phase in which the field is in classical slow-roll towards its minimum and a phase when the field oscillates around its minimum. Note that $V(\phi_*) = (m/M_{\rm Pl})M_{\rm Pl}^4/8 \lesssim 10^{-6}M_{\rm Pl}^4$ so that the first regime can happen at an energy scale for which the classical space-time description is legitimate.

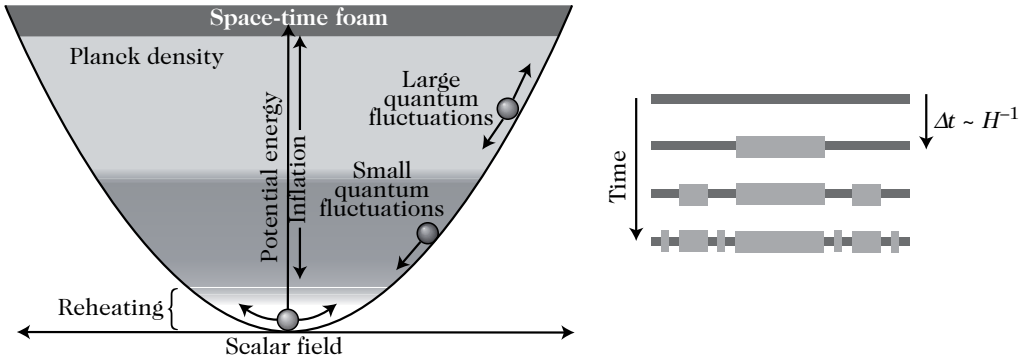


Figure 9.17: (left): The three phases of evolution of the inflaton depend on its value compared to ϕ_*. For $\phi > \phi_*$, the quantum fluctuations dominate the classical drift. (right): Evolution of an inflating domain. At each time interval $\Delta t \sim H^{-1}$, a domain of size H^{-1} gives rise to $\sim e^3 \sim 20$ domains of size H^{-1}, out of which a fraction $\sim 1/2$ has a field value $\phi > \phi_*$. The volume of the inflating domain thus grows by a factor of 10 during Δt. This process goes on indefinitely and produces a fractal structure.

In the regime where $\phi \gg \phi_*$, $\delta\phi \gg \Delta\phi$ and the probability that the field increases by $\delta\phi$ during $\Delta t = H^{-1}$ is of the order of 1/2. If the field increases, the expansion rate $H(\phi)$ also increases and the physical volume of this region grows as $\exp(3H\Delta t) = e^3 \approx 20$. So, after a time Δt, the region of initial volume H^{-1} contains ~ 20 regions of size H^{-1}, among which about 10 have $\phi > \phi_*$. The physical volume occupied by the field in the regime $\phi > \phi_*$ therefore evolves as

$$\mathcal{V}_{t+\Delta t}(\phi > \phi_*) \sim \frac{1}{2}\left(e^{H\Delta t}\right)^3 \mathcal{V}_t(\phi > \phi_*) \sim 10\,\mathcal{V}_t(\phi > \phi_*). \tag{9.6.2}$$

The physical volume where the field is in the self-reproduction regime grows exponentially and without limit. This means that inflation never stops globally.

Consequences for cosmology

This scenario has important consequences:

1. **Global structure of the Universe:** The global structure of the Universe is very inhomogeneous and has a fractal structure. The Universe is a collection of an infinite number of 'island Universes' (or 'pocket Universes') each having a scale much larger than the current Hubble radius. Each island Universe can experience different numbers of e-folds and thus possess different local parameters.

2. **The multiverse and anthropic reasoning:** In models where the fundamental theory possesses multiple vacuum states (such as the string theory landscape), quantum diffusion can populate all possible vacua. The physical laws in different pocket universes will correspond to different low-energy effective field theories. Our observable Universe is then merely one hospitable domain among an infinity of diverse cosmological regions.

3. **Loss of initial condition sensitivity:** Since self-reproduction continues indefinitely into the future, the stationary volume distribution of inflating domains becomes entirely decoupled from whatever initial state preceded inflation.

9.6.2 Stochastic approach

To give a rigorous mathematical description of this diffusion process, Starobinsky developed the **stochastic inflation formalism** [44].

Field decomposition

In the stochastic approach, we first introduce a smoothing function $W_R(\mathbf{x})$ where R is a comoving scale and the operator ϕ is then decomposed into two parts, $\bar{\phi}$ containing the long-wavelength modes and $\delta\phi$ containing the short-wavelength modes. Since the physical smoothing scale is of the order of H^{-1}, we set

$$R = (\epsilon a H)^{-1}, \tag{9.6.3}$$

where ϵ is a constant numerical coefficient ($\epsilon \ll 1$). If W_R is spherically symmetric and $w(kR)$ its Fourier transform, then

$$\phi(\mathbf{x}, t) = \bar{\phi}(\mathbf{x}, t) + \delta\phi(\mathbf{x}, t), \tag{9.6.4}$$

with

$$\bar{\phi}(\mathbf{x}, t) = \int \frac{\mathrm{d}^3 k}{(2\pi)^{3/2}} w(kR) \left[\phi_k(t)\hat{a}_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} + \phi_k^*(t)\hat{a}_{\mathbf{k}}^\dagger e^{-i\mathbf{k}\cdot\mathbf{x}}\right], \tag{9.6.5}$$

so that the Fourier components of the smoothed field are $\bar{\phi}_k = \phi_k w(kR)$ and $\delta\phi_k = [1 - w(kR)]\phi_k$.

The Langevin equation

In the slow-roll regime, the field $\bar{\phi}$ satisfies the Klein–Gordon equation

$$\dot{\bar{\phi}}(\mathbf{x}, t) = -\frac{1}{3H} V'(\bar{\phi}) + \xi_Q(\mathbf{x}, t), \tag{9.6.6}$$

obtained by inserting the decomposition (9.6.4) into the Klein–Gordon equation. ξ_Q is a quantum noise describing the effect of the short wavelengths on the evolution of the smoothed

field. The Klein–Gordon equation tells us that $\xi_Q = -\delta\dot{\phi} - V''(\bar{\phi})\delta\phi/(3H)$. Since $\delta\phi$ contains only short-wavelength modes, it is a quantum field. The quantum noise is thus given by

$$\xi_Q(\mathbf{x},t) = -\int \frac{\mathrm{d}^3k}{(2\pi)^{3/2}}(kHR)w'(kR)\left[\phi_k(t)\hat{a}_{\mathbf{k}}e^{i\mathbf{k}\cdot\mathbf{x}} + \text{h.c.}\right], \tag{9.6.7}$$

where we have used $\delta\dot{\phi}_k = -\dot{\phi}_k(1-w) + \dot{w}\phi_k \simeq \dot{w}\phi_k$ when $\epsilon \ll 1$. We have $\dot{w} = -kHRw'(kR)$ where the prime denotes a derivative with respect to the argument. The two-point correlation function of the noise operator is

$$\langle\xi_Q(\mathbf{x},t)\xi_Q(\mathbf{x}',t')\rangle = \int \frac{\mathrm{d}k}{2\pi^2}\frac{\sin kr}{r}[HR(t)]w'[kR(t)][HR(t')]w'[kR(t')]\phi_k(t)\phi_k^*(t'), \tag{9.6.8}$$

after integration over angles.

Although ϕ is a quantum operator, we can introduce a classical stochastic field, φ, the dynamics of which is described by the Langevin equation where the noise ξ_Q is replaced by a stochastic noise ξ with a correlation function of the form (9.6.8). The evolution of the inflaton, smoothed out on super-Hubble scales, is then dictated by

$$\dot{\varphi}(\mathbf{x},t) = -\frac{V'(\varphi)}{3H(\varphi)} + \xi(\mathbf{x},t), \tag{9.6.9}$$

$$\langle\xi(\mathbf{x},t)\xi(\mathbf{x}',t')\rangle = \langle\xi_Q(\mathbf{x},t)\xi_Q(\mathbf{x}',t')\rangle. \tag{9.6.10}$$

Evaluation of the noise

The general study of the Langevin equation requires us to know the function $H(\varphi)$ and to determine the properties of the noise, which themselves depend on the dynamics of inflation and on the choice of the function w. For example, in de Sitter space, the mode functions are $\phi_k(\eta) = \frac{H}{\sqrt{2k^3}}(1+ik\eta)e^{-ik\eta}$ and the correlation function (9.6.8) becomes

$$\langle\xi_Q(\mathbf{x},t)\xi_Q(\mathbf{x}',t')\rangle = \frac{H^2}{4\pi^2}\frac{\sin kr}{r}\int\frac{\mathrm{d}k}{k}w'\left(-\frac{k\eta}{\epsilon}\right)w'\left(-\frac{k\eta'}{\epsilon}\right)(1+ik\eta)(1-ik\eta')e^{-ik(\eta-\eta')}. \tag{9.6.11}$$

Note that when the window W_R is a Heaviside function then $w'(kR) = \delta(kR-1)$ so that the correlation function becomes proportional to $\delta[R(t)-R(t')] \propto \delta(t-t')$. For instance, in the de Sitter case, one can check that

$$\langle\xi(\mathbf{x},t)\xi(\mathbf{x}',t')\rangle = \frac{H^3}{4\pi^2}\frac{\sin(\epsilon aHr)}{\epsilon aHr}\delta(t-t'). \tag{9.6.12}$$

Restricting ourselves to this hypothesis for the window function, one can show in a general way that the Langevin equation takes the form

$$\dot{\varphi}(\mathbf{x},t) = -\frac{V'(\varphi)}{3H(\varphi)} + \frac{H^{3/2}(\varphi)}{2\pi}\xi(\mathbf{x},t), \qquad \langle\xi(\mathbf{x},t)\xi(\mathbf{x}',t')\rangle = \delta(t-t'). \tag{9.6.13}$$

This equation describes the field evolution as the combination of a random walk induced by non-correlated noise and a classical drift under the effect of the force $F(\varphi) = -V'(\varphi)/[3H(\varphi)]$. So during an interval Δt, $\langle(\Delta\varphi - F(\varphi)\Delta t)^2\rangle = \sigma^2(\varphi)\Delta t$ with $\sigma^2 = H^3/(4\pi^2)$.

Probability distribution of the inflaton

The stochastic nature of the Langevin equation (9.6.13) implies that we should establish an evolution equation for the probability distribution function, $P(\varphi,t) \equiv P[\varphi(\mathbf{x},t) = \varphi]$. The value

of this function at the time $t + \Delta t$ is related to its value at the time t through the Green's function, $G(\varphi, t|\varphi', t')$ as

$$P(\varphi, t + \Delta t) = \int G(\varphi, t + \Delta t|\varphi', t) P(\varphi', t)\, \mathrm{d}\varphi'. \tag{9.6.14}$$

The Green's function is the transition probability from state φ' at time t to φ at $t + \Delta t$. In the limit $\Delta t \to 0$, expanding G yields the **Fokker–Planck equation**

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial \varphi}\left[F(\varphi)P\right] + \frac{1}{2}\sigma^{2(1-\beta)}\frac{\partial}{\partial \varphi}\left[\sigma^{2\beta}P\right], \tag{9.6.15}$$

where β parametrizes the stochastic calculus prescription (e.g. $\beta = 0$ for Itô, $\beta = 1/2$ for Stratonovich, and $\beta = 1$ for kinetic prescription). In our cosmological context, taking $\beta = 1$, this equation takes the form

$$\frac{\partial P(\varphi, t)}{\partial t} = \frac{\partial}{\partial \varphi}\left[\frac{V'(\varphi)}{3H(\varphi)}P(\varphi, t)\right] + \frac{1}{8\pi^2}\frac{\partial^2}{\partial \varphi^2}\left[H^3(\varphi)P(\varphi, t)\right]. \tag{9.6.16}$$

In the form (9.6.15), the Fokker–Planck equation appears as a continuity equation that follows from the conservation of probability.

The mean value of any function of the quantum operator $\bar{\phi}$, $\langle f(\bar{\phi})\rangle$, can then be obtained as

$$\langle f(\bar{\phi})\rangle = \int f(\varphi)P(\varphi, t)\, \mathrm{d}\varphi, \tag{9.6.17}$$

where the mean value is now computed from the stochastic field φ.

Discussion For a massless field in de Sitter spacetime ($V = 0$, $H = \text{const}$), the Fokker–Planck equation reduces to pure diffusion, $\partial P/\partial t = \frac{H^3}{8\pi^2}\frac{\partial^2 P}{\partial \varphi^2}$. Starting from $P(\varphi, 0) = \delta(\varphi)$, the solution is a broadening Gaussian $P(\varphi, t) = \frac{1}{\sqrt{2\pi\sigma_t^2}}\exp\left[-\frac{\varphi^2}{2\sigma_t^2}\right]$ with variance $\langle\varphi^2\rangle = \frac{H^3}{4\pi^2}t$. For a massive field with potential $V = \frac{1}{2}m^2\varphi^2$, the competition between classical drift and diffusion establishes a stationary equilibrium distribution $\partial P/\partial t = 0$:

$$P_{\mathrm{eq}}(\varphi) = \mathcal{N}\exp\left[-\frac{8\pi^2 V(\varphi)}{3H^4}\right], \qquad \langle\varphi^2\rangle_{\mathrm{eq}} = \frac{3H^4}{8\pi^2 m^2}. \tag{9.6.18}$$

For a self-interacting field $V = \frac{\lambda}{4}\phi^4$, the equilibrium variance is $\langle\varphi^2\rangle_{\mathrm{eq}} = \sqrt{\frac{3}{2\lambda}}\frac{H^2}{2\pi}\frac{\Gamma(3/4)}{\Gamma(1/4)}$.

When considering volume weighting (since regions with larger $H(\varphi)$ expand faster), the physical probability distribution $P_{\mathrm{phys}}(\varphi, t|\chi)$ satisfies the modified equation

$$\frac{\partial P(\varphi, t|\chi)}{\partial t} = \frac{H^3(\chi)}{8\pi^2}\frac{\partial^2 P(\varphi, t|\chi)}{\partial \chi^2} - \frac{V'(\chi)}{3H(\chi)}\frac{\partial P(\varphi, t|\chi)}{\partial \chi}. \tag{9.6.19}$$

The stationary solution of this equation is

$$P_{\mathrm{eq}}(\varphi, \chi) \propto \exp\left[\frac{3M_{\mathrm{Pl}}^4}{8V(\chi)}\right]. \tag{9.6.20}$$

Unfortunately, if $V = 0$ at its minimum, this distribution function is not normalizable. This is due to the fact that a stationary state exists only if the stochastic diffusion, which increases φ, can compensate the classical drift towards the minimum of the potential.

In general, $P_{\mathrm{phys}}(\varphi, t|\chi) \sim \exp(cM_{\mathrm{Pl}}t)A(\varphi)B(\chi)$. If $V \ll M_{\mathrm{Pl}}^4$, we recover the result that $B(\chi) \propto \exp[-3M_{\mathrm{Pl}}^4/8V(\chi)]$ and the normalized probability to find a given volume in a state with a given value of φ is $\tilde{P}_{\mathrm{phys}}(\varphi, t|\chi) \sim A(\varphi)B(\chi)$.

This result has an important consequence. In the heuristic argument, we had only one region of size H^{-1} with $\phi \gg \phi_*$. However, if we start with a large number of regions, even with $\phi < \phi_*$, since the probability (9.6.20) is non-zero, some of them can enter a phase of eternal inflation.

This also implies that any hypothesis on the initial state of the Universe becomes completely disconnected from observations. If a region of the Universe enters a self-reproducing phase, it produces an infinite number of island Universes with properties approaching that of a stationary state, independently of the initial conditions. We stress the difficulty encountered in clearly defining probabilities and the notion of measure. Nonetheless, the probability for inflation to start becomes completely irrelevant as soon as it does not completely vanish.

Moreover, it seems that the Universe is not eternal in the past and that it still has an initial singularity [49], in the sense that space-time is geodesically incomplete in the past (a space is geodesically complete as soon as all the geodesics of this space can be continued indefinitely). This conclusion relies on the hypothesis that the weak energy condition is satisfied; this condition can, however, be violated by quantum fluctuations.

9.7 Extensions of the inflationary paradigm

The previous framework is the minimal framework for inflation. We now present some extensions to show the richness of the inflationary paradigm. These extensions allow us to address questions on the robustness of the predictions that we have established.

9.7.1 Multifield inflation

In single-field inflationary models, the inflaton controls both the expansion of the Universe and the generation of density perturbations and gravitational waves. This explains the existence of a consistency relation.

High-energy theories (Chapter 10) contain legions of scalar fields. If some of these fields are light, they will develop super-Hubble correlations. In the framework of multifield inflation one can extend the space of possibilities for inflation, simply through straightforward 'division of labour' between the inflaton and auxiliary fields. In particular, additional fields make it possible to conceive the generation of isocurvature perturbations, to extend the consistency relation, to modify the spectral index and to generate non-Gaussianities.

Friedmann and Klein–Gordon equations

To start with, let us look at how the field equations are modified when there are many scalar fields. Friedmann's equations become

$$H^2 = \frac{4\pi G_{\rm N}}{3}\left[\sum_{i=1}^{N}\dot{\phi}_i^2 + 2V(\phi_1,\ldots,\phi_N)\right], \tag{9.7.1}$$

$$\dot{H} = -4\pi G_{\rm N}\sum_{i=1}^{N}\dot{\phi}_i^2, \tag{9.7.2}$$

and the Klein–Gordon equation for each field takes the form

$$\ddot{\phi}_i + 3H\dot{\phi}_i = -\frac{\partial V}{\partial \phi_i}. \tag{9.7.3}$$

These N equations are coupled if V is not the sum of N decoupled potentials, i.e. of the form $V(\phi_1,\ldots,\phi_n) = \sum_{i=1}^{n} V_i(\phi_i)$. Note, however, that even in the absence of explicit interactions in the scalar field Lagrangian, the fields are still gravitationally coupled and, in particular, the Hubble damping that enters the individual Klein–Gordon equations involves the sum over all the fields.

Let us consider, for simplicity, the case of two fields (ϕ_1, ϕ_2). Both fields can be decomposed into a part tangent to the trajectory in field space, σ, and a part perpendicular, s,

$$\begin{pmatrix} \dot{\sigma} \\ \dot{s} \end{pmatrix} = M(\theta) \begin{pmatrix} \dot{\phi}_1 \\ \dot{\phi}_2 \end{pmatrix}, \qquad M(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \tag{9.7.4}$$

where the angle θ is defined by

$$\cos\theta = \frac{\dot{\phi}_1}{\sqrt{\dot{\phi}_1^2 + \dot{\phi}_2^2}}, \qquad \sin\theta = \frac{\dot{\phi}_2}{\sqrt{\dot{\phi}_1^2 + \dot{\phi}_2^2}}. \tag{9.7.5}$$

Relation (9.7.4) implies that $\dot{s} = 0$ by construction ($s = \text{const}$ along the background trajectory) and the Klein–Gordon equation reduces to

$$\ddot{\sigma} + 3H\dot{\sigma} + U_{,\sigma} = 0, \qquad \begin{pmatrix} U_{,\sigma} \\ U_{,s} \end{pmatrix} = M(\theta) \begin{pmatrix} V_{,1} \\ V_{,2} \end{pmatrix}. \tag{9.7.6}$$

The Friedmann equation takes the form

$$H^2 = \frac{4\pi G_{\rm N}}{3} \left[\dot{\sigma}^2 + 2V(\sigma, s = \text{const.}) \right], \tag{9.7.7}$$

since the value of s is a constant, by construction, along the trajectory. The fields σ and s are the adiabatic and isocurvature components of (ϕ_1, ϕ_2).

Examples

As an example of the dynamics of multifield inflation, let us mention **assisted inflation** [50] in which one considers n scalar fields with steep exponential potentials

$$V = \sum_i U_i \exp\left(-\sqrt{\frac{8\pi}{G_{\rm N}}} \frac{\lambda_i \phi_i}{M_{\rm Pl}} \right). \tag{9.7.8}$$

Each scalar field potential is too steep to drive inflation on its own if $\lambda_i^2 > 2$, but the additional damping effect due to the collective presence of the scalar fields in the Friedmann equation leads to a particular power-law solution, $a \propto t^p$ with $p = 2/\lambda^2$ where the combined fields have an effective potential $V \propto \exp\left(-\sqrt{8\pi}\lambda\sigma/M_{\rm Pl}\right)$ with $\lambda^{-2} = \sum_i \lambda_i^{-2}$. Thus $\lambda \to 0$ when the number of fields is increasing ($n \to \infty$) and we can have slow-roll inflation even though each individual $\lambda_i^2 > 2$.

Another topical example is **N-flation** [51] in which the effective potential is of the form

$$V = \frac{1}{2} \sum_i m_i^2 \phi_i^2. \tag{9.7.9}$$

The case $n = 1$ is indeed our standard chaotic inflationary model. But to get inflation, we have seen that the initial value of the scalar field should be several Planck masses. With many fields, the collective dynamics can yield inflation even for sub-Planckian initial values of each individual field.

Perturbation equations

Using the definitions of metric perturbations in the longitudinal gauge, the perturbed Einstein equations take the form

$$\dot{\Phi} + H\Phi = 4\pi G_{\rm N}(\dot{\phi}_1\chi_1 + \dot{\phi}_2\chi_2), \tag{9.7.10}$$

$$3\dot{\Phi} + (3H^2 + \dot{H})\Phi - \frac{\nabla^2}{a^2}\Phi = -4\pi G_{\rm N}(V_{,1}\chi_1 + \dot{\phi}_1\dot{\chi}_1 + V_{,2}\chi_2 + \dot{\phi}_2\dot{\chi}_2), \tag{9.7.11}$$

$$\ddot{\chi}_i + 3H\dot{\chi}_i - \frac{\nabla^2}{a^2}\chi_i + V_{,ij}\chi_j = -2V_{,i}\Phi + 4\dot{\phi}_i\dot{\Phi}, \tag{9.7.12}$$

where $\chi_i \equiv \delta\phi_i$ and we have used $\Psi = \Phi$ in the absence of anisotropic stress. By construction, δs is gauge invariant so that $\chi_s = Q_s = \delta s$, while $\dot{Q}_\sigma = \chi_\sigma + \dot{\sigma}\Phi/H$. Using the rotation matrix properties

$$^t\dot{M} = -\dot{\theta}\,{}^tMJ, \qquad {}^t\ddot{M} = -\ddot{\theta}\,{}^tMJ - \dot{\theta}^2\,{}^tM, \qquad J \equiv \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{9.7.13}$$

the Klein–Gordon equation (9.7.12) transforms to

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\begin{pmatrix}\chi_\sigma \\ \delta s\end{pmatrix} + 3H\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\chi_\sigma \\ \delta s\end{pmatrix} + \left[-\frac{\nabla^2}{a^2} - \dot{\theta}^2 + MV_{,ij}\,{}^tM\right]\begin{pmatrix}\chi_\sigma \\ \delta s\end{pmatrix} = 2\dot{\theta}J\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\chi_\sigma \\ \delta s\end{pmatrix} + (\ddot{\theta} + 3H\dot{\theta})J\begin{pmatrix}\chi_\sigma \\ \delta s\end{pmatrix} - 2\Phi\begin{pmatrix}U_{,\sigma} \\ U_{,s}\end{pmatrix} + 4\dot{\Phi}\begin{pmatrix}\dot{\sigma} \\ 0\end{pmatrix}, \tag{9.7.14}$$

and the Poisson equation becomes

$$-\frac{\nabla^2}{4\pi G_{\rm N}a^2}\Phi = 2\dot{\theta}\dot{\sigma}\delta s + \ddot{\sigma}\chi_\sigma - \dot{\sigma}(\dot{\chi}_\sigma - \dot{\sigma}\Phi), \tag{9.7.15}$$

with $\dot{\theta} = -U_{,s}/\dot{\sigma}$. Since the potential U satisfies

$$\begin{pmatrix}U_{,\sigma\sigma} & U_{,\sigma s} \\ U_{,s\sigma} & U_{,ss}\end{pmatrix} \equiv MV_{,ij}\,{}^tM, \tag{9.7.16}$$

we infer the coupled system for the perturbations:

$$\ddot{\chi}_\sigma + 3H\dot{\chi}_\sigma + \left(-\frac{\nabla^2}{a^2} - \dot{\theta}^2 + U_{,\sigma\sigma}\right)\chi_\sigma = -2\Phi U_{,\sigma} + 4\dot{\sigma}\dot{\Phi} + 2\frac{\mathrm{d}}{\mathrm{d}t}(\dot{\theta}\delta s) - 2\frac{\dot{\theta}}{\dot{\sigma}}U_{,\sigma}\delta s, \tag{9.7.17}$$

$$\ddot{\delta s} + 3H\dot{\delta s} + \left(-\frac{\nabla^2}{a^2} - \dot{\theta}^2 + U_{,ss}\right)\delta s = -\frac{\dot{\theta}}{\dot{\sigma}}\frac{\nabla^2}{2\pi G_{\rm N}a^2}\Phi. \tag{9.7.18}$$

Equation (9.7.17) translates into an evolution equation for the gauge-invariant Sasaki–Mukhanov variable Q_σ:

$$\ddot{Q}_\sigma + 3H\dot{Q}_\sigma - \left[\frac{\nabla^2}{a^2} + \dot{\theta}^2 - U_{,\sigma\sigma} + \frac{8\pi G_{\rm N}}{a^3}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{a^3\dot{\sigma}^2}{H}\right)\right]Q_\sigma = 2\frac{\mathrm{d}}{\mathrm{d}t}(\dot{\theta}\delta s) - 2\left(\frac{U_{,\sigma}}{\dot{\sigma}} + \frac{\dot{H}}{H}\right)\dot{\theta}\delta s. \tag{9.7.19}$$

Adiabatic and isocurvature modes

Defining the curvature perturbation as in (9.4.34), we find that

$$\mathcal{R} = \mathcal{C} + \frac{H}{\rho + P}(\dot{\phi}_1\delta\phi_1 + \dot{\phi}_2\delta\phi_2) = \mathcal{C} + \frac{H}{\dot{\sigma}}\delta\sigma = \frac{H}{\dot{\sigma}}Q_\sigma. \tag{9.7.20}$$

This equation can also take the form

$$\mathcal{R} = \cos^2\theta\,\mathcal{R}_{\phi_1} + \sin^2\theta\,\mathcal{R}_{\phi_2}. \tag{9.7.21}$$

So σ plays the role of the inflaton and its perturbations generate curvature perturbations. The fluctuations of s do not affect $\mathcal{R}$ on spatially flat hypersurfaces; s is an **isocurvature mode**.

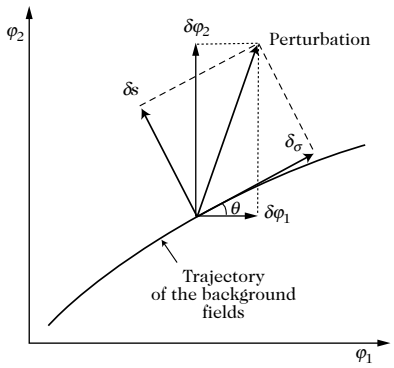


Figure 9.18: Illustration of the scalar field decomposition into an adiabatic perturbation, tangent to the trajectory, and an isocurvature component.

If we define the entropy perturbation as $\mathcal{S} \equiv H\delta s/\dot{\sigma}$, then for general modes, the perturbation theory developed in Chapter 5 teaches us that

$$\dot{\mathcal{R}} = -\frac{H}{\dot{H}}\frac{\nabla^2}{a^2}\Phi - 3Hc_s^2\mathcal{S}. \tag{9.7.22}$$

We therefore find that

$$\dot{\mathcal{R}} = -\frac{H}{\dot{H}}\frac{\nabla^2}{a^2}\Phi + \frac{2H}{\dot{\sigma}}\dot{\theta}\delta s. \tag{9.7.23}$$

In multifield inflation, the comoving curvature perturbation $\mathcal{R}$ is defined along the instantaneous background trajectory tangent σ by

$$\mathcal{R} = \frac{H}{\dot{\sigma}}Q_\sigma = -\frac{4\pi G_\mathrm{N}H}{\dot{H}}Q_\sigma, \tag{1}$$

where we have used the background Friedmann relation $\dot{H} = -4\pi G_\mathrm{N}\dot{\sigma}^2$ (with $\dot{\sigma}^2 = \dot{\phi}_1^2+\dot{\phi}_2^2$). Differentiating $\mathcal{R}$ with respect to cosmic time t gives

$$\dot{\mathcal{R}} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{H}{\dot{\sigma}}Q_\sigma\right) = \frac{H}{\dot{\sigma}}\dot{Q}_\sigma + \left(\frac{\dot{H}}{\dot{\sigma}} - \frac{H\ddot{\sigma}}{\dot{\sigma}^2}\right)Q_\sigma = \frac{H}{\dot{\sigma}}\left[\dot{Q}_\sigma + \left(\frac{\dot{H}}{H} - \frac{\ddot{\sigma}}{\dot{\sigma}}\right)Q_\sigma\right]. \tag{2}$$

The gauge-invariant Sasaki–Mukhanov variable along the trajectory is $Q_\sigma = \chi_\sigma + \frac{\dot{\sigma}}{H}\Phi$, whose time derivative evaluates to

$$\dot{Q}_\sigma = \dot{\chi}_\sigma + \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\dot{\sigma}}{H}\Phi\right) = \dot{\chi}_\sigma + \left(\frac{\ddot{\sigma}}{H} - \frac{\dot{\sigma}\dot{H}}{H^2}\right)\Phi + \frac{\dot{\sigma}}{H}\dot{\Phi}. \tag{3}$$

To express $\dot{\Phi}$, we use the perturbed $(0i)$ Einstein constraint equation (9.7.10) projected onto the adiabatic field direction:

$$\dot{\Phi} + H\Phi = 4\pi G_\mathrm{N}\dot{\sigma}\chi_\sigma = 4\pi G_\mathrm{N}\dot{\sigma}\left(Q_\sigma - \frac{\dot{\sigma}}{H}\Phi\right) = 4\pi G_\mathrm{N}\dot{\sigma}Q_\sigma + \frac{\dot{H}}{H}\Phi, \tag{4}$$

which yields

$$\dot{\Phi} = 4\pi G_\mathrm{N}\dot{\sigma}Q_\sigma - H\Phi + \frac{\dot{H}}{H}\Phi. \tag{5}$$

Next, we isolate the time derivative $\dot{\chi}_\sigma$ from the perturbed Poisson equation (9.7.15):

$$-\frac{\nabla^2}{4\pi G_\mathrm{N}a^2}\Phi = 2\dot{\theta}\dot{\sigma}\delta s + \ddot{\sigma}\chi_\sigma - \dot{\sigma}(\dot{\chi}_\sigma - \dot{\sigma}\Phi)$$

$$\Longrightarrow \quad \dot{\chi}_\sigma = \frac{\ddot{\sigma}}{\dot{\sigma}}\chi_\sigma + \dot{\sigma}\Phi + 2\dot{\theta}\delta s + \frac{\nabla^2\Phi}{4\pi G_{\rm N} a^2 \dot{\sigma}}. \tag{6}$$

Substituting $\chi_\sigma = Q_\sigma - \frac{\dot{\sigma}}{H}\Phi$ into this equation gives

$$\dot{\chi}_\sigma = \frac{\ddot{\sigma}}{\dot{\sigma}}Q_\sigma - \frac{\ddot{\sigma}}{H}\Phi + \dot{\sigma}\Phi + 2\dot{\theta}\delta s + \frac{\nabla^2\Phi}{4\pi G_{\rm N} a^2 \dot{\sigma}}. \tag{7}$$

Inserting $\dot{\chi}_\sigma$ and $\dot{\Phi}$ into $\dot{Q}_\sigma$, all non-gradient gravitational potential terms cancel identically:

$$\begin{aligned}
\dot{Q}_\sigma &= \left(\frac{\ddot{\sigma}}{\dot{\sigma}}Q_\sigma - \frac{\ddot{\sigma}}{H}\Phi + \dot{\sigma}\Phi + 2\dot{\theta}\delta s + \frac{\nabla^2\Phi}{4\pi G_{\rm N} a^2 \dot{\sigma}}\right) \\
&\quad + \left(\frac{\ddot{\sigma}}{H} - \frac{\dot{\sigma}\dot{H}}{H^2}\right)\Phi + \frac{\dot{\sigma}}{H}\left(4\pi G_{\rm N}\dot{\sigma}Q_\sigma - H\Phi + \frac{\dot{H}}{H}\Phi\right) \\
&= \left(\frac{\ddot{\sigma}}{\dot{\sigma}} + \frac{4\pi G_{\rm N}\dot{\sigma}^2}{H}\right)Q_\sigma + 2\dot{\theta}\delta s + \frac{\nabla^2\Phi}{4\pi G_{\rm N} a^2 \dot{\sigma}} \\
&= \left(\frac{\ddot{\sigma}}{\dot{\sigma}} - \frac{\dot{H}}{H}\right)Q_\sigma + 2\dot{\theta}\delta s - \frac{\dot{\sigma}}{\dot{H}}\frac{\nabla^2}{a^2}\Phi.
\end{aligned} \tag{8}$$

Substituting this result for $\dot{Q}_\sigma$ into $\dot{\mathcal{R}}$, the Q_σ terms cancel identically:

$$\begin{aligned}
\dot{\mathcal{R}} &= \frac{H}{\dot{\sigma}}\left[\left(\frac{\ddot{\sigma}}{\dot{\sigma}} - \frac{\dot{H}}{H}\right)Q_\sigma + 2\dot{\theta}\delta s - \frac{\dot{\sigma}}{\dot{H}}\frac{\nabla^2}{a^2}\Phi + \left(\frac{\dot{H}}{H} - \frac{\ddot{\sigma}}{\dot{\sigma}}\right)Q_\sigma\right] \\
&= -\frac{H}{\dot{H}}\frac{\nabla^2}{a^2}\Phi + \frac{2H}{\dot{\sigma}}\dot{\theta}\delta s.
\end{aligned} \tag{9}$$

In conformal time, the spatial Laplacian term is given by $-\frac{H}{\dot{H}}\frac{\nabla^2}{a^2}\Phi = \frac{H}{\mathcal{H}^2 - \mathcal{H}'}\Delta\Phi$, directly matching the single-field relation (9.4.31).

This result connects directly to the general thermodynamic formulation of perturbations. Defining the intrinsic non-adiabatic entropy perturbation of the fluid (not to be confused with the relative multi-fluid entropy perturbation S_{ab} of Eq. (4.3.64)) by

$$\mathcal{S} \equiv H\left(\frac{\delta P}{\dot{P}} - \frac{\delta\rho}{\dot{\rho}}\right) = \frac{H}{\dot{P}}\delta P_{\rm nad}, \tag{10}$$

where the background pressure evolves as $\dot{P} = c_s^2\dot{\rho} = -3Hc_s^2(\rho + P) = -3Hc_s^2\dot{\sigma}^2$, and the non-adiabatic pressure along curved trajectories evaluates to $\delta P_{\rm nad} = -2\dot{\theta}\dot{\sigma}\delta s$, the entropy term yields

$$-3Hc_s^2\mathcal{S} = -3Hc_s^2\left(\frac{H}{-3Hc_s^2\dot{\sigma}^2}\right)\left(-2\dot{\theta}\dot{\sigma}\delta s\right) = \frac{2H}{\dot{\sigma}}\dot{\theta}\delta s. \tag{11}$$

This rigorously establishes the full evolution equation (9.7.22)–(9.7.23):

$$\boxed{\dot{\mathcal{R}} = -\frac{H}{\dot{H}}\frac{\nabla^2}{a^2}\Phi - 3Hc_s^2\mathcal{S} = -\frac{H}{\dot{H}}\frac{\nabla^2}{a^2}\Phi + \frac{2H}{\dot{\sigma}}\dot{\theta}\delta s}. \tag{12}$$

On super-Hubble scales ($k/aH \ll 1$), the gradient term $-\frac{H}{\dot{H}}\frac{\nabla^2}{a^2}\Phi \to 0$, showing that the curvature perturbation is sourced on super-Hubble scales entirely by the isocurvature mode δs whenever the background trajectory turns in field space ($\dot{\theta} \neq 0$).

As long as $\dot{\theta} = 0$, the two modes of perturbations are decoupled: δs behaves as a test scalar field evolving in an unperturbed space-time and the metric perturbations are only coupled to σ. If $\dot{\theta} \neq 0$, the isocurvature and adiabatic modes can mix and $\mathcal{R}$ is sourced by δs on super-Hubble scales.

As seen earlier, a scalar field only develops super-Hubble perturbations if it is light. Equation (9.7.18) tells us that the effective mass of the field s is $m_s^2 = U_{,ss} + 3\dot{\theta}^2$, so that super-Hubble entropy perturbations will only be produced if

$$U_{,ss} + 3\dot{\theta}^2 < \frac{3}{2}H^2. \tag{9.7.24}$$

Assuming that both fields are light and that initially $\dot{\theta} = 0$, which is the case during the slow-roll regime, the fields will be decoupled and each of them will generate super-Hubble fluctuations, $Q_i|_{k=aH} = (H_k/2\pi)e_i$. During the later evolution of the super-Hubble modes, (9.7.17) has a source term proportional to δs if $\dot{\theta} \neq 0$. So, during this evolution, metric fluctuations develop a component proportional to e_s. We infer that the curvature and entropy perturbations can be correlated,

$$\langle \mathcal{R}\mathcal{S} \rangle \neq 0, \tag{9.7.25}$$

or in an equivalent way, $\langle Q_\sigma \delta s \rangle \neq 0$.

Extension to N fields and consistency relation

The transfer matrix formalism enables tracking the super-Hubble evolution from Hubble exit ($*$) to the end of inflation:

$$\begin{pmatrix} \mathcal{R} \\ \mathcal{S} \end{pmatrix} = \begin{pmatrix} 1 & T_{\mathcal{R}\mathcal{S}} \\ 0 & T_{\mathcal{S}\mathcal{S}} \end{pmatrix} \begin{pmatrix} \mathcal{R}_* \\ \mathcal{S}_* \end{pmatrix}. \tag{9.7.26}$$

In the transfer matrix, $T_{\mathcal{R}\mathcal{R}} = 1$ since the curvature perturbation is conserved for all strictly adiabatic perturbations and $T_{\mathcal{S}\mathcal{R}} = 0$ as adiabatic perturbations cannot generate any isocurvature perturbations.

Defining the correlation coefficient

$$r_c \equiv \frac{A_{\zeta\mathcal{S}}^2}{\sqrt{A_\zeta^2 A_{\mathcal{S}}^2}}, \tag{9.7.27}$$

with $A_{\zeta\mathcal{S}}^2$ defined by $A_{\zeta\mathcal{S}}^2 = \frac{2}{5}\langle \zeta\mathcal{S} \rangle$, the tensor-to-scalar consistency relation becomes

$$r_\zeta = -\frac{n_T}{2}(1 - r_c^2). \tag{9.7.28}$$

When $r_c = 0$, the modes are completely decoupled and we recover the usual relation. Note that this implies that $r_\zeta \leq -n_T/2$, with equality reached only in the case of single-field inflation.

The spectator field and the curvaton mechanism

In the **curvaton mechanism** [54], the inflaton ϕ is solely responsible for driving inflation while an auxiliary light scalar field χ (the curvaton, with $m_\chi \ll H$) is completely subdominant during inflation. The curvaton acquires scale-invariant quantum fluctuations $\delta\chi \approx H_*/(2\pi)$. After inflation ends and the inflaton decays into radiation, the Hubble parameter drops until $H \sim m_\chi$, at which point the curvaton begins oscillating and behaves as pressureless matter ($\rho_\chi \propto a^{-3}$). As radiation dilutes faster ($\rho_r \propto a^{-4}$), the curvaton energy fraction $r \equiv \rho_\chi/\rho_r$ grows. When the curvaton decays into radiation, it generates the primordial curvature perturbation:

$$\zeta_{\rm BST} = \frac{r}{4 + 3r}\delta_\chi, \qquad r \equiv \frac{\rho_\chi}{\rho_r}. \tag{9.7.29}$$

Depending on whether χ decays before or after dominating the energy content of the Universe, $\zeta = \delta_\chi/3$ or $\zeta = r\delta_\chi/4$.

Modulated reheating

In **modulated reheating** [58], the decay width of the inflaton $\Gamma_\phi = \Gamma_\phi(\chi)$ depends on the local expectation value of a light modulus/spectator field χ. Quantum fluctuations $\delta\chi$ generated during inflation induce spatial variations $\delta\Gamma_\phi$, causing reheating to occur at different times in different spatial regions. This spatial modulation converts isocurvature fluctuations of χ directly into adiabatic curvature perturbations ζ at the reheating surface:

$$\mathcal{R} = \frac{\epsilon}{1 + 4\pi\gamma^2}, \tag{9.7.30}$$

where γ parametrizes the coupling. Again we find that $\mathcal{R} < \epsilon$.

9.7.2 Primordial non-Gaussianity

Since the primordial curvature perturbations are related to the quantum fluctuations of a scalar field that is initially in its vacuum state, they have Gaussian statistics. As long as the evolution of the perturbations is linear, no non-Gaussianity can be generated. However, since general relativity is non-linear, we expect some deviations from Gaussianity. In particular, such deviations can be large in multifield inflation.

Single-field inflation

In single-field inflation, expanding the perturbation equation to second order yields:

$$\ddot{\delta\phi} + 3H\dot{\delta\phi} - \frac{\nabla^2}{a^2}\delta\phi = -V''\delta\phi - 3V'''(\delta\phi)^2 + \dots \tag{9.7.31}$$

In order for the field to generate non-Gaussianity during the inflationary phase, non-linear terms should be non-negligible. If the quadratic term dominates the expansion of the r.h.s., then $\delta\phi > V''/V'''$. As seen earlier, $\delta\phi \sim H$. While the wavelength band $[\lambda, \lambda + \Delta\lambda]$ associated with observable scales has become super-Hubble, the slow-roll parameter η_V has varied by $\Delta\eta_V \sim (8\pi G_\mathrm{N}\xi_V^2 - \eta_V)\sqrt{\epsilon_V}\Delta\phi/M_\mathrm{Pl}$. We infer that during a number of e-folds N, η_V varies by $10^5\eta_V N$, using for these modes $H/\sqrt{\epsilon} \sim 10^{-5}M_\mathrm{Pl}$.

So if the quadratic term dominates, the slow-roll regime ends rapidly. This is because the slow-roll regime imposes the potential to be very flat whereas it must have a significant curvature for the non-linear terms to dominate. This explains why initial perturbations remain Gaussian during the inflationary phase.

To second order in the perturbations, non-linearities can induce some non-Gaussianity, but the latter must be weak. A way to parameterize this effect is to introduce a deviation that characterizes the three-point function (the bispectrum). In Fourier space,

$$\Phi_\mathbf{k} = \Phi_\mathrm{lin}(\mathbf{k}) + \frac{1}{(2\pi)^{3/2}} f_\mathrm{NL}^{(\Phi)}(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k})\Phi_\mathrm{lin}(\mathbf{k}_1)\Phi_\mathrm{lin}(\mathbf{k}_2)\delta^{(3)}(\mathbf{k} - \mathbf{k}_1 - \mathbf{k}_2), \tag{9.7.32}$$

where $f_\mathrm{NL}^{(\Phi)}$ is a function that can be computed in specific models [61] and Φ_lin is the Bardeen potential in the linear regime. With this definition,

$$\langle\Phi_{\mathbf{k}_1}(\eta)\Phi_{\mathbf{k}_2}(\eta)\Phi_{\mathbf{k}_3}(\eta)\rangle = \frac{\delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3)}{(2\pi)^{3/2}}\left[f_\mathrm{NL}^{(\Phi)}(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)P_{\Phi_\mathrm{lin}}(k_1)P_{\Phi_\mathrm{lin}}(k_2) + (1 \to 2 \to 3) + (1 \to 3 \to 2)\right]. \tag{9.7.33}$$

Several f_NL parameters are introduced in the literature, depending on the variable used: Φ, $\mathcal{R}$, or $\tilde{\mathcal{R}}$ defined by $\exp\left(2\tilde{\mathcal{R}}\right) = 1 + 2\mathcal{R}$. These parameters are related by

$$f_\mathrm{NL}^{(\Phi)} = \frac{5}{3}f_\mathrm{NL}^{(\mathcal{R})}, \qquad f_\mathrm{NL}^{(\tilde{\mathcal{R}})} = f_\mathrm{NL}^{(\mathcal{R})} + 1. \tag{9.7.34}$$

In a model of single-field inflation, the non-Gaussianity is induced by the second-order dynamics and this coefficient is typically of the order of [62]

$$f_{\rm NL}^{(\tilde{\mathcal{R}})}(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) \sim \frac{1}{4}(n_s-1) + \epsilon\, g(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3), \tag{9.7.35}$$

where $g(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3)$ is a shape function such that $g \to 0$ in the squeezed limit ($k_1 \ll k_2 \approx k_3$). As expected, $f_{\rm NL}$ is suppressed by the slow-roll parameters ($\mathcal{O}(\epsilon,\eta) \ll 1$). Analysis from CMB data establishes strict bounds:

$$-9 < f_{\rm NL}^{(\Phi)} < 111 \quad (95\%\ \text{CL}). \tag{9.7.36}$$

Extensions generating large non-Gaussianity

Various extensions of the single-field model allow for the production of significant non-Gaussianity:

- **Topological defects and χ^2-seeds:** These models introduce scalar fields with negligible background energy. Since the energy density is quadratic in the field, it induces non-Gaussianity in the perturbations [67–69].
- **Multifield inflation:** Non-linear effects can be substantial in the isocurvature direction without destabilizing the slow-roll background. When the trajectory bends ($\dot{\theta} \neq 0$), this non-Gaussianity is transferred to the adiabatic mode [61, 70].
- **Curvaton mechanism:** In the intermediate regime where $H_* \sim \chi_*$, the curvature perturbation becomes non-linear:

$$\zeta_{\rm BST} = \frac{r}{2}\left[\frac{\delta\chi}{\chi} + \left(\frac{\delta\chi}{\chi}\right)^2\right], \qquad f_{\rm NL}^{(\Phi)} = -\frac{7}{6} - \frac{5r}{6} + \frac{5}{4r}. \tag{9.7.37}$$

  For $r \ll 1$, large non-Gaussianity $f_{\rm NL} \sim 5/(4r) \gg 1$ is naturally generated.
- **Non-canonical kinetic terms (DBI and k-inflation):** Generates non-Gaussianity with an **equilateral shape** ($k_1 \approx k_2 \approx k_3$) with amplitude $f_{\rm NL}^{\rm equil} \sim -1/c_s^2$ when the sound speed $c_s \ll 1$.
- **Excited initial states (Non-Bunch–Davies vacuum):** Generates non-Gaussianity peaking in the **folded/flattened shape** ($k_1 + k_2 \approx k_3$).

9.7.3 The trans-Planckian problem

The predictions of inflation are based on the use of quantum field theory in curved space-time. Even though the metric is treated in a quantum way, we have shown that at linear order, technically everything reduces to the quantization of a free scalar field in a Friedmann–Lemaître space-time.

Formulation of the problem

If we consider an inflationary model with potential $V = \frac{\lambda_4}{4}\phi^4$, the field value when observable scales become super-Hubble is $(\phi_*/M_{\rm Pl})^2 = (N_*+1)/\pi$, giving $(H_*/M_{\rm Pl})^2 = \lambda_4(N_*+1)^2/\pi^2$ and $\epsilon_* = 1/(N_*+1)$. We infer that

$$\frac{H_*}{\sqrt{\epsilon_*}M_{\rm Pl}} = \frac{\sqrt{\lambda_4}}{\pi}(N_*+1)^{3/2}. \tag{9.7.38}$$

Observational constraints impose this quantity to be of order 10^{-5}, fixing $\lambda_4 \sim 10^{-13}$ for $N_* \sim 60$. However, the initial condition of chaotic inflation $V(\phi_i) \sim M_{\rm Pl}^4$ requires a total

duration of $N \sim 5 \times 10^8$ e-folds. We infer that the mode corresponding to the Hubble radius today, $\ell_{H_0} \sim 10^{61} \ell_{\rm Pl}$, had a wavelength at the beginning of inflation of

$$\ell_H(t_i) \sim 10^{61} e^{-N} \ell_{\rm Pl} \ll \ell_{\rm Pl}. \tag{9.7.39}$$

So the modes observable today had a sub-Planckian length at the beginning of inflation.

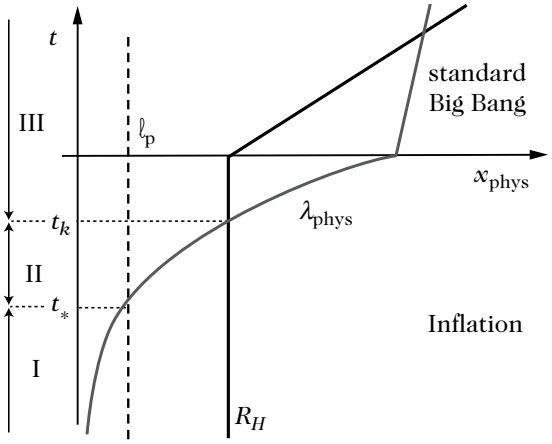


Figure 9.19: Evolution of the physical wavelength of a given mode and of the Hubble radius as a function of cosmic time. During inflation, R_H is constant and the physical wavelength grows exponentially. If inflation lasts long enough, a mode that has today a wavelength comparable to the Hubble radius was sub-Planckian at the beginning of inflation. The trans-Planckian problem addresses the question of knowing if the physics at sub-Planckian scales can modify the predictions of inflation.

As illustrated in Fig. 9.19, at the moment at which these modes become super-Planckian, the Hubble constant is well below the Planck scale, which implies that the description of space-time with a Friedmann–Lemaître metric remains justified. The trans-Planckian problem [71] therefore only involves quantum fluctuations.

Model with a modified dispersion relation

To phenomenologically parametrize ultraviolet modifications arising from quantum gravity, one modifies the dispersion relation:

$$k_{\rm phys} = \omega_{\rm phys}(k_{\rm phys}), \tag{9.7.40}$$

which, in terms of comoving quantities, amounts to

$$k^2 \to k_{\rm eff}^2(k,\eta) = a^2 F^2\left(\frac{k}{a}\right) \equiv \omega^2(k,\eta). \tag{9.7.41}$$

Figure 9.20 illustrates some dispersion relations used in the literature (Corley–Jacobson, Unruh, standard). The function F is linear as long as $k/a \ll k_C$, where k_C is a new ultraviolet cutoff scale. The Sasaki–Mukhanov mode equation becomes

$$v_k'' + \omega_T^2(k,\eta) v_k = 0, \qquad \omega_T^2(k,\eta) \equiv \omega^2(k,\eta) - \frac{z''}{z}. \tag{9.7.42}$$

To follow the standard procedure, we normalize v_k by the Wronskian condition $v_k v_k^{*\prime} - v_k' v_k^* = i$. Minimizing the energy density

$$\rho = \frac{1}{4\pi a^2} \int {\rm d}k\, k^2 \left[\left| \left(\frac{v_k}{a}\right)' \right|^2 + \omega^2 \left| \frac{v_k}{a} \right|^2 \right] \tag{9.7.43}$$

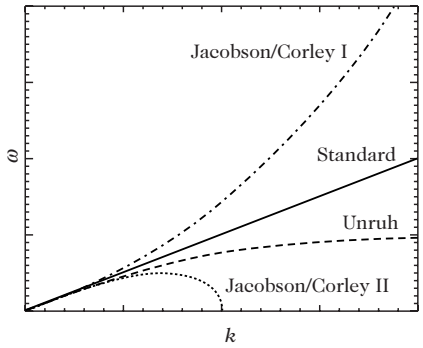


Figure 9.20: Some of the dispersion relations considered in the literature.

at an initial conformal time η_i yields the instantaneous adiabatic vacuum:

$$v_k(\eta_i) = \frac{1}{\sqrt{2\omega(k,\eta_i)}}, \qquad v_k'(\eta_i) = -i\sqrt{\frac{\omega(k,\eta_i)}{2}}. \tag{9.7.44}$$

This adiabatic vacuum corresponds to a positive-frequency WKB solution:

$$v_k(\eta) \simeq \frac{1}{\sqrt{2\omega(k,\eta)}} \exp\left[-i\int_{\eta_i}^{\eta} \omega(k,u)\,\mathrm{d}u\right]. \tag{9.7.45}$$

This solution is valid as long as the **adiabaticity condition**

$$\left|\frac{\omega'}{\omega^2}\right| = \left|\frac{H}{\omega_{\rm phys}}\left(1 - \frac{\mathrm{d}\ln\omega_{\rm phys}}{\mathrm{d}\ln k_{\rm phys}}\right)\right| \ll 1 \tag{9.7.46}$$

is satisfied.

If this adiabaticity condition is not violated between η_i and a time η_f for which the dispersion relation is linear ($k/a \ll k_C$) and the mode is sub-Hubble, then the solution for $\eta > \eta_f$ is

$$v_k(\eta) = \frac{1}{\sqrt{2k}} \exp\left[-ik\eta - i\int_{\eta_i}^{\eta_f} \omega(k,u)\,\mathrm{d}u\right]. \tag{9.7.47}$$

The accumulated phase has no effect on the absolute value of v_k or on observables.

So, in order for a trans-Planckian modification to affect the power spectrum, the dispersion relation must break the WKB adiabaticity condition. In this case, the mode at η_f develops a non-trivial Bogoliubov mixture:

$$v_k(\eta) = \frac{\alpha_k}{\sqrt{2\omega}} \exp\left[-i\int_{\eta_f}^{\eta} \omega\,\mathrm{d}u\right] + \frac{\beta_k}{\sqrt{2\omega}} \exp\left[+i\int_{\eta_f}^{\eta} \omega\,\mathrm{d}u\right], \tag{9.7.48}$$

with $|\alpha_k|^2 - |\beta_k|^2 = 1$. This excites the mode into a non-vacuum state, modulating the primordial power spectrum:

$$\mathcal{P}(k) = \mathcal{P}_0(k)\left[1 + 2\mathrm{Re}\left(\frac{\beta_k}{\alpha_k} e^{i\varphi_k}\right) + \mathcal{O}(\beta^2)\right], \tag{9.7.49}$$

where $\mathcal{P}_0$ is the standard scale-invariant spectrum and φ_k is a phase. This relation is valid if

$$\sigma \equiv \frac{k_C}{H} \gg 1. \tag{9.7.50}$$

Expanding the solution to second order in the WKB approximation, one can show that if (9.7.46) is satisfied between η_i and η_f, then $\beta_k \sim \exp(-\sigma)$. To conclude, if condition (9.7.46) is satisfied and $\sigma \gg 1$, then the standard predictions of inflation are exceptionally robust.

Minimal length and boundary state models

An alternative approach introduces a generalized uncertainty principle $[x, p] = i\hbar(1 + \beta p^2 + \dots)$, which implies the existence of a minimal physical distance $\ell_{\min} = \sqrt{\beta}\hbar$. Each Fourier mode then originates at its own initial time η_k defined by

$$\frac{2\pi}{k} a(\eta_k) = \ell_{\min}. \tag{9.7.51}$$

The most general initial state at η_k is parametrized as

$$v_k(\eta_k) = \frac{c_k + d_k}{\sqrt{2\omega(k, \eta_k)}}, \qquad v_k'(\eta_k) = -i(c_k - d_k)\sqrt{\frac{\omega(k, \eta_k)}{2}}, \tag{9.7.52}$$

with $|c_k|^2 - |d_k|^2 = 1$. Expanding $c_k = 1 + y\sigma + \dots$ and $d_k = x\sigma + \dots$ where $\sigma = H\ell_{\min}$, this induces oscillations in the primordial power spectrum:

$$\begin{aligned}\mathcal{P}_\zeta(k) = \mathcal{P}_\zeta(\sigma = 0)\Bigg[1 - 2|x|\sigma \cos\left(\frac{2}{\sigma}(1 + \epsilon + \epsilon \ln k\ell_{\min}) + \psi\right) \\ - \frac{\pi\epsilon M_{\rm Pl}^2}{H^2}|x|\sigma(2\epsilon - \delta)\sin\left(\frac{2}{\sigma}(1 + \epsilon + \epsilon \ln k\ell_{\min}) + \psi\right)\Bigg],\end{aligned} \tag{9.7.53}$$

where ψ is a constant phase. The logarithmic wavelength of the oscillations is $\Delta \ln k = \pi\sigma/\epsilon$.

Backreaction constraints

If the observable modes are initially in an excited non-vacuum state, their accumulated energy density can backreact gravitationally on the inflationary background. A mode that becomes super-Hubble at $N(k) \sim 56 - \ln(hk/a_0H_0)$ e-folds before the end of inflation possessed a physical momentum $k_{\rm phys} = \exp(N_i)H_{\rm inf}$ at the beginning of inflation ($N_i = N_{\rm tot} - N(k)$). The energy density carried by these excited modes is

$$\frac{\delta\rho_k}{\rho_{\rm crit}} \sim n_k \exp(4N_i)\left(\frac{H_{\rm inf}}{M_{\rm Pl}}\right)^2, \tag{9.7.54}$$

where $n_k = |\beta_k|^2$. To prevent backreaction from disrupting inflation ($\delta\rho_k/\rho_{\rm crit} < 1$) with $H_{\rm inf} \sim 10^{-6}M_{\rm Pl}$, we must have

$$n_k \lesssim \exp(28 - 4N_i) \iff N_i \lesssim 7. \tag{9.7.55}$$

This restriction indicates that for inflation of long duration, we strictly require $n_k \ll 1$. Gravitational backreaction is therefore only avoided if trans-Planckian modes originate in the adiabatic vacuum or an extremely close state, reaffirming the predictive power of the inflationary framework.

9.8 Status of the inflationary paradigm

9.8.1 Generic predictions

Single-field inflationary models have robust predictions that are independent of their specific implementation:

- **Homogeneity and isotropy:** Inflation erases any pre-existing classical inhomogeneities, dynamically relaxing the Universe into a smooth Friedmann–Lemaître background.

- **Spatial flatness:** During inflation, spatial curvature is exponentially diluted. Inflation predicts $\Omega = 1$ up to the amplitude of super-Hubble density perturbations ($\sim 10^{-5}$).
- **Scalar perturbations:** Generated by quantum fluctuations of the inflaton amplified and stretched to macroscopic super-Hubble scales. Inflation predicts that these perturbations are adiabatic, Gaussian, and have an almost scale-invariant power spectrum ($n_s \approx 1$). Their phase coherence produces the characteristic acoustic peak structure in the CMB temperature and polarization angular power spectra.
- **Absence of vector perturbations:** Primordial vector modes possess no growing mode and are exponentially damped during inflation.
- **Primordial gravitational waves:** Tensor perturbations are generated via parametric amplification of quantum vacuum metric fluctuations with an almost scale-invariant spectrum and satisfy the consistency relation $r = -8n_T$ (or $r_\zeta = -n_T/2$).
- **Spectator field fluctuations:** Any light scalar field ($m < H$) acquires super-Hubble quantum fluctuations with characteristic amplitude $\delta\phi \approx H/(2\pi)$.

More generally, inflationary models radically change our vision of cosmology in two fundamental respects:

1. **Particle production at reheating:** The non-perturbative decay of the inflaton condensate via parametric preheating explains the transition from a cold vacuum-like state to the hot primordial plasma.
2. **Eternal self-reproduction:** In the context of chaotic initial conditions, quantum diffusion naturally drives eternal inflation, producing a fractal multiverse of pocket universes.

9.8.2 Current observational constraints

Cosmological observations test the dynamics of inflation during the 8–10 e-folds when observable scales exited the Hubble radius. Precision measurements from CMB surveys (e.g. WMAP, Planck) confirm:

- Purely adiabatic, Gaussian initial conditions with no evidence for primordial non-Gaussianity ($f_{\rm NL}^{\rm local} = -0.9 \pm 5.1$).
- A red-tilted scalar spectral index $n_s = 0.9649 \pm 0.0042$, ruling out exact scale invariance at $> 8\sigma$.
- Tight upper limits on the tensor-to-scalar ratio ($r < 0.036$), constraining the inflationary energy scale to $V^{1/4} < 1.4 \times 10^{16}$ GeV.
- Spatial flatness $|\Omega_K| < 0.005$, perfectly consistent with $\Omega = 1$.

While current data strongly support spatial flatness, the detection of a positive spatial curvature ($K = +1$), even a tiny one, would pose a severe challenge to inflation [81]. Modeling inflation as a closed de Sitter space ($K = +1$), the scale factor evolves as $a(t) = a_i \cosh\!\left(\sqrt{\Lambda/3}\, t\right)$ and $\delta \equiv \Omega - 1 = 1/\sinh^2(\sqrt{\Lambda/3}\, t)$. The time t_* at which $k_0 = a_0 H_0$ becomes super-Hubble satisfies $\sqrt{\Lambda/3}\, t_* = \operatorname{arcsinh}(1/\sqrt{\delta_0})$. The maximum number of e-folds before this instant is bounded by:

$$\Delta N_{\rm max} = \frac{1}{2} \ln\left(1 + \frac{1}{\delta_0}\right) \sim 2 \quad (\text{for } \delta_0 \sim 0.02). \tag{9.8.1}$$

This demonstrates that a robust detection of spatial curvature would require a complete rethinking of the duration and initial conditions of inflation.

Finally, theorems by Borde, Guth, and Vilenkin (BGV) [49] establish that any spacetime with average expansion $H_{\rm avg} > 0$ is geodesically incomplete in the past, implying that inflation cannot be past-eternal and that an initial quantum-gravitational boundary is unavoidable.

9.8.3 Conclusions and open questions

Single-field inflation has robust, testable, and falsifiable predictions. It stands as a foundational pillar connecting general relativity, quantum field theory, and high-energy physics to observational cosmology. In this baseline framework, the inflationary sector is completely specified by 5 cosmological parameters: the scalar amplitude A_s, the scalar tilt n_s, its running $\alpha_s \equiv \mathrm{d}n_s/\mathrm{d}\ln k$, the tensor amplitude A_t, and the tensor tilt n_t (reduced from 6 by the consistency relation $r = -8n_T$).

Despite its remarkable predictive success, fundamental theoretical questions remain:

1. **The nature of the inflaton:** Is the inflaton a fundamental scalar field from grand unified or supersymmetric theories, a composite state, or a geometrical manifestation of modified gravity ($f(R)$ theories)?

2. **The trans-Planckian problem and quantum gravity:** How do ultraviolet quantum-gravitational degrees of freedom interface with cosmological perturbations?

3. **The initial singularity and past-completeness:** What physical mechanism initiates the inflationary epoch?

4. **The measure problem in eternal inflation:** How should probabilities and observational predictions be defined across an infinite multiverse?

These open questions define the vibrant frontier of modern theoretical cosmology.

Chapter 10

Non-Gaussianity in the Early Universe

Chapter 9 showed that single-field slow-roll inflation predicts an almost perfectly Gaussian spectrum of curvature perturbations, with a tiny calculable non-Gaussian residue $f_{\rm NL} \sim \mathcal{O}(\epsilon, \eta)$ (§9.7.2). This chapter takes that statement apart and rebuilds it from the ground up: *why* is the vacuum state of a quantum field Gaussian in the first place, what does "non-Gaussian" even mean for a stochastic field defined over all of space, and how do we go from an abstract statement about moments of a probability distribution to an actual number that the Planck satellite can report an error bar on? We will see that the search for primordial non-Gaussianity is really a search for *interactions*: a free (quadratic) field is exactly Gaussian at all orders, so any non-zero odd-point function is direct evidence that the field driving inflation was not free, and the detailed dependence of that function on the shape of the momentum triangle $(\vec{k}_1, \vec{k}_2, \vec{k}_3)$ encodes what interactions, and how many fields, were involved. The single-field consistency relation derived in §10.6 is one of the sharpest few-parameter theorems in cosmology: it states that decisively ruling out the entire class of single-field models requires nothing more than measuring one number, the squeezed-limit bispectrum.

This material follows Daniel Baumann's *Lectures on Cosmological Correlators* for the field-theoretic derivation of the origin of Gaussianity in §10.1 and the statistics formalism of §10.2, David Wands' review *Local non-Gaussianity from inflation* for the δN material of §10.5, and Komatsu & Spergel's WMAP bispectrum-estimator papers for §10.9. As in the rest of these notes, every intermediate step is written out explicitly.

10.1 The Wavefunction of the Universe and the Origin of Gaussianity

Before asking how much non-Gaussianity inflation produces, we should understand why the leading-order prediction is Gaussian at all. The statement "inflation predicts a Gaussian spectrum" is really the statement that the field fluctuations start out in the ground state of a (nearly) free quantum field, and the ground state of a free field is a Gaussian wavefunctional. We derive this in two complementary ways: canonically, by recognizing the mode equation as that of a harmonic oscillator, and covariantly, via the path integral.

10.1.1 Canonical argument: the field as a collection of oscillators

Work in the Schrödinger picture and ask: *what is the probability distribution for the value of a given perturbation mode?* We need the wavefunction, i.e. a solution of the Schrödinger equation for the field. Consider the eigenstates of the field operator $\hat{\Phi}$,

$$\hat{\Phi}(t, \vec{x})\,|\phi\rangle = \phi(t, \vec{x})\,|\phi\rangle\,, \tag{10.1.1}$$

where $\phi(t, \vec{x})$ is the (c-number) spatial profile picked out by the eigenstate. Recall first the single-particle case: for a particle of unit mass in a potential V, $H = \frac{1}{2}p^2 + V(x)$, and with

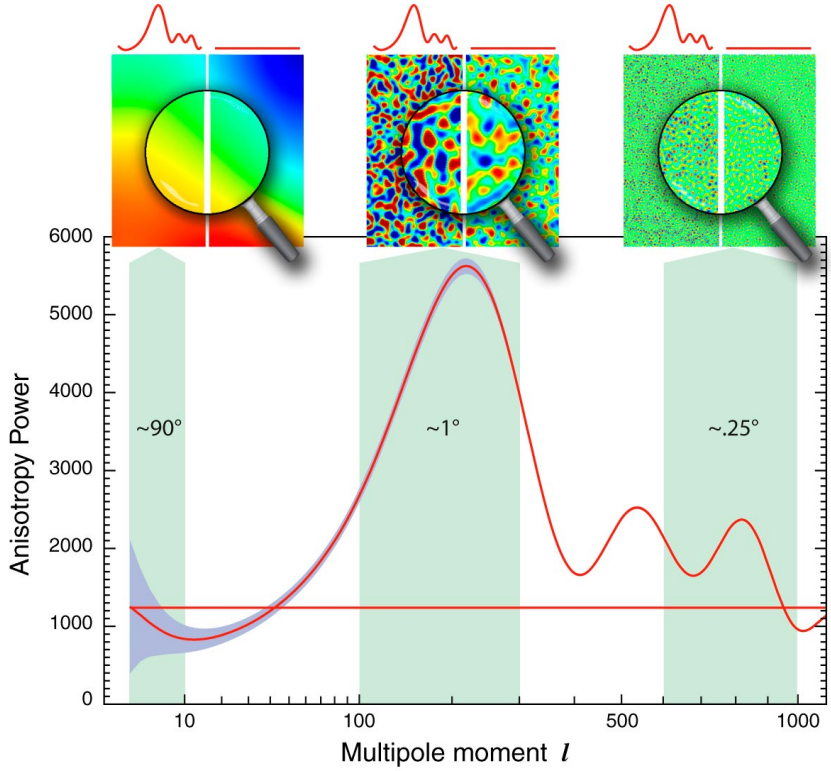


Figure 10.1: The angular power spectrum of the CMB, obtained by Fourier (spherical-harmonic) decomposing a temperature map and measuring the variance in each multipole ℓ. Only modes that have re-entered the horizon by last scattering imprint acoustic structure; both the power spectrum (2-point function) discussed in §10.4 and the bispectrum (3-point function) of §10.9 are extracted from maps of exactly this kind. Public codes exist for every stage of this pipeline: CAMB or CLASS for the linear Boltzmann hierarchy, CosmoMC for parameter inference, and CMB-BEST for bispectrum estimation.

$\psi(x) \equiv \langle x|\psi\rangle$, $\hat{x}\,|x\rangle = x\,|x\rangle$, $[\hat{x},\hat{p}] = i$ (we set $\hbar = 1$ throughout), the Schrödinger equation reads

$$i\frac{\partial\psi}{\partial t} = -\frac{\partial^2\psi}{\partial x^2} + V(x)\psi. \tag{10.1.2}$$

Quantum field theory promotes this to a functional Schrödinger equation for a wavefunctional $\psi[\phi]$.[1] A free scalar field with (possibly time-dependent) frequency ω per Fourier mode has Hamiltonian density $H = \frac{1}{2}\Pi^2 + \frac{1}{2}\omega^2\phi^2$, and defining $\psi[\phi(t,\vec{x})] \equiv \langle\phi|\psi\rangle$ with $[\hat{\Phi}(\vec{x}),\hat{\Pi}(\vec{y})] = i\delta(\vec{x}-\vec{y})$, the Schrödinger equation becomes

$$i\frac{\partial\psi[\phi(t,\vec{x})]}{\partial\eta} = \left(-\frac{1}{2}\frac{\partial^2}{\partial\phi^2} + \frac{1}{2}\omega^2\phi^2\right)\psi[\phi(t,\vec{x})]. \tag{10.1.3}$$

Mode by mode, this is exactly the Schrödinger equation for a harmonic oscillator (unit mass), so we already know the ground-state solution is a Gaussian:

$$\langle\phi|0\rangle = C\,e^{-\frac{\omega}{2}\phi(t,\vec{x})^2}\,e^{-iE_0\eta}, \tag{10.1.4}$$

with E_0 the (constant) ground-state energy. Every mode of the inflaton fluctuation obeys precisely this kind of mode equation deep inside the horizon (where the curvature of spacetime

[1] See §14.7.1 of *Gravitation* by Padmanabhan for the canonical field-theoretic setup, and §10.1 of *Quantum Field Theory of Point Particles and Strings* by Hatfield for the derivation of the functional Schrödinger equation quoted below.

is unimportant and the mode oscillates like a flat-space oscillator), so the Bunch–Davies vacuum – the initial condition adopted for inflationary perturbations – is, by this argument, mode by mode Gaussian.

As long as the primordial density fluctuation depends *linearly* on such a quantum field, it must inherit an exactly Gaussian distribution. This immediately suggests a diagnostic: any non-vanishing odd-order moment of the observed fluctuations is direct evidence either that the underlying field is not free, or that the map from field to observable is non-linear (or both). The lowest odd moment beyond the trivial vanishing mean is the three-point function, whose Fourier-space avatar is the **bispectrum**; a non-zero bispectrum guarantees non-Gaussianity (though the converse need not hold: a vanishing bispectrum does not by itself guarantee Gaussianity, since a distribution could be non-Gaussian only in its four-point function and higher). Since departures from Gaussianity are sourced by interactions or by particle production, measuring higher-order correlators is often called "cosmological collider physics": the analogue, for the sky, of measuring collisions in a particle accelerator.

10.1.2 Path-integral argument

The same conclusion follows from the path integral, and this version generalizes more transparently to interacting theories. The wavefunctional at some reference time (which we can set to $t_* = 0$ without loss of generality) is

$$\Psi[\phi] = \int \mathcal{D}\Phi \, e^{iS[\Phi]}\Big|_{\Phi(0)=\phi,\ \Phi(-\infty^+)=0} \approx \ e^{iS[\Phi_{\rm cl}]}, \tag{10.1.5}$$

where the last step uses perturbation theory at tree level: the path integral is dominated by the classical saddle point $\Phi_{\rm cl}$, corrections to which are the loop (interaction) contributions responsible for any non-Gaussianity.[2] Consider a toy oscillator with action

$$S[\Phi] = \int \mathrm{d}t \left(\frac{1}{2}\dot{\Phi}^2 - \frac{1}{2}\omega^2 \Phi^2 \right). \tag{10.1.6}$$

On shell, the action reduces to a pure boundary term,

$$S[\Phi_{\rm cl}] = \int_{t_i}^{t^*} \mathrm{d}t \left[\frac{1}{2}\partial_t(\dot{\Phi}_{\rm cl}\Phi_{\rm cl}) - \frac{1}{2}\Phi_{\rm cl}(\ddot{\Phi}_{\rm cl} + \omega^2 \Phi_{\rm cl}) \right] = \frac{1}{2}\dot{\Phi}_{\rm cl}\Phi_{\rm cl}\Big|_{t=t^*}, \tag{10.1.7}$$

having integrated by parts and used the classical equation of motion $\ddot{\Phi}_{\rm cl} + \omega^2 \Phi_{\rm cl} = 0$. Two boundary conditions fix $\Phi_{\rm cl}(t)$ uniquely: (1) $\Phi_{\rm cl}(t_* = 0) = \phi$, and (2) at early times the mode sits in its minimum-energy configuration, $\Phi_{\rm cl}(t) \sim e^{i\omega t}$ as $t \to -\infty^+$ (the Bunch–Davies condition). These fix $\Phi_{\rm cl}(t) = \phi \, e^{i\omega t}$, so that

$$S = \frac{i\omega}{2}\phi^2 \quad \Longrightarrow \quad \Psi[\phi] = \exp\left(-\frac{\omega}{2}\phi^2 \right) \quad \Longrightarrow \quad |\Psi[\phi]|^2 = e^{-\omega\phi^2}. \tag{10.1.8}$$

The ground-state wavefunctional of the (free) harmonic oscillator is Gaussian, with the width of the Gaussian fixing the size of the zero-point fluctuations,

$$\langle \phi^2 \rangle = \frac{1}{2\omega}. \tag{10.1.9}$$

This is the same statement as before, but phrased so that the generalization is obvious: turning on a cubic (or higher) self-interaction in $S[\Phi]$ adds a non-quadratic piece to $S[\Phi_{\rm cl}]$, which no longer exponentiates into a pure Gaussian in ϕ – this is precisely the mechanism by which interactions during inflation source a non-zero bispectrum, worked out concretely for the curvature perturbation in §10.5–§10.6.

[2] Beyond tree level, this path integral is evaluated using the in–in (Schwinger–Keldysh) formalism rather than the ordinary in–out S-matrix, because we want equal-time expectation values in an evolving state rather than transition amplitudes between asymptotic states. See Maldacena (2003) and Weinberg (2005) for the systematic treatment; we only need the tree-level (free-field) result here.

10.2 Statistics of Random Fields

10.2.1 Gaussian statistics of a single variable

Consider a random variable x drawn from a probability density function (PDF) $P(x)$, normalized as $\int_{-\infty}^{\infty} P(x)\mathrm{d}x = 1$. We say x obeys Gaussian statistics with zero mean if

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}, \tag{10.2.1}$$

the normalization constant $(2\pi\sigma^2)^{-1/2}$ being fixed by $\int P\,\mathrm{d}x = 1$. A Gaussian PDF is completely specified by a single number, σ: every moment can be computed explicitly,

$$\begin{aligned}
\langle x \rangle &= \int_{-\infty}^{\infty} x P(x)\mathrm{d}x = 0 && \text{(zero mean)} \\
\left\langle x^2 \right\rangle &= \int_{-\infty}^{\infty} x^2 P(x)\mathrm{d}x = \sigma^2 && \text{(variance)} \\
\left\langle x^3 \right\rangle &= \int_{-\infty}^{\infty} x^3 P(x)\mathrm{d}x = 0 && \text{(zero skewness)} \\
\left\langle x^4 \right\rangle &= \int_{-\infty}^{\infty} x^4 P(x)\mathrm{d}x = 3\sigma^4 && \text{(kurtosis} = 3\sigma^4\text{)} \\
\left\langle x^5 \right\rangle &= \int_{-\infty}^{\infty} x^5 P(x)\mathrm{d}x = 0 \\
&\vdots
\end{aligned}$$

so that all odd moments $\langle x^{2n-1} \rangle$ vanish while all even moments $\langle x^{2n} \rangle$ are fixed multiples of σ^{2n}; σ determines everything.

10.2.2 Multivariate Gaussian statistics and Wick's theorem

Now consider N random variables $\{x_1, \ldots, x_N\}$ distributed over space – for instance, the temperature measured at N different points in a room, or the density contrast sampled at N points across the Universe. These may obey a multivariate Gaussian distribution,

$$P(x_1, \ldots, x_N) = \frac{1}{(2\pi)^{N/2} |\xi|^{1/2}} e^{-\frac{1}{2}\sum_{ij} x_i (\xi^{-1})_{ij} x_j}, \tag{10.2.2}$$

normalized to $\int P\,\mathrm{d}x_1 \cdots \mathrm{d}x_N = 1$, where $\xi_{ij} = \langle x_i x_j \rangle$ is the **covariance matrix**, equivalently the 2**-point correlation function**. Its moments are

$$\begin{gathered}
\langle x_i \rangle = 0, \qquad \langle x_i x_j \rangle = \xi_{ij}, \qquad \langle x_i x_j x_k \rangle = 0, \\
\langle x_i x_j x_k x_l \rangle = \xi_{ij}\xi_{kl} + \xi_{ik}\xi_{jl} + \xi_{il}\xi_{jk}, \qquad \langle x_i x_j x_k x_l x_m \rangle = 0, \quad \ldots
\end{gathered}$$

The four-point (and all higher even-point) result is **Wick's theorem**: every even moment of a Gaussian is a sum over all ways of pairing up the variables ("contractions"), each pairing contributing the product of the corresponding two-point functions. Odd moments always vanish by the $x \to -x$ symmetry of a zero-mean Gaussian.

Why Wick's theorem holds

The cleanest proof uses the moment generating function. For the multivariate Gaussian above,

$$Z(J) \equiv \left\langle e^{J\cdot x} \right\rangle = \int \mathrm{d}^N x\, P(x)\, e^{\sum_i J_i x_i} = \exp\left(\frac{1}{2}\sum_{ij} J_i \xi_{ij} J_j\right), \tag{1}$$

obtained by completing the square in the Gaussian integral. Every moment is then a derivative of Z at $J = 0$,

$$\langle x_{i_1} \cdots x_{i_{2n}} \rangle = \frac{\partial}{\partial J_{i_1}} \cdots \frac{\partial}{\partial J_{i_{2n}}} Z(J)\bigg|_{J=0}. \tag{2}$$

Because $Z(J) = \exp\left(\frac{1}{2} J\xi J\right)$ is itself the exponential of a quadratic form, each derivative either brings down one more factor of ξJ (extending an existing pair) or acts on an existing (ξJ) factor to close a pair into a plain ξ_{ij}. Iterating, the only way to end up with no leftover powers of J after $2n$ derivatives is to have paired up all $2n$ indices into n factors of ξ; summing over all such pairings reproduces exactly Wick's theorem, with combinatorial coefficient 1 for each distinct pairing (for the single-variable case $x_1 = \cdots = x_{2n} = x$, the number of pairings is $(2n-1)!!$, reproducing $\langle x^4 \rangle = 3\sigma^4$, $\langle x^6 \rangle = 15\sigma^6$, etc.).

For example, for the multivariate case,

$$\langle x_1 x_2 x_3 x_4 \rangle = \langle x_1 x_2 \rangle \langle x_3 x_4 \rangle + \langle x_1 x_3 \rangle \langle x_2 x_4 \rangle + \langle x_1 x_4 \rangle \langle x_2 x_3 \rangle = \xi_{12}\xi_{34} + \xi_{13}\xi_{24} + \xi_{14}\xi_{23},$$

while for a single variable, all three pairings coincide,

$$\left\langle x^4 \right\rangle = \left\langle x^2 \right\rangle \left\langle x^2 \right\rangle + \left\langle x^2 \right\rangle \left\langle x^2 \right\rangle + \left\langle x^2 \right\rangle \left\langle x^2 \right\rangle = 3\left\langle x^2 \right\rangle^2 = 3\sigma^4.$$

10.2.3 Homogeneity, isotropy, and the power spectrum

Let $\vec{q}$ denote spatial coordinates and $X(\vec{q})$ a random field. The two-point function of X at points $\vec{q}_i, \vec{q}_j$ connected by $\vec{r}_{ij} = \vec{q}_j - \vec{q}_i$ is $\xi_{ij} = \langle X(\vec{q}_i) X(\vec{q}_i + \vec{r}_{ij}) \rangle$, which a priori depends both on the location $\vec{q}_i$ and on the separation $\vec{r}_{ij}$.

- **Homogeneity** (translational invariance): ξ_{ij} does not depend on $\vec{q}_i$, so $\xi_{ij} =: \xi(\vec{r}_{ij})$.
- **Isotropy** (rotational invariance): seen from a given point, ξ_{ij} does not depend on the direction of $\vec{r}_{ij}$, so $\xi_{ij} := \xi(\vec{q}_i, |\vec{r}_{ij}|)$.
- Together, if ξ_{ij} is isotropic as seen from *every* point, homogeneity follows automatically, and $\xi_{ij} \equiv \xi(|\vec{r}_{ij}|)$. We believe our Universe is statistically homogeneous and isotropic on large scales – a foundational enough assumption that it must be tested repeatedly rather than simply postulated.

Fourier transforming, $\tilde{X}(\vec{k}) \equiv \int \mathrm{d}^3 q\, e^{-i\vec{k}\cdot\vec{q}} X(\vec{q})$, translational invariance forces the covariance

$$C_{ij} \equiv \langle \tilde{X}(\vec{k}_i) \tilde{X}^*(\vec{k}_j) \rangle = (2\pi)^3 \delta_D^{(3)}(\vec{k}_i - \vec{k}_j)\, P(\vec{k}_j), \tag{10.2.3}$$

and rotational invariance further forces $P(\vec{k}_j) \to P(|\vec{k}_j|)$, defining the **power spectrum** $P(k)$.

Proof of Eq. (10.2.3)

$$C_{ij} = \int \mathrm{d}^3q \int \mathrm{d}^3q'\, e^{-i\vec{k}_i\cdot\vec{q}} e^{i\vec{k}_j\cdot\vec{q}'} \langle X(\vec{q})X(\vec{q}')\rangle = \int \mathrm{d}^3q \int \mathrm{d}^3r\, e^{-i(\vec{k}_i-\vec{k}_j)\cdot\vec{q}} e^{i\vec{k}_j\cdot\vec{r}} \underbrace{\langle X(\vec{q})X(\vec{q}+\vec{r})\rangle}_{\xi(\vec{r}) \text{ by homogeneity}}$$

$$= \underbrace{\int \mathrm{d}^3r\, e^{i\vec{k}_j\cdot\vec{r}}\xi(\vec{r})}_{P(\vec{k}_j)} \underbrace{\int \mathrm{d}^3q\, e^{-i(\vec{k}_i-\vec{k}_j)\cdot\vec{q}}}_{(2\pi)^3\delta^{(3)}(\vec{k}_i-\vec{k}_j)} = (2\pi)^3\delta^{(3)}(\vec{k}_i-\vec{k}_j)P(\vec{k}_j),$$

and isotropy, $\xi(\vec{r}) = \xi(|r|)$, immediately gives $P(\vec{k}_j) = P(|\vec{k}_j|)$.

We can equally write the PDF of the Fourier modes $\tilde{X}(\vec{k})$ themselves:

$$P(\tilde{X}(\vec{k}_1), \cdots, \tilde{X}(\vec{k}_N)) = \frac{1}{(2\pi)^{N/2}|C|^{1/2}} e^{-\frac{1}{2}\sum_{ij}\tilde{X}(\vec{k}_i)(C^{-1})_{ij}\tilde{X}^*(\vec{k}_j)}, \tag{10.2.4}$$

which for a translationally and rotationally invariant field simplifies dramatically, since C is then diagonal in $\vec{k}$:

$$P(\tilde{X}(\vec{k}_1), \cdots, \tilde{X}(\vec{k}_N)) = \frac{1}{(2\pi)^{N/2}\left(\prod_i P(|\vec{k}_i|)\right)^{1/2}} e^{-\frac{1}{2}\sum_i \frac{|\tilde{X}(\vec{k}_i)|^2}{P(|\vec{k}_i|)}}. \tag{10.2.5}$$

Different Fourier modes of a Gaussian, statistically homogeneous and isotropic field are statistically *independent* random variables. This is the deep reason the power spectrum is the natural language for describing cosmological data sets: for a Gaussian field it is not merely *a* useful summary statistic, it is a *complete* one.

A further remark worth making explicit here, since it underlies every observational estimator in §10.9: the expectation values $\langle\cdot\rangle$ above are *ensemble* averages over a hypothetical set of universes. We only ever observe *one* realization of the stochastic process that generated our own Universe. Replacing the ensemble average by a spatial (or, on the sky, angular) average over that one realization is only justified if the field is *ergodic*, which for a statistically homogeneous Gaussian field with a well-behaved (integrable) power spectrum is guaranteed. The unavoidable statistical scatter left over from having only one universe to look at – most severe on the largest scales, where there are only a handful of independent Fourier modes or low multipoles to average over – is known as **cosmic variance**, and we will meet it again as the fundamental noise floor of the f_{NL} estimators of §10.9.

The bispectrum

Extending the argument, define the **bispectrum** B by

$$\langle\tilde{X}(\vec{k}_1)\tilde{X}(\vec{k}_2)\tilde{X}(\vec{k}_3)\rangle = (2\pi)^3\delta^{(3)}(\vec{k}_1+\vec{k}_2+\vec{k}_3)\, B(k_1,k_2,k_3), \tag{10.2.6}$$

which vanishes identically for Gaussian $\tilde{X}$; translational and rotational invariance restrict B to depend only on the magnitudes k_1, k_2, k_3 (equivalently, since the Dirac delta enforces $\vec{k}_1+\vec{k}_2+\vec{k}_3 = 0$, on the shape and size of the momentum-space triangle these three vectors form). We return to this object, and to the physical significance of the shape of that triangle, in §10.4.

10.2.4 Weakly non-Gaussian statistics

Gaussian statistics are convenient precisely because the PDF is known in closed form. What if the PDF is not Gaussian? In general there is no universally optimal strategy when nothing is

known about $P(x)$; but cosmology hands us a favourable special case. Every observation of the primordial fluctuations (most directly, the CMB) is well described by a Gaussian distribution to high precision, so any departure from Gaussianity must be *small*. In that regime we can construct a controlled approximation to $P(x)$ by treating it as a small perturbation of a Gaussian, using the tool for systematically expanding an unknown function about a known one: the Taylor series, $f(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} \frac{\mathrm{d}^n f}{\mathrm{d}x^n}\Big|_{x=0}$. Applied to a Gaussian PDF, this becomes the Gram–Charlier and Edgeworth expansions below.

Gram–Charlier expansion

Let $G(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ be the unit-variance, zero-mean Gaussian PDF. Suppose $P(x)$ describes a weakly non-Gaussian variable that we wish to express as a perturbation of $G(x)$:

$$P(x) = \sum_{n=0}^{\infty} C_n \frac{\mathrm{d}^n G(x)}{\mathrm{d}x^n} = G(x)\left[C_0 - C_1 x + C_2(x^2-1) - C_3(x^3-3x) + \dots\right]. \tag{10.2.7}$$

This is the **Gram–Charlier expansion**. It is conventionally repackaged using the **Chebyshev–Hermite polynomials**

$$He_n(x) \equiv (-1)^n \frac{1}{G(x)} \frac{\mathrm{d}^n G}{\mathrm{d}x^n} = (-1)^n e^{x^2/2} \frac{\mathrm{d}^n e^{-x^2/2}}{\mathrm{d}x^n}, \qquad \begin{aligned} &He_0 = 1, && He_1 = x, \\ &He_2 = x^2 - 1, && He_3 = x^3 - 3x, \\ &He_4 = x^4 - 6x^2 + 3, \end{aligned} \tag{10.2.8}$$

so that

$$P(x) = G(x)\left[\sum_{n=0}^{\infty} C_n (-1)^n He_n(x)\right], \tag{10.2.9}$$

i.e. the ratio $P(x)/G(x)$ is expanded on the basis $(-1)^n He_n(x)$. Because He_n satisfies the orthogonality relation

$$\int_{-\infty}^{\infty} G(x) He_n(x) He_m(x) \mathrm{d}x = m!\, \delta_{mn}, \tag{10.2.10}$$

we can invert Eq. (10.2.9) for the coefficients C_n: multiply both sides by $He_n(x)$ and integrate,

$$\int_{-\infty}^{\infty} \mathrm{d}x\, P(x) He_n(x) = \sum_{m=0}^{\infty} C_m (-1)^m \int_{-\infty}^{\infty} \mathrm{d}x\, G(x) He_m(x) He_n(x) = (-1)^n n!\, C_n, \tag{10.2.11}$$

so that

$$\boxed{C_n = \frac{(-1)^n}{n!} \int_{-\infty}^{\infty} \mathrm{d}x\, P(x) He_n(x).} \tag{10.2.12}$$

Using Eq. (10.2.8), the first few coefficients are simply the moments of $P(x)$:

$$\begin{aligned}
C_0 &= \int P\, \mathrm{d}x = 1 && \text{(normalization)} \\
C_1 &= -\int xP\, \mathrm{d}x = 0 && \text{(zero mean)} \\
C_2 &= \tfrac{1}{2}\int (x^2-1)P\, \mathrm{d}x = \tfrac{1}{2}(\langle x^2 \rangle - 1) = 0 && \text{(unit variance)} \\
C_3 &= -\tfrac{1}{6}\int (x^3-3x)P\, \mathrm{d}x = -\tfrac{1}{6}\langle x^3 \rangle \equiv -\tfrac{1}{6}k_3 && (k_3 \text{ skewness}) \\
C_4 &= \tfrac{1}{24}\int (x^4-6x^2+3)P\, \mathrm{d}x = \tfrac{1}{24}(\langle x^4 \rangle - 3) \equiv \tfrac{1}{24}k_4 && (k_4 \text{ kurtosis}).
\end{aligned}$$

Collecting terms, the Gram–Charlier expansion of a zero-mean, unit-variance PDF is

$$P(x) = G(x)\left[1 + \frac{1}{6}k_3 He_3(x) + \frac{1}{24}k_4 He_4(x) + \dots\right]. \tag{10.2.13}$$

The expansion coefficients are literally the cumulants of $P(x)$: the entire machinery is nothing more than a change of basis, from moments to an orthogonal-polynomial expansion adapted to the Gaussian measure.

Edgeworth expansion

Extend this to $\sigma \neq 1$. The natural (generically dimensionful) expansion parameter turns out to be

$$S_n = \frac{k_n}{\sigma^{2n-2}}, \tag{10.2.14}$$

and for a random variable y with variance σ^2,

$$\begin{aligned} P(y) &= \frac{1}{\sigma}G(y/\sigma)\left[1 + \sigma\frac{S_3}{6}He_3(y/\sigma) + \sigma^2\frac{S_4}{24}He_4(y/\sigma) + \cdots\right] \\ &= \frac{1}{\sqrt{2\pi}\sigma}e^{-y^2/2\sigma^2}\left[1 + \frac{S_3}{6}\left(\frac{y^3}{\sigma^3} - 3\frac{y}{\sigma}\right) + \frac{S_4}{24}\left(\frac{y^4}{\sigma^4} - 6\frac{y^2}{\sigma^2} + 3\right) + \cdots\right], \end{aligned} \tag{10.2.15}$$

the **Edgeworth expansion**. A caveat worth stating plainly: unlike a genuine Taylor series, the Edgeworth expansion is only an *asymptotic* series in the small parameter σS_n – adding more terms does not necessarily improve the approximation once σS_n is not tiny, and the truncated series can even become negative in the tails, where it is manifestly not a good approximation to a true (positive) PDF. It is to be trusted only in a region around $y \sim \sigma$ and only for the first correction or two; this is precisely the regime relevant for a nearly-Gaussian primordial fluctuation field.

10.3 Application: Non-Gaussianity in Gravitational Clustering

Before returning to the primordial fluctuations themselves, it is instructive to see how the statistics of §10.2 feed directly into an observable: the abundance of collapsed structures (dark-matter haloes) as a function of mass, and how that abundance is modified if the initial density field is not exactly Gaussian.

In the standard picture, collapsed objects form at the locations of high peaks of the density field. To turn this into a mass function we go to a sufficiently early time that the density contrast δ is still linear, $\delta \ll 1$, and consider the mass enclosed in a sphere of comoving radius R,

$$M = \frac{4\pi}{3}\,\langle\rho\rangle\, R^3. \tag{10.3.1}$$

Smoothing the density field with a top-hat window of radius R (equivalently, binning it on scale R) defines a smoothed field $\delta_R(\vec{q})$ with variance $\sigma_R^2 \equiv \langle\delta_R^2\rangle$, a decreasing function of R because larger smoothing scales average away more small-scale power. As R shrinks (equivalently as we look at a fixed comoving patch at ever-later times, when σ_R has grown through linear structure growth), δ_R eventually exceeds the (spherical-collapse) threshold $\delta_c = 1.686$, at which point an object of mass $M = \frac{4\pi}{3}\,\langle\rho\rangle\, R^3$ is deemed to have formed.

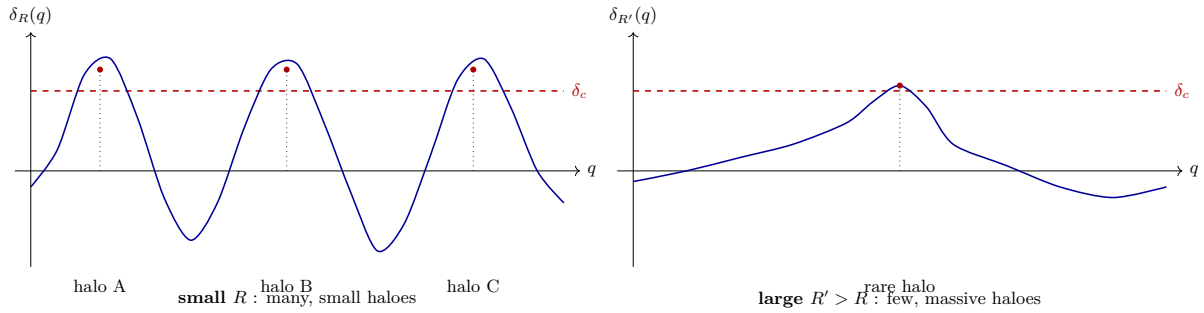


Figure 10.2: Left: a one-dimensional slice of the linear density contrast $\delta_R(\vec{q})$, smoothed on scale R with a top-hat window; every excursion above the collapse threshold $\delta_c = 1.686$ is identified with a virialized halo of mass $M = \frac{4\pi}{3}\langle\rho\rangle R^3$. Right: smoothing on a larger scale $R' > R$ suppresses $\sigma_{R'} < \sigma_R$, so fewer, broader regions cross threshold – these correspond to rarer, more massive haloes. This threshold-crossing statistic of a (nearly) Gaussian random field is the basis of the Press–Schechter mass function, Eq. (4).

Derivation of the Press–Schechter mass function

For Gaussian δ_R with variance σ_R^2, the fraction of *volume* with $\delta_R > \delta_c$ is

$$\frac{1}{2}\,\mathrm{erfc}\!\left(\frac{\delta_c}{\sqrt{2}\sigma_R}\right) = \int_{\delta_c}^{\infty} \mathrm{d}\delta_R\, \frac{1}{\sqrt{2\pi}\sigma_R} e^{-\delta_R^2/2\sigma_R^2}. \tag{1}$$

Press & Schechter (1974) identify this with the fraction of *mass* bound up in objects with $M(R') \geq M(R)$ for some $R' \geq R$ – but taken at face value this undercounts by exactly a factor of 2: a point with $\delta_R < \delta_c$ can still end up inside a larger collapsed region if $\delta_{R'} > \delta_c$ for some $R' > R$ (the "cloud-in-cloud" problem), and every point must belong to *some* halo once we smooth on a large enough scale. Press & Schechter simply multiply by hand by a fudge factor of 2,

$$F(>M) \equiv 2 \times \frac{1}{2}\,\mathrm{erfc}\!\left(\frac{\delta_c}{\sqrt{2}\sigma_R}\right) = \mathrm{erfc}\!\left(\frac{\delta_c}{\sqrt{2}\sigma_R}\right), \tag{2}$$

for the fraction of mass in haloes of mass $\geq M$; this happens to be exactly what is needed to normalize the mass function correctly, $\int_0^\infty M\,\frac{\mathrm{d}n}{\mathrm{d}M}\mathrm{d}M = \langle\rho\rangle$ (all the mass in the Universe ends up in haloes, once one integrates down to $M = 0$). The heuristic factor of 2 is derived properly, without ad-hoc doubling, in the excursion-set formalism of Bond, Cole, Efstathiou & Kaiser (1991), which reproduces the same result from a random-walk first-crossing argument.

The differential mass function follows from $M\,\frac{\mathrm{d}n}{\mathrm{d}M}\mathrm{d}M = \langle\rho\rangle\left|\frac{\mathrm{d}F}{\mathrm{d}M}\right|\mathrm{d}M$, i.e. $\frac{\mathrm{d}n}{\mathrm{d}M} = \frac{\langle\rho\rangle}{M}\left|\frac{\mathrm{d}F}{\mathrm{d}M}\right|$. Since $\frac{\mathrm{d}}{\mathrm{d}M}\,\mathrm{erfc}(u(M)) = -\frac{2}{\sqrt{\pi}}e^{-u^2}\frac{\mathrm{d}u}{\mathrm{d}M}$ with $u = \delta_c/(\sqrt{2}\sigma_R)$,

$$\left|\frac{\mathrm{d}F}{\mathrm{d}M}\right| = \sqrt{\frac{2}{\pi}}\,\delta_c\, e^{-\delta_c^2/2\sigma_R^2}\left|\frac{\mathrm{d}\sigma_R^{-1}}{\mathrm{d}M}\right|, \tag{3}$$

and therefore

$$\boxed{\frac{\mathrm{d}n}{\mathrm{d}M} = \sqrt{\frac{2}{\pi}}\,\frac{\langle\rho\rangle}{M}\,\delta_c\left|\frac{\mathrm{d}\sigma_R^{-1}}{\mathrm{d}M}\right| e^{-\delta_c^2/2\sigma_R^2}.} \tag{4}$$

This is the **Press–Schechter mass function**.

Now suppose instead that the primordial density field carries a small non-Gaussian skewness,

so that $P(\delta_R)$ is given by the Edgeworth expansion, Eq. (10.2.15) (dropping the kurtosis term for brevity, since S_3 dominates for local-type non-Gaussianity, see §10.7):

$$P(\delta_R) = \frac{1}{\sqrt{2\pi}\sigma_R} e^{-\delta_R^2/2\sigma_R^2} \left[1 + \frac{S_3}{6}\left(\frac{\delta_R^3}{\sigma_R^3} - 3\frac{\delta_R}{\sigma_R}\right) + \cdots\right]. \tag{10.3.2}$$

Repeating the derivation above with this modified $P(\delta_R)$ gives the non-Gaussian Press–Schechter mass function (Matarrese, Verde & Jimenez 2000; LoVerde et al. 2008):

$$\frac{\mathrm{d}n}{\mathrm{d}M} = \sqrt{\frac{2}{\pi}}\frac{\langle\rho\rangle}{M} e^{-\delta_c^2/2\sigma_R^2}\left[\delta_c\left|\frac{\mathrm{d}\sigma_R^{-1}}{\mathrm{d}M}\right|\left(1 + \frac{S_3}{6}\left\{\frac{\delta_c^3}{\sigma_R^3} - 2\frac{\delta_c}{\sigma_R} - \frac{\sigma_R}{\delta_c}\right\}\right) - \frac{\sigma_R}{6}\frac{\mathrm{d}S_3}{\mathrm{d}M}\left(\frac{\delta_c^2}{\sigma_R^2} - 1\right)\right]. \tag{10.3.3}$$

Since $\delta_c/\sigma_R \gg 1$ for the rare, massive haloes that dominate the exponential tail, a positive skewness $S_3 > 0$ exponentially *enhances* the abundance of the rarest, most massive objects relative to the pure Gaussian prediction – this is precisely why cluster counts and the abundance of high-redshift massive galaxies are, alongside the CMB bispectrum of §10.9, an independent observational handle on primordial non-Gaussianity.

10.4 Non-Gaussianity as a Test of Inflation

We now return to the primordial curvature perturbation itself. Inflation produces perturbations in the spatial metric δg_{ij}: the scalar part seeds the structures we see today, while the tensor part propagates as gravitational waves. Working in the uniform-density gauge, write[3]

$$g_{ij} = e^{2\zeta}\gamma_{ij}a^2(t), \qquad e^{2\zeta} = 1 + 2\zeta + \ldots, \qquad \gamma_{ij} = \delta_{ij} + h_{ij} + \ldots, \qquad i,j = 1,2,3, \tag{10.4.1}$$

where ζ is the curvature perturbation and h_{ij} the gravitational-wave amplitude. This section studies the statistics of ζ; it is worth recalling the two relations connecting ζ to observables established in Chapter 6:

1. On the very large angular scales where the Sachs–Wolfe approximation holds, $\dfrac{\delta T}{T} = -\dfrac{1}{5}\zeta(\hat{n}r_\star, z_\star)$, with $r_\star$ the comoving distance to last scattering.
2. In Fourier space during matter domination, the Newtonian potential is $\tilde{\Psi}_{\vec{k}} = -\dfrac{3}{5}\tilde{\zeta}_{\vec{k}}\,T(k)$, with $T(k)$ the transfer function.

10.4.1 Power spectrum and scale invariance

Defining $\langle\tilde{\zeta}_{\vec{k}}\tilde{\zeta}^\star_{\vec{k}'}\rangle = (2\pi)^3\delta^{(3)}(\vec{k}-\vec{k}')P_\zeta(k)$, inflation predicts a nearly power-law spectrum,

$$P_\zeta(k) \propto k^{n_s - 4}, \tag{10.4.2}$$

with $n_s \approx 1$ (a "scale-invariant" spectrum). To see why this label is appropriate, compute the real-space correlation function:

$$\xi_\zeta(r) \equiv \langle\zeta(\vec{q})\zeta(\vec{q}+\vec{r})\rangle = \int \frac{\mathrm{d}^3k}{(2\pi)^3}P_\zeta(k)e^{i\vec{k}\cdot\vec{r}} = \int_0^\infty \frac{\mathrm{d}k}{k}\frac{k^3P_\zeta(k)}{2\pi^2}\frac{\sin(kr)}{kr}, \tag{10.4.3}$$

[3]The more commonly quoted convention writes $g_{ij} = e^{2\Phi}\gamma_{ij}a^2(t)$ and then defines $\zeta = \Phi + \int \mathrm{d}\rho/[3(\rho+p)]$, the two constructions agreeing on super-horizon scales.

so that, evaluating the (slowly-varying) integrand at $k \approx 1/r$,

$$\boxed{\xi_\zeta(r) \approx \left.\frac{k^3 P_\zeta(k)}{2\pi^2}\right|_{k\approx 1/r} \propto r^{1-n_s}.} \tag{10.4.4}$$

For $n_s \approx 1$, $\xi_\zeta(r)$ is (nearly) independent of r – more precisely it depends on r only logarithmically – hence "scale-invariant". The latest full-mission Planck measurement gives $n_s = 0.9649 \pm 0.0042$ (§9.8.2), a red tilt excluding exact scale invariance at $> 8\sigma$.

While $P_\zeta(k)$ can be computed analytically, comparison with the sky requires the spherical-harmonic basis $Y_{\ell m}$, the natural basis for functions on S^2: $f(\hat{n}) = \sum_{\ell m} a_{\ell m} Y_{\ell m}(\hat{n})$, $a_{\ell m} = \int_{4\pi} f(\hat{n}) Y^*_{\ell m}(\hat{n}) \mathrm{d}\Omega$. Statistical isotropy translates the two-point function into the angular power spectrum C_ℓ via $C_\ell \delta_{\ell\ell'}\delta_{mm'} = \langle a_{\ell m} a^*_{\ell' m'}\rangle$; because we observe only one realization (cosmic variance again, §10.2.3), we cannot access C_ℓ directly and must *estimate* it from the $2\ell+1$ modes at that multipole.

The estimator $\hat{C}_\ell = \frac{1}{2\ell+1}\sum_m |a_{\ell m}|^2$, its derivation as the maximum-likelihood estimator of C_ℓ, and the (chi-squared, not Gaussian) distribution it obeys are given in Chapter 6, §6.4.1. What we need here is the next step: inflation predicts a definite theoretical value $C_\ell^{\rm th}(\boldsymbol{\lambda}) = \frac{2}{\pi}\int k^2 \mathrm{d}k\, P_\zeta(k)\, g^{(s)}_{T\ell}(k)^2$ for any choice of cosmological parameters $\boldsymbol{\lambda}$ (§10.9), and we want the likelihood that lets us fit $\boldsymbol{\lambda}$ by comparing this prediction to the measured $\hat{C}_\ell$.

The C_ℓ likelihood used by WMAP and Planck

Chapter 6, Eq. (6.4.8) gives the exact distribution of $\hat{C}_\ell$ for a sky with true power spectrum C_ℓ: a rescaled chi-squared distribution with $2\ell+1$ degrees of freedom,

$$\mathcal{P}(\hat{C}_\ell \mid C_\ell) = \frac{2\ell+1}{C_\ell} \frac{\left[(2\ell+1)\hat{C}_\ell / C_\ell\right]^{\ell-\frac{1}{2}}}{2^{\ell+\frac{1}{2}}\Gamma\left(\ell+\frac{1}{2}\right)} \exp\left(-\frac{(2\ell+1)\hat{C}_\ell}{2C_\ell}\right).$$

For parameter fitting this is read the other way around: $\hat{C}_\ell$ is now the fixed, already-measured number, and $C_\ell = C_\ell^{\rm th}(\boldsymbol{\lambda})$ is the unknown we vary. Taking $-2\ln$ of this expression as a function of $C_\ell^{\rm th}$ and dropping everything that does not depend on $C_\ell^{\rm th}$ (the powers and Γ-function prefactors are fixed once ℓ is fixed),

$$-2\ln \mathcal{P}(\hat{C}_\ell \mid C_\ell^{\rm th}) = (2\ell+1)\left[\ln C_\ell^{\rm th} + \frac{\hat{C}_\ell}{C_\ell^{\rm th}}\right] + \text{const.}$$

Different multipoles are statistically independent (for an idealized full-sky, noiseless map), so the likelihoods multiply and the log-likelihoods add. Adding and subtracting the $\boldsymbol{\lambda}$-independent constants $\ln \hat{C}_\ell + 1$ inside the bracket, purely so that each term vanishes exactly when $C_\ell^{\rm th} = \hat{C}_\ell$,

$$\boxed{-2\ln L(\boldsymbol{\lambda}) = \sum_\ell (2\ell+1)\left[\frac{\hat{C}_\ell}{C_\ell^{\rm th}(\boldsymbol{\lambda})} + \ln\frac{C_\ell^{\rm th}(\boldsymbol{\lambda})}{\hat{C}_\ell} - 1\right] + \text{const.}} \tag{1}$$

This is the (idealized) likelihood WMAP and Planck use to fit cosmological parameters to the observed power spectrum, in place of the ordinary Gaussian χ^2: it correctly captures the fact that $\hat{C}_\ell$ is not symmetrically scattered around $C_\ell^{\rm th}$, but instead follows the one-sided, positively-skewed chi-squared shape appropriate to a sum of $2\ell+1$ squared Gaussians.

The Gaussian χ^2 used in the amplitude fit of Chapter 7, §7.5.2 is recovered as the limit of Eq. (1) close to the best fit. Writing $C_\ell^{\rm th} = \hat{C}_\ell(1+\epsilon_\ell)$ for small ϵ_ℓ and expanding

$(1+\epsilon_\ell)^{-1} \approx 1-\epsilon_\ell+\epsilon_\ell^2$ and $\ln(1+\epsilon_\ell) \approx \epsilon_\ell - \frac{1}{2}\epsilon_\ell^2$,

$$\frac{\hat{C}_\ell}{C_\ell^{\rm th}} + \ln\frac{C_\ell^{\rm th}}{\hat{C}_\ell} - 1 \approx (1-\epsilon_\ell+\epsilon_\ell^2) + \left(\epsilon_\ell - \tfrac{1}{2}\epsilon_\ell^2\right) - 1 = \frac{1}{2}\epsilon_\ell^2 = \frac{(C_\ell^{\rm th}-\hat{C}_\ell)^2}{2\hat{C}_\ell^2},$$

so that

$$-2\ln L(\boldsymbol{\lambda}) \approx \sum_\ell (2\ell+1)\frac{(C_\ell^{\rm th}-\hat{C}_\ell)^2}{\hat{C}_\ell^2} = \sum_\ell \frac{(C_\ell^{\rm th}-\hat{C}_\ell)^2}{(\delta C_\ell)^2},$$

using the cosmic-variance formula $(\delta C_\ell)^2 = \frac{2}{2\ell+1}\hat{C}_\ell^2$ (Chapter 6, Eq. (6.4.10); with instrumental noise included, this is the Knox formula, Chapter 7, Eq. (7.5.25)). This is exactly the Gaussian χ^2 of the amplitude fit. The expansion is trustworthy precisely when $2\ell+1$ is large enough for the central limit theorem to make the chi-squared distribution look Gaussian – the same high-ℓ condition already noted in Chapter 6, Eq. (6.4.9) – which is why the low-ℓ multipoles of the real Planck and WMAP likelihoods are always treated with the exact form, Eq. (1), or its further refinements for a partial, noisy, non-uniform sky.

10.4.2 The bispectrum and its shapes

Any model satisfying all four of

- single field,
- canonical kinetic term,
- slow roll, and
- Bunch–Davies (vacuum) initial state,

produces a bispectrum too small to be observed by any conceivable experiment, as already anticipated in §9.7.2. Explicitly, defining B_ζ by Eq. (10.2.6) applied to ζ,

$$\frac{B_\zeta(k_1,k_2,k_3)}{P_\zeta(k_1)P_\zeta(k_2)+P_\zeta(k_2)P_\zeta(k_3)+P_\zeta(k_3)P_\zeta(k_1)} = \mathcal{O}\left(10^{-2}\right) \tag{10.4.5}$$

for this class of models, unobservably small.

If *any* of the four conditions is relaxed, however, inflation can produce a far larger bispectrum. Since $B_\zeta(k_1,k_2,k_3)$ depends on three numbers, imposing scale invariance leaves three independent shape parameters – conventionally k_2/k_1, k_3/k_1, and the angle between $\vec{k}_2$ and $\vec{k}_3$ – so B_ζ can take on many different **shapes**, each pointing to different underlying physics. Three limiting configurations, tied to the four assumptions above, are singled out in the literature:

1. **Squeezed limit**, $k_3 \ll k_2 \approx k_1$: detection would rule out *every* single-field model, regardless of the kinetic term or potential (§10.6).
2. **Equilateral limit**, $k_1 \approx k_2 \approx k_3$: peaks when field interactions are enhanced right at horizon crossing, e.g. from non-canonical kinetic terms (small sound speed $c_s \ll 1$, DBI inflation).
3. **Folded (flattened) limit**, $k_2 \approx k_3 \approx k_1/2$ (with $\vec{k}_1$ collinear with $\vec{k}_2, \vec{k}_3$): a detection would indicate that the initial state of the field fluctuations was *not* the Bunch–Davies vacuum.

4. **More general shapes**: violations of slow roll typically also break scale invariance, producing more intricate features such as an oscillating bispectrum (e.g. from features in the inflaton potential, or from resonance with a periodically modulated coupling).

In the language actually used to report observational constraints (§10.7), these correspond to *templates* $f_{\rm NL}^{\rm local}$, $f_{\rm NL}^{\rm equil}$, and $f_{\rm NL}^{\rm ortho}$ (the last a shape close to, but orthogonal to, the equilateral one, designed to have low correlation with it and to catch effects like a reduced sound speed combined with higher-derivative operators). For the remainder of this chapter we specialize to the squeezed limit and its local-form realization, since it is here that the sharpest, most model-independent theorem is available.

Question. Non-Gaussianity requires interactions (§10.1). But gravity is already non-linear even for a single free scalar field minimally coupled to Einstein gravity – so why is single-field slow-roll non-Gaussianity so absurdly small, $\mathcal{O}(10^{-2})$ in Eq. (10.4.5), rather than $\mathcal{O}(1)$?
Answer. The gravitational non-linearities are indeed always present, but slow roll forces them to be weak for two independent, compounding reasons. First, the strength of the relevant cubic self-coupling in the action for ζ is itself proportional to a slow-roll parameter (schematically $\dot\phi \sim \sqrt{\epsilon} H M_{\rm Pl}$ suppresses every vertex that converts one long-wavelength curvature mode into the next), so the interaction is perturbatively weak by construction: this is the field-theoretic version of the point made in §9.7.2, that the non-linear terms in the equation of motion for $\delta\phi$ only dominate once the potential is curved enough to end slow roll, at which point ζ has already been read off on super-horizon scales. Second, the observable bispectrum is itself proportional to the deviation from exact scale invariance, $B_\zeta/P_\zeta^2 \sim (1-n_s) = \mathcal{O}(\epsilon,\eta)$ (this is exactly the single-field consistency relation of §10.6): a perfectly scale-invariant spectrum would carry *zero* squeezed-limit bispectrum by a symmetry argument (a long-wavelength ζ_L is then locally indistinguishable from a coordinate rescaling and cannot affect any local, physically measurable short-wavelength observable), so the smallness of the bispectrum is tied directly to the smallness of $n_s - 1$, which is itself measured to be $\mathcal{O}(10^{-2})$. Any mechanism that evades either suppression – multiple fields, non-canonical kinetic terms, a non-Bunch–Davies state, or a sharp feature that transiently violates slow roll – can produce an $\mathcal{O}(1)$ or larger bispectrum, which is exactly why a detection of large non-Gaussianity in any shape would be so informative.

10.5 Multi-Field Inflation and the δN Formalism

The most direct way to rule out single-field models is to measure the squeezed bispectrum. To compute what multi-field inflation predicts for it, we use a remarkably powerful and physically transparent tool, the δN **formalism**, where N is the number of e-folds of expansion, $N \equiv \ln[a(t_{\rm obs})/a(t_{\rm ini})]$.

Consider the spatial metric, ignoring gravitational waves,

$$g_{ij} = a^2(t) e^{2\zeta(\vec{q})} \delta_{ij}, \tag{10.5.1}$$

with $\vec{q}$ comoving coordinates, and define an **inhomogeneous scale factor** $\hat{a}(t,\vec{q}) = a(t) e^{\zeta(\vec{q})}$. Choosing the initial time to lie on the spatially flat hypersurface, $g_{ij} = a^2(t)\delta_{ij}$, the local number of e-folds is

$$\hat{N}(\vec{q}) = \ln \frac{\hat{a}(t_{\rm obs})}{a(t_{\rm ini})} = \ln \frac{a(t_{\rm obs})}{a(t_{\rm ini})} + \zeta(\vec{q}) = N + \zeta(\vec{q}), \tag{10.5.2}$$

so that

$$\zeta(\vec{q}) = \delta N(\vec{q}) \equiv \hat{N}(t_{\rm obs}, t_{\rm ini}, \vec{q}) - N(t_{\rm obs}, t_{\rm ini}). \tag{10.5.3}$$

Here $t_{\rm ini}$ is on the flat hypersurface ($g_{ij} = a^2\delta_{ij}$) and $t_{\rm obs}$ on the uniform-density hypersurface ($\delta\rho = 0$). This is the δN formalism: the curvature perturbation on the final uniform-density slice equals the difference in the number of local expansion e-folds accumulated between two different, but everywhere spatially flat, initial patches.

To compute $\delta N(\vec{q})$ we use the **gradient expansion**: systematically drop terms involving spatial derivatives, valid as long as the perturbation wavelength is much larger than the local Hubble radius (corrections enter at relative order $(k/aH)^2$, the same expansion parameter that controls the stochastic-inflation formalism of §9.7.1–§8.7). At zeroth order in this gradient expansion, something remarkable happens: the Friedmann equation for the background, $3M_{\rm Pl}^2\langle H\rangle^2 = \langle\rho\rangle$, holds *unchanged* for the locally perturbed quantities,

$$3M_{\rm Pl}^2\hat{H}^2 = \hat{\rho}. \tag{10.5.4}$$

This is a highly non-trivial statement – it says that on super-Hubble scales, each local Hubble patch evolves *exactly as if it were its own separate, unperturbed FLRW universe* with slightly different initial conditions (Fig. 10.3), the "separate universe" picture. Under this approximation, computing $\hat{N}(\vec{q})$ reduces to solving the ordinary *background* Friedmann equation for different initial field values and reading off the difference:

$$\hat{N}(\vec{q}) = \hat{N}\Big(\phi^I_{\rm ini}(\vec{q}), \dot{\phi}^I_{\rm ini}(\vec{q})\Big) = \int \mathrm{d}t\, \hat{H}\Big(\phi^I_{\rm ini}(\vec{q}), \dot{\phi}^I_{\rm ini}(\vec{q}), t\Big)\,. \tag{10.5.5}$$

For ϕ^I to source ρ (and hence H) appreciably, it must be slow-rolling, so that $\dot{\phi}^I$ is itself a function of ϕ^I and N can be regarded as a function of the initial field values alone. Taylor expanding to second order,

$$\zeta(\vec{q}) = \delta N(\vec{q}) = \sum_I \frac{\partial N}{\partial\phi^I_{\rm ini}}\delta\phi^I_{\rm ini}(\vec{q}) + \frac{1}{2}\sum_{IJ}\frac{\partial^2 N}{\partial\phi^I_{\rm ini}\partial\phi^J_{\rm ini}}\delta\phi^I_{\rm ini}(\vec{q})\delta\phi^J_{\rm ini}(\vec{q}) + \ldots, \tag{10.5.6}$$

completing the formalism. We adopt the shorthand $N_I \equiv \partial N/\partial\phi^I_{\rm ini}$, $N_{IJ} \equiv \partial^2 N/\partial\phi^I_{\rm ini}\partial\phi^J_{\rm ini}$, etc., and drop the subscript "ini" from here on.

10.5.1 Single-field check

As a consistency check, apply δN to a single field $\phi \equiv \phi^1$ and recover the standard linear result $\zeta = -\frac{H}{\dot{\phi}}\Big|_{\rm ini}\delta\phi_{\rm ini}$. We have

$$\hat{N} = \int_{t_{\rm ini}}^{t_{\rm obs}} \mathrm{d}t\, \hat{H} = \int_{\phi_{\rm ini}}^{\phi_{\rm obs}} \mathrm{d}\phi\, \frac{H}{\dot{\phi}}. \tag{10.5.7}$$

Since $t_{\rm obs}$ lies on the uniform-density surface, $\phi_{\rm obs} = \langle\phi\rangle_{\rm obs}$ is unperturbed; since $t_{\rm ini}$ lies on the flat surface, $\phi_{\rm ini} = \langle\phi\rangle_{\rm ini} + \delta\phi_{\rm ini}$ carries the full perturbation. Expanding the lower limit of integration,

$$\hat{N} = \int_{\langle\phi\rangle_{\rm ini}+\delta\phi_{\rm ini}}^{\langle\phi\rangle_{\rm obs}} \mathrm{d}\phi\, \frac{H}{\dot{\phi}} \approx \int_{\langle\phi\rangle_{\rm ini}}^{\langle\phi\rangle_{\rm obs}} \mathrm{d}\phi\, \frac{\langle H\rangle}{\dot{\phi}} \underbrace{- \frac{\langle H\rangle}{\dot{\phi}}\Big|_{\rm ini}\delta\phi_{\rm ini} + \mathcal{O}\Big(\delta\phi^2\Big)}_{\delta N}, \tag{10.5.8}$$

so that

$$\zeta = \delta N = \hat{N} - \langle N\rangle = -\frac{H}{\dot{\phi}}\Big|_{\rm ini}\delta\phi_{\rm ini}, \tag{10.5.9}$$

exactly the standard linear-theory result.

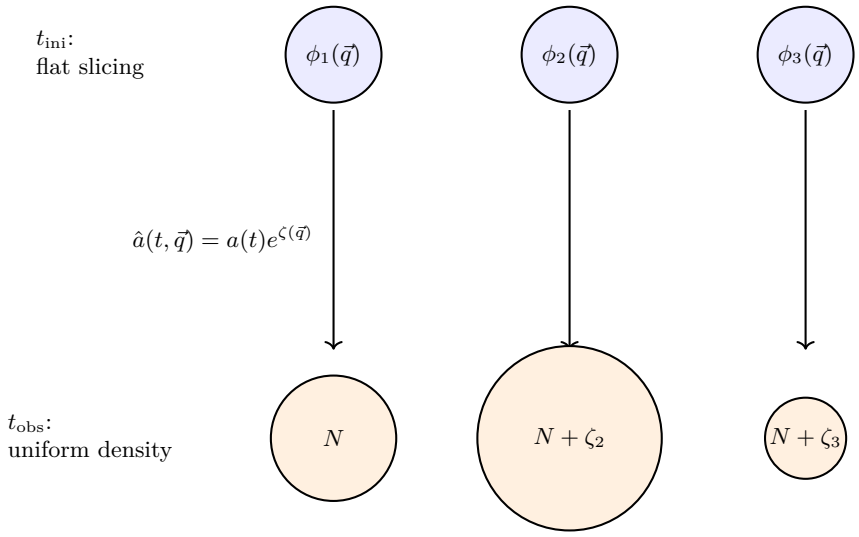


Figure 10.3: The separate-universe picture underlying the δN formalism. On scales far outside the Hubble radius, gradient terms in the Einstein equations are negligible and each comoving patch $\vec{q}$ evolves as its own unperturbed FLRW universe, with local scale factor $\hat{a}(t,\vec{q}) = a(t)e^{\zeta(\vec{q})}$ set by the (slightly different) local initial field values $\phi^I_{\rm ini}(\vec{q})$. The final curvature perturbation $\zeta(\vec{q}) = \delta N(\vec{q})$ is simply the difference in the number of e-folds elapsed in each patch by the time it reaches the same final density.

10.5.2 Local-form non-Gaussianity from δN

Reading Eq. (10.5.6) again: $\zeta(\vec{q})$ is built out of the *same-point* value of $\delta\phi^I(\vec{q})$, so this is called "local-form" non-Gaussianity. It implies immediately that ζ must be non-Gaussian even if every $\delta\phi^I(\vec{q})$ is individually Gaussian: squaring a Gaussian variable does not produce another Gaussian variable. To see this quantitatively, compute the real-space 3-point function,

$$\begin{aligned}\langle\zeta(\vec{q}_1)\zeta(\vec{q}_2)\zeta(\vec{q}_3)\rangle &= N_I N_J N_K \left\langle \delta\phi^I(\vec{q}_1)\delta\phi^J(\vec{q}_2)\delta\phi^K(\vec{q}_3)\right\rangle \\ &\quad + \frac{1}{4}N_I N_J N_{KL}\left\langle \delta\phi^I(\vec{q}_1)\delta\phi^J(\vec{q}_2)\delta\phi^K(\vec{q}_3)\delta\phi^L(\vec{q}_3)\right\rangle + 2 \text{ perms} + \dots \quad (10.5.10)\end{aligned}$$

The first term vanishes for Gaussian $\delta\phi^I$ (Wick's theorem gives no way to pair three variables); the second does not, by exactly the four-point Wick contraction worked out in §10.2.

Bispectrum from δN

In Fourier space, Eq. (10.5.6) reads

$$\zeta_{\vec{k}} = N_I\delta\phi^I_{\vec{k}} + \frac{1}{2}N_{IJ}\int\frac{\mathrm{d}^3p}{(2\pi)^3}\delta\phi^I_{\vec{p}}\delta\phi^J_{\vec{k}-\vec{p}} + \dots \quad (1)$$

Keeping every term at cubic order in $\delta\phi$ – i.e. expanding one of the three $\zeta_{\vec{k}_i}$ to quadratic order while leaving the other two linear, and summing over the three choices of which leg is expanded – gives

$$\langle\zeta_{\vec{k}_1}\zeta_{\vec{k}_2}\zeta_{\vec{k}_3}\rangle = N_I N_J N_K\left\langle\delta\phi^I_{\vec{k}_1}\delta\phi^J_{\vec{k}_2}\delta\phi^K_{\vec{k}_3}\right\rangle + \frac{1}{2}N_I N_J N_{KL}\int\frac{\mathrm{d}^3p}{(2\pi)^3}\left\langle\delta\phi^I_{\vec{k}_1}\delta\phi^J_{\vec{k}_2}\delta\phi^K_{\vec{p}}\delta\phi^L_{\vec{k}_3-\vec{p}}\right\rangle + 2\text{ perms}. \quad (2)$$

Assume the fields are Gaussian with $\left\langle \delta\phi^I_{\vec{k}} \delta\phi^J_{\vec{k}'} \right\rangle = (2\pi)^3 \delta^{(3)}(\vec{k}+\vec{k}') \delta^{IJ} P_\phi(k)$, so the first term drops out and we apply Wick's theorem to the four-point function in the second. Of its three pairings, one forces $\delta^{(3)}(\vec{k}_3)$ (the two $\vec{k}_1, \vec{k}_2$ legs pair with each other and the $\vec{p}, \vec{k}_3 - \vec{p}$ legs pair with each other, forcing their total momentum $\vec{k}_3$ to vanish) and is dropped as it does not contribute to the connected correlator at $\vec{k}_3 \neq 0$; the remaining two pairings both give, after the $\vec{p}$ integral localizes the momenta,

$$\left(\delta^{IK}\delta^{JL} + \delta^{IL}\delta^{JK}\right) (2\pi)^3 \delta^{(3)}(\vec{k}_1 + \vec{k}_2 + \vec{k}_3)\, P_\phi(k_1) P_\phi(k_2). \tag{3}$$

Contracting with $N_I N_J N_{KL}$ and using the symmetry $N_{KL} = N_{LK}$, the prefactor becomes $2 N_I N_J N^{IJ}$; the accompanying $\frac{1}{2}$ then cancels this 2 exactly, leaving, after summing the three permutations (which one of k_1, k_2, k_3 carries the quadratic piece),

$$\langle \zeta_{\vec{k}_1} \zeta_{\vec{k}_2} \zeta_{\vec{k}_3} \rangle = (2\pi)^3 \delta^{(3)}(\vec{k}_1+\vec{k}_2+\vec{k}_3)\, N_I N_J N^{IJ} \Big[P_\phi(k_1)P_\phi(k_2) + P_\phi(k_2)P_\phi(k_3) + P_\phi(k_3)P_\phi(k_1) \Big]. \tag{4}$$

The power spectrum of ζ follows the same way at quadratic order,

$$\langle \zeta_{\vec{k}_1} \zeta^\star_{\vec{k}_2} \rangle = N_I N_J \left\langle \delta\phi^I_{\vec{k}_1}\, \delta\phi^{J\star}_{\vec{k}_2} \right\rangle = (2\pi)^3 \delta^{(3)}(\vec{k}_1 - \vec{k}_2)\, N_I N^I P_\phi(k_1) \quad \Longrightarrow \quad P_\zeta(k) = N_I N^I P_\phi(k). \tag{10.5.11}$$

Combining Eqs. (4)–(10.5.11) with the bispectrum definition, Eq. (10.2.6),

$$\frac{B_\zeta(k_1,k_2,k_3)}{P_\zeta(k_1)P_\zeta(k_2) + \text{perm.}} = \frac{N_I N_J N^{IJ}}{(N_K N^K)^2}. \tag{10.5.12}$$

In the single-field limit ($\phi \equiv \phi^1$), this collapses to

$$\frac{B_\zeta(k_1,k_2,k_3)}{P_\zeta(k_1)P_\zeta(k_2) + \text{perm.}} = \frac{N''}{(N')^2}, \qquad N' \equiv \frac{\partial N}{\partial \phi_{\rm ini}} = -\frac{H}{\dot\phi}\Big|_{\rm ini} = -\frac{1}{M_{\rm Pl}\sqrt{2\epsilon}}, \tag{10.5.13}$$

using $\epsilon \equiv \frac{1}{2M_{\rm Pl}^2}(\dot\phi/H)^2$. A second differentiation with respect to $\phi_{\rm ini}$, together with the standard slow-roll definitions of ϵ and η used in Chapter 9, gives after some algebra the well-known result (Maldacena 2002)

$$\frac{N''}{(N')^2} = 2\epsilon - \eta, \tag{10.5.14}$$

so that for single-field slow-roll inflation with Gaussian $\delta\phi$, $B_\zeta/P_\zeta^2 = \mathcal{O}(2\epsilon-\eta) = \mathcal{O}(10^{-2})$: exactly the smallness already quoted in Eq. (10.4.5) and in §9.7.2, now derived from first principles via δN. A single-field detection of this quantity is therefore only possible in principle if some other mechanism (e.g. higher-derivative interactions in the equilateral or orthogonal shapes) enhances it far beyond this generic slow-roll suppression.

10.5.3 The squeezed bispectrum

Because $P_\zeta(k) \propto k^{-3}$ is nearly scale invariant, Eq. (10.5.12) peaks whenever one of the k_i is far smaller than the other two. Ordering $k_1 \geq k_2 \geq k_3$, the local-form bispectrum therefore peaks in the squeezed configuration $k_3 \ll k_2 \approx k_1$, where

$$B_\zeta(k_1,k_2,k_3 \to 0) \to \frac{2 N_I N_J N^{IJ}}{(N_K N^K)^2} P_\zeta(k_1) P_\zeta(k_3) = 2(2\epsilon - \eta)\, P_\zeta(k_1) P_\zeta(k_3) \qquad \text{(single field)}. \tag{10.5.15}$$

10.6 The Single-Field Consistency Theorem

The squeezed-limit bispectrum is far more powerful than ruling out single-field *slow-roll* models alone: it rules out *every* single-field model, regardless of the potential or the kinetic term. This **single-field consistency theorem** states

$$\lim_{k_3\to 0} B_\zeta(k_1,k_2,k_3) = (1-n_s)\,P_\zeta(k_1)P_\zeta(k_3) \tag{10.6.1}$$

for *all* single-field models. We derive it here by a slightly non-standard but physically transparent route.

Start again from the exact cubic-order expression,

$$\langle\zeta_{\vec{k}_1}\zeta_{\vec{k}_2}\zeta_{\vec{k}_3}\rangle = N_I N_J N_K\langle\delta\phi^I_{\vec{k}_1}\delta\phi^J_{\vec{k}_2}\delta\phi^K_{\vec{k}_3}\rangle + N_I N_J N^{IJ}(2\pi)^3\delta^{(3)}(\vec{k}_1+\vec{k}_2+\vec{k}_3)[P_\phi(k_1)P_\phi(k_2)+\text{perm.}]. \tag{10.6.2}$$

For a single field, the second (local-form) term is already $\mathcal{O}(\epsilon,\eta)$, but this is precisely the order at which we must now also keep the *first* term: unlike in the naive Gaussian-$\delta\phi$ treatment of §10.5.2, $\delta\phi$ itself develops a small intrinsic non-Gaussianity from its self-interactions during inflation, and that effect is of the same order in slow-roll as the local-form piece above. Its full, self-consistent computation requires genuine interacting quantum field theory in curved spacetime and was first carried out by Maldacena (2003); here we extract only its *squeezed* limit, $k_3 \equiv k_L \ll k_1 \approx k_2 \equiv k_S$, using the separate-universe picture of §10.5.

In this limit, the long mode $\zeta_L = \zeta_{\vec{k}_L}$ is far outside the horizon while the two short modes are still deep inside it, so from the point of view of the short-wavelength physics, the long mode acts purely as a background: it rescales the local spatial metric by $g_{ij} \to a^2(t)e^{2\zeta_L}\delta_{ij}$, i.e. it is locally indistinguishable from a constant rescaling of comoving coordinates, $\vec{x} \to e^{\zeta_L}\vec{x}$. We can therefore write the intrinsic three-point function as a two-point function of the short modes, computed in the presence of a background long mode, further correlated with that long mode:

$$\langle\delta\phi_{\vec{k}_1}\delta\phi_{\vec{k}_2}\delta\phi_{\vec{k}_L}\rangle = \left\langle\left\langle\delta\phi_{\vec{k}_S}\delta\phi_{-\vec{k}_S}\right\rangle_{\zeta_L}\delta\phi_{\vec{k}_L}\right\rangle. \tag{10.6.3}$$

A coordinate rescaling $\vec{x} \to e^{\zeta_L}\vec{x}$ sends a comoving wavenumber $k_S \to k_S e^{-\zeta_L}$, so the short-mode power spectrum measured in the presence of the long-wavelength background is simply the fiducial spectrum evaluated at the rescaled wavenumber. Because $d^3k \to e^{-3\zeta_L}d^3k$ under the same rescaling, it is the *dimensionless* combination $k^3P_\phi(k)$ that must be evaluated at the shifted argument; expanding to linear order in ζ_L,

$$\boxed{\left\langle\delta\phi_{\vec{k}_S}\delta\phi_{-\vec{k}_S}\right\rangle_{\zeta_L} = P_\phi(k_S) - \frac{\mathrm{d}\ln[k_S^3P_\phi(k_S)]}{\mathrm{d}\ln k_S}\,P_\phi(k_S)\,\zeta_L.} \tag{10.6.4}$$

For a nearly massless field in (quasi-)de Sitter space, $P_\phi(k) \propto H_*^2/k^3$ evaluated at horizon crossing $k = aH$, so $k^3P_\phi(k) \propto H_*(k)^2$ and, using $\mathrm{d}N = \mathrm{d}\ln k$ for a slowly varying H,

$$\frac{\mathrm{d}\ln[k^3P_\phi(k)]}{\mathrm{d}\ln k} = \frac{\mathrm{d}\ln H^2}{\mathrm{d}N} = \frac{\mathrm{d}\ln H^2}{\mathrm{d}\phi}\frac{\mathrm{d}\phi}{\mathrm{d}N} = -2\epsilon, \tag{10.6.5}$$

since $\dot{H} = -\epsilon H^2 \Rightarrow \mathrm{d}\ln H/\mathrm{d}t = -\epsilon H \Rightarrow \mathrm{d}\ln H^2/\mathrm{d}N = -2\epsilon$ ($N = \int H\mathrm{d}t$). Substituting Eq. (10.6.5) into Eq. (10.6.4) and converting back to ζ using $\zeta = N'\delta\phi$ (so that three powers of N' convert the $\delta\phi$ correlator into a ζ correlator, and the surviving ζ_L leg cancels one factor of N' against the definition of P_ζ),

$$\langle\zeta_{\vec{k}_1}\zeta_{\vec{k}_2}\zeta_{\vec{k}_3}\rangle_{\text{intrinsic}} \xrightarrow{k_3\to 0} (2\pi)^3\delta^{(3)}(\vec{k}_1+\vec{k}_2+\vec{k}_3)\times 2\epsilon\,P_\zeta(k_1)P_\zeta(k_3). \tag{10.6.6}$$

Adding this to the squeezed limit of the local-form term, Eq. (10.5.15),

$$\langle\zeta_{\vec{k}_1}\zeta_{\vec{k}_2}\zeta_{\vec{k}_3}\rangle \xrightarrow{k_3\to 0} (2\pi)^3\delta^{(3)}(\vec{k}_1+\vec{k}_2+\vec{k}_3)\times\Big\{2\epsilon\,P_\zeta(k_1)P_\zeta(k_3)+(4\epsilon-2\eta)\,P_\zeta(k_1)P_\zeta(k_3)\Big\}$$
$$=(2\pi)^3\delta^{(3)}(\vec{k}_1+\vec{k}_2+\vec{k}_3)\times(6\epsilon-2\eta)\,P_\zeta(k_1)P_\zeta(k_3). \tag{10.6.7}$$

Using the standard slow-roll relation for the tilt, $n_s-1=2\eta-6\epsilon$, the bracket is exactly $1-n_s$, reproducing Eq. (10.6.1). Because Eq. (10.6.1) was derived using *only* the separate-universe argument – valid for any single dynamical degree of freedom whatsoever, not just slow-roll potentials or canonical kinetic terms – the result holds for the entire class of single-field models. Any measurement of the squeezed-limit bispectrum in excess of $1-n_s=\mathcal{O}(10^{-2})$ would therefore rule out single-field inflation outright, independent of the potential, the sound speed, or any other model-building freedom. (Note this is exactly the same separate-universe logic that underlies the Fokker–Planck / stochastic-inflation treatment of super-horizon field diffusion in §9.7.1: both results follow from treating each Hubble patch as its own background FLRW universe.)

10.7 The Non-linearity Parameter $f_{\rm NL}$

It is conventional to characterize the amplitude of B_ζ/P_ζ^2 by a single **non-linearity parameter** $f_{\rm NL}$,

$$\frac{6}{5}f_{\rm NL}\equiv\frac{B_\zeta(k_1,k_2,k_3)}{P_\zeta(k_1)P_\zeta(k_2)+P_\zeta(k_1)P_\zeta(k_3)+P_\zeta(k_2)P_\zeta(k_3)}. \tag{10.7.1}$$

The awkward factor of 6/5 is a historical artifact: $f_{\rm NL}$ was originally defined for the Newtonian-gauge potential $\Phi=\frac{3}{5}\zeta$ during matter domination via $2f_{\rm NL}\equiv B_\Phi/(P_\Phi P_\Phi+\text{perm.})$ (this is Eq. (9.7.32) of Chapter 9), and the 5/3 rescaling between Φ and ζ combined with the definition of B as cubic in the field turns the front factor of 2 into 6/5.

The very first WMAP 7-year bispectrum analysis (Komatsu et al. 2011) gave $f_{\rm NL}=32\pm21$ (68% CL), i.e. $-10<f_{\rm NL}<74$ at 95% CL – a mild, ultimately statistically insignificant hint of non-Gaussianity that motivated a great deal of theoretical work on multi-field models. The final Planck full-mission (temperature + polarization) analysis subsequently pinned local-type non-Gaussianity down to (Planck 2018 IX; cf. Eq. (9.7.36) of Chapter 9)

$$f_{\rm NL}^{\rm local}=-0.9\pm5.1,\qquad f_{\rm NL}^{\rm equil}=-26\pm47,\qquad f_{\rm NL}^{\rm ortho}=-38\pm24\qquad(68\%\text{ CL}), \tag{10.7.2}$$

consistent with zero in every shape: no evidence for primordial non-Gaussianity of any kind has been found, and the single-field slow-roll prediction $f_{\rm NL}=\mathcal{O}(\epsilon,\eta)\sim10^{-2}$ remains many years away from being directly testable by CMB bispectrum measurements alone (though large-scale-structure surveys, via the scale-dependent halo bias sourced by $f_{\rm NL}^{\rm local}$, are expected to eventually close this gap).

10.8 Trispectrum and the Multi-Field Consistency Relation

If a large $f_{\rm NL}$ *were* found, single-field models would be excluded, but not necessarily every multi-field model along with them. Is there a further test that can distinguish among multi-field scenarios, or even rule some of them out? The four-point function (**trispectrum**) provides exactly this. Extending the local-form ansatz to cubic order,

$$\zeta=N_I\delta\phi^I+\frac{1}{2}N_{IJ}\delta\phi^I\delta\phi^J+\frac{1}{6}N_{IJK}\delta\phi^I\delta\phi^J\delta\phi^K+\ldots, \tag{10.8.1}$$

the connected four-point function receives two independent leading contributions, one from two insertions of the quadratic (N_{IJ}) term and one from a single insertion of the cubic (N_{IJK}) term paired against three linear legs:

$$\tau_{\rm NL}:\ (1\text{st})\times(1\text{st})\times(2\text{nd})\times(2\text{nd}), \tag{10.8.2}$$

$$g_{\rm NL}:\ \text{(1st)}\times\text{(1st)}\times\text{(1st)}\times\text{(3rd)}. \tag{10.8.3}$$

Explicitly (Byrnes, Sasaki & Wands 2006),

$$\langle\zeta_{\vec{k}_1}\zeta_{\vec{k}_2}\zeta_{\vec{k}_3}\zeta_{\vec{k}_4}\rangle = (2\pi)^3\delta^{(3)}(\vec{k}_1+\vec{k}_2+\vec{k}_3+\vec{k}_4)$$
$$\times\left[\tau_{\rm NL}P_\zeta(k_1)P_\zeta(k_2)\{P_\zeta(|\vec{k}_1+\vec{k}_3|)+P_\zeta(|\vec{k}_1+\vec{k}_4|)\}+10\text{ perm}+g_{\rm NL}\frac{54}{25}[P_\zeta(k_1)P_\zeta(k_2)P_\zeta(k_3)+3\text{ perm}]\right], \tag{10.8.4}$$

where $\tau_{\rm NL}$ peaks in the "small-diagonal" limit $|\vec{k}_1+\vec{k}_3|\to 0$ and $g_{\rm NL}$ peaks in the squeezed limit; the historical factor $54/25$ in front of $g_{\rm NL}$ plays the same role as the $6/5$ in front of $f_{\rm NL}$. To leading order,

$$\tau_{\rm NL}=\frac{(N_{IJ}N^J)(N^{IK}N_K)}{(N_LN^L)^3},\qquad \frac{54}{25}g_{\rm NL}=\frac{N_{IJK}N^IN^JN^K}{(N_LN^L)^3} \tag{10.8.5}$$

(if these leading-order combinations vanish by symmetry in a given model, one must go to the next order in the field-space Taylor expansion).

10.8.1 The Suyama–Yamaguchi inequality

At this order, both $\tau_{\rm NL}$ and $f_{\rm NL}$ are built purely from N_I and N_{IJ} (recall $\frac{6}{5}f_{\rm NL}=N_{IJ}N^IN^J/(N_LN^L)^{3/2}$ from Eq. (10.5.12)), so an inequality relating them should exist. Defining

$$a_I\equiv\frac{N_{IJ}N^J}{(N_LN^L)^{3/2}},\qquad b_I\equiv\frac{N_I}{(N_LN^L)^{1/2}}, \tag{10.8.6}$$

we find $\frac{6}{5}f_{\rm NL}=a_Ib^I$ and $\tau_{\rm NL}=(a_Ia^I)(b_Jb^J)$. The Cauchy–Schwarz inequality $(a_Ia^I)(b_Jb^J)\geq(a_Ib^I)^2$ then gives the (tree-level) **Suyama–Yamaguchi inequality** (Suyama & Yamaguchi, PRD **77**, 023505 (2008)):

$$\tau_{\rm NL}\geq\left(\frac{6}{5}f_{\rm NL}\right)^2. \tag{10.8.7}$$

Is Eq. (10.8.7) general?

The derivation above relies on the tree-level approximation, where the bispectrum comes from (1st) × (1st) × (2nd) and $\tau_{\rm NL}$ from (1st) × (1st) × (2nd) × (2nd). In models where these particular contractions vanish for symmetry reasons, the leading contributions may instead arise at the next order, e.g. $f_{\rm NL}$ from (2nd) × (2nd) × (2nd) and $\tau_{\rm NL}$ from (2nd) × (2nd) × (2nd) × (2nd), for which the simple Cauchy–Schwarz argument above no longer directly applies. Nonetheless, subsequent work (Sugiyama, Komatsu & Futamase 2011; Kehagias & Riotto 2012) has argued that $\tau_{\rm NL}\geq\left(\frac{6}{5}f_{\rm NL}\right)^2$ continues to hold on quite general grounds for multi-field inflation, making it a genuinely fundamental (not merely tree-level) test of the mechanism generating the curvature perturbation: a confirmed violation of Eq. (10.8.7) would point to physics beyond the standard multi-field inflationary paradigm entirely (e.g. non-attractor behaviour, or the fluctuations not being sourced purely during inflation).

Observationally, the trispectrum remains far less constrained than the bispectrum: the Planck 2018 local-shape analysis gives $\tau_{\rm NL}=(-5.8\pm6.5)\times10^4$ (68% CL statistical), consistent with zero but still allowing $\tau_{\rm NL}$ of order 10^4–10^5 – several orders of magnitude weaker than the bound $\left(\frac{6}{5}f_{\rm NL}\right)^2\sim(6)^2\approx36$ implied by saturating Eq. (10.7.2), so the Suyama–Yamaguchi inequality is not yet a competitive test in practice, though it will become one as future CMB and large-scale-structure surveys improve trispectrum sensitivity.

10.9 Measuring Non-Gaussianity from the CMB

How is a number like $f_{\rm NL} = -0.9 \pm 5.1$ actually extracted from a temperature map? It is instructive to work out the simplest possible analogue first: estimating the skewness $k_3 = \langle x^3 \rangle$ of a single, zero-mean, unit-variance random variable, before generalizing to the full $a_{\ell m}$ trispectrum estimator used by WMAP and Planck (Komatsu & Spergel 2001; Komatsu et al. 2005).

10.9.1 Toy model: estimating a skewness

Recall the Gram–Charlier PDF truncated at k_3,

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \left[1 + \frac{k_3}{6} \left(\frac{x^3}{\sigma^6} - 3\frac{x}{\sigma^4} \right) + \dots \right]. \tag{10.9.1}$$

The best-fit k_3 is found by maximizing the likelihood of the data with respect to k_3 (implicitly using Bayes' theorem to interpret the likelihood $P(x|k_3)$ as a posterior for k_3 given a flat prior), i.e. solving $\left\langle \frac{\mathrm{d}\ln P(x)}{\mathrm{d}k_3} \right\rangle = 0$. Expanding $\ln P(x)$ to $\mathcal{O}(k_3^2)$,

$$\ln P(x) = \text{const} + \frac{1}{6}\left(\frac{x^3}{\sigma^6} - 3\frac{x}{\sigma^4} \right) k_3 - \frac{1}{2}\left[\frac{1}{6}\left(\frac{x^3}{\sigma^6} - 3\frac{x}{\sigma^4} \right) \right]^2 k_3^2 + \dots, \tag{10.9.2}$$

so that $\frac{\mathrm{d}\ln P}{\mathrm{d}k_3} = 0$ yields

$$\left\langle \frac{1}{36}\left(\frac{x^6}{\sigma^{12}} - 6\frac{x^4}{\sigma^{10}} + 9\frac{x^2}{\sigma^8} \right) \right\rangle k_3 = \left\langle \frac{1}{6}\left(\frac{x^3}{\sigma^6} - 3\frac{x}{\sigma^4} \right) \right\rangle. \tag{10.9.3}$$

Using $\langle x^6 \rangle = 15\sigma^6$ and $\langle x^4 \rangle = 3\sigma^4$ (Wick's theorem again), the left-hand side coefficient reduces to $\frac{1}{6}\sigma^{-6}$, so that the estimator is simply

$$\boxed{k_3 = \left\langle x^3 \right\rangle - 3\sigma^2 \left\langle x \right\rangle,} \tag{10.9.4}$$

as one would naively expect, reducing to $k_3 = \langle x^3 \rangle$ for a zero-mean variable. Since the ensemble average is never actually available from a single data set, the practical estimator (weighting each of N independent measurements x_i by its own variance σ_i^2) is

$$k_3 = \frac{\frac{1}{6}\sum_i \left(\frac{x_i^3}{\sigma_i^6} - 3\frac{x_i}{\sigma_i^4} \right)}{\frac{1}{6}\sum_i \sigma_i^{-6}}. \tag{10.9.5}$$

This estimator, seemingly a toy problem, has an essentially one-to-one correspondence with the actual $f_{\rm NL}$ estimator used on WMAP and Planck data, as we now show.

10.9.2 PDF of the CMB temperature field

We need the multivariate generalization of the Gram–Charlier/Edgeworth expansion to the temperature anisotropy field $\delta T(\hat{n})$, obeying (Gaussian case)

$$P(\{\delta T(\hat{n}_i)\}) = \frac{1}{(2\pi)^{N_{\rm pix}/2}|\xi|^{1/2}} e^{-\frac{1}{2}\sum_{ij}\delta T(\hat{n}_i)(\xi^{-1})_{ij}\delta T(\hat{n}_j)}, \qquad \xi_{ij} = \langle \delta T(\hat{n}_i)\delta T(\hat{n}_j)\rangle. \tag{10.9.6}$$

Because we deal with a field on the two-sphere rather than flat space, the natural Fourier-type basis is the spherical harmonic transform, $\delta T(\hat{n}) = \sum_{\ell m} a_{\ell m} Y_{\ell m}(\hat{n})$, $a_{\ell m} = \int \mathrm{d}\hat{n}\, \delta T(\hat{n}) Y^\star_{\ell m}(\hat{n})$, in

terms of which the Gaussian PDF becomes

$$P(\{a_{\ell m}\}) = \frac{1}{(2\pi)^{N_{\rm harm}/2}|C|^{1/2}} \exp\left(-\frac{1}{2}\sum_{\ell m,\ell' m'} a^{\star}_{\ell m}(C^{-1})_{\ell m,\ell' m'} a_{\ell' m'}\right), \qquad C_{\ell m,\ell' m'} = \langle a^{\star}_{\ell m} a_{\ell' m'}\rangle\,. \tag{10.9.7}$$

For a translationally and rotationally invariant sky, $C_{\ell m,\ell' m'} = C_\ell \delta_{\ell\ell'}\delta_{mm'}$; however, since real instrumental noise and Galactic foregrounds are neither statistically isotropic nor homogeneous, the estimator below is kept general in $C_{\ell m,\ell' m'}$ without assuming this diagonal form until the final simplification.

Multivariate Edgeworth expansion and the $f_{\rm NL}$ estimator

Taylor-expanding the Gaussian PDF of $\{a_{\ell m}\}$ to include a skewness term,

$$P(\{a_{\ell m}\}) = \frac{1}{(2\pi)^{N_{\rm harm}/2}|C|^{1/2}} \exp\left(-\frac{1}{2}\sum_{\ell m,\ell' m'} a^{*}_{\ell m}(C^{-1})_{\ell m,\ell' m'} a_{\ell' m'}\right) \times\left\{1+\frac{1}{6}\sum_{\text{all }\ell m}\langle a_{\ell_1 m_1}a_{\ell_2 m_2}a_{\ell_3 m_3}\rangle\left[(C^{-1}a)_{\ell_1 m_1}(C^{-1}a)_{\ell_2 m_2}(C^{-1}a)_{\ell_3 m_3} - 3(C^{-1})_{\ell_1 m_1,\ell_2 m_2}(C^{-1}a)_{\ell_3 m_3}\right]+\dots\right\} \tag{10.9.8}$$

which is the exact multivariate analogue of the single-variable Edgeworth PDF, Eq. (10.2.15), truncated at cubic order:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}\left[1+\frac{k_3}{6}\left(\frac{x^3}{\sigma^6}-3\frac{x}{\sigma^4}\right)+\dots\right], \tag{10.9.9}$$

where the same combination $x^3/\sigma^6 - 3x/\sigma^4 = (C^{-1}x)^3 - 3(C^{-1})(C^{-1}x)$ appears with $C \to \sigma^2$. The full bispectrum $\langle a_{\ell_1 m_1}a_{\ell_2 m_2}a_{\ell_3 m_3}\rangle$, or its rotationally-averaged (m-summed) version, the "angle-averaged bispectrum"

$$B_{\ell_1\ell_2\ell_3} \equiv \sum_{\text{all } m}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3\\ m_1 & m_2 & m_3\end{pmatrix}\langle a_{\ell_1 m_1}a_{\ell_2 m_2}a_{\ell_3 m_3}\rangle\,, \tag{10.9.10}$$

could in principle be measured directly, but at the low signal-to-noise available for any single triangle (ℓ_1,ℓ_2,ℓ_3) it is far more powerful to fit the whole data set to a fixed theoretical **template shape** $\mathcal{B}_{\ell_1\ell_2\ell_3}$, and constrain only its overall amplitude, e.g. $B_{\ell_1\ell_2\ell_3} = f_{\rm NL}\mathcal{B}_{\ell_1\ell_2\ell_3}$.

Deriving the $f_{\rm NL}$ estimator

By statistical isotropy, the bispectrum in m-space must have the form $\langle a_{\ell_1 m_1}a_{\ell_2 m_2}a_{\ell_3 m_3}\rangle = \begin{pmatrix}\ell_1 & \ell_2 & \ell_3\\ m_1 & m_2 & m_3\end{pmatrix} B_{\ell_1\ell_2\ell_3}$, where the Wigner $3j$ symbol carries all the m-dependence and $B_{\ell_1\ell_2\ell_3}$ depends only on the multipoles. Squaring the $3j$ symbol and summing over m_1, m_2, m_3 gives 1, so multiplying both sides by the $3j$ symbol and summing over m reproduces the definition of $B_{\ell_1\ell_2\ell_3}$ above: it already is this reduced bispectrum. The model we want to fit is a single overall amplitude $f_{\rm NL}$ times a fixed template $\mathcal{B}_{\ell_1\ell_2\ell_3}$, i.e. $B_{\ell_1\ell_2\ell_3} = f_{\rm NL}\mathcal{B}_{\ell_1\ell_2\ell_3}$, so that

$$\langle a_{\ell_1 m_1}a_{\ell_2 m_2}a_{\ell_3 m_3}\rangle = f_{\rm NL}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3\\ m_1 & m_2 & m_3\end{pmatrix}\mathcal{B}_{\ell_1\ell_2\ell_3}. \tag{1}$$

To keep the algebra short, collect everything that multiplies $\langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3}\rangle$ inside the correction term of Eq. (10.9.8) into one symbol,

$$\mathcal{A}[a] \equiv \frac{1}{6}\sum_{\text{all }\ell m}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3\\ m_1 & m_2 & m_3\end{pmatrix}\mathcal{B}_{\ell_1\ell_2\ell_3}\Big[(C^{-1}a)_{\ell_1 m_1}(C^{-1}a)_{\ell_2 m_2}(C^{-1}a)_{\ell_3 m_3} - 3(C^{-1})_{\ell_1 m_1,\ell_2 m_2}(C^{-1}a)_{\ell_3 m_3}\Big], \tag{2}$$

which is a number computed purely from the data $a_{\ell m}$, the fixed template $\mathcal{B}$, and the fixed covariance C – it does not know about $f_{\rm NL}$ yet. With $B_{\ell_1\ell_2\ell_3} = f_{\rm NL}\mathcal{B}_{\ell_1\ell_2\ell_3}$ substituted in, Eq. (10.9.8) reads $P(\{a_{\ell m}\}) = G\,(1 + f_{\rm NL}\mathcal{A}[a])$, with G the Gaussian part. Using $\ln(1+x) \approx x - \tfrac{1}{2}x^2$ for small x,

$$\ln L(f_{\rm NL}) = \ln G + \ln\left(1 + f_{\rm NL}\mathcal{A}[a]\right) \approx \text{const} + f_{\rm NL}\,\mathcal{A}[a] - \frac{1}{2}f_{\rm NL}^2\,\mathcal{A}[a]^2. \tag{3}$$

This is a constant, plus a term linear in $f_{\rm NL}$, plus a term quadratic in $f_{\rm NL}$, so $\ln L(f_{\rm NL})$ is maximized where its derivative with respect to $f_{\rm NL}$ vanishes. Differentiating term by term,

$$\frac{\mathrm{d}\ln L}{\mathrm{d}f_{\rm NL}} = \mathcal{A}[a] - f_{\rm NL}\,\mathcal{A}[a]^2. \tag{4}$$

One difference from the earlier examples: $\mathcal{A}[a]^2$ is built from the same noisy sky map $a_{\ell m}$ as $\mathcal{A}[a]$ itself, so it fluctuates from sky to sky. Averaging this equation over many realizations of the noise (at fixed $f_{\rm NL}$) replaces $\mathcal{A}[a]^2$ by its average value $\langle\mathcal{A}[a]^2\rangle$, a fixed number that no longer depends on which particular sky we happened to observe:

$$\left\langle\frac{\mathrm{d}\ln L}{\mathrm{d}f_{\rm NL}}\right\rangle = \mathcal{A}[a] - f_{\rm NL}\left\langle\mathcal{A}[a]^2\right\rangle = 0, \tag{5}$$

so that

$$\hat{f}_{\rm NL} = \frac{\mathcal{A}[a]}{\langle\mathcal{A}[a]^2\rangle}. \tag{6}$$

It remains to work out $\langle\mathcal{A}[a]^2\rangle$. $\mathcal{A}[a]$ is built from products of three and of one factor of a; squaring it and taking the expectation value pairs up the a's using Wick's theorem, exactly as in §10.2. Pairing an a from one triangle $(\ell_1\ell_2\ell_3)$ with an a from the *same* triangle reproduces a piece that looks like $\mathcal{A}[a]$ itself evaluated with two of its own legs tied together; the built-in $-3(C^{-1})(C^{-1}a)$ term inside $\mathcal{A}[a]$ is there precisely to cancel every such term. What survives is only the pairing where each of the three legs of one triangle $(\ell_1\ell_2\ell_3)$ is tied to one leg of a second, independent triangle $(\ell_1'\ell_2'\ell_3')$, contributing one factor of C^{-1} per pair of legs and one further sum over $(\ell' m')$. Carrying this out gives

$$f_{\rm NL} = \frac{\frac{1}{6}\displaystyle\sum_{\text{all }\ell m}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3\\ m_1 & m_2 & m_3\end{pmatrix}\mathcal{B}_{\ell_1\ell_2\ell_3}\left[(C^{-1}a)_{\ell_1 m_1}(C^{-1}a)_{\ell_2 m_2}(C^{-1}a)_{\ell_3 m_3} - 3(C^{-1})_{\ell_1 m_1,\ell_2 m_2}(C^{-1}a)_{\ell_3 m_3}\right]}{\frac{1}{6}\displaystyle\sum_{\text{all }\ell m}\sum_{\text{all }\ell' m'}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3\\ m_1 & m_2 & m_3\end{pmatrix}\mathcal{B}_{\ell_1\ell_2\ell_3}(C^{-1})_{\ell_1 m_1,\ell_1' m_1'}(C^{-1})_{\ell_2 m_2,\ell_2' m_2'}(C^{-1})_{\ell_3 m_3,\ell_3' m_3'}\mathcal{B}_{\ell_1'\ell_2'\ell_3'}\begin{pmatrix}\ell_1' & \ell_2' & \ell_3'\\ m_1' & m_2' & m_3'\end{pmatrix}} \tag{10.9.11}$$

For diagonal C_ℓ, this collapses to

$$f_{\rm NL} = \frac{\frac{1}{6}\displaystyle\sum_{\text{all }\ell m}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3\\ m_1 & m_2 & m_3\end{pmatrix}\mathcal{B}_{\ell_1\ell_2\ell_3}\left[\frac{a_{\ell_1 m_1}}{C_{\ell_1}}\frac{a_{\ell_2 m_2}}{C_{\ell_2}}\frac{a_{\ell_3 m_3}}{C_{\ell_3}} - 3\delta_{\ell_1\ell_2}\delta_{m_1,-m_2}\frac{a_{\ell_3 m_3}}{C_{\ell_1}C_{\ell_2}}\right]}{\frac{1}{6}\displaystyle\sum_{\text{all }\ell m}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3\\ m_1 & m_2 & m_3\end{pmatrix}^2\frac{\mathcal{B}_{\ell_1\ell_2\ell_3}^2}{C_{\ell_1}C_{\ell_2}C_{\ell_3}}}, \tag{10.9.12}$$

which is exactly the toy skewness estimator of Eq. (10.9.5), with the sum over independent measurements i replaced by a sum over triangles of multipoles weighted by the theoretical template $\mathcal{B}_{\ell_1\ell_2\ell_3}$ and by the inverse variance $1/C_\ell$ of each leg. This is (up to the more sophisticated skew-C_ℓ and needlet-based refinements used in practice to make the numerator computationally tractable at Planck's $\ell_{\max}\sim 2000$) the estimator used to derive Eq. (10.7.2).

10.9.3 Relating $a_{\ell m}$ to $\zeta_{\vec{k}}$

To turn a theoretical prediction for $B_\zeta(k_1,k_2,k_3)$ into a template $\mathcal{B}_{\ell_1\ell_2\ell_3}$ that can be inserted into Eq. (10.9.11), we need the map from the observed $a_{\ell m}$ back to the primordial $\zeta_{\vec{k}}$. In the (flat-sky) Sachs–Wolfe limit, on a small patch of sky centered on $\hat{z}$, $\delta T(\hat{n})/T = -\frac{1}{5}\zeta(\hat{n}r_\star, z_\star)$ is manifestly just a slice through the three-dimensional field $\zeta(\vec{q})$. Fourier transforming on the two-dimensional patch, $\delta\tilde{T}(\vec{l})/T = \int \mathrm{d}^2\theta\, e^{-i\vec{l}\cdot\vec{\theta}}\,\delta T(\vec{\theta})/T$, and writing $\hat{n}\approx(\vec{\theta},1)$, $\vec{k}=(\vec{k}_\perp,k_\parallel)$,

$$\frac{\delta\tilde{T}}{T}(\vec{l}) = \int \mathrm{d}^2\theta\, e^{-i\vec{l}\cdot\vec{\theta}} \int \frac{\mathrm{d}^3k}{(2\pi)^3} e^{i\vec{k}_\perp\cdot\vec{\theta}r_\star} e^{ik_\parallel r_\star}\left(-\frac{1}{5}\tilde{\zeta}_{\vec{k}}\right) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}(2\pi)^2\delta^{(2)}(\vec{k}_\perp r_\star - \vec{l})\, e^{ik_\parallel r_\star}\left(-\frac{1}{5}\tilde{\zeta}_{\vec{k}}\right), \tag{10.9.13}$$

the $\vec{\theta}$ integral producing a two-dimensional Dirac delta, so that

$$\frac{\delta\tilde{T}}{T}(\vec{l}) = \frac{1}{r_\star^2}\int \frac{\mathrm{d}k_\parallel}{2\pi} e^{ik_\parallel r_\star}\left[-\frac{1}{5}\tilde{\zeta}\left(\vec{k}_\perp = \frac{\vec{l}}{r_\star}, k_\parallel\right)\right]. \tag{10.9.14}$$

Only the component of $\vec{k}$ perpendicular to the line of sight survives cleanly; $k_\parallel$ is integrated (smeared) over, which is the fundamental sense in which the CMB, being a two-dimensional projection, carries less statistical power than the full three-dimensional density field probed by large-scale structure.

Full-sky treatment

On the full sky we must use spherical harmonics throughout. In the Sachs–Wolfe limit,

$$a_{\ell m} = \int \mathrm{d}\hat{n}\, Y^\star_{\ell m}(\hat{n}) \int \frac{\mathrm{d}^3k}{(2\pi)^3} e^{i\vec{k}\cdot\hat{n}r_\star}\left(-\frac{1}{5}\tilde{\zeta}_{\vec{k}}\right). \tag{10.9.15}$$

Using the Rayleigh plane-wave expansion, $e^{i\vec{k}\cdot\hat{n}r_\star} = \sum_{\ell m} 4\pi\, i^\ell\, j_\ell(kr_\star) Y_{\ell m}(\hat{n}) Y^\star_{\ell m}(\hat{k})$, and orthonormality $\int \mathrm{d}\hat{n}\, Y_{\ell m} Y^*_{\ell' m'} = \delta_{\ell\ell'}\delta_{mm'}$,

$$a_{\ell m} = 4\pi i^\ell \int \frac{\mathrm{d}^3k}{(2\pi)^3} j_\ell(kr_\star) Y^\star_{\ell m}(\hat{k})\left(-\frac{1}{5}\tilde{\zeta}_{\vec{k}}\right). \tag{10.9.16}$$

Going beyond the Sachs–Wolfe approximation to include the full physics of the acoustic peaks, Silk damping, the integrated Sachs–Wolfe effect, etc. (Chapter 6), we simply replace $-\frac{1}{5}j_\ell(kr_\star) \to g^{(s)}_{T\ell}(k)$, the radiation transfer function computed numerically by a linear Boltzmann code such as CAMB or CLASS:

$$a_{\ell m} = 4\pi i^\ell \int \frac{\mathrm{d}^3k}{(2\pi)^3} g^{(s)}_{T\ell}(k)\, Y^\star_{\ell m}(\vec{k})\,\tilde{\zeta}_{\vec{k}}, \qquad C_\ell = \left\langle |a_{\ell m}|^2 \right\rangle = \frac{2}{\pi}\int k^2\mathrm{d}k\, P_\zeta(k)\, g^{(s)}_{T\ell}(k)^2. \tag{10.9.17}$$

The local-form CMB bispectrum template

To make contact with the observational literature, relate $a_{\ell m}$ to the matter-era Newtonian potential Φ rather than ζ directly,

$$a_{\ell m} = 4\pi i^\ell \int \frac{\mathrm{d}^3k}{(2\pi)^3}\Phi_{\vec{k}}\, g^{(s)}_{T\ell}(k)\, Y^*_{\ell m}(\hat{k}), \tag{10.9.18}$$

so that the local-form bispectrum $\left\langle \Phi_{\vec{k}_1}\Phi_{\vec{k}_2}\Phi_{\vec{k}_3}\right\rangle = (2\pi)^3\delta^{(3)}(\vec{k}_1+\vec{k}_2+\vec{k}_3)\times 2f_{\rm NL}[P_\Phi(k_1)P_\Phi(k_2)+$ perm.] (this is Eq. (9.7.32)–(9.7.36) of Chapter 9 for Φ, exactly the primordial B_ζ of §10.7 up to the $\Phi\leftrightarrow\zeta$ rescaling) maps onto the angular bispectrum

$$\langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3}\rangle = g^{\ell_1\ell_2\ell_3}_{m_1 m_2 m_3}\times 2f_{\rm NL}\int r^2{\rm d}r\,[\beta_{\ell_1}(r)\beta_{\ell_2}(r)\alpha_{\ell_3}(r)+\text{perm.}]\,, \tag{10.9.19}$$

with

$$g^{\ell_1\ell_2\ell_3}_{m_1 m_2 m_3}\equiv\int {\rm d}\hat{n}\,Y_{\ell_1 m_1}(\hat{n})Y_{\ell_2 m_2}(\hat{n})Y_{\ell_3 m_3}(\hat{n}), \tag{10.9.20}$$

$$\alpha_\ell(r)\equiv\frac{2}{\pi}\int k^2{\rm d}k\,g_{T\ell}(k)\,j_\ell(kr),\qquad \beta_\ell(r)\equiv\frac{2}{\pi}\int k^2{\rm d}k\,P_\Phi(k)\,g_{T\ell}(k)\,j_\ell(kr). \tag{10.9.21}$$

Reading off $\mathcal{B}_{\ell_1\ell_2\ell_3}\propto\int r^2{\rm d}r\,[\beta_{\ell_1}(r)\beta_{\ell_2}(r)\alpha_{\ell_3}(r)+\text{perm.}]$ and substituting into the estimator of Eq. (10.9.11) closes the loop from the primordial δN-formalism prediction of §10.5 all the way to the number quoted in Eq. (10.7.2).

Chapter 11

Cosmological Collider Physics: Poor man's accelerator

Cosmology is an observational science rather than an experimental one: the birth of the universe and the epochs that followed it cannot be recreated in a laboratory, so we have no direct access to events that took place during inflation. Inflation was an extremely high-energy process – the expansion rate could have been as large as 10^{14} GeV – and particles with masses up to that scale could have been produced during this phase. Given that we have no direct way of probing the inflationary era, how do we recognize the presence of these new particles?

The simplest version of this question is whether there were any new light or massless fields during inflation. Particles with mass comparable to the Hubble scale can be produced by quantum fluctuations during inflation and then decay back into ordinary inflatons. This process leaves a statistical imprint on the spectrum of primordial fluctuations – more precisely, it produces non-Gaussianity. Of course, self-interactions of the inflaton, or interactions of the inflaton with gravity, can also give rise to non-Gaussianity, so the real question is: *which specific non-Gaussianities are a signature of new particles during inflation, as opposed to signatures that arise from inflaton self-interactions?*

Mass draws a sharp line between two possible answers. A particle much heavier than the Hubble scale, $M \gg H$, can simply be integrated out: it leaves behind local, higher-dimension operators for the inflaton, and the non-Gaussianity these produce is bounded, $f_{\rm NL} \lesssim 1$ for a gravitational-strength coupling, and analytic in the squeezed limit[1] – exactly the kind of effect an inflaton self-interaction could also produce, so on its own it is not a clean signature of a new particle. A particle with $M \lesssim$ a few H cannot be integrated out this way: it is produced non-perturbatively by the expanding spacetime itself, and its later decay leaves a genuinely *non-local* imprint, a piece of the squeezed-limit bispectrum that is non-analytic in $k_{\rm long}/k_{\rm short}$ and that no local inflaton self-interaction can mimic. This non-local, non-analytic signal is what the rest of this chapter computes.

[1] See §9.2 of Baumann, "TASI Lectures on Primordial Cosmology," arXiv:1807.03098, for a worked example: integrating out a heavy scalar coupled as $\sigma(\partial\phi)^2$ generates $\mathcal{L} \supset (\partial\phi)^4/\Lambda^4$, whose bispectrum is bounded and analytic in $k_{\rm long}/k_{\rm short}$.

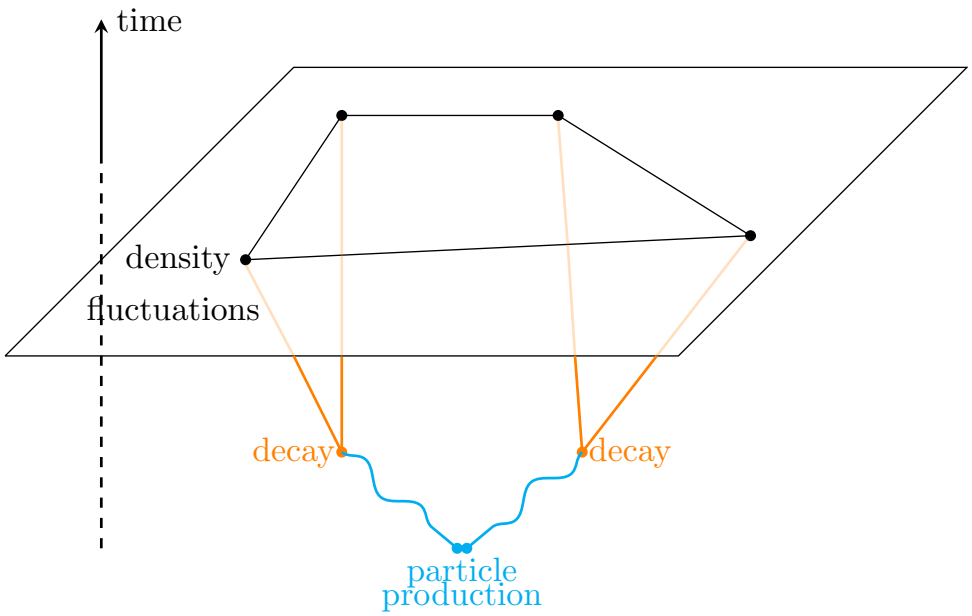


Figure 11.1: Is the intermediate particle here mediating an inflaton self-interaction, or was it produced by a quantum fluctuation and is now decaying? Only the second case counts as genuine particle production.

In particle physics, colliders like the Large Hadron Collider (LHC) probe the properties of fundamental particles at very high energies. The discipline of collider physics involves working backward from the direct collider observables to the underlying Lagrangian of the theory, and one of the simplest questions one can ask is how to recognize the presence of a new particle. In collider physics the answer lies in a peak in the cross section near the particle's mass; the spin can then be read off from the angular distribution of the scattering process. In the cosmological case there is no invariant-mass peak to look for. Instead, higher-order correlations oscillate as the distance between correlated points varies, with a frequency set by the mass of the new particle.

In the same spirit, cosmological collider physics uses the cosmic microwave background as a "detector" for non-Gaussianity in the early universe, effectively "colliding" new particles that arose from quantum fluctuations during inflation. These particles need not be ordinary matter – they are new fields that could have masses comparable to the Hubble scale, or even be massless. By studying the non-Gaussianity in the primordial fluctuation spectrum, we can reconstruct the history of particle production and decay during inflation, and learn about the properties and interactions of these particles.

Cosmological collider physics also gives an alternative way to test theories of inflation. The standard picture assumes the universe underwent an exponential expansion during inflation, but this expansion could have been driven by a variety of mechanisms, including the presence of new particles or fields. Analyzing the statistical signatures they leave behind in the CMB tells us which models are most consistent with observations and which predict the correct level of non-Gaussianity.

Chapter 10 (§10.6) showed that the squeezed-limit bispectrum vanishes for every single-field model of inflation, $\lim_{k_3\to 0} B_\zeta(k_1,k_2,k_3) = (1-n_s)P_\zeta(k_1)P_\zeta(k_3)$. So any measured excess above this floor is direct evidence for extra fields during inflation. The next question is: what does that excess actually look like, and what can we learn from it about the new particles?

The answer, worked out below, is that a field of mass m and spin s leaves a squeezed-limit

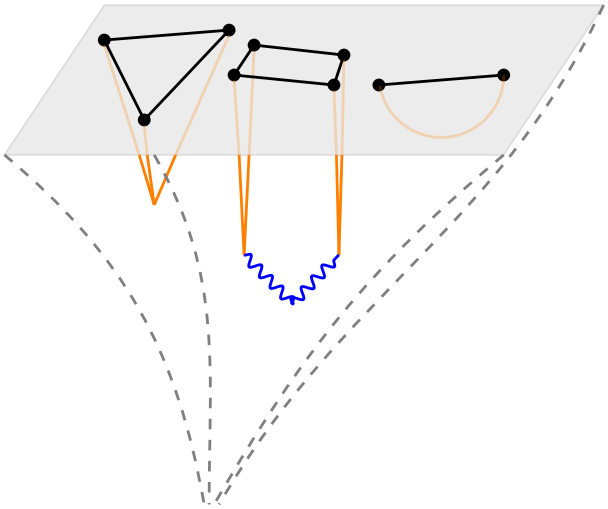


Figure 11.2: Massive particles produced from quantum fluctuations during inflation decay back into inflaton fluctuations, leaving their imprint on the primordial non-Gaussianity.

signal of the form

$$\frac{\langle\zeta\zeta\zeta\rangle}{\langle\zeta\zeta\rangle_{\text{short}}\langle\zeta\zeta\rangle_{\text{long}}} \sim \epsilon\, e^{-\pi\mu}|c(\mu)|\left[e^{i\delta(\mu)}\left(\frac{k_{\text{long}}}{k_{\text{short}}}\right)^{\frac{3}{2}+i\mu} + e^{-i\delta(\mu)}\left(\frac{k_{\text{long}}}{k_{\text{short}}}\right)^{\frac{3}{2}-i\mu}\right] P_s(\cos\theta), \quad (11.0.1)$$

where $\mu \equiv \sqrt{\frac{m^2}{H^2} - \frac{9}{4}}$, $\epsilon = -\dot{H}/H^2$ is the slow-roll parameter, and $\cos\theta = \hat{k}_{\text{long}} \cdot \hat{k}_{\text{short}}$. a power law in the squeezed ratio $k_{\text{long}}/k_{\text{short}}$. Once the field is heavy enough, $m > \frac{3}{2}H$, μ becomes real and this power law turns into an actual oscillation in $\log(k_{\text{long}}/k_{\text{short}})$. The oscillation frequency is set by the mass through μ, the angle dependence is set by the spin through the Legendre polynomial $P_s(\cos\theta)$, and there is a calculable phase $\delta(\mu)$. This is the cosmological version of a resonance peak at a collider: the LHC reads off a particle's mass and spin from a peak in the invariant-mass distribution and its angular spread; a cosmological collider reads off m and s from the frequency and angle dependence of this oscillation instead.

The oscillation is a genuinely quantum effect. It comes from interference between two processes: the ordinary evolution of the inflaton fluctuation, and the creation of a virtual pair of massive particles that later decay back into inflaton fluctuations. Measuring this phase is, quite literally, a cosmological double-slit experiment.

A word on whether this is actually measurable: for Planck-suppressed couplings to the inflaton, $\lambda\int(\nabla\phi)^2\sigma$ with $\lambda \sim 1/M_{\text{pl}}$, the signal is far too small for CMB temperature maps alone to see. §10.7 already showed that Planck's current bound, $f_{\text{NL}}^{\text{local}} = -0.9 \pm 5.1$, is consistent with zero, nowhere near the precision this effect needs. Large-scale structure surveys and future 21-cm maps are the more promising routes. So why bother with the squeezed limit at all, if it is not even the easiest place to look for non-Gaussianity? Because it is the one configuration where a detection has an unambiguous meaning: this is the same reason a collider experiment cares about rare, high-momentum events instead of the much more common low-energy ones – not because they happen often, but because they are the events that actually reveal new particles.

Outline. We build up to the inflationary squeezed limit one step at a time, following Arkani-Hamed and Maldacena's original argument. We start on a rigid de Sitter background with no dynamical gravity, where the de Sitter symmetries act just like the conformal group on the late-time boundary and almost completely fix the two- and three-point functions (§11.1–§11.3). That is why de Sitter comes first, before inflation. Among three-point functions, we do the case of two *conformally coupled* scalars, $m^2 = 2H^2$, before the physically relevant case of two

massless scalars (§11.4), simply because the conformally coupled answer is algebraically easier, and the massless result then comes from it by applying one differential operator, $\mathcal{O}_{12}$. The four-point function comes next (§11.6), since this is the first place a massive field appears as an internal exchanged line; taking two pairs of its four points close together (the OPE limit) turns it back into a product of three-point functions, tying the two calculations together. Only then do we get to the actual inflationary three-point function (§11.7), which just means swapping one of the three massless legs for the classical background $\dot{\phi}_0(t)$. If you only want the final squeezed-limit formula, Eq. (11.0.1), and what it means, you can skip straight there. The case of a massive field *with spin* is more technical throughout and can also be skipped on a first read; it is included because a detected spin-4 (or higher) signal would be close to a smoking gun for string theory during inflation – weakly-coupled towers of high-spin particles like that are essentially only known to arise in string theory.

11.1 Two-Point Functions in de Sitter Space

In de Sitter space, correlation functions of quantum fields are tightly constrained by the background isometry group $SO(4,1)$. In conformal coordinates $(\eta, \vec{x})$, the metric of d-dimensional de Sitter space is

$$ds^2 = \frac{-d\eta^2 + d\vec{x}^2}{H^2\eta^2}, \qquad \eta \in (-\infty, 0). \tag{11.1.1}$$

The two-point Green's function $\langle\phi(x)\phi(x')\rangle$ of a scalar field with mass m and non-minimal curvature coupling ξ obeys the curved-space wave equation:

$$\left[\Box - m^2 - \xi R\right]\langle\phi(x)\phi(x')\rangle = \frac{i}{\sqrt{-g}}\delta^d(x - x'). \tag{11.1.2}$$

Because de Sitter space is maximally symmetric, the Green's function depends on spacetime points x, x' solely through the de Sitter invariant geodesic distance $P(x,x')$, defined in $(d+1)$-dimensional embedding space by:

$$P(x,x') \equiv -H^2\eta_{AB}X^A X'^B = 1 + \frac{-(\eta-\eta')^2 + |\vec{x}-\vec{x}'|^2}{2\eta\eta'}, \tag{11.1.3}$$

satisfying the geometric identities $\nabla^\mu P\nabla_\mu P = H^2(P^2-1)$ and $\nabla_\mu\nabla_\nu P = g_{\mu\nu}H^2 P$. Away from coincident points ($P \neq 1$), the wave equation reduces to an ordinary differential equation in P:

$$(P^2-1)\frac{d^2G}{dP^2} + dP\frac{dG}{dP} + \left(\frac{m^2}{H^2} + \xi d(d-1)\right)G = 0. \tag{11.1.4}$$

Reducing the wave equation to a hypergeometric equation

We then have,

$$\begin{aligned}\Box G(P) &= g^{\mu\nu}\nabla_\mu\nabla_\nu G(P) = g^{\mu\nu}\nabla_\mu\left[G'(P)\nabla_\nu P\right] \\ &= G''(P)g^{\mu\nu}\partial_\mu P\partial_\nu P + G'(P)g^{\mu\nu}\nabla_\mu\nabla_\nu P\end{aligned}$$

Using now (11.1.3), we find

$$\Box G(P) = H^2(P^2-1)G''(P) + dH^2 PG'(P)$$

where d is the dimension of spacetime. It follows that the Wightman function for a scalar field satisfies,

$$(P^2-1)\partial_P^2 G + dP\partial_P G + \left(\frac{m^2}{H^2} + \xi d(d-1)\right) G = 0. \tag{1}$$

In terms of the variable $z = (1+P)/2$, this is the hypergeometric differential equation with singular points at $z=0$ (antipodal points $P=-1$), $z=1$ (coincident points $P=1$), and $z=\infty$.

Geodesic distance in conformal coordinates

$$\begin{aligned}
P(X,X') &= -\frac{H^2}{4\eta\eta'}\left[\left(n^2-\vec{x}^2-\frac{1}{H^2}\right)\left(n'^2-\vec{x}'^2-\frac{1}{H^2}\right)-\frac{4}{H^2}\vec{x}\cdot\vec{x}'\right.\\
&\qquad\left.-\left(n^2-\vec{x}^2+\frac{1}{H^2}\right)\left(n'^2-\vec{x}'^2+\frac{1}{H^2}\right)\right]\\
&= -\frac{H^2}{4\eta\eta'}\left[(n^2-\vec{x}^2)\left(\cancel{\eta'^2}-\cancel{\vec{x}'^2}-\frac{1}{H^2}-\cancel{\eta'^2}+\cancel{\vec{x}'^2}-\frac{1}{H^2}\right)-\frac{1}{H^2}\left(\eta'^2-\vec{x}'^2-\cancel{\frac{1}{H^2}}\right)\right.\\
&\qquad\left.-\frac{4}{H^2}\vec{x}\cdot\vec{x}'-\frac{1}{H^2}\left(\eta'^2-\vec{x}'^2+\cancel{\frac{1}{H^2}}\right)\right]\\
&= -\frac{H^2}{4\eta\eta'}\left[-\frac{2}{H^2}(\eta^2-\vec{x}^2)-\frac{4}{H^2}\vec{x}\cdot\vec{x}'-\frac{2}{H^2}(\eta'^2-\vec{x}'^2)\right]\\
&= \frac{1}{2\eta\eta}[\eta^2+\eta'^2-\vec{x}^2-\vec{x}'^2+2\vec{x}\cdot\vec{x}']\\
&= 1+\frac{(\eta-\eta')^2-(\vec{x}-\vec{x}')^2}{2\eta\eta'}
\end{aligned}$$

Requiring regularity at the antipodal point $P=-1$ (the Bunch-Davies vacuum condition) and properly normalizing the short-distance Hadamard singularity at $P=1$ fixes the two-point function in $d=4$ dimensions ($\xi=0$) to:

$$\langle\phi(x)\phi(x')\rangle = \frac{H^2\Gamma(h_+)\Gamma(h_-)}{(4\pi)^2}\,{}_2F_1\left(h_+,h_-,2;1-\frac{1}{w}\right), \tag{11.1.5}$$

where w is the conformal ratio

$$w \equiv \frac{4\eta\eta'}{-(\eta-\eta')^2+|\vec{x}-\vec{x}'|^2+i\epsilon}, \tag{11.1.6}$$

and the conformal scaling weights $h_\pm$ are:

$$h_\pm = \frac{3}{2}\pm\sqrt{\frac{9}{4}-\frac{m^2}{H^2}} = \frac{3}{2}\pm i\mu, \qquad \mu \equiv \sqrt{\frac{m^2}{H^2}-\frac{9}{4}}. \tag{11.1.7}$$

Solving for the de Sitter two point function

$$z = \frac{1+P}{2}$$

equation (1) turns into:

$$z(1-z)\partial_z^2 G + d\left(\frac{1}{2}-z\right)\partial_z G - \left(\frac{m^2}{H^2}+\xi d(d-1)\right)G = 0$$

This is hypergeometric equation and the solution is given as:

$$\begin{aligned}\langle\phi(x)\phi(x')\rangle &= c_d\ {}_2F_1\left(h_+, h_-, \frac{d}{2}, \frac{1+P}{2}\right)\\ &= H^{d-2}\frac{\Gamma(h_+)\Gamma(h_-)}{(4\pi)^{d/2}\Gamma(d/2)}\\ &\quad\times\ {}_2F_1\left(h_+, h_-, \frac{d}{2}, \frac{1+\left[1+\frac{(\eta-\eta')^2-(x-x')^2}{2\eta\eta'}\right]}{2}\right)\end{aligned}$$

In our case, $d = 4$

$$\begin{aligned}&= H^2\frac{\Gamma(h_+)\Gamma(h_-)}{(4\pi)^2}\\ &\quad\times\ {}_2F_1\left(h_+, h_-, 2, 1-\left[\frac{-(\eta-\eta')^2+(x-x')^2+i\epsilon}{4\eta\eta'}\right]\right)\end{aligned}$$

where

$$h_\pm = \frac{1}{2}\left[(d-1)\pm\sqrt{(d-1)^2-4\left\{\frac{m^2}{H^2}+\xi d(d-1)\right\}}\right]$$

for $d = 4$ and $\xi = 0$, we have

$$h_\pm = \frac{3}{2}\pm\sqrt{\left(\frac{3}{2}\right)^2-\left(\frac{m}{H}\right)^2} = \frac{3}{2}\pm i\mu$$

for $d = 4$ and $m = 0$ with $\xi = \frac{1}{6}$, we have[a]

$$h_\pm = \frac{3}{2}\pm\sqrt{\left(\frac{3}{2}\right)^2-\frac{12}{6}} = \frac{3}{2}\pm\frac{1}{2}$$

[a] This choice is equivalent to setting $m^2 = 2H^2$.

For massive particles with $m > \frac{3}{2}H$, μ is real (the *principal series*), while for light particles with $m < \frac{3}{2}H$, μ is imaginary (the *complementary series*). On the future conformal boundary ($\eta, \eta' \to 0$ with fixed comoving separation $|\vec{x}-\vec{x}'|$), applying the hypergeometric connection formula gives the asymptotic position-space correlator:

$$\langle\phi(\eta,\vec{x})\phi(\eta',\vec{0})\rangle \approx \frac{H^2}{4\pi^{5/2}|\vec{x}|^3}\left[\left(\frac{\eta\eta'}{|\vec{x}|^2}\right)^{i\mu}\Gamma(-i\mu)\Gamma\left(\frac{3}{2}+i\mu\right)+\left(\frac{\eta\eta'}{|\vec{x}|^2}\right)^{-i\mu}\Gamma(i\mu)\Gamma\left(\frac{3}{2}-i\mu\right)\right]. \tag{11.1.8}$$

In terms of proper time t ($\eta = -e^{-Ht}/H$), the two terms scale as $e^{-(\frac{3}{2}\mp i\mu)Ht}$, corresponding to boundary operators of conformal scaling dimensions $\Delta = \frac{3}{2}+i\mu$ and $\tilde{\Delta} = 3-\Delta = \frac{3}{2}-i\mu$.

Late-time limit of the de Sitter two point function

On the late conformal boundary for such massive particles, we have:

$$\langle\phi(\eta,\vec{x})\phi(\eta',\vec{0})\rangle' = H^2\frac{\Gamma\left(\frac{3}{2}+i\mu\right)\Gamma\left(\frac{3}{2}-i\mu\right)}{(4\pi)^2} \times {}_2F_1\left(\frac{3}{2}+i\mu,\frac{3}{2}-i\mu,2;1-\frac{-(\eta-\eta')^2+\vec{x}^2}{4\eta\eta'}\right)$$

Late-time, momentum-space limit via the connection formula

Using the duplication formula to simplify the Gamma functions and then taking the $\eta,\eta'\to 0$ limit with ${}_2F_1(a,b,c;0)=1$ gives

$$\begin{aligned}\langle\phi(\eta,\vec{x})\phi(\eta',\vec{0})\rangle' &= \frac{H^2}{4\pi^{5/2}}\left[\left(\frac{\eta\eta'}{|\vec{x}|^2}\right)^{\frac{3}{2}+i\mu}\Gamma(-i\mu)\Gamma\left(\frac{3}{2}+i\mu\right)\right.\\ &\qquad\left.+\left(\frac{\eta\eta'}{|\vec{x}|^2}\right)^{\frac{3}{2}-i\mu}\Gamma(i\mu)\Gamma\left(\frac{3}{2}-i\mu\right)\right]+\cdots\\ &= \mathcal{F}\left(\frac{H^2}{4\pi^{5/2}}\left[\left(\frac{\eta\eta'}{|\vec{x}|^2}\right)^{\frac{3}{2}+i\mu}\Gamma(-i\mu)\Gamma\left(\frac{3}{2}+i\mu\right)\right.\right.\\ &\qquad\left.\left.+\left(\frac{\eta\eta'}{|\vec{x}|^2}\right)^{\frac{3}{2}-i\mu}\Gamma(i\mu)\Gamma\left(\frac{3}{2}-i\mu\right)\right]\right)\\ \langle\phi_{\vec{k}}(\eta)\phi_{-\vec{k}}(\eta')\rangle' &= \frac{H^2}{4\pi}(\eta\eta')^{3/2}\left[\Gamma(\nu)^2\left(\frac{k^2\eta\eta'}{4}\right)^{-\nu}+\Gamma(-\nu)^2\left(\frac{k^2\eta\eta'}{4}\right)^{\nu}\right]\end{aligned}$$

Fourier transforming (11.1.8) to momentum space yields the late-time two-point correlator:

$$\left\langle\phi_{\vec{k}}(\eta)\phi_{-\vec{k}}(\eta')\right\rangle' = \frac{H^2}{4\pi}(\eta\eta')^{3/2}\left[\Gamma(-i\mu)^2\left(\frac{k^2\eta\eta'}{4}\right)^{i\mu}+\Gamma(i\mu)^2\left(\frac{k^2\eta\eta'}{4}\right)^{-i\mu}\right]+\text{local terms}. \tag{11.1.9}$$

Equivalently, this result follows from the bulk in-in formalism using the mode function for a massive scalar:

$$f_k(\eta) = \frac{H\sqrt{\pi}}{2}e^{-\frac{i\pi}{4}(1+2\nu)}(-\eta)^{3/2}H_\nu^{(2)}(-k\eta), \qquad \nu\equiv\sqrt{\frac{9}{4}-\frac{m^2}{H^2}} = i\mu. \tag{11.1.10}$$

Late-time limit from the mode function

if ν is real

$$\begin{aligned}\langle\phi_{\vec{k}}(\eta)\phi_{-\vec{k}}(\eta')\rangle' &= \frac{H^2\pi}{4}(\eta\eta')^{3/2}\lim_{k\eta\to 0}\left[H_\nu^{(2)}(-k\eta)\underbrace{H_\nu^{(2)}(-k\eta')^\star}_{H_\nu^{(1)}(-k\eta')}\right]\\ &= \frac{H^2\pi}{4}(\eta\eta')^{3/2}\frac{i}{\pi}\left[\Gamma(\nu)\left(\frac{-k\eta}{2}\right)^{-\nu}+e^{i\pi\nu}\Gamma(-\nu)\left(\frac{-k\eta}{2}\right)^{\nu}\right]\end{aligned}$$

$$\times \frac{-i}{\pi}\left[\Gamma(\nu)\left(\frac{-k\eta'}{2}\right)^{-\nu}+e^{i\pi\nu}\Gamma(-\nu)\left(\frac{-k\eta'}{2}\right)^{\nu}\right]$$
$$\approx \frac{H^2}{4\pi}(\eta\eta')^{3/2}\left[\Gamma(\nu)^2\left(\frac{k^2\eta\eta'}{4}\right)^{-\nu}+\Gamma(-\nu)^2\left(\frac{k^2\eta\eta'}{4}\right)^{\nu}\right]+\text{terms}$$

if ν is imaginary

$$\langle\phi_{\vec{k}}(\eta)\phi_{-\vec{k}}(\eta')\rangle' = \frac{H^2\pi}{4}(\eta\eta')^{3/2}e^{-i\pi\nu}\lim_{k\eta\to 0}\left[H_\nu^{(2)}(-k\eta)\underbrace{H_\nu^{(2)}(-k\eta')^\star}_{e^{i\pi\nu}H_\nu^{(1)}(-k\eta')}\right]$$
$$= \frac{H^2\pi}{4}(\eta\eta')^{3/2}e^{-i\pi\nu}\frac{i}{\pi}\left[\Gamma(\nu)\left(\frac{-k\eta}{2}\right)^{-\nu}+e^{i\pi\nu}\Gamma(-\nu)\left(\frac{-k\eta}{2}\right)^{\nu}\right]$$
$$\times\frac{-ie^{i\pi\nu}}{\pi}\left[\Gamma(\nu)\left(\frac{-k\eta'}{2}\right)^{-\nu}+e^{-i\pi\nu}\Gamma(-\nu)\left(\frac{-k\eta'}{2}\right)^{\nu}\right]$$
$$\approx \frac{H^2}{4\pi}(\eta\eta')^{3/2}\left[\Gamma(\nu)^2\left(\frac{k^2\eta\eta'}{4}\right)^{-\nu}+\Gamma(-\nu)^2\left(\frac{k^2\eta\eta'}{4}\right)^{\nu}\right]+\text{terms} \qquad (1)$$

For equal times $\eta=\eta'$, the late-time power spectrum $P_\phi(\eta,k)=|f_k(\eta)|^2$ evaluates to:

$$\lim_{k\eta\to 0}P_\phi(\eta,k)=\frac{H^2(-\eta)^3}{2\mu\sinh(\pi\mu)}\left[\cos\left(2\mu\log\left(\frac{-k\eta}{2}\right)-\delta(\mu)\right)+\cosh(\pi\mu)\right], \qquad (11.1.11)$$

where $\Gamma(\pm i\mu)\equiv|\Gamma(\pm i\mu)|e^{\pm i\delta(\mu)}$. For large masses $\mu\gg 1$, the oscillatory power spectrum is suppressed by the thermal Boltzmann factor:

$$\frac{1}{\sinh(\pi\mu)}\longrightarrow 2e^{-\pi\mu}=2e^{-\pi m/H}, \qquad (11.1.12)$$

reflecting the spontaneous quantum production of massive particle pairs in the expanding de Sitter spacetime with Gibbons-Hawking temperature $T_{\rm dS}=H/(2\pi)$.

Boltzmann suppression of massive-particle production

$$\langle\phi_{\vec{k}}(\eta)\phi_{-\vec{k}}(\eta')\rangle' = \underbrace{\cdots}_{\text{stuff}}(\eta\eta')^{3/2-\nu}+\underbrace{\cdots}_{\text{stuff}}(\eta\eta')^{3/2+\nu}$$

setting $H=1$

$$\sim e^{-(3/2-\nu)(t+t')}+e^{-(3/2+\nu)(t+t')}$$

For massive particles $m>\frac{3}{2}H$, we can have imaginary ν, which will show oscillatory behavior at late time signifying particle creation in de Sitter space. The complex conjugate of Hankel function is given as:

$$H_{i\mu}^{(2)\star}=e^{-\pi\mu}H_{i\mu}^{(1)}$$

Then,

$$\lim_{k\eta\to 0}\left|H_\nu^{(2)}(-k\eta)\right|^2=e^{-\pi\mu}\lim_{k\eta\to 0}H_\nu^{(2)}(-k\eta)H_\nu^{(1)}(-k\eta)$$

$$= \frac{e^{-\pi\mu}}{\pi^2}\left(\Gamma(i\mu)^2\left(\frac{-k\eta}{2}\right)^{-2i\mu} + \Gamma(-i\mu)^2\left(\frac{-k\eta}{2}\right)^{2i\mu} - \frac{2\pi\cot(\pi\mu)}{i\mu}\right),$$

and the late-time limit of the power spectrum becomes

$$\begin{aligned}\lim_{k\eta\to 0} P_\phi(\eta,k) &= \lim_{k\eta\to 0}\langle\phi_{\vec{k}}(\eta)\phi_{-\vec{k}}(\eta')\rangle' = \lim_{k\eta\to 0}\left|f_{\vec{k}}(\eta)\right|^2 \\ &= \frac{\pi}{4}e^{\pi\mu}H^2(-\eta)^3\lim_{k\eta\to 0}\left|H^{(2)}_\nu(-k\eta)\right|^2 \\ &= \frac{1}{4\pi}H^2(-\eta)^3\left(\Gamma(i\mu)^2\left(\frac{-k\eta}{2}\right)^{-2i\mu} + \Gamma(-i\mu)^2\left(\frac{-k\eta}{2}\right)^{2i\mu} + \frac{2\pi\coth(\pi\mu)}{\mu}\right).\end{aligned}$$

Writing $\Gamma(\pm i\mu) \equiv |\Gamma(\pm i\mu)|\, e^{\pm i\delta}$, and using

$$|\Gamma(\pm i\mu)|^2 = \frac{\pi}{\mu\sinh(\pi\mu)},$$

we can write this as

$$\begin{aligned}\lim_{k\eta\to 0} P_\phi(\eta,k) &= \frac{H^2(-\eta)^3}{4\mu\sinh(\pi\mu)}\left[e^{i\delta(\mu)}\left(\frac{-k\eta}{2}\right)^{-2i\mu} + e^{-i\delta(\mu)}\left(\frac{-k\eta}{2}\right)^{2i\mu} + 2\cosh(\pi\mu)\right] \\ &= \frac{H^2(-\eta)^3}{2\mu\sinh(\pi\mu)}\left[\cos\left(2\mu\log\left(\frac{-k\eta}{2}\right) - \delta\right) + \cosh(\pi\mu)\right].\end{aligned}$$

For large masses, $\mu \gg 1$, the power spectrum scales as $1/\sinh(\pi\mu) \to e^{-\pi\mu}$, corresponding to the Boltzmann suppression of the spontaneous production of massive particles.

Finally, the $i\epsilon$ vacuum prescription ensures that the physical correlator contains an unsuppressed particle propagation branch $e^{-i\mu(t-t')}$ (creating a particle at η' and destroying it at $\eta > \eta'$) alongside the exponentially suppressed pair-creation branch $e^{-2\pi\mu}$.

Particle creation from the $i\epsilon$ prescription

$$\begin{aligned}\langle\phi(\eta,\vec{x})\phi(\eta',\vec{0})\rangle' = H^2&\frac{\Gamma\left(\frac{3}{2}+i\mu\right)\Gamma\left(\frac{3}{2}-i\mu\right)}{(4\pi)^2} \\ &\times\ {}_2F_1\left(\frac{3}{2}+i\mu,\frac{3}{2}-i\mu,2;1-\frac{-(\eta-\eta')^2+\vec{x}^2}{4\eta\eta'}\right)\end{aligned}$$

Taking the limit:

$$\begin{aligned}&= H^2\frac{\Gamma\left(\frac{3}{2}+i\mu\right)\Gamma\left(\frac{3}{2}-i\mu\right)}{(4\pi)^2}\left[\left(\frac{4\eta\eta'}{\eta'^2}\right)^{\frac{3}{2}+i\mu}\frac{\Gamma(2)\Gamma(-2i\mu)}{\Gamma\left(\frac{3}{2}-i\mu\right)\Gamma\left(\frac{1}{2}-i\mu\right)}\right. \\ &\quad\left. + \left(\frac{4\eta\eta'}{\eta'^2}\right)^{\frac{3}{2}-i\mu}\frac{\Gamma(2)\Gamma(2i\mu)}{\Gamma\left(\frac{3}{2}+i\mu\right)\Gamma\left(\frac{1}{2}+i\mu\right)}\right] \\ &= \frac{H^2}{4\pi^{5/2}}\left[\left(\frac{\eta}{\eta'}\right)^{\frac{3}{2}+i\mu}\Gamma(-i\mu)\Gamma\left(\frac{3}{2}+i\mu\right) + \left(\frac{\eta}{\eta'}\right)^{\frac{3}{2}-i\mu}\Gamma(i\mu)\Gamma\left(\frac{3}{2}-i\mu\right)\right] + \cdots\end{aligned}$$

Keeping the $i\epsilon$ prescription inside the hypergeometric function instead gives

$$\langle\phi(\eta,0)\phi(\eta',\vec{0})\rangle' = \frac{H^2}{4\pi^{5/2}}\left[\left(\frac{\eta}{\eta'}e^{-i\pi}\right)^{\frac{3}{2}+i\mu}\Gamma(-i\mu)\Gamma\left(\frac{3}{2}+i\mu\right) + \left(\frac{\eta}{\eta'}e^{-i\pi}\right)^{\frac{3}{2}-i\mu}\Gamma(i\mu)\Gamma\left(\frac{3}{2}-i\mu\right)\right]+\cdots$$

This extra $e^{i\pi}$ gives factors of $e^{\pm\pi\mu}$, which imply that the first term in above is now not suppressed exponentially for large μ while the second term is suppressed by $e^{-2\pi\mu}$. The first term is the one expected where we create a particle at η_0 and destroy it at $\eta > \eta_0$.

11.2 Two-Point Functions of Fields with Spin

Having analyzed scalar fields in de Sitter space, we now turn to fields carrying spin $s \geq 1$. Particles with spin leave distinctive angular signatures in cosmological correlation functions, providing a direct window into the spin of massive particles produced during inflation.

On the late-time conformal boundary of de Sitter space ($\eta \to 0$), the de Sitter isometry group $SO(4,1)$ acts as the conformal group of Euclidean 3-space. Consequently, late-time correlation functions of spinning fields are constrained by boundary conformal symmetry.

11.2.1 Position-space correlation functions

For a spin-1 (vector) primary operator $J_\mu(\vec{x})$ of conformal dimension Δ, invariance under translations, dilatations, and spatial rotations restricts the two-point function to a tensor structure built out of $g_{\mu\nu}$ and $x_\mu x_\nu/x^2$. Invariance under special conformal transformations (or spatial inversion $\vec{x} \to \vec{x}/x^2$) completely fixes the relative coefficient:

$$\left\langle J_\mu(\vec{x})J_\nu(\vec{0})\right\rangle = \frac{C}{|\vec{x}|^{2\Delta}}\left[g_{\mu\nu} - 2\frac{x_\mu x_\nu}{x^2}\right]. \tag{11.2.1}$$

Vector two-point function: setup via inversion

$$\langle J_\mu(x_1)J_\nu(x_2)\rangle = \frac{C}{|x_1-x_2|^{2\Delta}}\left[g_{\mu\nu} + \delta\frac{(x_1-x_2)_\mu(x_1-x_2)_\nu}{(x_1-x_2)^2}\right] \tag{1}$$

The overall power of $|x_1 - x_2|$ is fixed exactly as in the scalar case, since it is controlled by dilatation alone; only the coefficient δ of the tensor structure is new. Since the correlation function is translation invariant, consider the following redefinition:

$$\langle J_\mu(x_1)J_\nu(x_2)\rangle = \langle J_\mu(x_1-x_2)J_\nu(0)\rangle \tag{2}$$
$$= \langle J_\mu(x_{12})J_\nu(0)\rangle = \langle J_\mu(x)J_\nu(0)\rangle \tag{3}$$

An SCT is just inversion→translation→inversion, so we can use this to our advantage: since the correlation function is already invariant under translations, it suffices to verify its invariance under inversion, and this will then imply invariance under the full SCT. The inversion transformation is[a]:

$$x'_\mu = \frac{x_\mu}{x^2} \qquad |x'|^2 = \frac{1}{|x|^2} \tag{4}$$

and

$$\frac{\partial x'_\nu}{\partial x^\mu} = \frac{\partial}{\partial x^\mu}\frac{x_\nu}{x^2} = \frac{1}{x^2}\left[g_{\mu\nu} - 2\frac{x_\mu x_\nu}{x^2}\right] = x'^2 \underbrace{\left[g_{\mu\nu} - 2\frac{x'_\mu x'_\nu}{x'^2}\right]}_{I_{\mu\nu}} \tag{5}$$

A primary operator of dimension Δ with a free vector index picks up, under $x \to x'$, one overall Jacobian-determinant factor $|\partial x/\partial x'|^{\Delta/d}$ for its scaling weight – exactly the factor already used to fix the scalar case above – together with one power of the orthogonal matrix $I^\mu{}_\nu$ built from the Jacobian for its free index; the two combine into the familiar tensor transformation law[b]. The vector field therefore transforms as

$$\langle J_\mu(x'_1)J_\nu(x'_2)\rangle = \left|\frac{\partial x_\alpha}{\partial x'^\mu}\right|^{\Delta/d}\left|\frac{\partial x_\beta}{\partial x'^\nu}\right|^{\Delta/d}\left|\frac{\partial x_\alpha}{\partial x'^\mu}\right|^{-1/d}\left|\frac{\partial x_\beta}{\partial x'^\nu}\right|^{-1/d}\frac{\partial x'_\alpha}{\partial x^\mu}\frac{\partial x'_\beta}{\partial x^\nu}\left\langle J^\alpha(x_1)J^\beta(x_2)\right\rangle \tag{6}$$

we see that the two overall scaling factors reproduce exactly what we already found for the scalar prefactor,

$$\underbrace{\left|\frac{\partial x_\alpha}{\partial x'^\mu}\right|^{\Delta/d}}_{x=x_1}\underbrace{\left|\frac{\partial x_\beta}{\partial x'^\nu}\right|^{\Delta/d}}_{x=x_2}\frac{1}{|x_{12}|^{2\Delta}} = |x'_1|^{-2\Delta}|x'_2|^{-2\Delta}\frac{1}{|x_{12}|^{2\Delta}} = \frac{1}{|x'_{12}|^{2\Delta}} \tag{7}$$

where we used

$$|x'_{12}|^2 = \left(\frac{x_1^\mu}{x_1^2} - \frac{x_2^\mu}{x_2^2}\right)^2 \tag{8}$$

$$= \frac{|x_1|^2}{x_1^4} + \frac{|x_2|^2}{x_2^4} - 2\frac{x_1^\mu}{x_1^2}\frac{x_{2\mu}}{x_2^2} \tag{9}$$

$$= \frac{1}{x_1^2} - 2\frac{x_1}{x_1^2}\cdot\frac{x_2}{x_2^2} + \frac{1}{x_2^2} \tag{10}$$

$$= \frac{(x_1-x_2)^2}{x_1^2x_2^2} = \frac{|x_{12}|^2}{|x_1|^2|x_2|^2} = \frac{|x_{12}|^2}{|x'_1|^{-2}|x'_2|^{-2}} \tag{11}$$

Then, we only have to ensure that $g^{\mu\nu} + \delta\frac{x'^\mu_{12}x'^\nu_{12}}{x'^2_{12}}$ is invariant under inversion, i.e. that it equals $\frac{\partial x'_\alpha}{\partial x^\mu}\frac{\partial x'_\beta}{\partial x^\nu}\left[g^{\alpha\beta} + \delta\frac{x^\alpha_{12}x^\beta_{12}}{x^2_{12}}\right]$ evaluated at $x = x'/|x'|^2$, i.e. after inverting once more:

$$\begin{aligned}
&g^{\mu\nu} + \delta\frac{(x'_{12})^\mu(x'_{12})^\nu}{(x'_{12})^2} \\
&= \left(\delta^\mu_\alpha - 2\frac{x_1^\mu x_{1\alpha}}{x_1^2}\right)\left(\delta^\nu_\beta - 2\frac{x_2^\nu x_{2\beta}}{x_2^2}\right)\left[g^{\alpha\beta} + \delta\frac{x^\alpha_{12}x^\beta_{12}}{x^2_{12}}\right]\Bigg|_{x^\mu=\frac{x'^\mu}{|x'|^2}} \\
&= \left(\delta^\mu_\alpha - 2\frac{x'^\mu_1 x'_{1\alpha}}{|x_1|'^2}\right)\left(\delta^\nu_\beta - 2\frac{x'^\nu_2 x'_{2\beta}}{|x'_2|^2}\right) \\
&\quad\times\left[g^{\alpha\beta} + \delta\frac{(x'_1|x'_2|^2 - x'_2|x'_1|^2)^\alpha(x'_1|x'_2|^2 - x'_2|x'_1|^2)^\beta}{x'^2_{12}|x'_1|^2|x'_2|^2}\right] \\
&= \Bigg[g^{\mu\beta} + \delta\frac{(x'_1|x'_2|^2 - x'_2|x'_1|^2)^\mu(x'_1|x'_2|^2 - x'_2|x'_1|^2)^\beta}{x'^2_{12}|x'_1|^2|x'_2|^2} - 2\frac{x'^\mu_1 x'^\beta_1}{|x_1|'^2}
\end{aligned}$$

$$
\begin{aligned}
&\quad - 2\delta\frac{x_1'^{\mu}(\cancel{|x_1'|^2}|x_2'|^2 - x_1'\cdot x_2'\cancel{|x_1'|^2})(x_1'|x_2'|^2 - x_2'|x_1'|^2)^{\beta}}{x_{12}'^2|x_1'|^2\cancel{|x_1'|^2}|x_2'|^2}\Bigg]\left(\delta^{\nu}_{\beta} - 2\frac{x_2'^{\nu}x_{2\beta}'}{|x_2'|^2}\right)\\
&= \left[g^{\mu\beta} - \delta\frac{(x_1'|x_2'|^2 + x_2'|x_1'|^2 - 2x_1'(x_1'\cdot x_2'))^{\mu}(x_1'|x_2'|^2 - x_2'|x_1'|^2)^{\beta}}{x_{12}'^2|x_1'|^2|x_2'|^2} - 2\frac{x_1'^{\mu}x_1'^{\beta}}{|x_1|'^2}\right]\\
&\quad\times\left(\delta^{\nu}_{\beta} - 2\frac{x_2'^{\nu}x_{2\beta}'}{|x_2'|^2}\right)\\
&= \left[g^{\mu\nu} - \delta\frac{(x_1'|x_2'|^2 + x_2'|x_1'|^2 - 2x_1'(x_1'\cdot x_2'))^{\mu}(x_1'|x_2'|^2 - x_2'|x_1'|^2)^{\nu}}{x_{12}'^2|x_1'|^2|x_2'|^2} - 2\frac{x_1'^{\mu}x_1'^{\nu}}{|x_1|'^2}\right]\\
&\quad - 2\frac{x_2'^{\nu}x_2'^{\mu}}{|x_2'|^2} + 2\delta\frac{(x_1'|x_2'|^2 + x_2'|x_1'|^2 - 2x_1'(x_1'\cdot x_2'))^{\mu}(x_2'\cdot x_1'\cancel{|x_2'|^2} - \cancel{|x_2'|^2}|x_1'|^2)x_2'^{\nu}}{x_{12}'^2|x_1'|^2|x_2'|^2\cancel{|x_2'|^2}}\\
&\quad + 4\frac{x_1'^{\mu}x_2'^{\nu}(x_1'\cdot x_2')}{|x_1|'^2}\\
&= g^{\mu\nu} - 2\frac{x_1'^{\mu}x_1'^{\nu}}{|x_1|'^2} - 2\frac{x_2'^{\nu}x_2'^{\mu}}{|x_2'|^2} + 4\frac{x_1'^{\mu}x_2'^{\nu}(x_1'\cdot x_2')}{|x_1|'^2}\\
&\quad - \delta\frac{\{x_1'|x_2'|^2 + x_2'|x_1'|^2 - 2x_1'(x_1'\cdot x_2')\}^{\mu}\{x_1'|x_2'|^2 + x_2'|x_1'|^2 - 2(x_1'\cdot x_2')x_2'\}^{\nu}}{x_{12}'^2|x_1'|^2|x_2'|^2}\\
&= g^{\mu\nu} - 2\frac{x_1'^{\mu}x_1'^{\nu}}{|x_1|'^2} - 2\frac{x_2'^{\nu}x_2'^{\mu}}{|x_2'|^2} + 4\frac{x_1'^{\mu}x_2'^{\nu}(x_1'\cdot x_2')}{|x_1|'^2}\\
&\quad - \delta\frac{\{x_2'|x_1'|^2 + |x_1' - x_2'|^2x_1' - |x_1'|^2x_1'\}^{\mu}\{x_1'|x_2'|^2 + |x_1' - x_2'|^2x_2' - |x_2'|^2x_2'\}^{\nu}}{x_{12}'^2|x_1'|^2|x_2'|^2}\\
&= g^{\mu\nu} - 2\frac{x_1'^{\mu}x_1'^{\nu}}{|x_1|'^2} - 2\frac{x_2'^{\nu}x_2'^{\mu}}{|x_2'|^2} + 4\frac{x_1'^{\mu}x_2'^{\nu}(x_1'\cdot x_2')}{|x_1|'^2}\\
&\quad + \delta\frac{x_{12}'^{\mu}x_{12}'^{\nu}}{x_{12}'^2} - \delta\frac{x_1'^{\mu}}{|x_1'|^2}x_{12}'^{\nu} + \delta\frac{x_2'^{\nu}}{|x_2'|^2}x_{12}'^{\mu} - \delta|x_{12}'|^2\frac{x_1'^{\mu}x_2'^{\nu}}{|x_1'|^2|x_2'|^2}\\
&= g^{\mu\nu} - (\delta+2)\frac{x_1'^{\mu}x_1'^{\nu}}{|x_1|'^2} - (\delta+2)\frac{x_2'^{\nu}x_2'^{\mu}}{|x_2'|^2} + 2(\delta+2)\frac{x_1'^{\mu}x_2'^{\nu}(x_1'\cdot x_2')}{|x_1|'^2} + \delta\frac{x_1'^{\mu}\cancel{x_2'^{\nu}}}{\cancel{|x_1'|^2}}\\
&\quad + \delta\frac{\cancel{x_1'^{\mu}}x_2'^{\nu}}{\cancel{|x_2'|^2}} - \delta\cancel{|x_1'|^2}\frac{x_1'^{\mu}x_2'^{\nu}}{\cancel{|x_1'|^2}|x_2'|^2} - \delta\cancel{|x_2'|^2}\frac{x_1'^{\mu}x_2'^{\nu}}{|x_1'|^2\cancel{|x_2'|^2}} + \delta\frac{x_{12}'^{\mu}x_{12}'^{\nu}}{x_{12}'^2}
\end{aligned}
$$

where in the last line the four cross terms proportional to $\delta\, x_1'^{\mu}x_2'^{\nu}/(|x_1'|^2|x_2'|^2)$ cancel pairwise (two from expanding $x_{12}' = x_1' - x_2'$ in $|x_1'|^2$, two from expanding it in $|x_2'|^2$). Comparing to $g^{\mu\nu} + \delta\, x_{12}'^{\mu}x_{12}'^{\nu}/|x_{12}'|^2$, the three terms not already of that tensor structure must cancel among themselves, which requires exactly $\delta + 2 = 0$, i.e.

$$\delta = -2. \tag{12}$$

Hence, the two point function is given as

$$\langle J_{\mu}(x)J_{\nu}(0)\rangle = \frac{C}{|x|^{2\Delta}}\left[g_{\mu\nu} - 2\frac{x_{\mu}x_{\nu}}{x^2}\right] \tag{13}$$

[a] See pp. 17–18 of *Quantum Gravity and Cosmology Based on Conformal Field Theory*, or §4.5 of Andrew Evans's *A Conformal Field Theory Primer in $D \geq 3$*.

[b] See eq. 3.18 of *Conformal Field Theory with Boundaries and Defects*, or eqs. 1.55 and 1.60 of the *EPFL Lectures on Conformal Field Theory in $D \geq 3$ Dimensions*: $O'^{\mu}(x') = |\partial x/\partial x'|^{(\Delta+1)/d} I^{\mu}_{\nu}(x')O^{\nu}(x)$,

i.e. $|\partial x/\partial x'|^{\Delta/d}$ for the scaling weight and one extra factor of $|\partial x/\partial x'|^{-1/d}\,\partial x'^{\mu}/\partial x^{\nu} = |\partial x/\partial x'|^{\Delta/d} I^{\mu}_{\ \nu}$ per vector index, using $\partial x'^{\mu}/\partial x^{\nu} = |\partial x/\partial x'|^{-1/d} I^{\mu}_{\ \nu}$.

Alternatively, the boundary tensor structure can be derived with far less algebra using the *embedding space formalism*, where the d-dimensional conformal group $SO(d+1,1)$ acts linearly as Lorentz transformations in $(d+2)$-dimensional Minkowski space.

Two-point function from the embedding-space formalism

Physical d-dimensional space is embedded as the null cone $X^2 \equiv X^A X_A = 0$ inside $\mathbb{R}^{d+1,1}$, with coordinates $X^A = (X^+, X^-, X^\mu)$ related to physical points via the Euclidean section

$$X^+ = 1, \qquad X^\mu = x^\mu, \qquad X^- = x^2$$

(this last relation is exactly the null condition $X^2 = -X^+X^- + X^\mu X_\mu = 0$ solved for X^-). Any other constant value of X^+ describes the very same physical point $x^\mu = X^\mu/X^+$, so rescaling $X \to \lambda X$ along a null ray is a gauge redundancy, not a new point. The entire point of going up to $d+2$ dimensions is that the unwieldy, non-linear d-dimensional conformal transformations of x^μ become ordinary linear Lorentz transformations $X \to \Lambda X$ of X^A, which is why conformally invariant objects are so much easier to write down upstairs.

A conformal primary $O(x)$ of dimension Δ (with n indices) is defined as the projection of a tensor $O_{A_1\cdots A_n}(X)$ on the null cone satisfying exactly the two non-trivial conditions listed above:

- *homogeneous of degree* $-\Delta$, $O(\lambda X) = \lambda^{-\Delta} O(X)$: since rescaling X leaves the physical point untouched, this is precisely what is needed for the projected object to carry a well-defined scaling dimension Δ under the physical dilatation $x \to \lambda x$;
- *transverse*, $X^{A_i} O_{A_1\cdots A_n}(X) = 0$: differentiating the defining condition $X^2 = 0$ with respect to x^μ gives $X_A\, \partial X^A/\partial x^\mu = 0$ identically, so the projection map below automatically annihilates any component of O_A along X^A itself – such components carry no physical information and may as well be set to zero from the start.

Projecting the indices down to $\mathbb{R}^d$ is then simply

$$\langle O^\mu(x_1) O^\nu(x_2)\rangle = \frac{\partial X_1^A}{\partial x_1^\mu}\frac{\partial X_1^B}{\partial x_2^\nu}\,\langle O_A(X_1) O_B(X_2)\rangle$$

For example, the most general form of the two-point function of two operators with spin-1 and dimension Δ can be derived as:[a]

$$\left\langle O^A(X_1) O^B(X_2)\right\rangle = \frac{C_{12}}{(X_1\cdot X_2)^\Delta}\left[\eta^{AB} + \alpha\frac{X_2^A X_1^B}{X_1\cdot X_2} + \beta\frac{X_1^A X_2^B}{X_1\cdot X_2}\right]$$

We drop the last term, since it projects to zero anyway.

$$\left\langle O^A(X_1) O^B(X_2)\right\rangle = \frac{C_{12}}{(X_1\cdot X_2)^\Delta}\left[\eta^{AB} + \alpha\frac{X_2^A X_1^B}{X_1\cdot X_2}\right]$$

According to the transverse condition

$$X_{A1}\left\langle O^A(X_1) O^B(X_2)\right\rangle = \frac{C_{12}}{(X_1\cdot X_2)^\Delta}[X_1^B + \alpha X_1^B] = 0 \implies \alpha = -1$$

we now use the projection to find the correlation function in $\mathbb{R}^d$:

$$\begin{aligned}\langle O_\mu(x_1)O_\nu(x_2)\rangle &= \frac{\partial X_1^A}{\partial x_1^\mu}\frac{\partial X_2^B}{\partial x_2^\nu}\langle O_A(X_1)O_B(X_2)\rangle \\ &= \frac{\partial X_1^A}{\partial x_1^\mu}\frac{\partial X_2^B}{\partial x_2^\nu}\frac{C_{12}}{(X_1\cdot X_2)^\Delta}\left[\eta_{AB}-\frac{X_{A2}X_{B1}}{X_1\cdot X_2}\right] \\ &= \frac{C_{12}}{(X_1\cdot X_2)^\Delta}\left[\frac{\partial X_1^A}{\partial x_1^\mu}\frac{\partial X_2^B}{\partial x_2^\nu}\eta_{AB}-\frac{\partial X_1^A}{\partial x_1^\mu}\frac{\partial X_2^B}{\partial x_2^\nu}\frac{X_{A2}X_{B1}}{X_1\cdot X_2}\right] \\ &= \frac{C_{12}}{(x_1-x_2)^{2\Delta}}\left[g_{\mu\nu}-2\frac{(x_1-x_2)_\mu(x_1-x_2)_\nu}{(x_1-x_2)^2}\right]\end{aligned}$$

where we used $X^A=(X^a,X^+,X^-)=(x^a,1,x^2)$, $X_B=\left(x_a,-\frac{1}{2}x^2,-\frac{1}{2}\right)$ and $\eta_{ab}=I_{d\times d}$ with $\eta_{+-}=\eta_{-+}=-1/2$

$$\frac{\partial X_1^A}{\partial x_1^\mu}\frac{\partial X_2^B}{\partial x_2^\nu}\eta_{AB}=g_{\mu\nu}$$

$$\begin{aligned}\frac{\partial X_1^A}{\partial x_1^\mu}X_{A2} &= \eta_{ab}\frac{\partial x_1^a}{\partial x_1^\mu}x_2^b-\frac{1}{2}\frac{\cancelto{0}{\partial 1}}{\partial x_1^\mu}x_2^2-\frac{1}{2}\frac{\partial x_1^2}{\partial x_1^\mu}1 \\ &= \eta_{ab}\delta_\mu^a x_2^b-x_{1\mu}=(x_2-x_1)_\mu=-(x_1-x_2)_\mu \\ \frac{\partial X_2^B}{\partial x_2^\nu}X_{B1} &= \eta_{ab}\frac{\partial x_2^a}{\partial x_2^\mu}x_1^b-\frac{1}{2}\frac{\cancelto{0}{\partial 1}}{\partial x_2^\mu}x_1^2-\frac{1}{2}\frac{\partial x_2^2}{\partial x_2^\mu}1=(x_1-x_2)_\nu\end{aligned}$$

$$(X_1-X_2)^A(X_1-X_2)_A=(x_1-x_2)^a(x_1-x_2)_a-\frac{1}{2}(1-1)(x_1^2-x_2^2)-(x_1^2-x_2^2)(\tfrac{1}{2}-\tfrac{1}{2})$$

$$\implies X_1\cdot X_2=-\frac{1}{2}(x_1-x_2)^2$$

[a] The terms in brackets are chosen to be invariant under $x\to\lambda x$; we are not using the transformation law for any of them, under which even the metric would change as $\eta_{AB}\to\lambda^{-2}\eta_{AB}$.

For a general symmetric, traceless primary operator $O_{i_1\cdots i_s}(\vec{x})$ of spin s, we contract the tensor indices with a null auxiliary polarization vector $\vec{\epsilon}$ (satisfying $\vec{\epsilon}\cdot\vec{\epsilon}=0$), defining $\epsilon^s\cdot O\equiv\epsilon^{i_1}\cdots\epsilon^{i_s}O_{i_1\cdots i_s}$. Conformal invariance generalizes the vector result to:

$$\langle\epsilon^s\cdot O(\vec{x})\,\tilde{\epsilon}^s\cdot O(\vec{0})\rangle=\frac{\left(\vec{\epsilon}\cdot\vec{\tilde{\epsilon}}-2(\vec{\epsilon}\cdot\hat{x})(\vec{\tilde{\epsilon}}\cdot\hat{x})\right)^s}{|\vec{x}|^{2\Delta}}=\frac{\left(\vec{\epsilon}\cdot R_x\vec{\tilde{\epsilon}}\right)^s}{|\vec{x}|^{2\Delta}},\tag{11.2.2}$$

where $\hat{x}\equiv\vec{x}/|\vec{x}|$ and R_x is the spatial reflection matrix:

$$R_x\vec{\tilde{\epsilon}}\equiv\vec{\tilde{\epsilon}}-2(\vec{\tilde{\epsilon}}\cdot\hat{x})\hat{x}.\tag{11.2.3}$$

The reflection matrix R_x has a direct physical meaning: the subgroup of the conformal group that fixes the two insertion points acts as an $SO(3)$ rotation around each point, with the rotation at one point conjugated by a reflection along the separation axis $\hat{x}$. Consequently, massive spinning particles pair-produced in de Sitter space have correlated spin orientations, ensuring that the total angular momentum along the separation axis vanishes.

11.2.2 Conformal Ward identities in momentum space

To calculate cosmological correlators and primordial non-Gaussianities, we need the spinning two-point function in momentum space. In momentum space, conformal invariance translates

into the boundary conformal Ward identities:

$$D\left\langle O_\Delta(\vec{k})O_\Delta(-\vec{k})\right\rangle' = 0, \qquad \mathbf{b}\cdot\mathbf{K}\left\langle O_\Delta(\vec{k})O_\Delta(-\vec{k})\right\rangle' = 0. \tag{11.2.4}$$

For scalar operators of scaling dimension Δ, the dilatation operator D and special conformal transformation generator $\mathbf{b}\cdot\mathbf{K}$ (with parameter $\vec{b}$) act in momentum space as:

$$D = -\left(\Delta - 3 - \vec{k}\cdot\partial_{\vec{k}}\right), \tag{11.2.5}$$

$$\mathbf{b}\cdot\mathbf{K} = -2(\Delta-3)\vec{b}\cdot\partial_{\vec{k}} - \left[(\vec{b}\cdot\vec{k})\partial_{\vec{k}}^2 - 2(\vec{k}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{k}})\right]. \tag{11.2.6}$$

The dilatation and SCT generators in momentum space

A boundary primary $O_\Delta(\vec{x})$ of dimension Δ transforms under dilatation and under a special conformal transformation with parameter $\vec{b}$ as (this is simply the $\eta\to 0$ limit of the de Sitter isometries acting on $\Phi = \sum_\Delta \eta^\Delta O_\Delta(\vec{x})$, once the overall power of η carried by each term is stripped off)

$$\begin{aligned} D\,O_\Delta(\vec{x}) &= -\left(\Delta + \vec{x}\cdot\partial_{\vec{x}}\right)O_\Delta(\vec{x}), \\ \mathbf{b}\cdot\mathbf{K}\,O_\Delta(\vec{x}) &= \left[-2(\vec{b}\cdot\vec{x})\Delta - 2(\vec{b}\cdot\vec{x})(\vec{x}\cdot\partial_{\vec{x}}) + x^2(\vec{b}\cdot\partial_{\vec{x}})\right]O_\Delta(\vec{x}). \end{aligned}$$

To find how these act on the momentum-space operator, write $O_\Delta(\vec{x}) = \int d^3k\, e^{i\vec{k}\cdot\vec{x}}O_\Delta(\vec{k})$ and integrate by parts in $\vec{k}$ until every explicit factor of $\vec{x}$ has migrated onto the exponential as a $\vec{k}$-derivative: this amounts to the replacement $x^\mu \to -i\partial_{k^\mu}$, $\partial_{x^\mu}\to -ik^\mu$, with the understanding that whichever operator used to act first (innermost) still acts first, so under the replacement the order gets reversed. For dilatation this is immediate: using the product rule identity $\partial_{k^j}k^j = 3 + k^j\partial_{k^j}$ (the 3 coming from the trace of the identity, in 3 spatial dimensions),

$$D\,O_\Delta(\vec{k}) = -\left[\Delta - i\partial_{k^j}\left(-ik^j\right)\right]O_\Delta(\vec{k}) = -\left(\Delta - 3 - k^j\partial_{k^j}\right)O_\Delta(\vec{k}),$$

and setting this to zero on the two point function is exactly the equation solved above by $O\propto k^{2\Delta-3}$. The same replacement on $\mathbf{b}\cdot\mathbf{K}$ needs two more product-rule identities of the same kind, $(\vec{b}\cdot\partial_{\vec{k}})(\vec{k}\cdot\partial_{\vec{k}}) = \vec{b}\cdot\partial_{\vec{k}} + (\vec{k}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{k}})$ and $\partial_{\vec{k}}^2(\vec{b}\cdot\vec{k}) = 2\vec{b}\cdot\partial_{\vec{k}} + (\vec{b}\cdot\vec{k})\partial_{\vec{k}}^2$, both of which just distribute the derivative over the $\vec{k}$ that it hits directly:

$$\begin{aligned} -i\,\mathbf{b}\cdot\mathbf{K} &= -2(\vec{b}\cdot\vec{x})\Delta - 2(\vec{b}\cdot\vec{x})(\vec{x}\cdot\partial_{\vec{x}}) + x^2(\vec{b}\cdot\partial_{\vec{x}}) \\ &= 2i\Delta\,\vec{b}\cdot\partial_{\vec{k}} - 2i\left[\vec{b}\cdot\partial_{\vec{k}} + 3\,\vec{b}\cdot\partial_{\vec{k}} + (\vec{k}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{k}})\right] + i\left[2\,\vec{b}\cdot\partial_{\vec{k}} + (\vec{b}\cdot\vec{k})\partial_{\vec{k}}^2\right] \\ &= i\left[2(\Delta-3)\vec{b}\cdot\partial_{\vec{k}} - 2(\vec{k}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{k}}) + (\vec{b}\cdot\vec{k})\partial_{\vec{k}}^2\right] \end{aligned}$$

Dividing out the common $-i$ – an overall normalization of the abstract charge $\mathbf{b}\cdot\mathbf{K}$ that drops out once we use it as a real differential operator, exactly as for D above – gives

$$\mathbf{b}\cdot\mathbf{K} = -2(\Delta-3)\vec{b}\cdot\partial_{\vec{k}} - \left[(\vec{b}\cdot\vec{k})\partial_{\vec{k}}^2 - 2(\vec{k}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{k}})\right],$$

11.2.3 Momentum-space two-point function with spin

For operators with spin s, the generator $\mathbf{b}\cdot\mathbf{K}$ acquires an additional spin piece acting on the polarization vector $\vec{\epsilon}$:[2]

$$\mathbf{b}\cdot\mathbf{K} = -2(\Delta-3)\vec{b}\cdot\partial_{\vec{k}} - \left[(\vec{b}\cdot\vec{k})\partial_{\vec{k}}^2 - 2(\vec{k}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{k}})\right] + 2\left[(\vec{\epsilon}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{\epsilon}}) - (\vec{b}\cdot\vec{\epsilon})(\partial_{\vec{k}}\cdot\partial_{\vec{\epsilon}})\right]. \tag{11.2.7}$$

Choosing the parameter $\vec{b}$ orthogonal to both the momentum $\vec{k}$ and polarization vector $\vec{\epsilon}$ ($\vec{b}\cdot\vec{k}=0$ and $\vec{b}\cdot\vec{\epsilon}=0$), this operator simplifies to:

$$\mathbf{b}\cdot\mathbf{K} = -2(\Delta-3)\vec{b}\cdot\partial_{\vec{k}} + 2(\vec{k}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{k}}) + 2(\vec{\epsilon}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{\epsilon}}). \tag{11.2.8}$$

Choosing coordinates with $\vec{k}=(0,0,k)$ along the $\hat{z}$-axis and specifying the transverse polarization vectors by phase angles ψ,ψ':

$$\vec{\epsilon} = (\cos\psi, \sin\psi, i), \qquad \vec{\tilde{\epsilon}} = (\cos\psi', \sin\psi', -i), \tag{11.2.9}$$

the Ward identity $\mathbf{b}\cdot\mathbf{K}\left\langle \epsilon^s\cdot O_{\Delta,s}(\vec{k})\,\tilde{\epsilon}^s\cdot O_{\Delta,s}(-\vec{k})\right\rangle' = 0$ uniquely determines the momentum-space two-point function:

$$\left\langle \epsilon^s\cdot O_{\Delta,s}(\vec{k})\,\tilde{\epsilon}^s\cdot O_{\Delta,s}(-\vec{k})\right\rangle' = k^{2\Delta-3} I_2(\vec{\epsilon},\vec{\tilde{\epsilon}},\hat{k}), \tag{11.2.10}$$

where the angular spin factor I_2 is given by the helicity sum:

$$I_2(\vec{\epsilon},\vec{\tilde{\epsilon}},\hat{k}) \equiv \sum_{m=-s}^{s} e^{im\chi}\frac{(2s)!}{(s-m)!(s+m)!}\frac{\Gamma(\Delta-1+m)\Gamma(2-\Delta+s)}{\Gamma(2-\Delta+m)\Gamma(\Delta-1+s)} \tag{11.2.11}$$

$$= e^{-is\chi}\,{}_2F_1\left(\Delta-s-1,-2s,2-s-\Delta;-e^{i\chi}\right), \tag{11.2.12}$$

with $\chi\equiv\psi-\psi'$. Here $m\in\{-s,\ldots,s\}$ labels the helicity (angular momentum component along the $\vec{k}$ axis) of the intermediate mode.

Solving the spinning conformal Ward identity

We begin by operating the SCT generator on the two point correlation function. We consider[a]

$$\epsilon^i\mathbf{b}\cdot\mathbf{K}O_i = \mathbf{b}\cdot\mathbf{K}(O_i\epsilon^i) - O_i\mathbf{b}\cdot\mathbf{K}\epsilon^i$$

whose generalization

$$\mathbf{b}\cdot\mathbf{K} = -(\Delta-3)2\vec{b}\cdot\partial_{\vec{k}} - [(\vec{b}\cdot\vec{k})\partial_{\vec{k}}^2 - 2(\vec{k}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{k}})] + 2[(\vec{\epsilon}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{\epsilon}}) - (\vec{b}\cdot\vec{\epsilon})(\partial_{\vec{k}}\cdot\partial_{\vec{\epsilon}})]$$

where the last term comes from the product rule and using $O(\varepsilon)=\varepsilon^i O_i$ together with the action of the spin generator on an arbitrary tensor[b]:

$$(\Sigma^{\mu\nu}O)_i = \delta_i^\mu O_\nu - \delta_i^\nu O_\mu \quad\Rightarrow\quad \varepsilon^i(\Sigma^{\mu\nu}O)_i = \varepsilon^\mu O_\nu - \varepsilon^\nu O_\mu.$$

This can also be generated by the action of the following differential operator:

$$(\varepsilon^\mu\partial_{\varepsilon_\nu} - \varepsilon^\nu\partial_{\varepsilon_\mu})(\varepsilon^\alpha O_\alpha) = \varepsilon^\mu O_\nu - \varepsilon^\nu O_\mu,$$

[2] Acting on primed correlators $\langle\cdots\rangle'$, momentum-conserving delta function terms cancel once dilatation invariance is imposed.

For the general case, the same steps give[c]

$$\begin{aligned}-i(-2b_\mu x_\beta L_\epsilon^{\mu\beta}) &= 2b_\mu i[\epsilon^\mu\partial_\epsilon^\beta - \epsilon^\beta\partial_\epsilon^\mu](-i\partial_{k^\beta}) \\ &= 2b_\mu\{\epsilon^\mu\partial_\epsilon^\beta - \epsilon^\beta\partial_\epsilon^\mu\}\partial_{k^\beta} \\ &= -2b_\mu\{\epsilon^\beta\partial_\epsilon^\mu - \epsilon^\mu\partial_\epsilon^\beta\}\partial_{k^\beta} \\ &= -2[(\vec{\epsilon}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{\epsilon}}) - (\vec{b}\cdot\vec{\epsilon})(\partial_{\vec{k}}\cdot\partial_{\vec{\epsilon}})]\end{aligned}$$

assuming

$$\vec{b}\cdot\vec{k} = 0 \qquad\qquad \vec{b}\cdot\vec{\epsilon} = 0$$

we have:

$$\mathbf{b}\cdot\mathbf{K} = -(\Delta-3)2\vec{b}\cdot\partial_{\vec{k}} + 2(\vec{k}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{k}}) + 2(\vec{\epsilon}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{\epsilon}})$$

As a check, if we ignore the spin-dependent piece for a moment, the first two terms alone repeat the exact argument that fixed the scalar two-point function above, giving back $O \propto k^{2\Delta-3}$ (writing p for the radial derivative direction here, as in that scalar argument). The ansatz for the spinning two-point function must reduce to this when the spin dependence is stripped away:

$$\langle\vec{\epsilon}^{\,s}\cdot O_s\bar{\epsilon}^{\bar{s}}O_s\rangle' = k^{2\Delta-3}h \qquad h = \sum_{i=0}^{s} c_i\rho^i(\alpha\beta)^{s-i},$$

$$\alpha \equiv \frac{\vec{\epsilon}\cdot\vec{k}}{k}, \qquad \beta \equiv \frac{\vec{\bar{\epsilon}}\cdot\vec{k}}{k}, \qquad \rho \equiv \vec{\epsilon}\cdot\vec{\bar{\epsilon}}$$

Therefore,

$$\mathbf{b}\cdot\mathbf{K}\,\langle\epsilon\cdot O\bar{\epsilon}\cdot O\rangle' = \mathbf{b}\cdot\mathbf{K}\left[k^{2\Delta-3}\sum_{i=0}^{s} c_i\rho^i(\alpha\beta)^{s-i}\right] = 0 \tag{1}$$

The part of the operator acting purely on $k^{2\Delta-3}$ vanishes by the usual conformal invariance condition for scalar operators[d]:

$$\begin{aligned}\mathbf{b}\cdot\mathbf{K}\,\langle O(k)O(-k)\rangle &= \mathbf{b}\cdot\mathbf{K}(k^{2\Delta-3}) \\ &= 0\end{aligned}$$

we will be using the following properties for simplification

$$\partial_{\vec{k}}\alpha = \frac{1}{k}(\vec{\epsilon} - \hat{k}\alpha), \qquad \partial_{\vec{k}}\beta = \frac{1}{k}(\vec{\bar{\epsilon}} - \hat{k}\beta), \qquad \partial_{\vec{k}}\rho = 0$$

$$\partial_{\vec{\epsilon}}\alpha = \vec{k}, \qquad \partial_{\vec{\epsilon}}\beta = 0, \qquad \partial_{\vec{\epsilon}}\rho = \vec{\bar{\epsilon}}$$

The only term in (1) which will survive are

$$k^{2\Delta-3}(\mathbf{b}\cdot\mathbf{K}h) + 2(\vec{k}\partial_{\vec{k}}k^{2\Delta-3})(\vec{b}\cdot\partial_{\vec{k}}h) + 2(\vec{\epsilon}\cdot\partial_{\vec{k}}k^{2\Delta-3})(\vec{b}\cdot\partial_{\vec{\epsilon}}h) + h(\mathbf{b}\cdot\mathbf{K}k^{2\Delta-3}) = 0$$

$$[(\Delta-1)\vec{b}\cdot\partial_{\vec{k}} + (2\Delta-3)\frac{\epsilon\cdot\vec{k}}{k^2}\vec{b}\cdot\partial_{\vec{k}} + (\vec{\epsilon}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{\epsilon}})]h = 0$$

where the first term is because of

$$-(\Delta-3)2\vec{b}\cdot\partial_{\vec{k}} + 2\underbrace{(\vec{k}\cdot\partial_{\vec{k}})(\vec{b}\cdot\partial_{\vec{k}}h)}_{-\vec{b}\cdot\partial_{\vec{k}}h} + 2(\vec{k}\partial_{\vec{k}}k^{2\Delta-3})(\vec{b}\cdot\partial_{\vec{k}}h)$$

$$= [-(\Delta-3)2\vec{b}\cdot\partial_{\vec{k}} - 2\vec{b}\cdot\partial_{\vec{k}} + 2(2\Delta-3)\vec{b}\cdot\partial_{\vec{k}}]h$$
$$= 2[\Delta\vec{b}\cdot\partial_{\vec{k}} - \vec{b}\cdot\partial_{\vec{k}}]h$$

in the second line we used the following:

$$\vec{b}\cdot\partial_{\vec{k}}\alpha = \frac{1}{k}(\vec{b}\cdot\vec{\epsilon} - \vec{b}\cdot\hat{k}\alpha) = 0$$
$$\vec{b}\cdot\partial_{\vec{k}}\beta = \frac{1}{k}(\vec{b}\cdot\vec{\tilde{\epsilon}} - \vec{b}\cdot\hat{k}\beta) = \frac{\vec{b}\cdot\vec{\tilde{\epsilon}}}{k}$$

$$\vec{b}\cdot\partial_{\vec{k}}h = \sum_i (s-i)(\alpha\beta)^{s-i-1}\left[\cancelto{0}{(\vec{b}\cdot\partial_{\vec{k}}\alpha)}\beta + \alpha(\vec{b}\cdot\partial_{\vec{k}}\beta)\right]$$
$$= \sum_i (s-i)(\alpha\beta)^{s-i-1}\frac{\vec{b}\cdot\vec{\tilde{\epsilon}}}{k}\alpha$$

since the above only has k^{-1} dependence on momentum, the eigenvalue for $\vec{k}\partial_{\vec{k}} = \vec{k}\cdot\hat{k}\partial_k = k\partial_k$ is -1. We now replace h by $(\alpha\beta)^s F(\rho/\alpha\beta)$ and solve the new ODE in $y \equiv \rho/\alpha\beta$:

$$y(y-2)F'' + [(\Delta-s)(2-y)-1]F' + s(\Delta-1)F = 0$$

The solution is given as:

$${}_2F_1(\Delta-1, -s, \Delta-s-1/2, y/2)$$

This is a hypergeometric differential equation, and we are interested in the solution that resembles a Jacobi polynomial, as expected for the polarization sum at integer spin s. This follows from the symmetry in the first two arguments of the hypergeometric function,

$${}_2F_1(a,b,c;z) = {}_2F_1(b,a,c;z),$$

together with eq. 15.4.6 of Abramowitz and Stegun's *Handbook of Mathematical Functions*.

$${}_2F_1(-n, n+\alpha+\beta+1, \alpha+1; z) = \frac{n!}{(\alpha+1)_n}P_n^{(\alpha,\beta)}(1-2z)$$
$$= \frac{s!}{\left(\Delta-s-\frac{1}{2}\right)_s}P_s^{\left(\Delta-s-\frac{3}{2},-\frac{1}{2}\right)}(1-y)$$

For

$$\epsilon = (\cos\psi, \sin\psi, i)$$
$$\tilde{\epsilon} = (\cos\psi', \sin\psi', -i)$$
$$\vec{k} = (0,0,k)$$

we get $y = 1 + \cos\underbrace{(\psi-\psi')}_{\chi} = 1 + \frac{e^{i\chi}+e^{-i\chi}}{2}$.

Before quoting the general result, it is worth seeing exactly where it comes from, since the ODE only fixes h up to an overall constant – exactly as the overall coefficient C of the two-point function was never fixed by the Ward identity to begin with. We remove that ambiguity by a normalization convention: we fix the overall scale of h by demanding that its highest-helicity Fourier mode, $m = \pm s$, has coefficient exactly 1. With that convention

fixed, the lowest spin can be checked entirely by hand. For $s = 1$ the hypergeometric series in (1) terminates after two terms,

$$F(y) = {}_2F_1\left(\Delta - 1, -1; \Delta - \tfrac{3}{2}; \tfrac{y}{2}\right) = 1 - \frac{\Delta - 1}{2\Delta - 3}\, y,$$

and substituting $y = 1 + \cos\chi$ gives

$$F = \frac{\Delta - 2}{2\Delta - 3} - \frac{\Delta - 1}{2\Delta - 3}\cos\chi = \frac{\Delta - 2}{2\Delta - 3} - \frac{\Delta - 1}{2(2\Delta - 3)}\left(e^{i\chi} + e^{-i\chi}\right).$$

Rescaling by $-2(2\Delta - 3)/(\Delta - 1)$ so that the $e^{\pm i\chi}$ coefficient becomes 1, as our convention demands, turns the constant term into $-2(\Delta - 2)/(\Delta - 1)$, so

$$h\big|_{s=1} = e^{i\chi} + e^{-i\chi} - \frac{2(\Delta - 2)}{\Delta - 1}.$$

This is exactly the $m = -1, 0, 1$ terms of the general formula below, evaluated by hand rather than quoted: the $m = \pm 1$ coefficients are 1 by the normalization we imposed, and the $m = 0$ coefficient works out to $-2(\Delta - 2)/(\Delta - 1)$ once the Gamma functions are simplified using $\Gamma(\Delta - 1) = (\Delta - 2)\Gamma(\Delta - 2)$ and $\Gamma(\Delta) = (\Delta - 1)\Gamma(\Delta - 1)$. The same rescale-and-Fourier-expand pattern holds at any spin, order by order in Δ – it is the general statement that this particular family of Jacobi polynomials, evaluated at $\cos\chi$, admits an exact finite Fourier expansion (equivalently, this is the explicit closed form for Wigner's "small" d-matrix).[e] Fourier expanding the Jacobi polynomial this way gives the closed form used throughout the rest of this chapter:

$$h = \sum_{m=-s}^{s} e^{im\chi} \frac{(-1)^{s-m}(2s)!}{(s-m)!(s+m)!} \frac{\Gamma(\Delta - 1 - m)\Gamma(\Delta - 1 + m)}{\Gamma(\Delta - 1 - s)\Gamma(\Delta - 1 + s)} \tag{2}$$

The second, more compact form quoted later in this chapter is just this same finite sum repackaged as a hypergeometric series, which is worth seeing explicitly since it is a purely mechanical rewriting rather than a new physical input. Shift the summation index to $n \equiv m + s$, so that n runs over the $2s + 1$ integers $0, 1, \ldots, 2s$ instead of $m = -s, \ldots, s$. Then $s - m = 2s - n$, $s + m = n$, $(-1)^{s-m} = (-1)^{2s-n} = (-1)^n$ (since s is an integer), and

$$\Delta - 1 - m = (\Delta - 1 + s) - n, \qquad \Delta - 1 + m = (\Delta - 1 - s) + n.$$

Each Gamma-function ratio in (2) turns into a Pochhammer symbol: the upward shift is the ordinary definition, $\Gamma(w + n)/\Gamma(w) = (w)_n$, while the downward shift picks up an extra sign, $\Gamma(z - n)/\Gamma(z) = (-1)^n/(1 - z)_n$ (both follow directly from $\Gamma(x + 1) = x\Gamma(x)$ applied n times), so

$$\frac{\Gamma((\Delta - 1 + s) - n)}{\Gamma(\Delta - 1 + s)} = \frac{(-1)^n}{(2 - \Delta - s)_n}, \qquad \frac{\Gamma((\Delta - 1 - s) + n)}{\Gamma(\Delta - 1 - s)} = (\Delta - 1 - s)_n.$$

The binomial coefficient converts the same way, $\dfrac{(2s)!}{(s-m)!(s+m)!} = \dbinom{2s}{n} = \dfrac{(2s)!}{n!(2s-n)!} = \dfrac{(-1)^n(-2s)_n}{n!}$. Collecting all three pieces, the two explicit factors of $(-1)^n$ cancel against each other, leaving

$$h = e^{-is\chi} \sum_{n=0}^{2s} e^{in\chi} \binom{2s}{n} \frac{(\Delta - 1 - s)_n}{(2 - \Delta - s)_n} = e^{-is\chi} \sum_{n=0}^{2s} \frac{(-2s)_n(\Delta - 1 - s)_n}{(2 - \Delta - s)_n\, n!} \left(-e^{i\chi}\right)^n,$$

and the sum on the right is, by the definition of the hypergeometric series, exactly ${}_2F_1(-2s, \Delta-1-s; 2-\Delta-s; -e^{i\chi})$, i.e.

$$h = e^{-is\chi}{}_2F_1(\Delta-s-1, -2s, 2-s-\Delta; -e^{i\chi}) \tag{3}$$

(the first two arguments may be swapped freely, since ${}_2F_1(a,b;c;z) = {}_2F_1(b,a;c;z)$). As a further, independent check, this closed form agrees with the companion paper's own version of the same coefficient, written using $\Gamma(2-\Delta\pm m)$ instead of $\Gamma(\Delta-1\mp m)$: the two are related by the reflection formula $\Gamma(z)\Gamma(1-z) = \pi/\sin(\pi z)$ together with $\sin(\pi(\Delta-1-s)) = (-1)^{s-m}\sin(\pi(\Delta-1-m))$, which holds because $s-m$ is always an integer.

[a] An alternative implementation of this idea is discussed in arXiv:2212.13135.
[b] See arXiv:2312.05728.
[c] See §A.22 of arXiv:2112.12540, which leads to eq. 3.12 of arXiv:2212.13135.
[d] See §2.2.1 of the literature on four-point functions in momentum space and topological terms in gravity, or eq. 5.45 of Baumann's *Lectures on Cosmological Correlators*.
[e] See eq. A.11 of *Momentum Space Approach to Crossing Symmetric CFT Correlators* for the general-spin resummation, which we do not reproduce here.

11.2.4 Mass-dimension relation, positivity, and partial masslessness

The closed-form result (11.2.12) carries profound physical content regarding the particle spectrum and unitary bounds of de Sitter space:

Mass-Dimension Relation. For a massive spin-s field in 4-dimensional de Sitter space, the boundary conformal dimensions $\Delta_\pm$ are determined by the mass m and spin s through[3]:

$$\Delta_\pm = \frac{3}{2} \pm \sqrt{\left(s-\frac{1}{2}\right)^2 - \frac{m^2}{H^2}} = \frac{3}{2} \pm i\mu. \tag{11.2.13}$$

For heavy fields satisfying $m > (s-\frac{1}{2})H$, the parameter μ is real, and the Gamma function ratios in (11.2.11) become pure phases, producing the characteristic oscillatory clock signal in the squeezed bispectrum. For massless gauge fields ($m=0$), $\Delta_+ = 1+s$, which corresponds to the conformal dimension of a conserved boundary current of spin s.

Positivity and Partial Masslessness. For light fields with $m < (s-\frac{1}{2})H$, the scaling dimensions $\Delta_\pm$ are real, and the late-time field profile is dominated by the more slowly decaying mode with dimension Δ_-.

As a concrete example, consider a massive spin-2 field ($s=2$). Expanding the hypergeometric polynomial in (11.2.12) with $\chi \equiv \psi - \psi'$ gives:

$$\left\langle \epsilon^2 \cdot O_{\vec{k}}\, \tilde{\epsilon}^2 \cdot O_{-\vec{k}} \right\rangle' \propto k^{2\Delta-3}\left[e^{-2i\chi} + \frac{4(3-\Delta)}{\Delta}e^{-i\chi} + \frac{6(3-\Delta)(2-\Delta)}{(\Delta-1)\Delta} + \frac{4(3-\Delta)}{\Delta}e^{i\chi} + e^{2i\chi}\right]. \tag{11.2.14}$$

Starting at $\Delta = 3/2$ where all helicity coefficients are positive, as Δ decreases towards 1, the longitudinal ($m=0$) coefficient diverges at $\Delta = 1$ (corresponding to $m^2 = 2H^2$). This divergence signals the appearance of a new gauge invariance in de Sitter space known as *partial masslessness*, where the helicity-0 mode becomes pure gauge and decouples.

For $\Delta < 1$ (i.e. $m^2 < 2H^2$), the $m=0$ coefficient flips sign and becomes negative. Since this term represents the expectation value of a field and its Hermitian conjugate, $\left\langle OO^\dagger \right\rangle$, a negative

[3] See e.g. Deser and Waldron (2001), arXiv:hep-th/0103255.

coefficient violates reflection positivity and unitarity in de Sitter space. Unitarity of spinning representations in de Sitter spacetime thus enforces the celebrated *Higuchi bound*:

$$m^2 \geq s(s-1)H^2 \qquad (s \geq 1). \tag{11.2.15}$$

Conformal symmetry on the late-time boundary therefore directly encodes both the mass spectrum and the fundamental unitarity constraints of cosmological spacetimes.

11.3 Three-Point Functions in de Sitter Space

Conformal symmetry on the future boundary also constrains three-point functions. When one of the inserted operators carries spin s, the tensor structure of the correlation function is uniquely determined by conformal invariance.

For two scalar operators $O_1(x_1), O_2(x_2)$ of dimensions Δ_1, Δ_2 and a spin-s operator contracted with auxiliary polarization vector $\vec{\epsilon}$ at x_3, the position-space three-point function is fixed to:

$$\langle O_1(x_1)O_2(x_2)\epsilon^s \cdot O(x_3)\rangle = \frac{(\vec{\epsilon}\cdot W)^s}{|x_{13}|^{\Delta_1+\Delta_3-\Delta_2+s}|x_{23}|^{\Delta_2+\Delta_3-\Delta_1+s}|x_{12}|^{\Delta_1+\Delta_2-\Delta_3+s}}, \tag{11.3.1}$$

where the conformally invariant spatial vector W_μ is given by:

$$W_\mu \equiv \frac{|x_{23}|^2(x_{13})_\mu - |x_{13}|^2(x_{23})_\mu}{|x_{12}||x_{13}||x_{23}|}. \tag{11.3.2}$$

Three-point function with spin from the embedding space

On the null cone we will have

$$\langle \phi_1(X_1)\phi_2(X_2)J_M(X_3)\rangle = \frac{W_M}{(-2X_1\cdot X_2)^{\alpha_{123}}(-2X_1\cdot X_3)^{\alpha_{132}}(-2X_2\cdot X_3)^{\alpha_{231}}}$$

where the powers α_{ijk} of the scalar factor are determined by the dilatation as in case of scalar operators, and the tensor structure W_M – invariant under dilatation, and transverse to all three points, $(X_1)^M W_M = (X_2)^M W_M = (X_3)^M W_M = 0$ – equals to

$$W_M = \frac{(-2X_2\cdot X_3)X_{1M} - (-2X_1\cdot X_3)X_{2M} - (-2X_1\cdot X_2)X_{3M}}{(-2X_1\cdot X_2)^{\frac{1}{2}}(-2X_1\cdot X_3)^{\frac{1}{2}}(-2X_2\cdot X_3)^{\frac{1}{2}}}.$$

A few comments on this tensor structure: the relative sign is, as before, fixed by transversality. We drop the term proportional to X_{3M}: this leg will eventually be projected down with $\partial X_3^M/\partial x_3^\mu$, and we already showed in §11.4 that $X_{3M}\,\partial X_3^M/\partial x_3^\mu = 0$ identically, so any piece of W_M along X_{3M} contributes nothing to the physical correlator and can simply be dropped now. The scaling behavior under dilatation is completely fixed by the scalar prefactor, so the tensor structure itself has scaling weight 0 in every variable ($X \to \lambda X \implies W_\mu \to \lambda^0 W_\mu$), and it is immediate to check that it is transverse, i.e. $(X_3)_M W_M = 0$. Projecting to physical space,

$$\langle \phi_1(x_1)\phi_2(x_2)J_\mu(x_3)\rangle = \frac{\partial X_3^M}{\partial x_3^\mu}\langle \phi_1(X_1)\phi_2(X_2)J_M(X_3)\rangle$$

we find, as explicitly computed before,

$$\frac{\partial X_3^M}{\partial x_3^\mu}X_{iM} = (x_i - x_3)_\mu, \quad i = 1,2$$
$$-2X_i\cdot X_j = (x_i - x_j)^2, \quad i = 1,2,3\ (i<j),$$

so that we end up with the tensor structure

$$W_\mu = \frac{|x_2-x_3|^2(x_1-x_3)_\mu - |x_1-x_3|^2(x_2-x_3)_\mu}{|x_1-x_2||x_1-x_3||x_2-x_3|} = \frac{|x_{23}|^2(x_{13})_\mu - |x_{13}|^2(x_{23})_\mu}{|x_{12}||x_{13}||x_{23}|}$$

Therefore, the three-point function of two scalars and one spin 1 operators in physical space is given by

$$\begin{aligned}\langle\phi_1(x_1)\phi_2(x_2)J_\mu(x_3)\rangle &= \frac{\frac{|x_{23}|^2(x_{13})_\mu-|x_{13}|^2(x_{23})_\mu}{|x_{12}||x_{13}||x_{23}|}}{|x_{12}|^{\Delta_1+\Delta_2-\Delta_3}|x_{13}|^{\Delta_1+\Delta_3-\Delta_2}|x_{23}|^{\Delta_2+\Delta_3-\Delta_1}} \\ &= \frac{|x_{23}|^2(x_{13})_\mu - |x_{13}|^2(x_{23})_\mu}{|x_{12}|^{\Delta_1+\Delta_2-\Delta_3+1}|x_{13}|^{\Delta_1+\Delta_3-\Delta_2+1}|x_{23}|^{\Delta_2+\Delta_3-\Delta_1+1}}\end{aligned}$$

For spin-1 ($s=1$), this yields:

$$\langle\phi_1(x_1)\phi_2(x_2)J_\mu(x_3)\rangle = \frac{|x_{23}|^2(x_{13})_\mu - |x_{13}|^2(x_{23})_\mu}{|x_{12}|^{\Delta_1+\Delta_2-\Delta_3+1}|x_{13}|^{\Delta_1+\Delta_3-\Delta_2+1}|x_{23}|^{\Delta_2+\Delta_3-\Delta_1+1}}. \tag{11.3.3}$$

Higher-spin operators $J_{\mu_1\ldots\mu_s}$ are built simply by taking s symmetric copies of W_μ, which is the unique conformally invariant vector for three boundary points.

11.4 Two Conformally Coupled Fields and a General Scalar

We now calculate three-point correlation functions in momentum space, starting with two conformally coupled scalars ($\Delta_1=\Delta_2=2$, corresponding to $m^2=2H^2$) exchanging a general massive scalar field σ_Δ of conformal dimension $\Delta=\frac{3}{2}+i\mu$.

By spatial translation invariance, the three momenta form a closed triangle, $\vec{k}_1+\vec{k}_2+\vec{k}_3=0$. Dilatation invariance allows us to trade the magnitudes k_1,k_2,k_3 for the scale-invariant kinematic variables:

$$p\equiv\frac{k_1+k_2}{k_3}, \qquad q\equiv\frac{k_1-k_2}{k_3}, \tag{11.4.1}$$

where $p\geq 1$ and $q\in[-1,1]$ parameterize the shape of the momentum triangle. The momentum-space correlator is then written as:

$$\left\langle\phi(\vec{k}_1)\phi(\vec{k}_2)\sigma_\Delta(\vec{k}_3)\right\rangle' = k_3^{\Delta-2}F(p,q) = k_3^{\Delta-2}\left[G(p)+G(q)\right]. \tag{11.4.2}$$

Derivatives in the q,p variables

$$\frac{\partial}{\partial k_1} = \frac{\partial p}{\partial k_1}\frac{\partial}{\partial p} = \frac{1}{k_3}\frac{\partial}{\partial p}$$

$$\frac{\partial}{\partial k_2} = \frac{\partial p}{\partial k_2}\frac{\partial}{\partial p} = \frac{1}{k_3}\frac{\partial}{\partial p}$$

$$\frac{\partial}{\partial k_3} = \frac{\partial p}{\partial k_3}\frac{\partial}{\partial p} = \frac{k_1+k_2}{-k_3^2}\frac{\partial}{\partial p}$$

eliminating k_3

$$= -\frac{p^2}{k_1+k_2}\frac{\partial}{\partial p}$$

also

$$\frac{\partial^2}{\partial k_1^2} = \left(\frac{1}{k_3}\right)^2 \frac{\partial^2}{\partial p^2}$$

$$\frac{\partial^2}{\partial k_2^2} = \left(\frac{1}{k_3}\right)^2 \frac{\partial^2}{\partial p^2}$$

$$\begin{aligned}
\frac{\partial^2}{\partial k_3^2} &= -\frac{p^2}{k_1+k_2}\frac{\partial}{\partial p}\left(-\frac{p^2}{k_1+k_2}\frac{\partial}{\partial p}\right) \qquad \text{or} \qquad \frac{\partial}{\partial k_3}\left(-\frac{k_{12}}{k_3^2}\frac{\partial}{\partial p}\right) \\
&= \left(\frac{p}{k_1+k_2}\right)^2 \frac{\partial}{\partial p}\left(p^2\frac{\partial}{\partial p}\right) \qquad \text{or} \qquad \frac{\partial}{\partial k_3}\left(-\frac{k_{12}}{k_3^2}\right)\frac{\partial}{\partial p} + \frac{-k_{12}}{k_3^2}\frac{\partial}{\partial k_3}\frac{\partial}{\partial p} \\
&= \frac{p^2}{k_{12}^2}\left(2p\frac{\partial}{\partial p} + p^2\frac{\partial^2}{\partial p^2}\right) \qquad \text{or} \qquad \frac{2k_{12}}{k_3^3}\frac{\partial}{\partial p} + \left(-\frac{k_{12}}{k_3^2}\right)^2\frac{\partial^2}{\partial p^2} \\
&= \frac{1}{k_3^2}\left(2p\frac{\partial}{\partial p} + p^2\frac{\partial^2}{\partial p^2}\right)
\end{aligned}$$

Imposing the special conformal Ward identity $\mathbf{b}\cdot\mathbf{K}\,\langle\phi\phi\sigma\rangle' = 0$ reduces the determination of $G(p)$ to the Legendre differential equation:

$$\left[(p^2-1)\frac{d^2}{dp^2} + 2p\frac{d}{dp} + (\Delta-2)(1-\Delta)\right] G(p) = 0. \tag{11.4.3}$$

Solving the CWI for the conformally-coupled three-point function

$$\begin{aligned}
k_\mu &= -\left[2(\Delta-d)\frac{\partial}{\partial k^\mu} - 2k^\alpha\frac{\partial^2}{\partial k^\mu \partial k^\alpha} + k_\mu\frac{\partial^2}{\partial k^\alpha \partial k_\alpha}\right] && (1)\\
&= -\left[2(\Delta-d)\frac{k_\mu}{k}\frac{\partial}{\partial k} - 2k^\alpha\frac{\partial}{\partial k^\mu}\left(\frac{k_\alpha}{k}\frac{\partial}{\partial k}\right) + k_\mu\frac{\partial}{\partial k^\alpha}\left(\frac{k^\alpha}{k}\frac{\partial}{\partial k}\right)\right] && (2)\\
&= -\left[2(\Delta-d)\frac{k_\mu}{k}\frac{\partial}{\partial k} - 2k^\alpha\left(\frac{g_{\mu\alpha}}{k}\frac{\partial}{\partial k} + \frac{k_\alpha k_\mu}{k^2}\frac{\partial^2}{\partial k^2}\right)\right. && (3)\\
&\qquad\qquad \left. k_\mu\left(\frac{\delta^\alpha_\alpha}{k}\frac{\partial}{\partial k} + \frac{k^\alpha k_\alpha}{k^2}\frac{\partial^2}{\partial k^2}\right)\right] && (4)\\
&= -\left[2(\Delta-d)\frac{k_\mu}{k}\frac{\partial}{\partial k} - 2\left(\frac{k_\mu}{k}\frac{\partial}{\partial k} + k_\mu\frac{\partial^2}{\partial k^2}\right)\right. && (5)\\
&\qquad\qquad \left.\left(\frac{4k_\mu}{k}\frac{\partial}{\partial k} + k_\mu\frac{\partial^2}{\partial k^2}\right)\right] && (6)\\
&= -\left[2(\Delta-d)\frac{k_\mu}{k}\frac{\partial}{\partial k} + 2\frac{k_\mu}{k}\frac{\partial}{\partial k} - k_\mu\frac{\partial^2}{\partial k^2}\right] && (7)
\end{aligned}$$

$$= k_\mu \left[-2(\Delta - d + 1)\frac{1}{k}\frac{\partial}{\partial k} + \frac{\partial^2}{\partial k^2} \right]$$

In our case $d = 3$, so the CWI takes the form

$$\vec{b}\cdot\vec{k}\left[-2(\Delta-2)\frac{1}{k}\frac{\partial}{\partial k} + \frac{\partial^2}{\partial k^2} \right] \langle \phi(\vec{k}_1)\phi(\vec{k}_2)\sigma_\Delta(\vec{k}_3)\rangle$$

acting on each leg. Choosing $\vec{b}\cdot\vec{k}_3 = 0$ removes the σ_Δ leg, leaving

$$\begin{aligned}
(\vec{b}\cdot\vec{K}_1 + \vec{b}\cdot\vec{K}_2)k_3^{\Delta-2}G = 0 \\
\left(\vec{b}\cdot\vec{k}_1\frac{\partial^2}{\partial k_1^2} + \vec{b}\cdot\vec{k}_2\frac{\partial^2}{\partial k_3^2}\right)k_3^{\Delta-2}G = 0 \\
\left(\vec{b}\cdot\vec{k}_1\frac{\partial^2}{\partial k_1^2} - \vec{b}\cdot\vec{k}_1\frac{\partial^2}{\partial k_3^2}\right)k_3^{\Delta-2}G = 0 \\
k_3^{\Delta-2}\vec{b}\cdot\vec{k}_1(\partial_{k_1}^2 - \partial_{k_2}^2)G = 0 \implies \partial_q\partial_p G = 0
\end{aligned}$$

As explained above, it suffices to solve for the piece with $G(q,p) \equiv G(p)$, so we set that from here on.

$$\begin{aligned}
\left[\vec{b}\cdot\vec{K}_1 + \vec{b}\cdot\vec{K}_2 + \vec{b}\cdot\vec{K}_3\right]k_3^{\Delta-2}G = 0 \\
\left[\vec{b}\cdot\vec{k}_1\partial_{k_1}^2 + \vec{b}\cdot\vec{k}_2\partial_{k_2}^2 + \vec{b}\cdot\vec{k}_3\left\{-2(\Delta-3)\frac{1}{k}\partial_{k_3} + \partial_{k_3}^2\right\}\right]k_3^{\Delta-2}G = 0 \\
\left[\underbrace{(\vec{b}\cdot\vec{k}_1 + \vec{b}\cdot\vec{k}_2)}_{=-\vec{b}\cdot\vec{k}_3}\frac{1}{k_3^2}\partial_p^2 + \vec{b}\cdot\vec{k}_3\left\{\frac{(\Delta-2)(1-\Delta)}{k_3^2} + \frac{2p\partial_p + p^2\partial_p^2}{k_3^2}\right\}\right]k_3^{\Delta-2}G = 0 \\
k_3^{\Delta-2}[(p^2-1)\partial_p^2 + 2p\partial_p + (\Delta-2)(1-\Delta)]G = 0
\end{aligned}$$

where we used

$$\begin{aligned}
\left(-2(\Delta-3)\frac{1}{k}\frac{\partial}{\partial k_3} + \frac{\partial^2}{\partial k_3^2}\right)k_3^{\Delta-2}G &= -\frac{2(\Delta-2)}{k_3}\frac{\partial}{\partial k_3}\left(k_3^{\Delta-3}G\right) + \frac{\partial^2}{\partial k_3^2}\left(k_3^{\Delta-2}G\right) \\
&= -\frac{2(\Delta-2)^2}{k_3}k_3^{\Delta-3}G + \left[-\frac{2(\Delta-2)}{k_3}k_3^{\Delta-2}\frac{\partial G}{\partial k_3}\right] \\
&\quad + \frac{\partial}{\partial k_3}\left[(\Delta-2)k_3^{\Delta-3}G + k_3^{\Delta-2}\frac{\partial G}{\partial k_3}\right] \\
&= -\frac{2(\Delta-2)^2}{k_3}k_3^{\Delta-3}G - \cancel{\frac{2(\Delta-2)}{k_3}k_3^{\Delta-2}\frac{\partial G}{\partial k_3}} \\
&\quad + (\Delta-2)(\Delta-3)k_3^{\Delta-4}G + \cancel{(\Delta-2)k_3^{\Delta-3}\frac{\partial G}{\partial k_3}} \\
&\quad + \cancel{(\Delta-2)k_3^{\Delta-3}\frac{\partial G}{\partial k_3}} + k_3^{\Delta-2}\frac{\partial^2 G}{\partial k_3^2} \\
&= \frac{\Delta-2}{k_3^2}[\Delta-3-2(\Delta-2)]k_3^{\Delta-2}G + k_3^{\Delta-2}\frac{\partial^2 G}{\partial k_3^2} \\
&= -\frac{(\Delta-2)(1-\Delta)}{k_3^2}k_3^{\Delta-2}G + k_3^{\Delta-2}\frac{\partial^2 G}{\partial k_3^2} \\
&= \frac{(\Delta-2)(1-\Delta)}{k_3^2}k_3^{\Delta-2}G + \frac{k_3^{\Delta-2}}{k_3^2}\left(2p\frac{\partial G}{\partial p} + p^2\frac{\partial^2 G}{\partial p^2}\right)
\end{aligned}$$

The unique solution regular at the collinear boundary $p = 1$ ($k_3 = k_1 + k_2$) is the hypergeometric function:

$$G(p) = {}_2F_1\left(2-\Delta, \Delta-1, 1; \frac{1-p}{2}\right) = \frac{\pi}{2\sin(\pi\Delta)} P_{\Delta-1}(p), \tag{11.4.4}$$

where $P_\alpha(p)$ is the Legendre function.

In the bulk in-in formalism, this correlator arises from the cubic vertex $\mathcal{L}_{\rm int} = \frac{\lambda}{2}\phi^2\sigma$ integrated across the conformal time contour:

$$\langle\phi(\vec{k}_1)\phi(\vec{k}_2)\sigma_\Delta(\vec{k}_3)\rangle' = \frac{i\lambda\eta_*^2}{4k_1k_2}\left[\bar{g}_{k_3}(\eta_*)\int_{-\infty}^{\eta_*}\frac{d\eta}{\eta^2}e^{ik_{12}\eta}g_{k_3}(\eta) - g_{k_3}(\eta_*)\int_{-\infty}^{\eta_*}\frac{d\eta}{\eta^2}e^{-ik_{12}\eta}\bar{g}_{k_3}(\eta)\right], \tag{11.4.5}$$

where $k_{12} \equiv k_1 + k_2$ and $g_k(\eta) = H\sqrt{\frac{\pi}{4}}e^{-\frac{i\pi}{4}(1+2\nu)}(-\eta)^{3/2}H_\nu^{(2)}(-k\eta)$ is the massive field's bulk mode function.

Bulk (in-in) verification via conformally coupled mode functions

For conformally coupled scalars $m^2 = 2H^2 \quad (\Delta_+ = 2)$, thus the mode function looks like

$$f_k(\eta) = H \times (-\eta)\frac{e^{ik\eta}}{\sqrt{2k}}$$

the three point correlation function is given as[a]:

$$\begin{aligned}\langle\phi(\vec{k}_1)\phi(\vec{k}_2)\sigma_\Delta(\vec{k}_3)\rangle' &= i\lambda\int_{-\infty}^{\eta_*} D\eta W_{k_1}(\eta_*,\eta)W_{k_2}(\eta_*,\eta)\tilde{W}_{k_3}(\eta_*,\eta)\\ &\quad - i\lambda\int_{-\infty}^{\eta_*} D\eta W_{k_1}(\eta,\eta_*)W_{k_2}(\eta,\eta_*)\tilde{W}_{k_3}(\eta,\eta_*)\end{aligned}$$

since, $W_k(\eta,\eta_*) = \bar{f}_k(\eta)f_k(\eta_*)$

$$\begin{aligned}&= i\lambda\bar{f}_{k_1}(\eta_\star)\bar{f}_{k_2}(\eta_\star)\bar{g}_{k_3}(\eta_\star)\int_{-\infty}^{\eta^\star} D\eta f_{k_1}(\eta)f_{k_2}(\eta)g_{k_3}(\eta)\\ &\quad - i\lambda f_{k_1}(\eta_\star)f_{k_2}(\eta_\star)g_{k_3}(\eta_\star)\int_{-\infty}^{\eta_\star} D\eta\bar{f}_{k_1}(\eta)\bar{f}_{k_2}(\eta)\bar{g}_{k_3}(\eta)\\ &= i\lambda\frac{H^2\eta_\star^2}{4k_1k_2}\bar{g}_{k_3}(\eta_\star)\int_{-\infty}^{\eta^\star} D\eta H^2\eta^2 e^{i(k_1+k_2)\eta}g_{k_3}(\eta)\\ &\quad - i\lambda\frac{H^2\eta_\star^2}{4k_1k_2}g_{k_3}(\eta_\star)\int_{-\infty}^{\eta_\star} D\eta H^2\eta^2 e^{-i(k_1+k_2)\eta}\bar{g}_{k_3}(\eta)\\ &= i\lambda\frac{\eta_\star^2}{4k_1k_2}\bar{g}_{k_3}(\eta_\star)\int_{-\infty}^{\eta^\star}\frac{d\eta}{\eta^2}e^{i(k_1+k_2)\eta}g_{k_3}(\eta)\\ &\quad - i\lambda\frac{\eta_\star^2}{4k_1k_2}g_{k_3}(\eta_\star)\int_{-\infty}^{\eta_\star}\frac{d\eta}{\eta^2}e^{-i(k_1+k_2)\eta}\bar{g}_{k_3}(\eta)\end{aligned}$$

where

$$g_k(\eta) = H\sqrt{\frac{\pi}{4}}e^{-\frac{i\pi}{4}(1+2\nu)}(-\eta)^{3/2}H_\nu^{(2)}(-k\eta)$$

[a] here we assume the interaction term is being given as $\mathcal{L}_{\rm int} = (\lambda/3!)\phi^2\sigma$

11.5 Three-Point Functions of Massless Inflaton Fluctuations

In inflationary cosmology, the fluctuations of the inflaton field ξ are massless ($\Delta = 3$). Shift symmetry protects the derivative coupling to the massive field σ:

$$S_{\rm int} = \lambda \int \frac{d\eta d^3x}{H^4\eta^4}(\nabla\xi)^2\sigma. \tag{11.5.1}$$

The massless mode function is $f_k(\eta) = \frac{H}{\sqrt{2k^3}}(1-ik\eta)e^{ik\eta}$. Because $(1-ik\eta)$ is a linear polynomial in η, factors of η inside the time integral can be traded for momentum derivatives $\eta e^{ik_{12}\eta} = -i\partial_{k_{12}}e^{ik_{12}\eta}$. Consequently, the three-point function of two massless scalars is generated from the conformally coupled result by acting with a differential operator $\mathcal{O}_{12}$:

$$\left\langle \xi(\vec{k}_1)\xi(\vec{k}_2)\sigma_\Delta(\vec{k}_3)\right\rangle' = \frac{1}{2\sqrt{k_1^3k_2^3}}\mathcal{O}_{12}\left[k_3^{\Delta-2}G(p)\right], \tag{11.5.2}$$

where $\mathcal{O}_{12}$ in terms of (p,q) variables is:

$$\mathcal{O}_{12} = \frac{k_3^2}{8}\left[(q^2-p^2)(1-p^2)\partial_p^2 + 2(p^2+q^2-2)(1-p\partial_p)\right] \tag{11.5.3}$$

$$= \Delta(\Delta-3) - (q^2-1)\partial_q^2 - 2q\partial_q. \tag{11.5.4}$$

The massless three-point function from a differential operator $\mathcal{O}_{12}$

Using these mode functions, the integral producing the three point vertex is given as:

$$\begin{aligned}
\int(\nabla\xi)^2\sigma &= \int\frac{d\eta}{H^2\eta^2}[-(\partial_\eta\xi)^2 + (\partial\xi)^2]\sigma \\
&= \int\frac{d\eta}{H^2\eta^2}[-\partial_\eta f_{k_1}\partial_\eta f_{k_2} + (i\vec{k}_1)\cdot(i\vec{k}_2)f_{k_1}f_{k_2}]\sigma(\eta,k_3) \\
&= \frac{1}{2\sqrt{k_1^3k_2^3}}\int\frac{d\eta}{\eta^2}[-k_1^2k_2^2\eta^2 - \vec{k}_1\cdot\vec{k}_2(1-ik_1\eta)(1-ik_2\eta)]e^{i(k_1+k_2)\eta}\sigma(\eta,k_3) \\
&= \frac{1}{2\sqrt{k_1^3k_2^3}}\mathcal{O}_{12}\int\frac{d\eta}{\eta^2}e^{ik_{12}\eta}\sigma(\eta,k_3), \quad k_{12}\equiv k_1+k_2
\end{aligned} \tag{1}$$

where we defined (by using momentum conservation):

$$\begin{aligned}
\mathcal{O}_{12} &= k_1^2k_2^2\partial_{k_{12}}^2 + \frac{(k_1^2+k_2^2-k_3^2)}{2}[1+k_1k_2\partial_{k_{12}}^2 - k_{12}\partial_{k_{12}}] \\
&= -\frac{1}{2}k_1k_2(k_3^2-k_{12}^2)\partial_{k_{12}}^2 + \frac{1}{2}(k_1^2+k_2^2-k_3^2)(1-k_{12}\partial_{k_{12}}) \\
&= \frac{k_3^2}{8}[(q^2-p^2)(1-p^2)\partial_p^2 + 2(p^2+q^2-2)(1-p\partial_p)]
\end{aligned}$$

and used

$$\frac{\partial f_k}{\partial\eta} = H\frac{-ike^{ik\eta} + ik(1-ik\eta)e^{ik\eta}}{\sqrt{2k^3}} = \frac{H}{\sqrt{2k^3}}k^2\eta e^{ik\eta} \tag{2}$$

11.6 Four-Point Functions

Four-point functions play a central role in cosmological collider physics: they are the simplest correlation functions that contain an internal exchanged particle line. Unlike two- and three-point

functions, four-point correlators are not uniquely determined by conformal symmetry alone, because four spatial points admit two independent, conformally invariant cross-ratios:

$$u \equiv \frac{x_{12}^2 x_{34}^2}{x_{13}^2 x_{24}^2} = z\bar{z}, \qquad v \equiv \frac{x_{14}^2 x_{23}^2}{x_{13}^2 x_{24}^2} = (1-z)(1-\bar{z}). \tag{11.6.1}$$

In position space, local contact interactions and analytic exchange contributions produce local terms involving dilogarithms:

$$\langle \varphi(\vec{x}_1)\cdots\varphi(\vec{x}_4)\rangle \propto \frac{1}{x_{13}^2 x_{24}^2}\frac{1}{(z-\bar{z})}\left[2\mathrm{Li}_2(z) - 2\mathrm{Li}_2(\bar{z}) + \log(z\bar{z})\log\frac{1-z}{1-\bar{z}}\right]. \tag{11.6.2}$$

These local terms have no branch cuts in the physical region and represent interactions of existing fluctuations. To identify the non-local, oscillatory signals of new massive particles, we compute the four-point function in momentum space using the in-in formalism.

11.6.1 Contact four-point function ($\lambda\phi^4$ interaction)

For a contact interaction $\mathcal{L}_{\rm int} = -\frac{\lambda}{4!}\phi^4$, the late-time four-point function at conformal time $\eta_* \to 0$ evaluates to:

$$\langle \phi_{k_1}\phi_{k_2}\phi_{k_3}\phi_{k_4}\rangle' = -\frac{\lambda(H\eta_*)^4}{8k_1k_2k_3k_4}\frac{1}{k_t}, \tag{11.6.3}$$

where $k_t \equiv k_1+k_2+k_3+k_4$ is the total energy.

Contact four-point function from $\lambda\phi^4$

We begin by stating the Wightman propagator,

$$W_k(\eta,\eta') = \frac{H^2}{2k}(\eta\eta')e^{-ik(\eta-\eta')}$$

Then, the four point function at later time η_* is given by:

$$\begin{aligned} I_+ &= -i\lambda\int_{-\infty}^{\eta_*} D\eta W_{k_1}(\eta_*,\eta)W_{k_2}(\eta_*,\eta)W_{k_3}(\eta_*,\eta)W_{k_4}(\eta_*,\eta) \\ &= -i\frac{\lambda(H\eta_*)^4}{16k_1k_2k_3k_4}\int_{-\infty}^{\eta_*} d\eta\, e^{i(k_1+k_2+k_3+k_4)\eta} \\ &= -\frac{\lambda(H\eta_*)^4}{16k_1k_2k_3k_4(k_1+k_2+k_3+k_4)}. \end{aligned}$$

This contribution is again purely real ($I_+ = I_-$), so the full correlator is just twice the above answer:

$$\langle \phi_{k_1}\phi_{k_2}\phi_{k_3}\phi_{k_4}\rangle' = -\frac{\lambda(H\eta_*)^4}{8k_1k_2k_3k_4}\frac{1}{k_t}$$

11.6.2 Exchange four-point function ($\lambda\phi^3$ interaction)

Now consider two conformally coupled scalars exchanging a field through cubic vertices $\frac{\lambda}{2}\phi^2\sigma$. In flat Minkowski space, the time-ordered exchange amplitude I_{++} evaluated by contour integration gives:

$$I_{++}^{\rm flat} = \frac{g^2}{32k_I\, k_1k_2k_3k_4(k_{12}+k_{34})}\left(\frac{1}{k_{12}+k_I}+\frac{1}{k_{34}+k_I}\right), \tag{11.6.4}$$

where $k_{12} \equiv k_1+k_2$, $k_{34} \equiv k_3+k_4$, and $k_I \equiv |\vec{k}_1+\vec{k}_2|$ is the exchanged momentum.

Flat-space warm-up: the exchange integral via contour integration

We begin by considering a conformally coupled scalar. We can write the Wightman and Feynman propagators explicitly

$$W_k(\eta,\eta') = H^2\eta\eta'\frac{1}{2k}e^{-ik(\eta-\eta')},$$
$$G_F(k;\eta,\eta') = H^2\eta\eta'\underbrace{\frac{1}{2k}\left[e^{-ik(\eta-\eta')}\theta(\eta-\eta')+e^{ik(\eta-\eta')}\theta(\eta'-\eta)\right]}_{G_F^{(\text{flat})}}$$
$$= H^2\eta\eta'\frac{1}{2k}(-2k)\int_{-\infty}^{+\infty}\frac{d\omega}{2\pi i}\frac{e^{i\omega(\eta-\eta')}}{(\omega^2-k^2+i\epsilon)} = -H^2\eta\eta'\int_{-\infty}^{+\infty}\frac{d\omega}{2\pi i}\frac{e^{i\omega(\eta-\eta')}}{(\omega^2-k^2+i\epsilon)}$$

Notice that these are exactly $H^2\eta\eta'$ times their flat-space counterparts. (This is another manifestation of the conformal invariance of the conformally coupled scalar.) In flat space we will perform the analysis using the fourier transform. We start with the definition

$$I_{++} = (-ig)^2\int_{-\infty}^{0}dt'dt''\,W_{k_1}(0,t')\,W_{k_2}(0,t')\,G_F(k_I;t',t'')\,W_{k_3}(0,t'')\,W_{k_4}(0,t''). \quad (1)$$

Using

$$W_k(0,t) = \frac{1}{2k}e^{ikt},$$

we obtain

$$W_{k_1}(0,t')W_{k_2}(0,t') = \frac{1}{4k_1k_2}\,e^{ik_{12}t'}, \quad (2)$$
$$W_{k_3}(0,t'')W_{k_4}(0,t'') = \frac{1}{4k_3k_4}\,e^{ik_{34}t''}, \quad (3)$$

with the shorthand

$$k_{12}\equiv k_1+k_2, \qquad k_{34}\equiv k_3+k_4.$$

Thus

$$I_{++} = -\frac{g^2}{16k_1k_2k_3k_4}\int_{-\infty}^{0}dt'\,dt''\,e^{ik_{12}t'}\,G_F(k_I;t',t'')\,e^{ik_{34}t''}. \quad (4)$$

The Feynman propagator is written as

$$G_F(k_I;t',t'') = -\int_{-\infty}^{+\infty}\frac{d\omega}{2\pi i}\,\frac{e^{i\omega(t'-t'')}}{\omega^2-k_I^2+i\epsilon}. \quad (5)$$

Substituting this into the expression for I_{++} gives

$$I_{++} = \frac{g^2}{16k_1k_2k_3k_4}\int_{-\infty}^{+\infty}\frac{d\omega}{2\pi i}\frac{1}{\omega^2-k_I^2+i\epsilon}\int_{-\infty}^{0}dt'\,e^{i(k_{12}+\omega)t'}\int_{-\infty}^{0}dt''\,e^{i(k_{34}-\omega)t''}. \quad (6)$$

Both integrals converge with the prescription $t\to t(1-i\epsilon)$:

$$\int_{-\infty}^{0}dt'\,e^{i(k_{12}+\omega)t'} = \frac{1}{i(k_{12}+\omega)+\epsilon}, \quad (7)$$
$$\int_{-\infty}^{0}dt''\,e^{i(k_{34}-\omega)t''} = \frac{1}{i(k_{34}-\omega)+\epsilon}. \quad (8)$$

Therefore

$$I_{++} = \frac{g^2}{16k_1k_2k_3k_4}\int_{-\infty}^{+\infty}\frac{d\omega}{2\pi i}\,\frac{1}{(\omega^2 - k_I^2 + i\epsilon)(\omega + k_{12} - i\epsilon)(\omega - k_{34} + i\epsilon)}. \tag{9}$$

The integrand has simple poles at

$$\omega = \pm k_I \quad \text{from the propagator,} \qquad \omega = -k_{12}, \qquad \omega = k_{34}.$$

With the $i\epsilon$ prescription:

- $\omega = +k_I - i\epsilon$ lies slightly *below* the real axis,
- $\omega = -k_I + i\epsilon$ lies slightly *above* the real axis,
- $\omega = -k_{12} + i\epsilon$ lies above,
- $\omega = +k_{34} - i\epsilon$ lies below.

Closing the contour in the upper half-plane picks up the poles at

$$\omega = -k_{12}, \qquad \omega = -k_I.$$

Residue at $\omega = -k_{12}$:

$$\begin{aligned}\text{Res}_{\omega=-k_{12}} &= \lim_{\omega\to -k_{12}}(\omega + k_{12})f(\omega)\\ &= -\frac{1}{(k_{12}^2 - k_I^2)(k_{12} + k_{34})}.\end{aligned}$$

At $\omega = -k_I$:

$$\begin{aligned}\text{Res}_{\omega=-k_I} &= \lim_{\omega\to -k_I}(\omega + k_I)f(\omega)\\ &= \frac{1}{-2k_I}\cdot\frac{1}{(-k_I + k_{12})(-k_I - k_{34})}\\ &= \frac{1}{2k_I}\cdot\frac{1}{(k_{12} - k_I)(k_I + k_{34})}\end{aligned}$$

Therefore

$$\begin{aligned}I_{++} &= \frac{g^2}{16k_1k_2k_3k_4}\left[\frac{1}{2k_I(k_{12} - k_I)(k_{34} + k_I)} - \frac{1}{(k_{12} - k_I)(k_{12} + k_I)(k_{12} + k_{34})}\right]\\ &= \frac{g^2}{16k_1k_2k_3k_4}\left[\frac{(k_{12} + k_I)(k_{12} + k_{34}) - 2k_I(k_{34} + k_I)}{2k_I(k_{12} - k_I)(k_{34} + k_I)(k_{12} + k_I)(k_{12} + k_{34})}\right]\\ &= \frac{g^2}{16k_1k_2k_3k_4}\frac{1}{2k_I}\frac{1}{k_{12} + k_{34}}\left[\frac{(k_{12} - k_I)(k_{12} + k_{34} + 2k_I)}{(k_{12} - k_I)(k_{34} + k_I)(k_{12} + k_I)}\right]\\ &= \frac{g^2}{16k_1k_2k_3k_4}\frac{1}{2k_I}\frac{1}{k_{12} + k_{34}}\left(\frac{1}{k_{12} + k_I} + \frac{1}{k_{34} + k_I}\right)\end{aligned}$$

Storing the time ordering inside the pole is sensitive to the sign on $i\epsilon$ prescription. A slight miscalculation can completely ruin the result. This is precisely why the direct method is often preferred. In our context, it is also the shortest way to calculate the result:

$$I_{++} = -\frac{g^2}{16k_1k_2k_3k_4}\frac{1}{2k_I}\int_{-\infty}^{0}dt\int_{-\infty}^{0}dt'\,e^{ik_{12}t}e^{-ik_I(t-t')}\theta(t-t')e^{ik_{34}t'} + (k_{12}\leftrightarrow k_{34})$$

$$
\begin{aligned}
&= -\frac{g^2}{16k_1k_2k_3k_4}\frac{1}{2k_I}\int_{-\infty}^{0} dt\, e^{i(k_{12}-k_I)t}\int_{-\infty}^{t} dt'\, e^{i(k_{34}+k_I)t'} + (k_{12}\leftrightarrow k_{34})\\
&= -\frac{g^2}{16k_1k_2k_3k_4}\frac{1}{2k_I}\int_{-\infty}^{0} dt\, \frac{e^{i(k_{12}+k_{34})t}}{i(k_{34}+k_I)} + (k_{12}\leftrightarrow k_{34})\\
&= \frac{g^2}{16k_1k_2k_3k_4}\frac{1}{2k_I}\frac{1}{k_{12}+k_{34}}\left(\frac{1}{k_{12}+k_I}+\frac{1}{k_{34}+k_I}\right).
\end{aligned}
$$

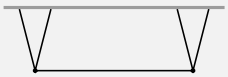
In de Sitter space, the exchange diagram receives contributions from all four in-in branches $(I_{++}, I_{+-}, I_{-+}, I_{--})$. The connected piece evaluates to:

$$
I^{(c)}_{++} = \frac{\lambda^2 H^2 \eta_*^4}{32k_1k_2k_3k_4k_I}\left[\mathrm{Li}_2\left(\frac{E-E_L}{E}\right)+\mathrm{Li}_2\left(\frac{E-E_R}{E}\right)+\log\left(\frac{E_L}{E}\right)\log\left(\frac{E_R}{E}\right)-\frac{\pi^2}{6}\right], \tag{11.6.5}
$$

where $E_L \equiv k_{12}+k_I$, $E_R \equiv k_{34}+k_I$, and $E \equiv k_{12}+k_{34} = k_t$.

The de Sitter exchange integral: connected piece

The I_{++} contribution is[a]

$I_{++} \equiv$ 

$$
\begin{aligned}
&= (-i\lambda)^2\int_{-\infty}^{\eta_*} D\eta' D\eta'' W_{k_1}(\eta_*,\eta')W_{k_2}(\eta_*,\eta')G_F(k_I;\eta',\eta'')W_{k_3}(\eta_*,\eta'')W_{k_4}(\eta_*,\eta'')\\
&= -\frac{\lambda^2H^2\eta_*^4}{16k_1k_2k_3k_4}\int_{-\infty}^{\eta_*}\frac{\mathrm{d}\eta'}{\eta'}\int_{-\infty}^{\eta_*}\frac{\mathrm{d}\eta''}{\eta''}e^{ik_{12}\eta'}G_F^{(\mathrm{flat})}(k_I;\eta',\eta'')e^{ik_{34}\eta''},
\end{aligned}
$$

where, we defined $k_{12} = k_1 + k_2$, $k_{34} = k_3 + k_4$ and we expressed the second line in terms of the flat-space Feynman propagator. This integral is divergent, and the divergence is exactly what will cancel against similar divergences from I_{-+} and I_{+-} once we add all four in-in branches together; in order to isolate that divergent piece cleanly now, it is useful to add and subtract a finite reference term from the integrand:

$$
\begin{aligned}
I_{++} = -\frac{\lambda^2H^2\eta_*^4}{16k_1k_2k_3k_4}\int_{-\infty}^{\eta_*}\frac{\mathrm{d}\eta'}{\eta'}\int_{-\infty}^{\eta_*}\frac{\mathrm{d}\eta''}{\eta''}e^{ik_{12}\eta'}\Bigg[&\underbrace{G_F^{(\mathrm{flat})}(k_I;\eta',\eta'')-\frac{1}{2k_I}e^{ik_I(\eta'+\eta'')}}_{G_B^{(\mathrm{flat})}}\\
&+\frac{1}{2k_I}e^{ik_I(\eta'+\eta'')}\Bigg]e^{ik_{34}\eta''}
\end{aligned}
$$

where we have defined

$$
G_B^{(\mathrm{flat})}(k_I;\eta',\eta'') = G_F^{(\mathrm{flat})}(k_I;\eta',\eta'') - \frac{1}{2k_I}e^{ik_I(\eta'+\eta'')}
$$

We can now perform the integral involving G_B via a trick. We note that

$$
-i\int_a^\infty \mathrm{d}x e^{ix\eta} = \frac{1}{\eta}e^{ia\eta},
$$

So we can write

$$
I_{++} = \frac{\lambda^2H^2\eta_*^4}{16k_1k_2k_3k_4}\left[\int_{k_{12}}^\infty \mathrm{d}x\int_{k_{34}}^\infty \mathrm{d}y\int_{-\infty}^{\eta_*}\mathrm{d}\eta'\mathrm{d}\eta'' e^{ix\eta'}e^{iy\eta''}G_B^{(\mathrm{flat})}(k_I;\eta',\eta'')\right.
$$

$$-\frac{1}{2k_I}\int_{-\infty}^{\eta_*}\frac{\mathrm{d}\eta'}{\eta'}\int_{-\infty}^{\eta_*}\frac{\mathrm{d}\eta''}{\eta'''}e^{i(k_{12}+s)\eta'}e^{i(k_{34}+s)\eta''}\Bigg]$$

The time integrals in the first line now only involve flat-space quantities and can therefore easily be computed

$$I(x,y)\equiv\int_{-\infty}^{0}\mathrm{d}\eta'\mathrm{d}\eta''e^{ix\eta'}e^{iy\eta''}G_B^{(\mathrm{flat})}(k_I;\eta',\eta'')=-\frac{1}{(x+y)(x+k_I)(y+k_I)}$$

The integral was performed using the fact that the limits of the integrals are decided by the theta function

$$\theta(\eta'-\eta'')=\begin{cases}1 & \eta''<\eta'\\ 0 & \eta''>\eta'\end{cases}$$

$$\begin{aligned}\int_{-\infty}^{\eta'}d\eta''\int_{-\infty}^{0}d\eta' e^{ix\eta'}e^{iy\eta''}e^{-ik_I(\eta'-\eta'')}&=-\frac{1}{(k_I+y)(x+y)}\\ \int_{-\infty}^{0}d\eta''\int_{-\infty}^{\eta''}d\eta' e^{ix\eta'}e^{iy\eta''}e^{ik_I(\eta'-\eta'')}&=-\frac{1}{(k_I+x)(x+y)}\\ \int_{-\infty}^{0}d\eta''\int_{-\infty}^{0}d\eta' e^{ix\eta'}e^{iy\eta''}e^{ik_I(\eta'+\eta'')}&=-\frac{1}{(k_I+x)(k_I+y)}\end{aligned}$$

It is then useful to split I_{++} into two pieces: a "connected" component, $I_{++}^{(c)}$, which is the energy integral of $I(x,y)$, and a "disconnected" component, $I_{++}^{(d)}$, which involves integral of terms without time ordering. Performing the integrals of the connected piece, we get:

$$\begin{aligned}I_{++}^{(c)}&\equiv\frac{\lambda^2H^2\eta_*^4}{16k_1k_2k_3k_4}\int_{k_{12}}^{\infty}\mathrm{d}x\int_{k_{34}}^{\infty}\mathrm{d}y\ I(x,y)\\ &=\frac{\lambda^2H^2\eta_*^4}{32k_1k_2k_3k_4k_I}\left[\mathrm{Li}_2\left(\frac{E-E_L}{E}\right)+\mathrm{Li}_2\left(\frac{E-E_R}{E}\right)+\log\left(\frac{E_L}{E}\right)\log\left(\frac{E_R}{E}\right)-\frac{\pi^2}{6}\right]\end{aligned}$$

where Li_2 is the dilogarithm and we have **defined** the variables $E_L\equiv k_{12}+k_I$ (the energy flowing into the left vertex), $E_R\equiv k_{34}+k_I$ (same for the right vertex) and the total energy $E\equiv k_{12}+k_{34}=k_t$.

[a]Vertices and propagators below the conformal boundary are time-ordered, and those above the conformal boundary are anti-time-ordered.

The logarithmic divergences of the disconnected pieces $I_{\pm\pm}^{(d)}$ cancel precisely against $I_{\pm\mp}$, leaving a finite boundary remainder:

$$I_{++}^{(d)}+I_{+-}+I_{--}^{(d)}+I_{-+}=\frac{\pi^2\lambda^2H^2\eta_*^4}{32k_1k_2k_3k_4k_I}. \tag{11.6.6}$$

Summing the connected and disconnected contributions yields the total four-point correlator:

$$\langle\phi_{\vec{k}_1}\phi_{\vec{k}_2}\phi_{\vec{k}_3}\phi_{\vec{k}_4}\rangle'=\frac{\lambda^2H^2\eta_*^4}{4k_1k_2k_3k_4k_I}\left[\mathrm{Li}_2\left(\frac{k_{12}-k_I}{k_t}\right)+\mathrm{Li}_2\left(\frac{k_{34}-k_I}{k_t}\right)+\log\left(\frac{k_{12}+k_I}{k_t}\right)\log\left(\frac{k_{34}+k_I}{k_t}\right)+\frac{\pi^2}{3}\right]. \tag{11.6.7}$$

In the squeezed limit $k_I \to 0$, this correlator exhibits the non-local pole:

$$\langle \phi_{\vec{k}_1}\phi_{\vec{k}_2}\phi_{\vec{k}_3}\phi_{\vec{k}_4}\rangle' \approx \frac{\lambda^2 H^2 \eta_*^4}{8k_1k_2k_3k_4}\left[\frac{\pi^2}{k_I} - 2\frac{k_I}{k_{12}k_{34}} + \cdots\right], \tag{11.6.8}$$

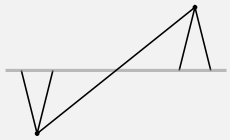
whose $1/k_I$ singularity is the distinct hallmark of an exchanged intermediate state.

The de Sitter exchange integral: disconnected piece and divergence cancellation

In order to compute the disconnected piece, it is simplest to do the time integrals directly, rather than trying to write them as energy integrals of the flat-space result. This is because we have to be careful about the $i\epsilon$ prescriptions for the contours

$$\begin{aligned} I_{++}^{(d)} &\equiv -\frac{\lambda^2 H^2 \eta_*^4}{32k_1k_2k_3k_4k_I}\int_{-\infty}^{\eta_*}\frac{\mathrm{d}\eta'}{\eta'}e^{i(k_{12}+s)\eta'}\int_{-\infty}^{\eta_*}\frac{\mathrm{d}\eta''}{\eta''}e^{i(k_{34}+s)\eta''} \\ &= -\frac{\lambda^2 H^2 \eta_*^4}{32k_1k_2k_3k_4k_I}\log(iE_L\eta_*)\log(iE_R\eta_*) \end{aligned} \tag{1}$$

As advertised, this contribution is divergent as $\eta_* \to 0$, which is why we split it off from $I_{++}^{(c)}$ to begin with. These divergences cancel with the $I_{\mp\pm}$ pieces, which we now want to calculate. Using the Feynman rules, we find

$I_{+-} \equiv$ 

$$\begin{aligned} &= \lambda^2\int_{-\infty}^{\eta_*} D\eta' D\eta'' W_{k_1}(\eta_*,\eta')W_{k_2}(\eta_*,\eta')W_{k_I}(\eta'',\eta')W_{k_3}(\eta'',\eta_*)W_{k_4}(\eta'',\eta_*) \\ &= \frac{\lambda^2 H^2 \eta_*^4}{32k_1k_2k_3k_4k_I}\int_{-\infty}^{\eta_*}\frac{\mathrm{d}\eta'}{\eta'}e^{i(k_{12}+k_I)\eta'}\int_{-\infty}^{\eta_*}\frac{\mathrm{d}\eta''}{\eta''}e^{-i(k_{34}+k_I)\eta''} \\ &= \frac{\lambda^2 H^2 \eta_*^4}{32k_1k_2k_3k_4k_I}\log(iE_L\eta_*)\log(-iE_R\eta_*) \end{aligned} \tag{2}$$

In the last line, we evaluated the integrals, noting that the different $i\epsilon$ prescriptions required for convergence of the integrals lead to different signs of the arguments of the log for the η' and η''integrals. Adding (1) to (2)we get,

$$I_{++}^{(d)} + I_{+-} = \frac{\lambda^2 H^2 \eta_*^4}{32k_1k_2k_3k_4k_I}\left(-\pi\log(iE_L\eta_*)\right) \tag{3}$$

The sum of the $I_{--}^{(d)}$ and I_{-+} contributions gives the complex conjugate of (3). Putting everything together, we therefore find

$$I_{++}^{(d)} + I_{+-} + I_{--}^{(d)} + I_{-+} = \frac{\lambda^2 H^2 \eta_*^4}{32k_1k_2k_3k_4k_I}\pi^2 \tag{4}$$

11.6.3 Four-point function with massive intermediate state

When the exchanged field σ possesses a mass $m > \frac{3}{2}H$, its internal propagator is governed by Bessel/Hankel functions. Acting on the in-in time integral with σ's bulk equation of motion operator converts the four-point function into ordinary differential equations in $p \equiv k_{12}/k_I$ and

$p' \equiv k_{34}/k_I$:

$$\left[(p^2-1)\partial_p^2 + 2p\partial_p + \frac{m^2}{H^2} - 2\right] G_{++}(p,p') = \frac{1}{p+p'}, \tag{11.6.9}$$

$$\left[(p^2-1)\partial_p^2 + 2p\partial_p + \frac{m^2}{H^2} - 2\right] G_{+-}(p,p') = 0. \tag{11.6.10}$$

The exchanged massive field: an equation-of-motion trick

We can now consider conformally coupled scalars φ coupled to a general massive scalar field σ with an interaction vertex $\int \lambda\varphi^2\sigma$. The late time expectation value is given by the following expression:

$$\langle \varphi_{\vec{k}_1}(\eta_0)\cdots\varphi_{\vec{k}_4}(\eta_0)\rangle' = \frac{\eta_0^4 2^2\lambda^2}{16k_1k_2k_3k_4} I_E(k_{12},k_{34},k_I) + \text{two other diagrams}$$

$$I_E(k_{12},k_{34},k_I) \equiv I_{++} + I_{+-} + I_{-+} + I_{--}$$

$$I_{\pm\pm} = (\pm i)(\pm i)\int_{-\infty}^{0}\frac{d\eta}{\eta^2}e^{\pm ik_{12}\eta}\int_{-\infty}^{\eta}\frac{d\eta'}{\eta'^2}e^{\pm ik_{34}\eta'}\langle\sigma_{k_I}(\eta)\sigma_{-k_I}(\eta')\rangle'_{\pm\pm} \tag{1}$$

where the $\pm$ signs indicate the type of branch in the integration contour, and consequently the sign in subscript of $\langle\sigma_{k_I}(\eta)\sigma_{-k_I}(\eta')\rangle'_{\pm\pm}$ represents the propagator along that branch. The equations of motion for $\langle\sigma_{k_I}(\eta)\sigma_{-k_I}(\eta')\rangle'_{\pm\pm}$ are given as:

$$\left(\eta^2\frac{\partial^2}{\partial\eta^2} - 2\eta\frac{\partial}{\partial\eta} + k_I^2\eta^2 + \frac{m^2}{H^2}\right)\langle\sigma_{k_I}(\eta)\sigma_{-k_I}(\eta')\rangle'_{++} = -iH^2\eta^2\eta'^2\delta(\eta-\eta') \tag{2}$$

Now, consider the following auxiliary object:

$$\tilde{F} = -\int_{-\infty}^{0}\frac{d\eta}{\eta^2}\frac{d\eta'}{\eta'^2}e^{ik_{12}\eta}e^{ik_{34}\eta'}\left(\eta^2\frac{\partial^2}{\partial\eta^2} - 2\eta\frac{\partial}{\partial\eta} + k_I^2\eta^2 + \frac{m^2}{H^2}\right)\langle\sigma_{k_I}(\eta)\sigma_{-k_I}(\eta')\rangle'_{++} \tag{3}$$

and evaluate it in two ways. First, we use (2) to write

$$\tilde{F} = iH^2\int_{-\infty}^{0}d\eta e^{i(k_{12}+k_{34})\eta} = \frac{H^2}{k_{12}+k_{34}},$$

Next, we integrate the time derivatives by parts[a] in (3) and trade factors of η for $-i\partial_{k_{12}}$ to obtain

$$\begin{aligned}\tilde{F} &= -\int_{-\infty}^{0}\frac{d\eta}{\eta^2}\frac{d\eta'}{\eta'^2}\left[-(k_{12}^2-k_I^2)\eta^2 + 2ik_{12}\eta + \frac{m^2}{H^2} - 2\right]e^{ik_{12}\eta}e^{ik_{34}\eta'}\langle\sigma_{k_I}(\eta)\sigma_{-k_I}(\eta')\rangle'_{++}\\ &= -\int_{-\infty}^{0}\frac{d\eta}{\eta^2}\frac{d\eta'}{\eta'^2}\left[(k_{12}^2-k_I^2)\partial_{k_{12}}^2 + 2k_{12}\partial_{k_{12}} + \frac{m^2}{H^2} - 2\right]e^{ik_{12}\eta}e^{ik_{34}\eta'}\langle\sigma_{k_I}(\eta)\sigma_{-k_I}(\eta')\rangle'_{++}\\ &= -\left[(k_{12}^2-k_I^2)\partial_{k_{12}}^2 + 2k_{12}\partial_{k_{12}} + \frac{m^2}{H^2} - 2\right]\int_{-\infty}^{0}\frac{d\eta}{\eta^2}\frac{d\eta'}{\eta'^2}e^{ik_{12}\eta}e^{ik_{34}\eta'}\langle\sigma_{k_I}(\eta)\sigma_{-k_I}(\eta')\rangle'_{++}\end{aligned}$$

using (1)

$$\left[(k_{12}^2-k_I^2)\partial_{k_{12}}^2 + 2k_{12}\partial_{k_{12}} + \frac{m^2}{H^2} - 2\right]I_{++} = \frac{H^2}{k_{12}+k_{34}}$$

under the redefinition:

$$p = \frac{k_{12}}{k_I} \qquad p' = \frac{k_{34}}{k_I} \qquad I_{\pm\pm} = \frac{G_{\pm\pm}(p,p')}{k_I}$$

thus,

$$\frac{\partial}{\partial p} = \frac{\partial k_{12}}{\partial p}\frac{\partial}{\partial k_{12}} = k_I \frac{\partial}{\partial k_{12}}$$

hence,

$$k_I^2\left[(p^2-1)\partial_p^2 + 2p\partial_p + \frac{m^2}{H^2} - 2\right] I_{++} = H^2 \frac{k_I}{p+p'} \tag{4}$$

If we used

$$\left(\eta'^2\frac{\partial^2}{\partial\eta'^2} - 2\eta'\frac{\partial}{\partial\eta'} + k_I^2\eta'^2 + \frac{m^2}{H^2}\right)\langle\sigma_{k_I}(\eta)\sigma_{-k_I}(\eta')\rangle'_{++} = -iH^2\eta^2\eta'^2\delta(\eta-\eta')$$

then by redoing all the same steps we arrive at:

$$k_I^2\left[(p'^2-1)\partial_{p'}^2 + 2p'\partial_{p'} + \frac{m^2}{H^2} - 2\right] I_{++} = H^2 \frac{k_I}{p+p'} \tag{5}$$

from (4) and (5), we have:

$$\left[(p^2-1)\partial_p^2 + 2p\partial_p - (p'^2-1)\partial_{p'}^2 - 2p'\partial_{p'}\right] I_{++} = 0$$

Similarly, we can use

$$\left(\eta'^2\frac{\partial^2}{\partial\eta'^2} - 2\eta'\frac{\partial}{\partial\eta'} + k_I^2\eta'^2 + \frac{m^2}{H^2}\right)\langle\sigma_{k_I}(\eta)\sigma_{-k_I}(\eta')\rangle'_{+-} = 0$$

to show that

$$\left[(k_{12}^2-k_I^2)\partial_{k_{12}}^2 + 2k_{12}\partial_{k_{12}} + \frac{m^2}{H^2} - 2\right] I_{+-} = 0.$$

If we used

$$\left(\eta'^2\frac{\partial^2}{\partial\eta'^2} - 2\eta'\frac{\partial}{\partial\eta'} + k_I^2\eta'^2 + \frac{m^2}{H^2}\right)\langle\sigma_{k_I}(\eta)\sigma_{-k_I}(\eta')\rangle'_{++} = 0$$

in (3) then, we could show:

$$\left[(k_{34}^2-k_I^2)\partial_{k_{34}}^2 + 2k_{34}\partial_{k_{34}} + \frac{m^2}{H^2} - 2\right] I_{+-} = 0.$$

let's define the following new coordinate

$$p = \frac{k_{12}}{k_I} \qquad p' = \frac{k_{34}}{k_I} \qquad I_{\pm\pm} = \frac{G_{\pm\pm}(p,p')}{k_I}$$

and then,

$$\left[(p^2-1)\partial_p^2 + 2p\partial_p + \frac{m^2}{H^2} - 2\right] G_{+-} = 0$$
$$\left[(p'^2-1)\partial_{p'}^2 + 2p'\partial_{p'} + \frac{m^2}{H^2} - 2\right] G_{+-} = 0.$$

[a] we are dropping the boundary term at $\eta = 0$.

The homogeneous solution G_{+-} factorizes into a product of two three-point building blocks regular at $p, p' = 1$:

$$G_{+-}(p,p') = \frac{\pi^2}{2\cosh^2(\pi\mu)}\,{}_2F_1\left(\frac{1}{2}+i\mu,\frac{1}{2}-i\mu,1;\frac{1-p}{2}\right)\,{}_2F_1\left(\frac{1}{2}+i\mu,\frac{1}{2}-i\mu,1;\frac{1-p'}{2}\right). \tag{11.6.11}$$

Fixing the normalization of the factorized four-point function

The proportionality constant in (11.6.11) is determined by imposing the normalization condition at $p=-1$:

$$\begin{aligned} I_{+-} &= \lim_{p,p'\to -1}\int_{-\infty}^{0}\frac{d\eta}{\eta^2}e^{ik_{12}\eta}\int_{-\infty}^{\eta}\frac{d\eta'}{\eta'^2}e^{-ik_{34}\eta'}\frac{\eta\eta'}{2k_I}e^{ik_I(\eta-\eta')} \\ &= \frac{1}{2k_I}\ln(1+p)\ln(1+p') \implies G_{+-} = \frac{1}{2}\ln(1+p)\ln(1+p') \end{aligned}$$

Using the reflection formula for the Gamma function,

$$\begin{aligned} \Gamma\left(\frac{1}{2}+i\mu\right)\Gamma\left(1-\frac{1}{2}-i\mu\right) &= \frac{\pi}{\sin\left[\pi\left(\frac{1}{2}+i\mu\right)\right]} \\ &= -\frac{\pi}{\cos(i\pi\mu)}, \end{aligned}$$

we get[a]:

$$\begin{aligned} \lim_{p\to -1}{}_2F_1\left(\frac{1}{2}+i\mu,\frac{1}{2}-i\mu,1,\frac{1-p}{2}\right) &\to -\frac{\ln\left(1-\frac{1-p}{2}\right)}{\Gamma\left(\frac{1}{2}+i\mu\right)\Gamma\left(\frac{1}{2}-i\mu\right)} \\ &\approx \frac{\cosh(\pi\mu)}{\pi}\ln(1+p) \end{aligned}$$

[a] Using eq. F.3 on p. 96 of Baumann's *Cosmological Bootstrap*.

Squeezed OPE limit and cosmological clock oscillations

In the squeezed limit $k_I \to 0$ ($p, p' \to \infty$), the four-point function factorizes via the Operator Product Expansion (OPE) into the product of two mode integrals:

$$I_E = \frac{1}{2\pi\sqrt{k_{12}k_{34}}}\left[\left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{i\mu}(1+i\sinh\pi\mu)\Gamma(-i\mu)^2\Gamma\left(\frac{1}{2}+i\mu\right)^2+\text{c.c.}\right]. \tag{11.6.12}$$

The squeezed-limit signal from the OPE

$$\begin{aligned} \langle\sigma_{k_I}(\eta)\sigma_{-k_I}(\eta')\rangle'_{\pm\pm} &\approx \frac{H^2}{4\pi}(\eta\eta')^{3/2}\left[\Gamma(\nu)^2\left(\frac{k_I^2\eta\eta'}{4}\right)^{-\nu}+\Gamma(-\nu)^2\left(\frac{k_I^2\eta\eta'}{4}\right)^{\nu}\right] \\ &= (\eta\eta')^{3/2+\nu}\frac{H^2}{4\pi}\left[\Gamma(-\nu)^2\left(\frac{k_I^2}{4}\right)^{\nu}\right]+(\eta\eta')^{3/2-\nu}\frac{H^2}{4\pi}\left[\Gamma(\nu)^2\left(\frac{k_I^2}{4}\right)^{-\nu}\right] \end{aligned}$$

$$= (\eta\eta')^{\Delta}\frac{H^2}{4\pi}\left[\Gamma(-i\mu)^2\left(\frac{k_I^2}{4}\right)^{i\mu}\right] + \underbrace{(\eta\eta')^{\tilde{\Delta}}\frac{H^2}{4\pi}\left[\Gamma(i\mu)^2\left(\frac{k_I^2}{4}\right)^{-i\mu}\right]}_{c.c.}$$

using $\nu = i\mu$ and defining a new variable $J_\pm$:

$$\begin{aligned} J_\pm(k_{12}) &= \pm i\int_{-\infty}^{0}\frac{d\eta}{\eta^2}e^{\pm ik_{12}\eta}(-\eta)^{\Delta} \\ &= -(\mp i)^{\Delta}(k_{12})^{1-\Delta}\Gamma(\Delta-1) \\ &= -(\mp e^{i\frac{\pi}{2}})^{\Delta}(k_{12})^{1-\Delta}\Gamma(\Delta-1) \end{aligned}$$

using $\Delta = \frac{3}{2} + i\mu$, we get from (1):

$$\begin{aligned} I_{++} + I_{+-} + I_{-+} + I_{--} = &\Big(J_+(k_{12})J_+(k_{34}) + J_+(k_{12})J_-(k_{34}) + J_-(k_{12})J_+(k_{34}) \\ &+ J_-(k_{12})J_-(k_{34})\Big)\frac{H^2}{4\pi}\left[\Gamma(-i\mu)^2\left(\frac{k_I^2}{4}\right)^{i\mu}\right] + (\mu \leftrightarrow -\mu) \end{aligned}$$

setting $H = 1$

$$= [J_+(k_{12}) + J_-(k_{12})][J_+(k_{34}) + J_-(k_{34})]\frac{1}{4\pi}\left[\Gamma(-i\mu)^2\left(\frac{k_I^2}{4}\right)^{i\mu}\right] + (\mu \leftrightarrow -\mu)$$

$$I_E = \frac{1}{2\pi\sqrt{k_{12}k_{34}}}\left[\left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{i\mu}(1 + i\sinh\pi\mu)\Gamma(-i\mu)^2\Gamma\left(\frac{1}{2}+i\mu\right)^2 + c.c.\right]$$

For large mass $\mu \gg 1$, this generates the oscillatory cosmological collider clock signal:

$$I_E \sim \frac{\pi e^{-\pi\mu}}{2\mu\sqrt{k_{12}k_{34}}}\sin\left[\mu\ln\left(\frac{k_I^2}{4k_{12}k_{34}}\right) + \delta(\mu)\right]. \tag{11.6.13}$$

Amplitude of the oscillation: Boltzmann suppression and subleading terms

We can see the overall amplitude scales as $e^{-\pi\mu}$ for large μ by using the following:

$$\Gamma(x)\Gamma(-x) = -\frac{\pi}{x\sin(\pi x)}, \qquad \Gamma(x)\Gamma(1-x) = \frac{\pi}{\sin(\pi x)}$$

$$\begin{aligned} \Gamma(\pm i\mu)^2\Gamma\left(\frac{1}{2}\mp i\mu\right)^2 &= |\Gamma(\pm i\mu)|^2\left|\Gamma\left(\frac{1}{2}\mp i\mu\right)\right|^2 e^{\pm i\delta} \\ &= \frac{\pi}{\mu\sinh(\pi\mu)}\frac{\pi}{\cosh(\pi\mu)}e^{\pm i\delta} \end{aligned}$$

We can simplify the I_E for large mass:

$$\begin{aligned} I_E &= \frac{1}{2\pi\sqrt{k_{12}k_{34}}}\frac{\pi^2}{\mu\sinh(\pi\mu)\cosh(\pi\mu)}\left[\left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{i\mu}(1 + i\sinh(\pi\mu))e^{i\delta} + c.c\right] \\ &= \frac{2\pi e^{-2\pi\mu}}{\mu\sqrt{k_{12}k_{34}}}\left[e^{i\mu\ln\left(\frac{k_I^2}{4k_{12}k_3}\right)}(1 + i\sinh(\pi\mu))e^{i\delta} + c.c\right] \end{aligned}$$

$$= \frac{\pi}{\mu}\frac{e^{-2\pi\mu}}{\sqrt{k_{12}k_{34}}}\left[\cos\left[\mu\ln\left(\frac{k_I^2}{4k_{12}k_3}\right)+\delta\right]+\sinh(\pi\mu)\sin\left[\mu\ln\left(\frac{k_I^2}{4k_{12}k_3}\right)+\delta\right]\right]$$
$$\sim \frac{\pi e^{-\pi\mu}}{2\mu\sqrt{k_{12}k_{34}}}\sin\left[\mu\ln\left(\frac{k_I^2}{4k_{12}k_3}\right)+\delta\right]$$

The leading contribution scaling as $e^{-\pi\mu}$ arises from the I_{++} and I_{--} branches, which satisfy the analytic continuation identity $I_{+\pm}(k_{12},k_{34},k_I) = -I_{-\pm}(e^{i\pi}k_{12},k_{34},k_I)$.

Leading vs. subleading terms, and a cross-check via the OPE factorization

As we can check directly using the same identity:

$$\begin{aligned}
I_{++} &= J_+(k_{12})J_+(k_{34})\frac{H^2}{4\pi}\left[\Gamma(-i\mu)^2\left(\frac{k_I^2}{4}\right)^{i\mu}\right]+(\mu\leftrightarrow-\mu)\\
&= (-i)^{2\Delta}\frac{H^2}{4\pi\sqrt{k_{12}k_{34}}}\left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{i\mu}\Gamma(-i\mu)^2\Gamma\left(\frac{1}{2}+i\mu\right)^2+(\mu\leftrightarrow-\mu)\\
&= ie^{\pi\mu}\frac{H^2}{4\pi\sqrt{k_{12}k_{34}}}\left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{i\mu}\Gamma(-i\mu)^2\Gamma\left(\frac{1}{2}+i\mu\right)^2+(\mu\leftrightarrow-\mu)\\
I_{-+} &= J_-(k_{12})J_+(k_{34})\frac{H^2}{4\pi}\left[\Gamma(-i\mu)^2\left(\frac{k_I^2}{4}\right)^{i\mu}\right]+(\mu\leftrightarrow-\mu)\\
&= i^{\Delta}(-i)^{\Delta}\frac{H^2}{4\pi\sqrt{k_{12}k_{34}}}\left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{i\mu}\Gamma(-i\mu)^2\Gamma\left(\frac{1}{2}+i\mu\right)^2+(\mu\leftrightarrow-\mu)\\
&= \frac{H^2}{4\pi\sqrt{k_{12}k_{34}}}\left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{i\mu}\Gamma(-i\mu)^2\Gamma\left(\frac{1}{2}+i\mu\right)^2+(\mu\leftrightarrow-\mu)
\end{aligned}$$

It satisfies the analytic continuation condition:

$$I_{+\pm}(k_{12},k_{34},k_I) = -I_{-\pm}(e^{i\pi}k_{12},k_{34},k_I)$$

We can derive the same using (11.6.11) by taking the limit $p,p'\to\infty$.

$$\lim_{|z|\to\infty}{}_2F_1(a,b,c;z) \approx \frac{\Gamma(b-a)\Gamma(c)}{\Gamma(b)\Gamma(c-a)}(-z)^{-a}+\frac{\Gamma(a-b)\Gamma(c)}{\Gamma(a)\Gamma(c-b)}(-z)^{-b}$$

and duplication formula

$$\begin{aligned}
\Gamma(2i\mu) &= \frac{2^{2i\mu}}{2\sqrt{\pi}}\Gamma(i\mu)\Gamma\left(\frac{1}{2}+i\mu\right)\\
\Gamma(-2i\mu) &= \frac{2^{-2i\mu}}{2\sqrt{\pi}}\Gamma(-i\mu)\Gamma\left(\frac{1}{2}-i\mu\right)\\
\left|\Gamma\left(\frac{1}{2}+i\mu\right)\right|^2 &= \frac{\pi}{\cosh(\pi\mu)}
\end{aligned}$$

Then

$$G_{+-}\sim\frac{1}{2\sqrt{\pi p}}\left[(2p)^{-i\mu}\Gamma(-i\mu)\Gamma\left(\frac{1}{2}+i\mu\right)+\text{c.c.}\right]$$

$$\times \frac{1}{2\sqrt{\pi p'}}\left[(2p')^{-i\mu}\Gamma(-i\mu)\Gamma\left(\frac{1}{2}+i\mu\right)+\text{c.c.}\right]$$

upon simplification, we get

$$G_{+-} \sim \frac{k_I}{4\pi(k_{12}k_{34})^{\frac{1}{2}}}\left[(4pp')^{i\mu}\Gamma(i\mu)^2\Gamma\left(\frac{1}{2}-i\mu\right)^2+c.c.\right]+\dots$$

$$I_{+-} \sim \frac{1}{4\pi(k_{12}k_{34})^{\frac{1}{2}}}\left[(4pp')^{i\mu}\Gamma(i\mu)^2\Gamma\left(\frac{1}{2}-i\mu\right)^2+c.c.\right]+\underbrace{\text{terms independent of } k_I}_{\text{analytic}}$$

In the small-k_I limit ($p, p' \to \infty$), the other integrals give the same result: analytically continuing them and adding them up reproduces I_E, keeping only the terms non-analytic in k_I.

11.6.4 Four-point function with a spinning particle intermediate state

For the scalar exchange above, the squeezed-limit OPE argument said something simple: pulling the two vertices far apart in position space, the exchanged field's propagator reduces to its late-time two-point function, so the whole diagram factorizes into (three-point function at the left vertex) × (something built from the two-point function) × (three-point function at the right vertex). When the exchanged particle carries spin s, the same picture holds, but the exchanged operator $O_{\Delta,s}$ has $2s+1$ independent components – one for each helicity $m = -s, \dots, s$ along the line joining the two vertices – and a general exchange has to sum over all of them, exactly as computing $\langle A|\hat{1}|B\rangle$ requires summing over a complete set of intermediate states. The subtlety is that a completeness sum of this kind is only literally $\sum_m |m\rangle\langle m|$ if the states are unit-normalized; here the natural states are normalized instead by the helicity-m two-point function of §11.2, $\left\langle O_{\Delta,m}(\vec{k}_I)O_{\Delta,-m}(-\vec{k}_I)\right\rangle' = k_I^{2\Delta-3}I_2(\Delta,m)$, i.e. the single term inside the sum (11.2.11) that carries phase $e^{im\chi}$. Dividing each three-point-function product by that overlap is precisely what removes the double-counting of the normalization, giving:

$$I_{+-} = \sum_{m=-s}^{s}\langle O_2O_2O_{\Delta,m}\rangle'\frac{1}{k_I^{2\Delta-3}I_2(\Delta,m)}\langle O_{\Delta,-m}O_2O_2\rangle'. \tag{11.6.14}$$

The sum over helicities packages the angular dependence into the functions $\Theta_\pm(\theta)$, yielding the spinning four-point exchange signal:

$$I_E = \frac{1}{(k_{12}k_{34})^{\frac{1}{2}}}\left[\left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{i\mu}(1+i\sinh\pi\mu)\Gamma(-i\mu)^2\Gamma\left(\frac{1}{2}+s+i\mu\right)^2\Theta_+(\theta)+\text{c.c.}\right]. \tag{11.6.15}$$

Spin-s exchange from the OPE factorization

$$\begin{aligned} I_{+-} &= \sum_m \langle O_{\Delta_e=2}O_{\Delta_e=2}O_{\Delta,m}\rangle'\frac{1}{k_I^{2\Delta-3}I_2(\Delta,m)}\langle O_{\Delta,-m}O_{\Delta_e=2}O_{\Delta_e=2}\rangle' \\ &= \sum_m \langle O_2O_2O_{\Delta,m}\rangle'\frac{1}{k_I^{2\Delta-3}I_2(\Delta,m)}\langle O_{\Delta,-m}O_2O_2\rangle' \end{aligned} \tag{1}$$

where the sum is over all helicity components m. In the small-k_I limit, the Fourier transform of the spinning three-point function (11.3.1), specialized to $\Delta_1 = \Delta_2 = 2$, gives[a]

$$\langle O_2O_2O_{\Delta,m}\rangle'_{k_I\to 0} \sim A_m k_{12}^{1-\Delta}k_I^{2\Delta-3}I_2(m) + B_m k_{12}^{\Delta-2}$$

We have discarded the helicity independent coefficient A_s and B_s for the moment. Then the I_{+-} term becomes:

$$\begin{aligned} I_{+-} &\sim \sum_{m=-s}^{s} A_m A_{-m} (k_{12}k_{34})^{1-\Delta} k_I^{(2\Delta-3)} I_2(m) \\ &\quad + \sum_{m=-s}^{s} B_m B_{-m} (k_{12}k_{34})^{\Delta-2} \frac{1}{k_I^{(2\Delta-3)} I_2(m)} \\ &= (k_{12}k_{34})^{1-\Delta} k_I^{(2\Delta-3)} \sum_{m=-s}^{s} A_m A_{-m} I_2(m) \\ &\quad + (k_{12}k_{34})^{\Delta-2} k_I^{-(2\Delta-3)} \sum_{m=-s}^{s} B_m B_{-m} \frac{1}{I_2(m)} \end{aligned}$$

Without the factor of $I_2(m)$, the sum $\sum_{m=-s}^{s} A_m A_{-m}$ would just be the Legendre polynomial that appears in the spinning two-point function.[b] Since $I_2(m)$ is present, we instead package the sum into a new quantity $\Theta_\pm$:

$$\begin{aligned} I_{+-} &\sim (k_{12}k_{34})^{1-\Delta} k_I^{(2\Delta-3)} \underbrace{\sum_{m=-s}^{s} A_m A_{-m} I_2(m)}_{\Theta_+} \\ &\quad + (k_{12}k_{34})^{\Delta-2} k_I^{-(2\Delta-3)} \underbrace{\sum_{m=-s}^{s} B_m B_{-m} \frac{1}{I_2(m)}}_{\Theta_-} \end{aligned}$$

we also have helicity independent term which depend purely on the spin and are calculated by looking at the highest helicity component. They are given as:

$$\begin{aligned} A_s &\propto \Gamma\left(\frac{1}{2} + s + i\mu\right) \Gamma(-i\mu) 2^{-i\mu} \\ B_s &\propto \Gamma\left(\frac{1}{2} + s - i\mu\right) \Gamma(i\mu) 2^{i\mu} \end{aligned} \tag{2}$$

Therefore, with $\Delta = 3/2 + i\mu$

$$\begin{aligned} I_{+-} &\propto (k_{12}k_{34})^{1-\Delta} k_I^{(2\Delta-3)} A_s^2 \Theta_+ + (k_{12}k_{34})^{\Delta-2} k_I^{-(2\Delta-3)} B_s^2 \Theta_- \\ &= \frac{1}{(k_{12}k_{34})^{\frac{1}{2}}} \left[\left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{i\mu} \Gamma(-i\mu)^2 \Gamma\left(\frac{1}{2} + s + i\mu\right)^2 \Theta_{++} \right. \\ &\qquad \left. + \left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{-i\mu} \Gamma(i\mu)^2 \Gamma\left(\frac{1}{2} + s - i\mu\right)^2 \Theta_{--} \right] \end{aligned}$$

Rest of the diagrams can be evaluated using analytic continuation given as $I_{+\pm}(k_{12}, k_{34}, k_I) = -I_{-\pm}(e^{i\pi}k_{12}, k_{34}, k_I)$ and corresponding complex conjugates. The final answer can be expressed as

$$\begin{aligned} I_E &= I_{++} + I_{+-} + I_{-+} + I_{--} \\ &= \frac{1}{(k_{12}k_{34})^{\frac{1}{2}}} \left[\left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{i\mu} (1 + i\sinh\pi\mu) \Gamma(-i\mu)^2 \Gamma\left(\frac{1}{2} + s + i\mu\right)^2 \Theta_+ \right. \end{aligned}$$

$$+\left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{-i\mu}(1-i\sinh\pi\mu)\Gamma(i\mu)^2\Gamma\left(\frac{1}{2}+s-i\mu\right)^2\Theta_-\Bigg]$$

substituting this back into (1), we will get the four point function with massive intermediate state. The angular dependence is all contained within the factors $\Theta_\pm$.

[a] This is $\Delta_1 = \Delta_2 = 2$ substituted into eq. 4.63 of Arkani-Hamed and Maldacena's *Cosmological Collider Physics*; the helicity-dependent and helicity-independent coefficients A_m, B_m are not fixed by that equation alone and are computed separately below.

[b] Arkani-Hamed and Maldacena, *Cosmological Collider Physics*, arXiv:1503.08043.

11.7 Three Point in inflation

We now turn to correlation functions in an inflationary scenario, where fluctuations of the scalar field are well approximated by a massless Goldstone mode ξ. In single-field slow-roll inflation, the de Sitter isometries are weakly broken by the classical background field $\phi_0(t)$ with non-zero time derivative $\dot{\phi}_0$.

Decomposing the inflaton into its classical background and quantum fluctuations, $\phi(t,\vec{x}) = \phi_0(t)+\xi(t,\vec{x})$, the derivative coupling $\mathcal{L}_{\rm int} = \lambda\int\frac{d\eta d^3x}{H^4\eta^4}(\nabla\xi)^2\sigma$ that produced four-point functions in rigid de Sitter space generates an effective cubic interaction in inflation by replacing one external leg with the rolling background:

$$\lambda\int\frac{d\eta d^3x}{H^4\eta^4}(\nabla\xi)^2\sigma \longrightarrow \frac{\lambda k_3^2\dot{\phi}_0}{\sqrt{2k_3^3}}\int\frac{d\eta d^3x}{H^2\eta^2}e^{ik_3\eta}\sigma(\eta,\vec{k}_3). \tag{11.7.1}$$

The inflaton vertex from the de Sitter one, via $\dot{\phi}_0$

$$\begin{aligned}\lambda\int\frac{d\eta}{H^4\eta^4}(\nabla\xi)^2\sigma &= \lambda\int\frac{d\eta}{H^2\eta^2}(-\partial_\eta\xi\partial_\eta\xi)\sigma = \lambda\int\frac{d\eta}{H^2\eta^2}(-\partial_\eta\xi)\partial_\eta\phi_0\sigma\\ &= \frac{\lambda}{\sqrt{2k_3^3}}\int\frac{d\eta}{H^2\eta^2}(-Hk_3^2\eta e^{ik_3\eta})\partial_\eta\phi_0\sigma\\ &= \frac{\lambda}{\sqrt{2k_3^3}}\int\frac{d\eta}{H^2\eta^2}k_3^2(-H\eta\partial_\eta\phi_0)e^{ik_3\eta}\sigma\\ &= \frac{\lambda k_3^2\dot{\phi}_0}{\sqrt{2k_3^3}}\int\frac{d\eta}{H^2\eta^2}e^{ik_3\eta}\sigma\end{aligned}$$

In the second line, we used (2). However, in the last step, we assumed that $\dot{\phi}_0(\eta)$ is a slowly varying function of η, allowing us to factor it out of the integral. The $\mathcal{L}_{\rm interaction}$ at the two vertices is then:

$$\text{vertex 1:}\quad \lambda\int\frac{d\eta}{H^4\eta^4}(\nabla\xi)^2\sigma = \lambda\int\frac{d\eta}{\eta^2}(-k_1^2\eta e^{ik_1\eta})(-k_2^2\eta e^{ik_2\eta})\sigma$$

$$\text{vertex 2:}\quad \lambda\int\frac{d\eta}{H^4\eta^4}(\nabla\xi)^2\sigma = \lambda k_3^2\dot{\phi}_0\int\frac{d\eta}{H^2\eta^2}e^{ik_3\eta}\sigma$$

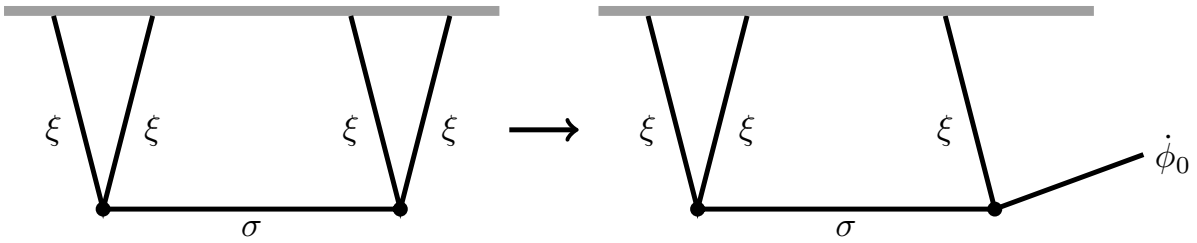
All modes cross the Hubble radius as time proceeds and become "super-Hubble." Unlike "sub-Hubble" modes, $k \gg a(t)H \equiv |\eta|^{-1}$, which oscillate, super-Hubble modes ($k\eta \to 0$) freeze out and approach a constant value.[a] So $\dot{\phi}_0(0)$ is essentially the same as its value at

horizon crossing, $\dot\phi_0(\eta_\star)$ with $k\eta_\star = 1$. We can evaluate the first vertex using the results for the conformally coupled scalar with $\mathcal{L} = \int \xi^2\sigma$, by acting on it with $\mathcal{O}_{12}$ from (1):

$$\int \frac{d\eta}{H^4\eta^4}(\nabla\xi)^2\sigma = \frac{1}{\sqrt{2k_1^3}\sqrt{2k_2^3}}\mathcal{O}_{12}\int \frac{d\eta}{\eta^2}e^{ik_{12}\eta}\sigma(\eta,k_3)$$

Doing so has a direct consequence: at second order, such derivative coupling interaction term in de Sitter space yields four-point functions. However, when used in an inflationary background, it will give rise to three-point functions instead.

[a] See p. 34 of Sayantan Choudhury's *Large Fluctuations and Primordial Black Holes.*



In the squeezed limit $k_3 \ll k_1 \approx k_2$, setting $k_4 \to 0$ in the four-point exchange amplitude ($k_{34} = k_I = k_3$, $p' \equiv k_{34}/k_I \to 1$) and acting with the differential operator $\mathcal{O}_{12}$ gives the three-point bispectrum of inflaton fluctuations:

$$\langle\xi_{\vec{k}_1}\xi_{\vec{k}_2}\xi_{\vec{k}_3}\rangle'\Big|_{k_3\to 0} = H^3\frac{\dot\phi_0(\eta_*)\lambda^2}{2k_1^3k_3^3}J\left(\mu,\frac{k_3}{k_1}\right), \tag{11.7.2}$$

where the cosmological collider shape function $J(\mu, k_3/k_1)$ is:

$$J\left(\mu,\frac{k_3}{k_1}\right) \equiv \frac{\pi^{3/2}}{4\cosh^2(\pi\mu)}\left(\frac{k_3}{k_1}\right)^{3/2}\left[\left(\frac{k_3}{4k_1}\right)^{-i\mu}\frac{(1-i\sinh\pi\mu)\left(\frac{5}{2}-i\mu\right)\left(\frac{3}{2}-i\mu\right)\Gamma(i\mu)}{\Gamma\left(\frac{1}{2}+i\mu\right)} + \text{c.c.}\right]. \tag{11.7.3}$$

Relating the inflaton fluctuation ξ to the gauge-invariant comoving curvature perturbation $\zeta = -\frac{H}{\dot\phi_0}\xi = \frac{1}{\sqrt{2\epsilon}M_{\rm pl}}\xi$, the dimensionless ratio measuring primordial non-Gaussianity evaluates to:

$$\frac{\langle\zeta_{\vec{k}_1}\zeta_{\vec{k}_2}\zeta_{\vec{k}_3}\rangle'}{4\langle\zeta_{\vec{k}_1}\zeta_{-\vec{k}_1}\rangle'\langle\zeta_{\vec{k}_3}\zeta_{-\vec{k}_3}\rangle'} = -\frac{\dot\phi_0^2\lambda^2}{2H^2}J\left(\mu,\frac{k_3}{k_1}\right). \tag{11.7.4}$$

The inflationary squeezed-limit three-point function

The three point function using in-in formalism can be written as:

$$\begin{aligned}\langle\xi_{k_1}\xi_{k_2}\xi_{k_3}\rangle &= (+i)(+i)\langle 0|\,\xi_{k_1}\xi_{k_2}\xi_{k_3}\left(\lambda\int\frac{d\eta d^3x}{H^4\eta^4}(\nabla\xi)^2\sigma\right)\\ &\quad\times\left(\frac{\lambda k_3^2\dot\phi_0}{\sqrt{2k_3^3}}\int\frac{d\eta' d^3x}{H^2\eta'^2}e^{ik_3\eta'}\sigma\right)|0\rangle_{++} + \dots\\ &= \frac{(+i)(+i)H^3}{(2k_1^3)(2k_2^3)(2k_3^3)}\left(\lambda\mathcal{O}_{12}\int_{-\infty}^{0}\frac{d\eta}{\eta^2}e^{ik_{12}\eta}\langle 0|\,\sigma(\eta,k_3)\right)\end{aligned}$$

$$
\begin{aligned}
&\times \left(\lambda k_3^2 \dot{\phi}_0 \int \frac{d\eta'}{H^2\eta'^2} e^{ik_3\eta'} \sigma(\eta', k_3) \left|0\right\rangle_{++}\right) + \dots \\
&= (+i)(+i) H^3 \frac{\lambda^2 k_3^2 \dot{\phi}_0}{8 k_1^3 k_2^3 k_3^3} \mathcal{O}_{12} \int_{-\infty}^{0} \frac{d\eta}{\eta^2} e^{ik_{12}\eta} \\
&\times \int_{-\infty}^{0} \frac{d\eta'}{H^2\eta'^2} e^{ik_3\eta'} \left\langle 0\right| \sigma(\eta, k_3)\sigma(\eta', k_3) \left|0\right\rangle_{++} + \dots \\
&= 2^2 H^3 \frac{\lambda^2 k_3^2 \dot{\phi}_0}{8 k_1^3 k_2^3 k_3^3} \mathcal{O}_{12} \underbrace{(I_{++} + I_{+-} + I_{-+} + I_{--})}_{I_E} + 2 \text{ other diagrams}
\end{aligned}
$$

where d^3x integrals were converted into $\delta^3(k_1 + k_2 + k_3)$ at vertex 1 and vertex 2. We also used

$$\lim_{\eta\to 0} f_{k_1} f_{k_2} f_{k_3} = \frac{H^3}{\sqrt{2k_1^3}\sqrt{2k_2^3}\sqrt{2k_3^3}}$$

for the mode function of ξ. We quickly find that I_E is same as (1) with the key difference being $k_4 = 0$. Following relationship is used to see how $\mathcal{O}_{12}$ acts on I_E:

$$
\begin{aligned}
\Gamma\left(\frac{1}{2} - i\mu\right)\Gamma\left(1 - \frac{1}{2} + i\mu\right) &= \Gamma\left(\frac{1}{2} - i\mu\right)\Gamma\left(\frac{1}{2} + i\mu\right) \\
&= \frac{\pi}{\sin\pi\left(\frac{1}{2} - i\mu\right)} = \frac{\pi}{\cosh(\pi\mu)}
\end{aligned}
$$

and, recalling $q = \frac{k_1 - k_2}{k_I}$, $p = \frac{k_1 + k_2}{k_I}$ (in the limit $q << p$ and $p \to \infty$)

$$
\begin{aligned}
\mathcal{O}_{12}(k_{12})^a &= \frac{k_I^2}{8}\left[(q^2 - p^2)(1 - p^2)\frac{\partial^2}{\partial p^2} + 2(p^2 + q^2 - 2)\left(1 - p\frac{\partial}{\partial p}\right)\right](k_{12})^a \\
&= \frac{k_I^2}{8}\left[-p^2(1 - p^2)\frac{\partial^2}{\partial p^2} + 2(p^2 - 2)\left(1 - p\frac{\partial}{\partial p}\right)\right] k_I^a p^a \\
&= \frac{k_I^{a+2}}{8}[-p^2(1 - p^2)a(a-1)p^{a-2} + 2(p^2 - 2)(1 - a)p^a] \\
&= \frac{k_I^{a+2}}{8}[(p^2 - 1)a(a-1)p^a + 2(p^2 - 2)(1 - a)p^a] \\
&= \frac{k_I^{a+2}}{8}(a-1)p^a[ap^2 - a - 2(p^2 - 2)] \\
&= \frac{k_I^{a+2}}{8}(a-1)p^a[ap^2 - a - 2p^2 + 4] \\
&= \frac{k_I^{a+2}}{8}(a-1)(a-2)p^{a+2} + \mathcal{O}(p^a) \\
&\sim \frac{1}{8}(a-1)(a-2)k_{12}^{a+2} \qquad (1)
\end{aligned}
$$

In the long wavelength limit:

$$
\begin{aligned}
I_E(k_{12}, k_3, k_3)\Big|_{k_3\to 0} = \frac{1}{\sqrt{k_3 k_{12}}} \frac{\pi}{\sqrt{2\pi}\cosh\pi\mu} \Bigg[&\left(\frac{k_3}{2k_{12}}\right)^{i\mu} (1 + i\sinh\pi\mu)\Gamma(-i\mu)\Gamma\left(\frac{1}{2} + i\mu\right) \\
&+ \left(\frac{k_3}{2k_{12}}\right)^{-i\mu} (1 - i\sinh\pi\mu)\Gamma(i\mu)\Gamma\left(\frac{1}{2} - i\mu\right)\Bigg]
\end{aligned}
$$

We can now act with the operator O_{12} on I_E and use (1) to obtain

$$\langle \xi_{k_1}\xi_{k_2}\xi_{k_3}\rangle' \big|_{k_3\to 0} = H^3 \frac{\dot\phi_0(\eta_*)\lambda^2}{2k_1^3k_3^3} J\left(\mu, \frac{k_3}{k_1}\right)$$

where, using $\frac{k_3^2}{\sqrt{k_3}} = k_3^{3/2}$, $\frac{k_{12}^{3/2}}{k_2^3} \approx 2^{3/2}k_1^{-3/2}$, and $k_{12} = \left|\vec{k}_1\right| + \left|\vec{k}_2\right| = 2k_1$,

$$\begin{aligned} J\left(\mu, \frac{k_3}{k_1}\right) &\equiv \frac{1}{8}\frac{1}{\sqrt{2\pi}}\left(\frac{\pi^2}{\cosh^2\pi\mu}\right)\frac{k_3^2}{k_2^3}\frac{k_{12}^{3/2}}{\sqrt{k_3}}\times \\ &\quad \left[\left(\frac{k_3}{2k_{12}}\right)^{-i\mu}\frac{(1-i\sinh\pi\mu)\left(\frac{5}{2}-i\mu\right)\left(\frac{3}{2}-i\mu\right)\Gamma(i\mu)}{\Gamma\left(\frac{1}{2}+i\mu\right)} + c.c.\right] \\ &= \frac{1}{4\sqrt{\pi}}\left(\frac{\pi^2}{\cosh^2\pi\mu}\right)\left(\frac{k_3}{k_1}\right)^{3/2}\times \\ &\quad \left[\left(\frac{k_3}{4k_1}\right)^{-i\mu}\frac{(1-i\sinh\pi\mu)\left(\frac{5}{2}-i\mu\right)\left(\frac{3}{2}-i\mu\right)\Gamma(i\mu)}{\Gamma\left(\frac{1}{2}+i\mu\right)} + c.c.\right] \end{aligned}$$

Using these correlators, we can obtain correlator of curvature perturbation given by:

$$\zeta = -\frac{H}{\dot\phi_0}\xi$$

Using $\zeta = -H/\dot\phi_0\,\xi$ above and the slow-roll parameter ϵ defined below, we can also express it as:

$$\zeta = \frac{2\sqrt{\pi}}{M_{\rm pl}\sqrt{\epsilon}}\xi \qquad \text{or} \qquad \zeta = \frac{1}{m_{\rm p}\sqrt{2\epsilon}}\xi$$

where the second equality holds when we define the slow-roll parameter using the reduced Planck mass, $\epsilon = \frac{m_{\rm p}^2}{2}\left(\frac{V'}{V}\right)^2$, via $M_{\rm pl} = 2\sqrt{2\pi}m_{\rm p}$. The mode function for the curvature perturbation using the second definition is[a]:

$$\begin{aligned} f_{\vec{k}} &= \frac{1}{M_{\rm pl}\sqrt{2\epsilon}}\frac{H}{\sqrt{2k^3}}(1-ik\eta)e^{ik\eta} \\ &= \frac{H}{2\sqrt{\epsilon}M_{\rm pl}}\frac{1}{\sqrt{k^3}}(1-ik\eta)e^{ik\eta} \end{aligned}$$

We are interested in evaluating the following ratio:

$$\begin{aligned} \frac{\langle\zeta_{\vec{k}_1}\zeta_{\vec{k}_2}\zeta_{\vec{k}_3}\rangle'}{4\langle\zeta_{\vec{k}_1}\zeta_{-\vec{k}_1}\rangle'\langle\zeta_{\vec{k}_3}\zeta_{-\vec{k}_3}\rangle'} &= \frac{\frac{H^3}{\dot\phi_0^3}\langle\xi_{\vec{k}_1}\xi_{\vec{k}_2}\xi_{\vec{k}_3}\rangle'}{4\frac{H^2}{\dot\phi_0^2}\langle\xi_{\vec{k}_1}\xi_{-\vec{k}_1}\rangle\frac{H^2}{\dot\phi_0^2}\langle'\xi_{\vec{k}_3}\xi_{-\vec{k}_3}\rangle'} \\ &= \frac{-\frac{H^3}{\dot\phi_0^3}H^3\frac{\dot\phi_0\lambda^2}{2k_1^3k_3^3}J\left(\mu,\frac{k_3}{k_1}\right)}{4\frac{H^4}{\dot\phi_0^4}\frac{H^2}{2k_1^3}\frac{H^2}{2k_1^3}} \\ &= -\frac{\dot\phi_0^2\lambda^2}{2H^2}J\left(\mu,\frac{k_3}{k_1}\right) \end{aligned}$$

[a] See eq. B.6 of *Gravitational Production of Superheavy Dark Matter and Associated Cosmological Signatures*; the only difference is that they use $f_{\vec{k}}^\star$.

The ratio in (11.7.4) provides the fundamental observable for local primordial non-Gaussianity, parameterized by $\frac{3}{5}f_{\rm NL}^{\rm local}$. In single-field slow-roll inflation, the Maldacena consistency condition dictates that a long-wavelength mode acts purely as a coordinate transformation on short-wavelength modes, leaving physical observables unperturbed at leading order and suppressing the squeezed limit as $(k_3/k_1)^2$ (or fixing $f_{\rm NL}^{\rm local} = \frac{5}{12}(1-n_s) \ll 1$). In sharp contrast, the exchange of an extra massive scalar field produces a physical signal scaling as $(k_3/k_1)^{3/2\pm i\mu}$, which is parametrically larger and features distinct clock oscillations in $\ln(k_3/k_1)$.

11.7.1 Intermediate massive field with spin

When the exchanged massive particle carries spin $s \geq 1$, the OPE factorization in the inflationary background selects only the longitudinal helicity-zero ($m = 0$) mode in the squeezed limit $k_3 \to 0$:

$$I_{+-} = \langle O_3 O_3 O_{\Delta,0}\rangle' \frac{1}{k_3^{2\Delta-3} I_2(\Delta,0)} \langle O_{\Delta,0} O_3\rangle'. \tag{11.7.5}$$

Helicity modes with $m > 0$ involve contractions of the exchanged polarization tensor with external momenta $(k_3 - k_4)_i \epsilon_{ij}(k_I)$, which vanish identically by transversality in the limit $k_4 \to 0$.[4]

The linear mixing between the rolling background inflaton and the massive spinning operator is given by:

$$\langle O_3 \epsilon' \cdot O\rangle' = \dot\phi_0 \frac{\Gamma(\Delta-1)}{\Gamma(\Delta-1+s)} \frac{G_s(1)}{[\Delta(\Delta-3) - s(s-3)]} k^\Delta (\vec\epsilon \cdot \vec k)^s, \tag{11.7.6}$$

where $G_s(1) \propto \Gamma(\frac{1}{2}+s+i\mu)\Gamma(\frac{1}{2}+s-i\mu)$. The resulting squeezed bispectrum for spin-s exchange takes the factorized form:

$$\langle \xi_{\vec k_1} \xi_{\vec k_2} \xi_{\vec k_3}\rangle'\Big|_{k_3\to 0} = H^3 \frac{\dot\phi_0 \lambda_s^2}{2k_1^3 k_3^3} P_s(\cos\theta) J_s\left(\mu, \frac{k_3}{k_1}\right), \tag{11.7.7}$$

where $P_s(\cos\theta)$ is the Legendre polynomial of order s with $\cos\theta = \hat k_1 \cdot \hat k_3$, and the angular dependence uniquely isolates the spin s of the intermediate particle.

The squeezed-limit signal for spin-s exchange

$$\begin{aligned} I_{+-} &= \sum_m \langle O_{\Delta_e=3} O_{\Delta_e=3} O_{\Delta,m}\rangle' \frac{1}{k_I^{2\Delta-3} I_2(\Delta,m)} \langle O_{\Delta,-m} O_{\Delta_e=3} O_{\Delta_e=3}\rangle' \\ &= \langle O_3 O_3 O_{\Delta,0}\rangle' \frac{1}{k_I^{2\Delta-3} I_2(\Delta,0)} \langle O_{\Delta,0} O_3\rangle' \end{aligned} \tag{1}$$

where only the helicity-zero part contributes. The parts of the four-point function with helicity greater than zero involve contractions of the polarization tensor for the exchanged particle (with intermediate momentum k_I) with the external momenta; in particular, there is always at least one contraction of the form $(k_3 - k_4)_i \epsilon_{ij}(k_I)$, and by transversality of the polarization tensor this contraction vanishes in the limit $k_4 \to 0$.[a] As in the previous section, the two-point function $\langle O_{\Delta,0} O_3\rangle'$ in the inflationary background can be expressed in terms of the de Sitter three-point function as:

$$\langle O_3 \epsilon'.O\rangle' = \dot\phi_0 \frac{\Gamma(\Delta-1)}{\Gamma(\Delta-1+s)} \frac{G_s(1)}{[\Delta(\Delta-3) - s(s-3)]} k^\Delta (\vec\epsilon.\vec k)^s \tag{2}$$

[4] See p. 52 of Baumann's *Cosmological Bootstrap*.

where, from the normalization condition that $I_{+-} \propto {}^{1}/_{(1+p)^s}$ we have:

$$G_s(1) \propto \Gamma\left(\frac{1}{2}+s+i\mu\right)\Gamma\left(\frac{1}{2}+s-i\mu\right)$$

We now go back to (1) and look at the answer for small $\vec{k}_1$. We only need to expand the first factor of (1) for large p. The expansion of the first factor for large p can be given as:

$$\langle O_3 O_3 O_{\Delta,m=0}\rangle' \propto P_s(\cos\theta)\left[A_s\frac{\Delta+1+s}{\Delta-3+s}k_1^{3-\Delta}k_3^{2\Delta-3}I_2(0)+B_s\frac{\Delta-s-4}{\Delta-s}k_1^{\Delta}\right]$$

with A_s and B_s given in (2). Here $\cos\theta = \hat{k}_1.\hat{k}_3$. Together with (2) we can write the small k_3 expansion as

$$\begin{aligned} I_{+-} \propto P_s(\cos\theta)(k_1k_3)^{\frac{3}{2}}&\frac{\Gamma(\frac{1}{2}+s+i\mu)^2\Gamma(\frac{1}{2}+s-i\mu)^2}{\Delta(\Delta-3)-s(s-3)} \\ &\times\left[\left(\frac{k_3}{4k_1}\right)^{i\mu}\frac{5}{3}\frac{\frac{5}{2}+i\mu+s}{\frac{5}{2}-i\mu-s}\frac{\Gamma(\frac{1}{2}+i\mu)}{\Gamma(-i\mu)}\frac{\Gamma(-i\mu+s)\Gamma(\frac{1}{2}+i\mu+s)}{\Gamma(\frac{1}{2}-i\mu+s)}\right. \\ &\left.+\left(\frac{k_3}{4k_1}\right)^{-i\mu}\frac{5}{3}\frac{\frac{5}{2}-i\mu+s}{\frac{5}{2}+i\mu-s}\frac{\Gamma(\frac{1}{2}-i\mu)}{\Gamma(i\mu)}\frac{\Gamma(i\mu+s)\Gamma(\frac{1}{2}-i\mu+s)}{\Gamma(\frac{1}{2}+i\mu+s)}\right] \end{aligned}$$

The rest of the steps are the same as in the previous section.

[a] See p. 52 of Baumann's *Cosmological Bootstrap* for this.

11.8 Comments on Loop correction

When the inflaton couples to composite operators, such as the Standard Model Higgs doublet bilinear through $\int(\nabla\xi)^2 hh^\dagger$, loop diagrams contribute to cosmological correlators. More generally, consider the coupling of the inflaton to a general four-dimensional operator $\mathcal{O}^{(4)}$:

$$S_{\rm int} = \lambda\int d^4x\sqrt{-g}\,\phi^2\mathcal{O}^{(4)}. \tag{11.8.1}$$

On the future conformal boundary ($\eta\to 0$), the operator $\mathcal{O}^{(4)}$ decomposes into a sum of boundary conformal primaries $O_{\Delta_j}(\vec{x})$ with scaling dimensions Δ_j:

$$\mathcal{O}^{(4)}(\eta,\vec{x}) \sim \sum_j (-\eta)^{\Delta_j} O_{\Delta_j}(\vec{x}), \qquad \langle O_{\Delta_j}(\vec{x})O_{\Delta_j}(\vec{0})\rangle = \frac{c_j}{|\vec{x}|^{2\Delta_j}}. \tag{11.8.2}$$

The four-point loop exchange amplitude evaluates to:

$$I_{jE} = k_I^{2\Delta_j-3}8\pi\sin(\pi\Delta_j)(k_{12}k_{34})^{1-\Delta_j}(1+\cos(\pi\Delta_j))\Gamma(\Delta_j-1)^2\Gamma(2-2\Delta_j)c_j, \tag{11.8.3}$$

yielding the late-time four-point function in the squeezed limit $k_I\to 0$:

$$\langle\xi_{\vec{k}_1}\xi_{\vec{k}_2}\xi_{\vec{k}_3}\xi_{\vec{k}_4}\rangle'_{k_I\to 0} \sim \frac{\lambda^2}{4k_1^3k_2^3k_3^3}\sum_j I_{jE}. \tag{11.8.4}$$

Loop correction as a tree-level exchange of boundary operators

The two-point function of the three-dimensional operator is

$$\langle O_{\Delta_j}(\vec{x})O_{\Delta_j}(\vec{0})\rangle = \frac{c_j}{|x|^{2\Delta_j}}$$

we can use this to find the four-point function:

$$I_{j\pm\pm} = (\pm i)(\pm i)\int_{-\infty}^{0}\frac{d\eta}{\eta^2}e^{ik_{12}\eta}\int_{-\infty}^{0}\frac{d\eta'}{\eta'^2}e^{ik_{34}\eta'}\int d^3x e^{i\vec{k}_I\cdot\vec{x}}(-\eta)^{\Delta_j}(-\eta')^{\Delta_j}\langle O_{\Delta_j}(\vec{x})O_{\Delta_j}(\vec{0})\rangle$$

we evaluate the x integral using:

$$\int d^3x e^{-\vec{k}\cdot\vec{x}}|x|^{-2\alpha} = 2^{2(3/2-\alpha)}\pi^{3/2}\frac{\Gamma(3/2-\alpha)}{\Gamma(\alpha)}(k^2)^{\alpha-3/2}$$

Thus,

$$\begin{aligned} I_{j\pm\pm} &= (\pm i)(\pm i)2^{3-2\Delta_j}\pi^{3/2}\frac{\Gamma\left(3/2-\Delta_j\right)}{\Gamma(\Delta_j)}k_I^{2\Delta_j-3} \\ &\quad\times\int_{-\infty}^{0}\frac{d\eta}{\eta^2}e^{ik_{12}\eta}\int_{-\infty}^{0}\frac{d\eta'}{\eta'^2}e^{ik_{34}\eta'}(-\eta)^{\Delta_j}(-\eta')^{\Delta_j}c_j \end{aligned}$$

which can be simplified by using:

$$\begin{aligned} \frac{\Gamma\left(3/2-\Delta\right)}{\Gamma(\Delta)} &= \frac{\Gamma\left(3/2-\Delta\right)\Gamma(1-\Delta)}{\Gamma(\Delta)\Gamma(1-\Delta)} = \frac{\sin(\pi\Delta)}{\pi}\Gamma\left(3/2-\Delta\right)\Gamma(1-\Delta) \\ &= \frac{\sin(\pi\Delta)}{\pi}\Gamma\left(1/2+1-\Delta\right)\Gamma(1-\Delta) \end{aligned}$$

using duplication formula:

$$\begin{aligned} &= \frac{\sin(\pi\Delta)}{\pi}2^{1-2(1-\Delta)}\sqrt{\pi}\Gamma(2(1-\Delta)) \\ &= \frac{\sin(\pi\Delta)}{\pi}2^{2\Delta-1}\sqrt{\pi}\Gamma(2-2\Delta) \end{aligned}$$

hence,

$$I_{j\pm\pm} = (\pm i)(\pm i)2^2\pi\sin(\pi\Delta_j)k_I^{2\Delta_j-3}\int_{-\infty}^{0}\frac{d\eta}{\eta^2}e^{ik_{12}\eta}\int_{-\infty}^{0}\frac{d\eta'}{\eta'^2}e^{ik_{34}\eta'}(-\eta)^{\Delta_j}(-\eta')^{\Delta_j}c_j$$

performing the η and η' integrals using

$$\begin{aligned} J_{\pm}(k_{12}) &= \pm i\int_{-\infty}^{0}\frac{d\eta}{\eta^2}e^{\pm ik_{12}\eta}(-\eta)^{\Delta} = -(\mp i)^{\Delta}(k_{12})^{1-\Delta}\Gamma(\Delta-1) \\ [J_+(k_{12})+J_-(k_{12})][J_+(k_{34})+J_-(k_{34})] &= 2(1+\cos(\pi\Delta))(k_{12}k_{34})^{1-\Delta}\Gamma(\Delta-1)^2 \end{aligned}$$

we get,

$$\begin{aligned} I_{jE} &= \sum I_{j\pm\pm} \\ &= k_I^{2\Delta_j-3}8\pi\sin(\pi\Delta_j)(k_{12}k_{34})^{1-\Delta_j}(1+\cos(\pi\Delta_j))\Gamma(\Delta_j-1)^2\Gamma(2-2\Delta_j)c_j \end{aligned}$$

The final answer can be expressed as:

$$\langle\xi_{\vec{k}_1}\xi_{\vec{k}_2}\xi_{\vec{k}_3}\xi_{\vec{k}_4}\rangle_{k_I\to 0} \sim \frac{\lambda^2}{4k_1^3k_2^3k_3^3}\sum_j I_{jE}$$

The analytic structure of (11.8.3) reveals a fundamental physical diagnostic distinguishing light from heavy loop states:

- **Loops of Light Fields:** For light particles ($m < \frac{3}{2}H$), the composite double-trace operators $[\mathcal{O}\mathcal{O}]_{n,\ell}$ have real conformal dimensions $\Delta_j = 2\Delta_{\text{light}} + 2n + \ell \in \mathbb{R}$. The factor $k_I^{2\Delta_j - 3}$ produces purely real power laws and local analytic terms with no oscillations in $\ln k_I$. A loop of light particles is therefore completely invisible to the oscillatory clock signature.
- **Loops of Heavy Fields:** For heavy particles ($m > \frac{3}{2}H$), the composite operator $[\sigma\sigma]$ carries complex scaling dimension $\Delta \approx 2\left(\frac{3}{2} + i\mu\right) = 3 + 2i\mu$. Consequently, a loop of heavy particles produces an oscillatory signal with *double frequency* 2μ:

$$\left(\frac{k_I^2}{4k_{12}k_{34}}\right)^{2i\mu} = e^{2i\mu \ln\left(\frac{k_I^2}{4k_{12}k_{34}}\right)}, \tag{11.8.5}$$

  suppressed by the two-particle Boltzmann factor $e^{-2\pi\mu}$.

The frequency of the squeezed-limit oscillation therefore directly counts the number of massive particles exchanged in the internal channel: single-particle tree exchange oscillates at frequency μ, while two-particle loop exchange oscillates at frequency 2μ.

11.9 Conclusion and Cosmological Horizons

Inflation provides the highest-energy physical environment accessible to human observation, with Hubble expansion rates reaching $H \lesssim 10^{14}$ GeV. Cosmological collider physics establishes a precise dictionary that translates the primordial correlation functions of the curvature perturbation ζ into particle-detector outputs from this ultra-high-energy epoch.

11.9.1 Summary of the Particle Spectrogram

Throughout this chapter, we have traced how the symmetries of de Sitter space and their controlled breaking during inflation constrain the analytic structure of primordial correlators. A new particle of mass m and spin s leaves an unmistakable, non-analytic imprint on the squeezed limit of the bispectrum:

$$\lim_{k_{\text{long}}\to 0} \frac{\langle \zeta_{\vec{k}_1}\zeta_{\vec{k}_2}\zeta_{\vec{k}_3}\rangle'}{4\langle \zeta_{\vec{k}_1}\zeta_{-\vec{k}_1}\rangle'\langle \zeta_{\vec{k}_3}\zeta_{-\vec{k}_3}\rangle'} \propto \epsilon\, e^{-\pi\mu}\left(\frac{k_{\text{long}}}{k_{\text{short}}}\right)^{3/2} \cos\left[\mu \ln\left(\frac{k_{\text{long}}}{4k_{\text{short}}}\right) + \delta(\mu)\right] P_s(\cos\theta), \tag{11.9.1}$$

where $\mu \equiv \sqrt{m^2/H^2 - 9/4}$, $\epsilon \equiv -\dot{H}/H^2$, and $\cos\theta \equiv \hat{k}_{\text{long}} \cdot \hat{k}_{\text{short}}$. This signal encodes the complete physical profile of the intermediate state:

1. **Mass from the Clock Frequency:** The logarithmic frequency μ directly measures the particle mass in Hubble units. Heavy particles in the principal series ($m > \frac{3}{2}H$) act as standard clocks that oscillate during horizon exit.
2. **Spin from the Angular Profile:** The angular dependence is governed uniquely by the Legendre polynomial $P_s(\cos\theta)$, reflecting the representation of the intermediate state under the spatial rotation group $SO(3)$ that stabilizes the long momentum vector.
3. **Quantum Interference and Phase:** The phase shift $\delta(\mu) = \arg[\Gamma(i\mu)^2/\Gamma(\frac{1}{2} + i\mu)]$ and the Boltzmann factor $e^{-\pi\mu}$ reflect spontaneous pair production and quantum interference between virtual particle creation and classical background evolution.

4. **Parametric Violation of Single-Field Consistency:** While single-field inflation strictly suppresses the physical squeezed limit as $(k_{\rm long}/k_{\rm short})^2$, the cosmological collider signal scales as $(k_{\rm long}/k_{\rm short})^{3/2}$, providing an unambiguous, smoking-gun signature of multi-field particle production.

Figure 11.3 plots the oscillatory factor itself, $\cos[\mu \ln(k_{\rm long}/4k_{\rm short}) + \delta(\mu)]$, for two representative masses – the power-law envelope $(k_{\rm long}/k_{\rm short})^{3/2}$ and the overall Boltzmann suppression $e^{-\pi\mu}$ are left out, exactly as in the analogous plots in the literature,[5] since otherwise the exponentially decaying envelope would visually swamp the oscillation on a linear scale. A heavier particle ($\mu = 3$) oscillates about three times as fast as a lighter one ($\mu = 1$) over the same range of squeezing, which is the entire content of "reading off the mass from the frequency."

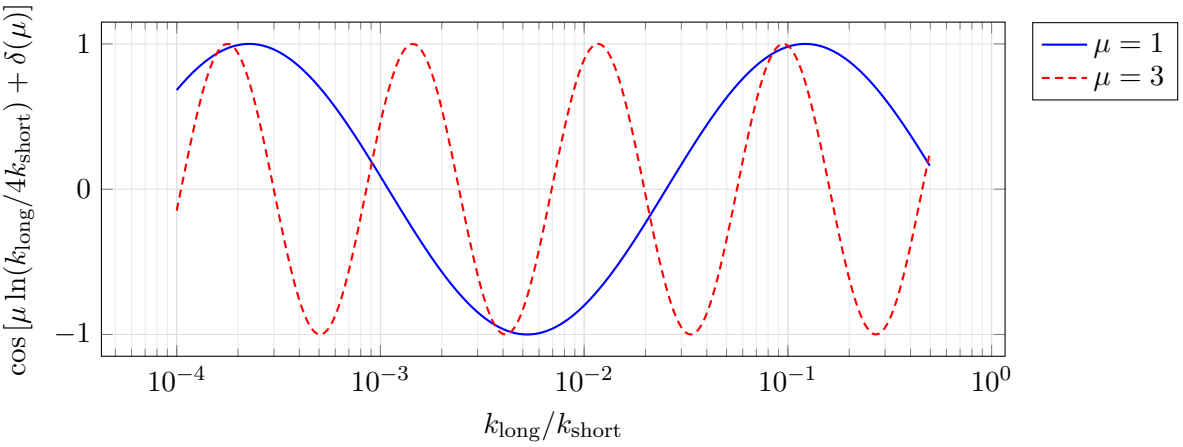


Figure 11.3: The oscillatory part of the squeezed-limit signal, Eq. (11.0.1), as a function of the squeezing ratio $k_{\rm long}/k_{\rm short}$, for a field with $\mu = 1$ (blue) and $\mu = 3$ (red dashed), using $\delta(\mu) = \arg[\Gamma(i\mu)^2/\Gamma(\frac{1}{2} + i\mu)]$. The power-law envelope $(k_{\rm long}/k_{\rm short})^{3/2}$ and the Boltzmann factor $e^{-\pi\mu}$, which multiply this oscillation in the actual physical signal, are omitted here so the frequency and phase are visible on a common scale; the heavier field oscillates three times as fast, exactly as $\mu = 3$ versus $\mu = 1$ predicts.

11.9.2 Scale-Dependent vs. Ratio-Dependent Oscillations

It is essential to distinguish the cosmological collider signal from other oscillatory features in primordial cosmology:

- **Scale-Dependent Oscillations:** Oscillations of the form $\cos(\omega \ln(k_1 + k_2 + k_3) + \varphi)$ arise from sharp features in the inflationary potential (such as steps, inflection points, or periodic axion monodromy potentials) that break the time-translation and scale invariance of de Sitter space. These oscillations depend on the *overall scale* $k_t = k_1 + k_2 + k_3$.

- **Ratio-Dependent Oscillations:** In contrast, cosmological collider oscillations preserve de Sitter scale invariance and depend strictly on momentum *ratios* $k_{\rm long}/k_{\rm short}$. They arise entirely from the internal dynamics and on-shell propagation of massive quantum fields in an unperturbed, scale-invariant background.

[5] See e.g. Fig. 11 of Arkani-Hamed and Maldacena, *Cosmological Collider Physics*, arXiv:1503.08043, or Fig. 31 of Baumann and Joyce's Cargese lectures, *Quantum Fields in Curved Spacetime*, which further compares this shape side by side with the Z-resonance line shape measured at LEP – the collider-physics object this whole chapter has been building an analogue of.

11.9.3 Observational Prospects: CMB, LSS, and 21-cm Tomography

Because of the thermal Boltzmann suppression $e^{-\pi m/H}$, detecting cosmological collider signals requires probing a vast statistical sample of primordial perturbation modes:

- **Cosmic Microwave Background (CMB):** Current measurements from the *Planck* satellite constrain local non-Gaussianity to $f_{\rm NL}^{\rm local} = -0.9 \pm 5.1$. While the CMB is cosmic-variance limited to $N_{\rm modes} \sim 10^6$ multipoles, it provides strong upper bounds on light and moderately coupled states.
- **Large-Scale Structure (LSS):** Galaxy clustering surveys (such as DESI, *Euclid*, the Vera Rubin Observatory, and the Roman Space Telescope) probe 3D matter distributions, mapping up to $N_{\rm modes} \sim 10^8$ modes. Squeezed non-Gaussianity induces a distinctive scale-dependent halo bias:

$$\Delta b(k) \propto f_{\rm NL}\, k^{-(3/2-\nu)}, \tag{11.9.2}$$

  which scales non-trivially with wavenumber and provides a sensitive test for light massive fields ($m < \frac{3}{2}H$).
- **21-cm Tomography:** Neutral hydrogen tomography during the Cosmic Dawn and Epoch of Reionization ($z \sim 10 - 30$) represents the ultimate observational frontier, potentially accessing up to $N_{\rm modes} \sim 10^{12}$ independent Fourier modes. This unprecedented statistical power will enable cosmologists to detect heavily Boltzmann-suppressed signals ($m \sim 3{-}5H$) and map out the particle spectrum of the early universe.

11.9.4 Frontiers: The Cosmological Bootstrap and Quantum Gravity

The methods developed in this chapter form the bedrock of the modern *Cosmological Bootstrap* program. By bypassing explicit bulk time integrals and instead exploiting boundary conformal Ward identities, locality, and cosmological cutting rules (analyticity and unitarity of the wave-function of the universe), one can reconstruct cosmological correlators purely from boundary data and flat-space scattering amplitudes.

Furthermore, extending this framework to spinning gauge fields, the graviton (spin-2), and supersymmetric multiplets opens the possibility of testing string theory during inflation. Because string theory predicts infinite towers of weakly-coupled higher-spin particles ($s \geq 4$) along Regge trajectories $m_s^2 \sim s/\alpha'$, observing higher-spin oscillatory signatures ($P_{s\geq 4}(\cos\theta)$) in primordial non-Gaussianity would provide direct observational evidence for stringy physics at the Hubble scale of the infant universe.

A more speculative frontier concerns the soft ($k \to 0$) gravitons that appear at the opposite, spin-2-and-massless end of this chapter's mass spectrum: it has been proposed that these are the Goldstone modes of a spontaneously broken asymptotic symmetry of de Sitter space, in direct analogy with how soft gravitons in flat space are the Goldstone modes of spontaneously broken BMS supertranslations, with the single-field consistency relation of §10.6 playing the role of the de Sitter soft theorem.[6] This connects the squeezed-limit machinery built in this chapter to the broader program of asymptotic symmetries and soft theorems, but the specific claims about observer-dependent decoherence between symmetric and broken vacuum states on Page-time scales are recent, single-author, and not yet settled community wisdom, unlike the well-established results this chapter otherwise builds on.

[6] M. S. Sloth, "Soft gravitons as Goldstone modes of spontaneously broken asymptotic symmetries in de Sitter spacetimes," arXiv:2502.03520.

Part IV

Appendices

Appendix A

Numerical values

This appendix provides a compilation of recommended physical constants, astronomical parameters, conversion factors between the International System of Units (SI) and natural Planck systems, characteristic scales of subatomic particles, present-day cosmological parameters, and the spectral energy distribution of the cosmic electromagnetic background radiation.

A.1 Physical constants

The fundamental physical constants listed in Table A.1 are based on the internationally recommended CODATA values [1,2].

Table A.1: Fundamental physical constants in SI and practical energy units (CODATA recommended values [1]). Parentheses indicate standard uncertainties in the last digits.

Constant	Symbol / Definition	Numerical value
Speed of light in vacuum	c	$2.997\,924\,58 \times 10^{8}$ m $\cdot$ s^{-1} (exact)
Newtonian constant of gravitation	G_{N}	$6.674\,28(67) \times 10^{-11}$ m^{3} $\cdot$ kg^{-1} $\cdot$ s^{-2}
Planck constant	h	$6.626\,068\,96(33) \times 10^{-34}$ J $\cdot$ s
Reduced Planck constant	$\hbar \equiv h/(2\pi)$	$1.054\,571\,628(53) \times 10^{-34}$ J $\cdot$ s
		$= 6.582\,118\,99(16) \times 10^{-22}$ MeV $\cdot$ s
Boltzmann constant	k_{B}	$1.380\,650\,4(24) \times 10^{-23}$ J $\cdot$ K^{-1}
Fermi coupling constant	$G_{\mathrm{F}}/(\hbar c)^3$	$1.166\,37(1) \times 10^{-5}$ GeV^{-2}
Magnetic permeability of vacuum	μ_0	$4\pi \times 10^{-7}$ N $\cdot$ A^{-2} (exact)
Electric permittivity of vacuum	$\varepsilon_0 \equiv 1/(\mu_0 c^2)$	$8.854\,187\,817 \times 10^{-12}$ F $\cdot$ m^{-1}
Fine-structure constant	$\alpha \equiv \frac{e^2}{4\pi\varepsilon_0\hbar c}$	$1/137.035\,999\,679(94)$
Rydberg constant	$R_\infty \equiv m_{\mathrm{e}} c^2 \alpha^2/2$	13.605 691 93(34) eV
Bohr radius	$a_0 \equiv \hbar/(m_{\mathrm{e}} c\alpha)$	$5.291\,772\,0859(36) \times 10^{-11}$ m
Thomson scattering cross-section	$\sigma_{\mathrm{T}} \equiv \frac{e^4}{6\pi\varepsilon_0^2 m_{\mathrm{e}}^2 c^4}$	$6.652\,458\,73(13) \times 10^{-29}$ m^{2}

A.2 Astrophysical quantities

The characteristic solar parameters compiled in Table A.2 serve as primary benchmarks across stellar and galactic astrophysics [3].

A.3 Units

Table A.2: Astrophysical reference quantities for the Sun [3].

Quantity	Symbol / Definition	Value
Solar mass	$M_\odot$	$1.988\,44(30) \times 10^{30}$ kg $= 1.189 \times 10^{57} M_\mathrm{P}$
Solar equatorial radius	$R_\odot$	6.961×10^{8} m
Solar Schwarzschild radius	$R_{\mathrm{s}\odot} \equiv 2G_\mathrm{N}M_\odot/c^2$	2.953 250 08 km
Solar luminosity	$L_\odot$	$3.846(8) \times 10^{26}$ W

A.3.1 Natural units

In the International System of Units (SI), physical equations are frequently encumbered by dimensional conversion factors such as the speed of light c, the reduced Planck constant $\hbar$, the Boltzmann constant k_B, and the gravitational constant G_N [1,2].

To simplify analytical calculations, relativistic and high-energy physics adopts a natural unit system wherein fundamental velocity and action scales are set equal to unity:

$$\text{`}\hbar = c = 1\text{'} \iff [L] = [T] = [M^{-1}] = [E^{-1}]. \tag{A.3.1}$$

Setting $c = 1$ identifies length and time dimensions ($[L] = [T]$), expressing durations in meters or distances in seconds. Setting $\hbar = 1$ equates action to unity, causing energy and momentum to possess inverse length dimensions ($[E] = [P] = [L^{-1}] = [T^{-1}]$), while mass is identified with energy ($[M] = [E]$).

Selecting the mega-electronvolt (MeV $= 10^6$ eV) or giga-electronvolt (GeV $= 10^9$ eV) as the base unit of energy, the energy conversion factor is 1 GeV $\simeq 1.6022 \times 10^{-10}$ J. Practical SI dimensions are readily restored by multiplying by powers of the conversion constant:

$$\hbar c \simeq 197.326\,968 \text{ MeV} \cdot \text{fm} \approx 200 \text{ MeV} \cdot \text{fm}.$$

To fully nondimensionalize the system of units, a third dimensionful scale must be fixed. Setting Newton's gravitational constant to unity ($G_\mathrm{N} = 1$) defines *Planck units*. In this system, length, time, and mass are expressed in terms of the fundamental Planck scales:

$$\ell_\mathrm{P} \equiv \sqrt{\frac{\hbar G_\mathrm{N}}{c^3}}, \quad t_\mathrm{P} \equiv \sqrt{\frac{\hbar G_\mathrm{N}}{c^5}}, \quad M_\mathrm{P} \equiv \sqrt{\frac{\hbar c}{G_\mathrm{N}}}.$$

The thermal energy scale is linked to temperature through Boltzmann's constant k_B, which acts solely as a conversion factor between thermal degrees of freedom and energy. Defining the Planck temperature:

$$T_\mathrm{P} \equiv \sqrt{\frac{\hbar c^5}{G_\mathrm{N} k_\mathrm{B}^2}},$$

one can set $k_\mathrm{B} = 1$ without physical consequence, identifying $[T_\mathrm{therm}] = [E]$. In these complete Planck units ($\hbar = c = G_\mathrm{N} = k_\mathrm{B} = 1$), all dimensionful constants drop out of the fundamental dynamical equations.

The dimensions of the fundamental constants in terms of length $[L]$, time $[T]$, and mass $[M]$ are:

$$[c] = LT^{-1}, \quad [\hbar] = ML^2T^{-1}, \quad [G_\mathrm{N}] = M^{-1}L^3T^{-2}. \tag{1}$$

Setting $c = 1$ implies $[L] = [T]$. Enforcing $\hbar = 1$ yields $[M][L]^2[T]^{-1} = [M][L] = 1$, which requires $[M] = [L]^{-1}$. Since energy has dimensions $[E] = [ML^2T^{-2}] = [M]$, it follows

directly that:

$$[L] = [T] = [M^{-1}] = [E^{-1}], \tag{2}$$

establishing Eq. (A.3.1).
To construct a length scale $\ell_{\rm P} \propto \hbar^{\alpha} G_{\rm N}^{\beta} c^{\gamma}$, dimensional matching requires:

$$L = (ML^2T^{-1})^{\alpha}(M^{-1}L^3T^{-2})^{\beta}(LT^{-1})^{\gamma} = M^{\alpha-\beta}L^{2\alpha+3\beta+\gamma}T^{-\alpha-2\beta-\gamma}. \tag{3}$$

Setting the exponents: $\alpha - \beta = 0 \implies \alpha = \beta$; $-\alpha - 2\beta - \gamma = -3\alpha - \gamma = 0 \implies \gamma = -3\alpha$; and $2\alpha + 3\beta + \gamma = 5\alpha - 3\alpha = 2\alpha = 1 \implies \alpha = 1/2$. Thus:

$$\ell_{\rm P} = \sqrt{\frac{\hbar G_{\rm N}}{c^3}}, \quad t_{\rm P} = \frac{\ell_{\rm P}}{c} = \sqrt{\frac{\hbar G_{\rm N}}{c^5}}, \quad M_{\rm P} = \frac{\hbar}{\ell_{\rm P} c} = \sqrt{\frac{\hbar c}{G_{\rm N}}}, \tag{4}$$

confirming the exact algebraic expressions for the Planck hierarchy.

A.3.2 Conversion factors

Table A.3 summarizes the numerical values of the Planck scales and practical conversion equivalences between natural energy units and SI units.

Table A.3: Planck units and SI conversion equivalents [1,2].

Quantity	Analytic form	Value
Planck mass	$M_{\rm P} = \sqrt{\hbar c/G_{\rm N}}$	$1.220\,892(61) \times 10^{19}$ GeV/c^2
		$= 2.176\,44(11) \times 10^{-8}$ kg
Planck time	$t_{\rm P} = \sqrt{\hbar G_{\rm N}/c^5}$	$5.391\,24(27) \times 10^{-44}$ s
Planck length	$\ell_{\rm P} = \sqrt{\hbar G_{\rm N}/c^3}$	$1.616\,252(81) \times 10^{-35}$ m
Planck temperature	$T_{\rm P} = \sqrt{\hbar c^5/(G_{\rm N} k_{\rm B}^2)}$	$1.416\,785(71) \times 10^{32}$ K
Energy Equivalents in SI and Natural Units:		
1 GeV		$= 1.6022 \times 10^{-10}$ J
		$= 1.1605 \times 10^{13}$ K
		$= 1.7827 \times 10^{-27}$ kg
1 GeV^{-1}		$= 1.9733 \times 10^{-16}$ m
		$= 6.6522 \times 10^{-25}$ s
1 cm		$= 5.068 \times 10^{13}$ GeV$^{-1}\hbar$
1 s		$= 1.519 \times 10^{24}$ GeV$^{-1}\hbar/c$
1 g		$= 5.608 \times 10^{23}$ GeV$\hbar/c^2$
1 erg		$= 6.242 \times 10^{2}$ GeV
1 J		$= 6.242 \times 10^{9}$ GeV
1 K		$= 8.618 \times 10^{-14}$ GeV/$k_{\rm B}$
Astronomical and Laboratory Units:		
Astronomical unit (AU)		$1.495\,978\,706\,60(20) \times 10^{11}$ m
Parsec (pc)		$3.085\,677\,5807(4) \times 10^{16}$ m $= 3.2616$ ly
Sidereal year (yr)		$3.155\,815 \times 10^{7}$ s
Erg		10^{-7} J
Gauss		10^{-4} T

A.4 Particle physics

Fundamental properties of standard subatomic particles and nuclear mass differences are detailed in Table A.4 [2].

Table A.4: Subatomic particle parameters and nuclear properties [2].

Quantity	Symbol	Value
Elementary charge	e	$1.602\,176\,487(40) \times 10^{-19}$ C
Electron mass	m_e	$0.510\,998\,910(13)$ MeV$/c^2$ $= 9.109\,389 \times 10^{-31}$ kg
Proton mass	m_p	$938.271\,013(23)$ MeV$/c^2$ $= 1.672\,622 \times 10^{-27}$ kg
Proton-to-electron mass ratio	$\mu \equiv m_\mathrm{p}/m_\mathrm{e}$	$1836.152\,672\,47(80)$
Proton–neutron mass difference	$Q_\mathrm{np} \equiv m_\mathrm{n} - m_\mathrm{p}$	$1.293\,318(9)$ MeV$/c^2$
Neutron-to-proton mass ratio	$m_\mathrm{n}/m_\mathrm{p}$	$1.001\,378\,419\,18(46)$
Deuteron mass	m_d	$1875.612\,793(47)$ MeV$/c^2$
Free neutron lifetime	τ_n	$885.7(0.8)$ s

A.5 Cosmological quantities

Table A.5 gathers the standard cosmological parameters governing the background FLRW dynamics, where the dimensionless Hubble parameter is $h \equiv H_0/(100\ \mathrm{km \cdot s^{-1} \cdot Mpc^{-1}}) \approx 0.72$ and $\Theta_{2.7} \equiv T_{\gamma 0}/(2.725\ \mathrm{K})$.

Table A.5: Cosmological parameters for the standard ΛCDM model.

Quantity	Symbol / Relation	Value
Hubble constant	H_0	$100h$ km $\cdot$ s^{-1} $\cdot$ Mpc^{-1}
Dimensionless Hubble parameter	h	0.72 ± 0.05
Hubble time	$t_{H0} \equiv H_0^{-1}$	$9.7776h^{-1} \times 10^9$ yr $= 3.09h^{-1} \times 10^{17}$ s
Hubble distance	$D_{H0} \equiv cH_0^{-1}$	$2997.9h^{-1}$ Mpc
Critical density today	$\rho_\mathrm{crit} \equiv \frac{3H_0^2}{8\pi G_\mathrm{N}}$	$1.878\,37(28) \times 10^{-29}h^2$ g $\cdot$ cm^{-3} $= 8.0992 \times 10^{-47}h^2$ GeV^4 $= 1.0540 \times 10^4 h^2$ eV $\cdot$ cm^{-3}
CMB temperature	$T_{\gamma 0}$	$2.725\Theta_{2.7}$ K $= 2.348 \times 10^{-4}\Theta_{2.7}$ eV
CMB photon number density	$n_{\gamma 0}$	$410.44\Theta_{2.7}^3$ cm^{-3}
CMB photon energy density	$\rho_{\gamma 0}$	$4.6408 \times 10^{-34}\Theta_{2.7}^4$ g $\cdot$ cm^{-3} $= 0.2604\Theta_{2.7}^4$ eV $\cdot$ cm^{-3}
CMB density parameter	$\Omega_{\gamma 0}h^2$	$2.4697 \times 10^{-5}\Theta_{2.7}^4$
Relic neutrino temperature	$T_{\nu 0} = (4/11)^{1/3}T_{\gamma 0}$	$1.9272\Theta_{2.7}$ K
Neutrino number density (per flavor)	$n_{\nu 0} = (3/11)n_{\gamma 0}$	$108.89\Theta_{2.7}^3$ cm^{-3}
Neutrino energy density (per flavor)	$\rho_{\nu 0}$	$0.2271\rho_{\gamma 0}$
Comoving entropy density	$s_0 = 7.0394n_{\gamma 0}$	$2889.2\Theta_{2.7}^3$ cm^{-3}
Total radiation energy density	$\rho_{\mathrm{r}0}$	$7.8042 \times 10^{-34}\Theta_{2.7}^4$ g $\cdot$ cm^{-3}
Total radiation density parameter	$\Omega_{\mathrm{r}0}h^2$	$4.1534 \times 10^{-5}\Theta_{2.7}^4$
Redshift at equality	$1 + z_\mathrm{eq}$	$2.4 \times 10^4 \Omega_{\mathrm{m}0}h^2$
Temperature at equality	$T_\mathrm{eq} = T_{\gamma 0}(1 + z_\mathrm{eq})$	$6.55 \times 10^4 \Omega_{\mathrm{m}0}h^2$ K
Redshift at decoupling	$1 + z_\mathrm{dec}$	1100
CMB temperature at decoupling	T_dec	$0.255\Theta_{2.7}$ eV

A.6 Electromagnetic spectrum

The cosmic microwave background represents the overwhelmingly dominant radiative energy component in the Universe today. The diffuse extragalactic background radiation comprises the spatial volume average of all electromagnetic radiation throughout the Universe, defined as the

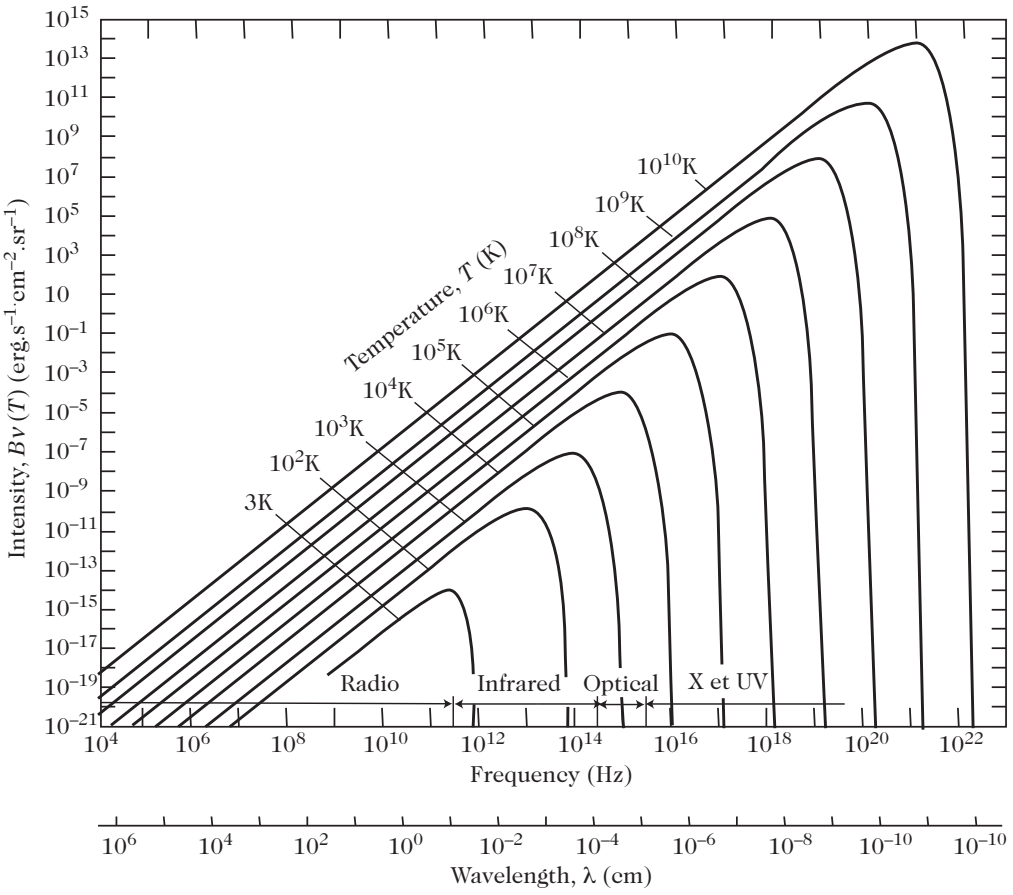


Figure A.1: The spectral radiance $B_\nu(T)$ of blackbody radiation as a function of frequency ν (bottom axis) and wavelength λ (top axis) across a wide range of temperatures T, from the CMB ($T \approx 3$ K) up to 10^{10} K. (Adapted from Ref. [3]).

isotropic background remaining after subtracting the localized anisotropic emissions of the Solar System (zodiacal light) and the Milky Way (interstellar dust and synchrotron emission).

The spectral energy distribution of this cosmic radiation has been observed across more than twenty decades of frequency, spanning from the long-wavelength radio band to ultra-high-energy γ-rays. The cosmic microwave background accounts for approximately 93% of the total electromagnetic energy density in the Universe (as discussed in Section 4.4.2).

The extragalactic radio background arises from the cumulative emission of discrete extragalactic radio galaxies, active galactic nuclei (AGN), and starburst systems. While historically the first diffuse component discovered, it contributes negligibly to the overall cosmic energy budget.

In contrast, the cosmic X-ray background (XRB), originally detected by Giacconi and collaborators, contains substantial energy. Its spectral energy distribution peaks near 30 keV and is produced by the integrated emission of obscured and unobscured active galactic nuclei throughout cosmic history. The diffuse γ-ray background has been mapped between 0.1 MeV and 1 GeV; it is dominated by blazars (radio-loud AGNs whose relativistic jets point toward Earth), star-forming galaxies, and annihilation or decay of high-energy cosmic rays.

In the optical and near-infrared domains, radiation originates from the integrated starlight of all galaxies throughout cosmic time. A substantial portion of the ultraviolet and optical radiation emitted by stars in dense regions is absorbed by interstellar dust grains and re-radiated as thermal emission in the far-infrared (the Cosmic Infrared Background, CIB). Hard X-rays,

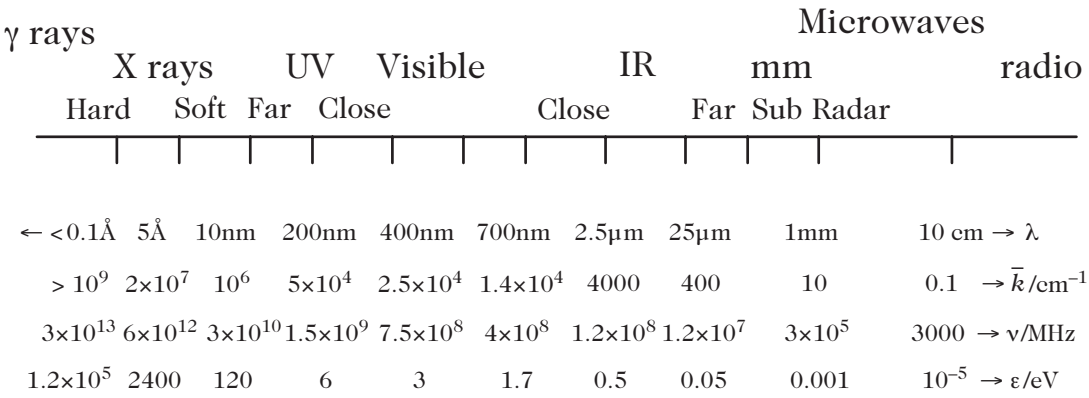


Figure A.2: The full electromagnetic radiation spectrum and its conventional designations across wavelength (λ), wavenumber ($\bar{k}$), frequency (ν), and photon energy (ε).

however, penetrate interstellar dust unattenuated.

Table A.6: Relative contributions of different wavelength regimes to the total extragalactic electromagnetic background intensity.

Frequency band	Intensity ($\mathrm{W \cdot m^{-2} \cdot sr^{-1}}$)	Fractional contribution (%)
Radio	1.2×10^{-12}	1.1×10^{-4}
Microwave (CMB)	9.96×10^{-6}	93
Infrared	$(4.6 \pm 0.6) \times 10^{-8}$	4.5 ± 0.5
Optical	$(3 \pm 1) \times 10^{-8}$	3 ± 1
X-ray	2.7×10^{-10}	3.5×10^{-2}
γ-ray	3×10^{-11}	2.5×10^{-3}

Table A.6 and Fig. A.3 quantify the comparative energetic inventory across all spectral windows. The cosmic microwave background blackbody distribution in the millimeter-to-centimeter regime dominates all other components combined by more than an order of magnitude.

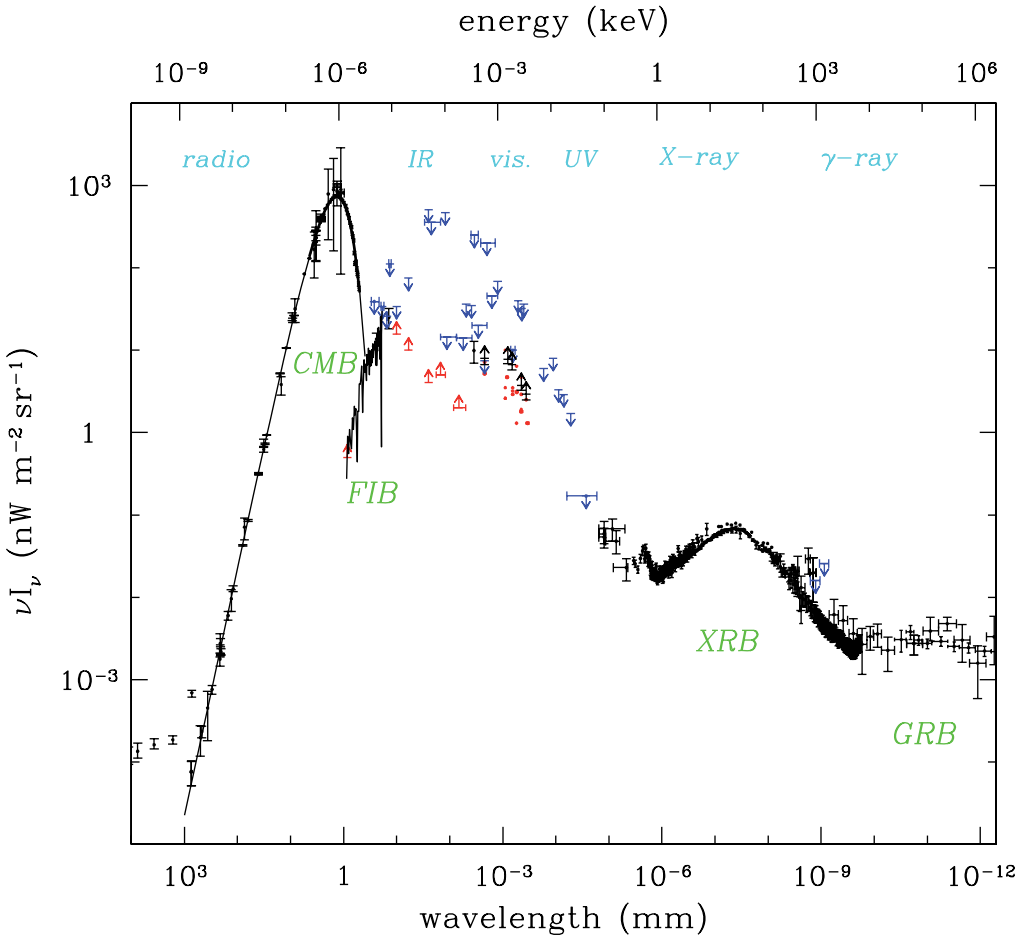


Figure A.3: Broad-band spectral energy distribution (νI_ν) of the cosmic background radiation compiled from observations across all electromagnetic bands. The energy density contained within each logarithmic frequency interval is proportional to νI_ν, showing the pronounced dominance of the cosmic microwave background (CMB). Also visible are the far-infrared background (FIB), the X-ray background (XRB), and the γ-ray background (GRB). (Adapted from Ref. [4]).

Appendix B

Special Functions

This appendix gathers the definitions and general properties of the special functions appearing in the core of the book. More details can be found in dedicated monographs, such as those in Refs. [1–3].

B.1 Euler Functions

The Euler function of the second kind, also frequently called the Euler Γ function, is defined as

$$\Gamma(z) = \int_0^\infty \mathrm{e}^{-t} t^{z-1}\,\mathrm{d}t\,, \tag{B.1.1}$$

for $\mathrm{Re}(z) > 0$. The Γ function is analytic with simple poles in $z = -p$, with $p \in \mathbb{N}$. For $n \in \mathbb{N}$, one recovers the well-known factorials,

$$\Gamma(n) = (n-1)!, \tag{B.1.2}$$

while for half-integer arguments one has

$$\Gamma\left(n + \frac{1}{2}\right) = \frac{\sqrt{\pi}}{2^n}(2n-1)!!, \tag{B.1.3}$$

and

$$\Gamma\left(-n + \frac{1}{2}\right) = (-1)^n \frac{2^n\sqrt{\pi}}{(2n-1)!!}. \tag{B.1.4}$$

For large z, Γ behaves asymptotically as

$$\ln\Gamma(z) = \left(z - \frac{1}{2}\right)\ln z - z + \frac{1}{2}\ln(2\pi) + \mathcal{O}(z^{-1}), \tag{B.1.5}$$

which is simply the Weierstrass expansion formula (also known as Stirling's series for the logarithm of the gamma function).

The Euler function of the first kind, also called the Beta function, is defined by

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t\,, \tag{B.1.6}$$

and is related to the Γ function through

$$B(x,y) = B(y,x) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \tag{B.1.7}$$

To establish the fundamental relation (B.1.7), consider the product of two Gamma functions $\Gamma(x)$ and $\Gamma(y)$ with $\mathrm{Re}(x) > 0$ and $\mathrm{Re}(y) > 0$. Using the substitution $t = u^2$ in the definition (B.1.1), we have

$$\Gamma(x) = 2\int_0^\infty \mathrm{e}^{-u^2} u^{2x-1}\,\mathrm{d}u\,, \qquad \Gamma(y) = 2\int_0^\infty \mathrm{e}^{-v^2} v^{2y-1}\,\mathrm{d}v\,. \tag{1}$$

Their product can be written as a double integral over the first quadrant of the (u, v) plane:

$$\Gamma(x)\Gamma(y) = 4\int_0^\infty \mathrm{d}u \int_0^\infty \mathrm{d}v\ \mathrm{e}^{-(u^2+v^2)} u^{2x-1} v^{2y-1}. \tag{2}$$

Transforming to polar coordinates $u = r\cos\theta$, $v = r\sin\theta$ with Jacobian $\mathrm{d}u\,\mathrm{d}v = r\,\mathrm{d}r\,\mathrm{d}\theta$, where $r \in [0, \infty)$ and $\theta \in [0, \pi/2]$, the integral factors into radial and angular parts:

$$\Gamma(x)\Gamma(y) = 4\left(\int_0^\infty \mathrm{e}^{-r^2} r^{2(x+y)-1}\,\mathrm{d}r\right)\left(\int_0^{\pi/2} (\cos\theta)^{2x-1}(\sin\theta)^{2y-1}\,\mathrm{d}\theta\right). \tag{3}$$

The radial integral is recognized as $\frac{1}{2}\Gamma(x+y)$ by substituting $w = r^2$. For the angular integral, make the change of variable $t = \cos^2\theta$, so that $\mathrm{d}t = -2\cos\theta\sin\theta\,\mathrm{d}\theta$, $\sin^2\theta = 1-t$, and the limits map to $t = 1$ (for $\theta = 0$) and $t = 0$ (for $\theta = \pi/2$):

$$2\int_0^{\pi/2} (\cos^2\theta)^{x-1/2}(\sin^2\theta)^{y-1/2}\,\mathrm{d}\theta = \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t = B(x, y). \tag{4}$$

Combining these two results yields

$$\Gamma(x)\Gamma(y) = \Gamma(x+y)B(x, y) \implies B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{5}$$

which proves (B.1.7).
Setting $x = y = 1/2$ in this relation gives

$$B\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{\Gamma(1/2)^2}{\Gamma(1)} = \Gamma\left(\frac{1}{2}\right)^2. \tag{6}$$

On the other hand, from the integral definition:

$$B\left(\frac{1}{2}, \frac{1}{2}\right) = \int_0^1 \frac{\mathrm{d}t}{\sqrt{t(1-t)}} = \Big[2\arcsin\sqrt{t}\Big]_0^1 = \pi. \tag{7}$$

Since $\Gamma(1/2) > 0$, we conclude that $\Gamma(1/2) = \sqrt{\pi}$. Integration by parts on (B.1.1) yields the recurrence relation

$$\Gamma(z+1) = z\Gamma(z). \tag{8}$$

Repeated application of this identity for $z = n - 1/2$ gives

$$\Gamma\left(n+\frac{1}{2}\right) = \left(n-\frac{1}{2}\right)\left(n-\frac{3}{2}\right)\cdots\frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{2n-1}{2}\frac{2n-3}{2}\cdots\frac{1}{2}\sqrt{\pi} = \frac{(2n-1)!!}{2^n}\sqrt{\pi}, \tag{9}$$

which is Eq. (B.1.3). Finally, Euler's reflection formula $\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin(\pi z)}$ evaluated at $z = n + 1/2$ yields

$$\Gamma\left(-n+\frac{1}{2}\right) = \frac{\pi}{\sin\left[\pi\left(n+\frac{1}{2}\right)\right]\Gamma\left(n+\frac{1}{2}\right)} = \frac{(-1)^n\pi}{\frac{(2n-1)!!}{2^n}\sqrt{\pi}} = (-1)^n\frac{2^n\sqrt{\pi}}{(2n-1)!!}, \tag{10}$$

proving Eq. (B.1.4).

B.2 Spherical Harmonics

B.2.1 Definition

The spherical harmonics Y_ℓ^m are functions defined on the 2-sphere S^2. They are the eigenfunctions of the 2-dimensional spherical Laplacian. In other words, they satisfy

$$\left[\frac{\partial^2}{\partial\theta^2} + \cot\theta\frac{\partial}{\partial\theta} + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right] Y_\ell^m = -\ell(\ell+1)Y_\ell^m, \tag{B.2.1}$$

where θ and ϕ are the angular coordinates on the surface of the sphere ($\theta \in [0,\pi]$, $\phi \in [0,2\pi)$). The Y_ℓ^m form a complete orthonormal set of functions, so they can be used as a basis for functions on the sphere.

B.2.2 Expressions

Spherical harmonics are related to associated Legendre polynomials P_ℓ^m through—(5.2.1) in Ref. [1]—

$$Y_\ell^m(\theta,\phi) = \sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-m)!}{(\ell+m)!}}P_\ell^m(\cos\theta)\mathrm{e}^{im\phi}. \tag{B.2.2}$$

They can also be expressed in terms of the Gegenbauer polynomials C_n^α according to—(5.2.6.39c) of Ref. [1]—

$$Y_\ell^m(\theta,\phi) = \zeta_m \mathrm{e}^{im\phi}\sqrt{\frac{2\ell+1}{4\pi}}\sqrt{\frac{(\ell-|m|)!}{(\ell+|m|)!}}(2|m|-1)!!\,(\sin\theta)^{|m|}C_{\ell-|m|}^{|m|+1/2}(\cos\theta), \tag{B.2.3}$$

ζ_m being defined as

$$\zeta_m = \begin{cases}(-1)^m & m > 0,\\ 1 & m \le 0.\end{cases} \tag{B.2.4}$$

The Gegenbauer polynomials C_n^α are solutions of the second-order differential equation

$$(1-x^2)y'' - (2\alpha+1)xy' + n(n+2\alpha)y = 0, \tag{B.2.5}$$

and satisfy the normalization relation—(7.313) of Ref. [2]—

$$\int_{-1}^{1}(1-x^2)^{\alpha-1/2}\left[C_n^\alpha(x)\right]^2\mathrm{d}x = \frac{\pi 2^{1-2\alpha}\Gamma(2\alpha+n)}{n!(n+\alpha)[\Gamma(\alpha)]^2}, \tag{B.2.6}$$

if $\mathrm{Re}(\alpha) > -1/2$.

B.2.3 Properties

The complex conjugate of any given spherical harmonic is given by—(5.4.1) of Ref. [1]—

$$Y_\ell^{m*}(\theta,\phi) = (-1)^m Y_\ell^{-m}(\theta,\phi) = Y_\ell^m(\theta,-\phi). \tag{B.2.7}$$

Spherical harmonics are normalized according to—(5.6.1) of Ref. [1]—

$$\int_0^{2\pi}\mathrm{d}\phi\int_0^\pi \sin\theta\,\mathrm{d}\theta\, Y_{\ell m}^*(\theta,\phi)Y_{\ell' m'}(\theta,\phi) = \delta_{\ell\ell'}\delta_{mm'}, \tag{B.2.8}$$

leading to the closure relation—(5.2.2) of Ref. [1]—

$$\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} Y_{\ell m}(\theta,\phi)Y_{\ell m}^*(\theta',\phi') = \delta(\cos\theta-\cos\theta')\delta(\phi-\phi'), \tag{B.2.9}$$

as well as to the addition property—(5.17.2.9) of Ref. [1]—

$$\sum_{m=-\ell}^{\ell} Y_{\ell m}(\theta,\phi)Y^*_{\ell m}(\theta',\phi') = \frac{2\ell+1}{4\pi}P_\ell(\cos\alpha), \tag{B.2.10}$$

where α is the angle between the two directions (θ,ϕ) and (θ',ϕ'), and P_ℓ is the Legendre polynomial of order ℓ.

Any real function on the sphere is thus expanded on the spherical harmonics basis according to

$$f(\theta,\phi) = \sum_{\ell,m} f_{\ell m}Y_{\ell m}(\theta,\phi), \qquad f_{\ell m} = (-1)^m f^*_{\ell,-m}. \tag{B.2.11}$$

B.2.4 Physical Interpretation: Standing Waves on the Sphere

In cosmological analysis, most notably when expanding the temperature fluctuations $\Delta T/T(\hat{\boldsymbol{n}})$ and polarization of the Cosmic Microwave Background (CMB) across the sky, the spherical harmonics $Y_\ell^m(\theta,\phi)$ play the precise role on the compact 2-sphere S^2 that Fourier plane waves $\mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{x}}$ play on flat Euclidean space $\mathbb{R}^n$.

Multipole Order ℓ and Spatial Wavelengths. On a 1D circle S^1 of circumference L, the eigenfunctions of $-\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ are sinusoidal modes e^{ikx}, where the wavenumber $k = 2\pi n/L$ requires that an integer number n of full wavelengths fit around the closed circumference ($\lambda = L/n$). On the two-dimensional unit sphere S^2 (with great-circle circumference 2π), the multipole index $\ell \in \{0,1,2,\dots\}$ plays the role of the total angular wavenumber:

- **Angular scale:** The multipole ℓ governs the overall spatial oscillation frequency across the sphere, corresponding to an angular wavelength

$$\theta \sim \frac{2\pi}{\ell} \qquad \left(\text{characteristic lobe size } \theta_{\text{lobe}} \sim \frac{180^\circ}{\ell}\right). \tag{B.2.12}$$

  On a spherical shell of comoving radius χ_* (such as the last scattering surface), this translates into a physical comoving wavenumber $k \approx \ell/\chi_*$.

- **Nodal count:** The multipole order ℓ equals the total number of nodal lines (curves on the sphere where $Y_\ell^m = 0$). Traversing any great circle on the sphere intersects ℓ full wavelengths (2ℓ nodal crossings).

Azimuthal Index m and Nodal Partitioning. On a 1D circle, a wave propagates along a single dimension, giving a 2-fold degeneracy ($\mathrm{e}^{\pm ikx}$). On the 2-sphere (θ,ϕ), the SO(3) rotational invariance of the spherical Laplacian Δ_{S^2} produces a $(2\ell+1)$-fold degeneracy: rotating any standing wave pattern of order ℓ in 3D space yields another valid eigenmode with the identical eigenvalue $-\ell(\ell+1)$. The $(2\ell+1)$ harmonics $\{Y_\ell^m\}_{m=-\ell}^{\ell}$ form the canonical $(2\ell+1)$-dimensional irreducible representation of SO(3).

Within this $(2\ell+1)$-dimensional eigenspace, the azimuthal index $m \in \{-\ell,-\ell+1,\dots,\ell\}$ determines how the ℓ total nodal lines are partitioned between *longitudinal meridians* (passing through the poles, oscillating as $\cos m\phi$ or $\sin m\phi$) and *latitudinal circles* (parallels of constant θ, given by zeros of $P_\ell^m(\cos\theta)$), as summarized in Table B.1 and illustrated in Fig. B.1.

Rotational Transformation vs. Orientation

It is essential to distinguish the geometric roles of ℓ, m, and the expansion coefficients $a_{\ell m}$:

Table B.1: Classification of spherical harmonics $Y_\ell^m(\theta,\phi)$ by azimuthal index m.

Type	**Azimuthal Index**	**Nodal Lines (Lat./Long.)**	**Geometry on S^2**
Zonal	$m = 0$	ℓ / 0	Azimuthally symmetric bands between poles
Tesseral	$0 < \lvert m\rvert < \ell$	$\ell - \lvert m\rvert$ / $\lvert m\rvert$	Checkerboard grid of $(\ell - \lvert m\rvert + 1) \times 2\lvert m\rvert$ lobes
Sectorial	$\lvert m\rvert = \ell$	0 / ℓ	2ℓ longitudinal wedges around the equator

- ℓ fixes the *general geometrical structure* (monopole $\ell = 0$, dipole $\ell = 1$, quadrupole $\ell = 2$, etc.).
- m fixes the *transformation behaviour under rotation about a chosen axis*, conventionally the z-axis. Specifically, under an azimuthal rotation $\phi \to \phi + \alpha$ about the z-axis, each harmonic transforms as an eigenfunction of $L_z = -i\partial_\phi$:

$$Y_{\ell m}(\theta, \phi + \alpha) = \mathrm{e}^{im\alpha} Y_{\ell m}(\theta, \phi). \tag{1}$$

Together, (ℓ, m) specify a particular basis harmonic. Physically, m labels how the pattern transforms under rotation about the z-axis rather than an orientation angle, and this transformation property constrains its alignment relative to that axis:

- For $m = 0$, the harmonic is invariant under rotations about the z-axis ($\mathrm{e}^{i\cdot 0\cdot\alpha} = 1$). For a dipole ($\ell = 1$), this invariance forces the dipole axis to lie along the z-axis ($Y_{1,0} \propto \cos\theta$).
- For $m = \pm 1$, the modes pick up phases $\mathrm{e}^{\pm i\alpha}$ under z-rotations and combine into dipole patterns in the equatorial (xy) plane.

To orient a multipole ℓ arbitrarily in three dimensions, one superposes the different m modes at fixed ℓ:

$$T_\ell(\theta, \phi) = \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\theta, \phi). \tag{2}$$

The $(2\ell + 1)$ coefficients $a_{\ell m}$ collectively encode the actual orientation, overall amplitude, and, for higher multipoles ($\ell \geq 2$), the detailed shape of the angular pattern.

B.2.5 Fourier Transform

The Fourier transform of spherical harmonics is given by—(5.9.2.6) of Ref. [1]—

$$\int_0^\infty r^2\,\mathrm{d}r \int_0^{2\pi} \mathrm{d}\phi \int_0^\pi \sin\theta\,\mathrm{d}\theta\, \frac{\mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{r}}}{(2\pi)^{3/2}} j_\ell(k'r) Y_{\ell m}(\theta,\phi) = \sqrt{\frac{2}{\pi}}\, i^\ell \frac{\delta(k'-k)}{k^2} Y_{\ell m}(\theta_k, \phi_k), \tag{B.2.13}$$

where j_ℓ is a spherical Bessel function (see below). From (B.2.13), one deduces the expansion of the exponential as—(5.17.3.14) of Ref. [1]—

$$\mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{r}} = \sum_{\ell=0}^{\infty} (2\ell + 1) i^\ell j_\ell(kr) P_\ell(\cos\theta_{\boldsymbol{k},\boldsymbol{r}}), \tag{B.2.14}$$

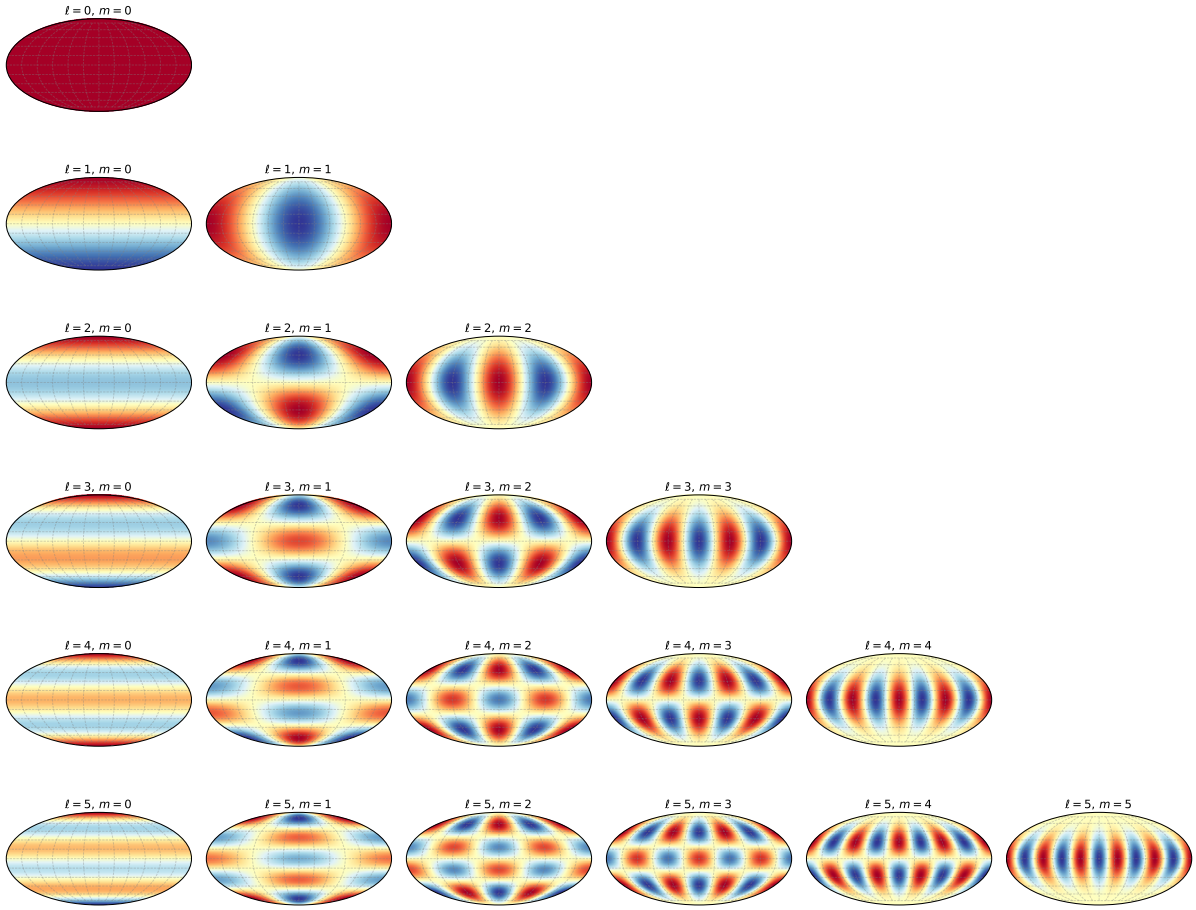


Figure B.1: Mollweide equal-area projections of real spherical harmonics $Y_\ell^m(\theta,\phi)$ for multipoles $\ell = 0, 1, 2, 3, 4, 5$ and azimuthal orders $m = 0, \ldots, \ell$. Amplitude is color-coded: warm red indicates positive regions, cool blue indicates negative regions, and pale yellow denotes the nodal lines ($Y_\ell^m = 0$). The multipole ℓ determines the total spatial frequency (ℓ wavelengths fitting around the sphere), while the azimuthal index m determines how those nodal lines are partitioned between latitudinal parallels ($\ell - m$) and longitudinal meridians (m), producing zonal ($m = 0$), tesseral ($0 < m < \ell$), and sectorial ($m = \ell$) lobe geometries.

or, in other words,

$$e^{i\boldsymbol{p}\cdot\boldsymbol{x}} = 4\pi \sum_{\ell,m} i^\ell j_\ell(pr) Y^*_{\ell m}(\theta_p,\phi_p) Y_{\ell m}(\theta_x,\phi_x), \tag{B.2.15}$$

where (θ_p,ϕ_p) and (θ_x,ϕ_x) represent the coordinates of the unit vectors parallel to $\boldsymbol{p}$ and $\boldsymbol{x}$, respectively, and—(5.17.4.18) of Ref. [1]—

$$\delta^{(3)}(\boldsymbol{r}_1 - \boldsymbol{r}_2) = \frac{\delta(r_1 - r_2)}{r_1^2} \sum_{\ell=0}^{\infty} \frac{2\ell+1}{4\pi} P_\ell(\cos\theta_{12}). \tag{B.2.16}$$

The addition theorem (B.2.10) follows from the rotation invariance of functions on the 2-sphere S^2. Let $\hat{\boldsymbol{n}}_1 = (\theta_1,\phi_1)$ and $\hat{\boldsymbol{n}}_2 = (\theta_2,\phi_2)$ be two unit vectors separated by angle α, so that $\cos\alpha = \hat{\boldsymbol{n}}_1 \cdot \hat{\boldsymbol{n}}_2$. Consider the kernel

$$K(\hat{\boldsymbol{n}}_1,\hat{\boldsymbol{n}}_2) \equiv \sum_{m=-\ell}^{\ell} Y_{\ell m}(\hat{\boldsymbol{n}}_1) Y^*_{\ell m}(\hat{\boldsymbol{n}}_2). \tag{1}$$

Under any three-dimensional rotation $\mathcal{R} \in \mathrm{SO}(3)$, the spherical harmonics transform under the irreducible representation D^ℓ:

$$Y_{\ell m}(\mathcal{R}\hat{\boldsymbol{n}}) = \sum_{m'=-\ell}^{\ell} D^{\ell}_{m'm}(\mathcal{R}) Y_{\ell m'}(\hat{\boldsymbol{n}}). \tag{2}$$

Since the rotation matrices D^ℓ are unitary ($D^\dagger D = \mathbb{I}$), the bilinear sum satisfies

$$\begin{aligned} K(\mathcal{R}\hat{\boldsymbol{n}}_1, \mathcal{R}\hat{\boldsymbol{n}}_2) &= \sum_{m,m',m''} D^{\ell}_{m'm} D^{\ell *}_{m''m} Y_{\ell m'}(\hat{\boldsymbol{n}}_1) Y^*_{\ell m''}(\hat{\boldsymbol{n}}_2) \\ &= \sum_{m'} Y_{\ell m'}(\hat{\boldsymbol{n}}_1) Y^*_{\ell m'}(\hat{\boldsymbol{n}}_2) = K(\hat{\boldsymbol{n}}_1, \hat{\boldsymbol{n}}_2). \end{aligned} \tag{3}$$

Therefore, $K(\hat{\boldsymbol{n}}_1, \hat{\boldsymbol{n}}_2)$ depends only on the invariant scalar product $\cos\alpha = \hat{\boldsymbol{n}}_1 \cdot \hat{\boldsymbol{n}}_2$. Rotate the coordinate system so that $\hat{\boldsymbol{n}}_2$ lies along the z-axis ($\theta_2 = 0$). Since $P^m_\ell(1) = 0$ for all $m \neq 0$ and $P_\ell(1) = 1$, the only surviving term in the sum is $m = 0$:

$$Y_{\ell 0}(\hat{\boldsymbol{n}}_2) = \sqrt{\frac{2\ell+1}{4\pi}} P_\ell(1) = \sqrt{\frac{2\ell+1}{4\pi}}. \tag{4}$$

With $\hat{\boldsymbol{n}}_1$ at polar angle α relative to $\hat{\boldsymbol{n}}_2$, we have $Y_{\ell 0}(\hat{\boldsymbol{n}}_1) = \sqrt{\frac{2\ell+1}{4\pi}} P_\ell(\cos\alpha)$. Substituting into the sum yields

$$K(\hat{\boldsymbol{n}}_1, \hat{\boldsymbol{n}}_2) = Y_{\ell 0}(\hat{\boldsymbol{n}}_1) Y^*_{\ell 0}(\hat{\boldsymbol{n}}_2) = \frac{2\ell+1}{4\pi} P_\ell(\cos\alpha), \tag{5}$$

which establishes the addition theorem (B.2.10).
To prove the Rayleigh plane-wave expansion (B.2.14), expand the plane wave $\mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{r}} = \mathrm{e}^{ikr\cos\theta}$ in Legendre polynomials:

$$\mathrm{e}^{ikr\cos\theta} = \sum_{\ell=0}^{\infty} c_\ell(kr) P_\ell(\cos\theta). \tag{6}$$

Using the orthogonality of Legendre polynomials $\int_{-1}^{1} P_\ell(x) P_{\ell'}(x)\,\mathrm{d}x = \frac{2}{2\ell+1}\delta_{\ell\ell'}$, the coefficients are

$$c_\ell(kr) = \frac{2\ell+1}{2} \int_{-1}^{1} \mathrm{e}^{ikrx} P_\ell(x)\,\mathrm{d}x\,. \tag{7}$$

Using Rodrigues' formula $P_\ell(x) = \frac{1}{2^\ell \ell!} \frac{\mathrm{d}^\ell}{\mathrm{d}x^\ell}^{\ell}$ and integrating by parts ℓ times:

$$c_\ell(kr) = \frac{2\ell+1}{2} \frac{(-ikr)^\ell}{2^\ell \ell!} \int_{-1}^{1} \mathrm{e}^{ikrx} (x^2-1)^\ell\,\mathrm{d}x = (2\ell+1) i^\ell \frac{(kr)^\ell}{2^{\ell+1}\ell!} \int_{-1}^{1} \mathrm{e}^{ikrx} (1-x^2)^\ell\,\mathrm{d}x\,. \tag{8}$$

Poisson's integral representation for the spherical Bessel function gives $j_\ell(z) = \frac{z^\ell}{2^{\ell+1}\ell!} \int_{-1}^{1} \mathrm{e}^{izx}(1-x^2)^\ell\,\mathrm{d}x$. Hence,

$$c_\ell(kr) = (2\ell+1) i^\ell j_\ell(kr), \tag{9}$$

proving (B.2.14). Finally, substituting the addition theorem (B.2.10) into (B.2.14) directly produces the spherical harmonic expansion (B.2.15).

B.2.6 Useful Integrals

Integrating three spherical harmonics requires the so-called $3j$-Wigner symbols,

$$\int \mathrm{d}^2\boldsymbol{\gamma}\, Y_{\ell_1 m_1}(\boldsymbol{\gamma}) Y_{\ell_2 m_2}(\boldsymbol{\gamma}) Y_{\ell_3 m_3}(\boldsymbol{\gamma}) = \sqrt{\frac{(2\ell_1+1)(2\ell_2+1)(2\ell_3+1)}{4\pi}} \times \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix}. \tag{B.2.17}$$

These $3j$-Wigner symbols,

$$\begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix},$$

are related to the Clebsch–Gordan coefficients (see Ref. [1] for their properties) and are non-zero provided

$$|\ell_1 - \ell_2| \leq \ell_3 \leq \ell_1 + \ell_2, \qquad m_1 + m_2 + m_3 = 0. \tag{B.2.18}$$

Furthermore, they satisfy the property

$$\sum_{m_1,m_2} \begin{pmatrix} \ell_1 & \ell_2 & \ell \\ m_1 & m_2 & m \end{pmatrix} \begin{pmatrix} \ell_1 & \ell_2 & \ell' \\ m_1 & m_2 & m' \end{pmatrix} = \frac{\delta_{\ell\ell'}\delta_{mm'}}{2\ell+1}. \tag{B.2.19}$$

With the above properties at hand, the integrals we have encountered in Chapter 6 are easily calculated. We have

$$B^{\pm 1}_{\ell m} = \int \mathrm{d}^2\boldsymbol{\gamma}\, Y^*_{\ell m}(\boldsymbol{\gamma}) \mathrm{e}^{i\boldsymbol{\gamma}\cdot\boldsymbol{k}} \sin\theta \mathrm{e}^{\pm i\phi}, \tag{B.2.20}$$

where $\boldsymbol{\gamma}$ is the unit vector pointing in the direction (θ, ϕ). Relation (8.733) in Ref. [2] tells us that

$$B^{\pm 1}_{\ell m} = (-1)^{\ell+1}\delta_{m,\pm 1}\sqrt{4\pi(2\ell+1)}\sqrt{\frac{\ell(\ell+1)}{2}}\frac{j_\ell(k)}{k}. \tag{B.2.21}$$

Analogously, the relation

$$A^\lambda_{\ell m} = \int \mathrm{d}^2\boldsymbol{\gamma}\, Y^*_{\ell m}(\boldsymbol{\gamma}) \mathrm{e}^{i\boldsymbol{\gamma}\cdot\boldsymbol{k}} \sin^2\theta \left(\delta^\lambda_\times \sin 2\phi + \delta^\lambda_+ \cos 2\phi\right) \tag{B.2.22}$$

reduces to

$$A^\lambda_{\ell m} = \pi\sqrt{\frac{2\ell+1}{4\pi}}\sqrt{\frac{(\ell+2)!}{(\ell-2)!}}\alpha^\lambda_{\ell m}\sum_n i^n(2n+1)j_\ell(k)\left(c_{-2}\delta^n_{\ell-2} - 2c_0\delta^n_\ell + c_{+2}\delta^n_{\ell+2}\right), \tag{B.2.23}$$

with coefficients

$$\alpha^\lambda_{\ell m} = \delta^2_m\left(\delta^\lambda_+ - i\delta^\lambda_\times\right) + \delta^{-2}_m\left(\delta^\lambda_+ + i\delta^\lambda_\times\right),$$

and

$$c_n^{-1} = \frac{1}{2}(2\ell - 1 + n)(2\ell + 1 + n)(2\ell + 3 + n).$$

B.3 Bessel Functions

B.3.1 Definition

Bessel functions, generically denoted by $Z_\nu(z)$, are solutions of the differential equation

$$\frac{\mathrm{d}^2 Z_\nu}{\mathrm{d}z^2} + \frac{1}{z}\frac{\mathrm{d}Z_\nu}{\mathrm{d}z} + \left(1 - \frac{\nu^2}{z^2}\right)Z_\nu = 0. \tag{B.3.1}$$

It is common to distinguish between Bessel functions of the first (called J_ν) and the second (N_ν) kind, the latter being also often called the Neumann functions (or Y_ν). They are defined, respectively, through

$$J_\nu(z) = \left(\frac{z}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,\Gamma(\nu+k+1)} \left(\frac{z}{2}\right)^{2k}, \tag{B.3.2}$$

if $|\arg z| < \pi$, and

$$N_\nu(z) = \frac{1}{\sin(\nu\pi)} \left[\cos(\nu\pi) J_\nu(z) - J_{-\nu}(z)\right], \tag{B.3.3}$$

if ν is not an integer. For integer orders $\nu = n \in \mathbb{Z}$, the limit is taken, and Bessel functions of the first and second kind enjoy the symmetry

$$J_{-n}(z) = (-1)^n J_n(z), \qquad N_{-n}(z) = (-1)^n N_n(z). \tag{B.3.4}$$

The two functions J_ν and N_ν are two linearly independent solutions of (B.3.1).

It is sometimes useful to introduce the Bessel functions of the third kind, also called Hankel functions, defined by

$$H_\nu^{(1)}(z) = J_\nu(z) + iN_\nu(z), \qquad H_\nu^{(2)}(z) = J_\nu(z) - iN_\nu(z). \tag{B.3.5}$$

Generic Bessel functions satisfy the recursion relations

$$zZ_{\nu-1} + zZ_{\nu+1} = 2\nu Z_\nu, \tag{B.3.6}$$

$$Z_{\nu-1} - Z_{\nu+1} = 2Z'_\nu. \tag{B.3.7}$$

B.3.2 Asymptotic Properties

One can expand the functions J_ν as series according to

$$J_\nu(z) = \left(\frac{z}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,\Gamma(\nu+k+1)} \left(\frac{z}{2}\right)^{2k}. \tag{B.3.8}$$

From this, one gets their behaviour at the origin, i.e. for small values of the argument (around $z \sim 0$),

$$J_\nu(z) \sim \frac{1}{\Gamma(\nu+1)} \left(\frac{z}{2}\right)^\nu, \qquad N_\nu(z) \sim -\frac{\Gamma(\nu)}{\pi} \left(\frac{2}{z}\right)^\nu = \frac{1}{\Gamma(-\nu+1)\sin(\nu\pi)} \left(\frac{z}{2}\right)^{-\nu}. \tag{B.3.9}$$

At infinity, on the other hand, Bessel functions have the following asymptotic behaviours:

$$J_{\pm\nu}(z) \sim \sqrt{\frac{2}{\pi z}} \cos\left(z \mp \frac{\pi}{2}\nu - \frac{\pi}{4}\right), \qquad N_{\pm\nu}(z) \sim \sqrt{\frac{2}{\pi z}} \sin\left(z \mp \frac{\pi}{2}\nu - \frac{\pi}{4}\right), \tag{B.3.10}$$

and hence those of the Hankel functions are

$$H_\nu^{(1)}(z) \sim \sqrt{\frac{2}{\pi z}} \mathrm{e}^{i\left(z - \frac{\pi}{2}\nu - \frac{\pi}{4}\right)}, \qquad H_\nu^{(2)}(z) \sim \sqrt{\frac{2}{\pi z}} \mathrm{e}^{-i\left(z - \frac{\pi}{2}\nu - \frac{\pi}{4}\right)}. \tag{B.3.11}$$

B.3.3 Special Cases

Some Bessel functions can be expressed in terms of known simple functions for some values of the index ν. For instance, one has

$$J_{1/2}(z) = \sqrt{\frac{2}{\pi z}} \sin z, \qquad N_{1/2}(z) = -\sqrt{\frac{2}{\pi z}} \cos z, \tag{B.3.12}$$

$$H^{(1)}_{1/2}(z) = \sqrt{\frac{2}{\pi z}}\frac{\mathrm{e}^{iz}}{i}, \qquad H^{(2)}_{-1/2}(z) = \sqrt{\frac{2}{\pi z}}\frac{\mathrm{e}^{-iz}}{-i}, \tag{B.3.13}$$

and many other similar simple forms for a half-integer ν.

It is of interest to write down at this point the solution to some differential equations frequently met in cosmology in terms of Bessel functions. For instance, we have

$$u'' + \left[(\beta\gamma z^{\gamma-1})^2 - \left(\frac{\nu\gamma}{z}\right)^2\right] u = 0 \implies u = zZ_\nu(\beta z^\gamma), \tag{B.3.14}$$

$$u'' + \frac{1-2\alpha}{z}u' + \left[(\beta\gamma z^{\gamma-1})^2 + \frac{\alpha^2-\nu^2\gamma^2}{z^2}\right] u = 0 \implies u = z^\alpha Z_\nu(\beta z^\gamma), \tag{B.3.15}$$

and it should be emphasized that a particularly important case in cosmology, as the reader undoubtedly remarked, is that for which $\alpha = -1/2$ and $\gamma = 1$.

B.3.4 Spherical Bessel Functions

Spherical Bessel functions are defined through the relation

$$j_\nu(z) \equiv \sqrt{\frac{\pi}{2z}} J_{\nu+1/2}(z), \tag{B.3.16}$$

and satisfy the recursion relations

$$j_\ell = \frac{z}{2\ell+1}(j_{\ell-1} + j_{\ell+1}), \qquad j'_\ell = \frac{1}{2\ell+1}\left[\ell j_{\ell-1} - (\ell+1)j_{\ell+1}\right], \qquad j'_\ell = j_{\ell-1} - \frac{\ell+1}{z}j_\ell, \tag{B.3.17}$$

from which one deduces

$$\frac{j_\ell}{z^2} = \frac{j_{\ell-2}}{(2\ell-1)(2\ell+1)} + \frac{2j_\ell}{(2\ell-1)(2\ell+3)} + \frac{j_{\ell+2}}{(2\ell+1)(2\ell+3)}. \tag{B.3.18}$$

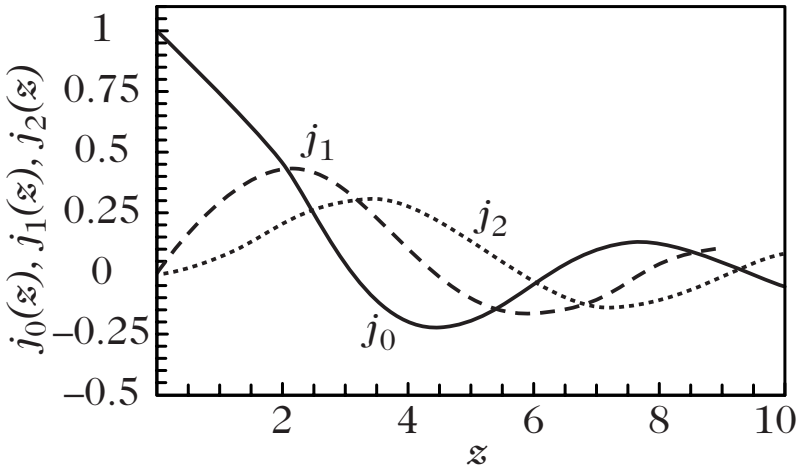


Figure B.2: Spherical Bessel functions for $\ell = 0, 1, 2$.

It can be checked from (B.3.15) that they satisfy the second-order differential equation

$$j''_\ell(x) + \frac{2}{x}j'_\ell(x) + \left[1 - \frac{\ell(\ell+1)}{x^2}\right] j_\ell(x) = 0. \tag{B.3.19}$$

The first three spherical Bessel functions are illustrated in Fig. B.2:

$$j_0(z) = \frac{\sin z}{z}, \qquad j_1(z) = \frac{\sin z}{z^2} - \frac{\cos z}{z}, \qquad j_2(z) = \left(\frac{3}{z^3} - \frac{1}{z}\right)\sin z - \frac{3}{z^2}\cos z.$$

B.3.5 Some Useful Integrals

The Bessel functions satisfy

$$\int_0^\infty J_\nu(t)J_\mu(t)t^{-\lambda}\,\mathrm{d}t = \frac{2^{-\lambda}\Gamma(\lambda)\Gamma\left(\frac{\nu+\mu-\lambda+1}{2}\right)}{\Gamma\left(\frac{-\nu+\mu+\lambda+1}{2}\right)\Gamma\left(\frac{\nu+\mu+\lambda+1}{2}\right)\Gamma\left(\frac{\nu-\mu+\lambda+1}{2}\right)}, \tag{B.3.20}$$

if $\mathrm{Re}(\nu+\mu+1) > \mathrm{Re}(\lambda) > 0$. This means that

$$\int_0^\infty j_\ell(t)j_\ell(t)t^{-k}\,\mathrm{d}t = \frac{\pi}{2}\frac{\Gamma(k+1)\Gamma\left(\frac{2\ell+1-k}{2}\right)}{2^{k+1}\Gamma^2\left(\frac{k+2}{2}\right)\Gamma\left(\frac{2\ell+k+3}{2}\right)}, \tag{B.3.21}$$

with $-1 < k < 2\ell+1$. This also leads to

$$\int_0^\infty j_\ell(pr)j_{\ell'}(p'r)r^2\,\mathrm{d}r = \frac{\pi}{2p^2}\delta(p-p'), \tag{B.3.22}$$

defining the normalization of the spherical Bessel functions.

Consider the general linear differential equation (B.3.15):

$$u'' + \frac{1-2\alpha}{z}u' + \left[(\beta\gamma z^{\gamma-1})^2 + \frac{\alpha^2-\nu^2\gamma^2}{z^2}\right]u = 0. \tag{1}$$

To solve this equation, let $u(z) = z^\alpha v(\xi)$, where $\xi(z) \equiv \beta z^\gamma$. Computing derivatives:

$$u'(z) = \alpha z^{\alpha-1}v(\xi) + \beta\gamma z^{\alpha+\gamma-1}\frac{\mathrm{d}v}{\mathrm{d}\xi}, \tag{2}$$

$$u''(z) = \alpha(\alpha-1)z^{\alpha-2}v(\xi) + 2\alpha\beta\gamma z^{\alpha+\gamma-2}\frac{\mathrm{d}v}{\mathrm{d}\xi} + \beta\gamma(\gamma-1)z^{\alpha+\gamma-2}\frac{\mathrm{d}v}{\mathrm{d}\xi} + (\beta\gamma)^2 z^{\alpha+2\gamma-2}\frac{\mathrm{d}^2v}{\mathrm{d}\xi^2}. \tag{3}$$

Substitute these expressions into the differential equation and multiply by $z^{2-\alpha}$:

$$\begin{aligned}(\beta\gamma)^2 z^{2\gamma}\frac{\mathrm{d}^2v}{\mathrm{d}\xi^2} + \beta\gamma z^\gamma\left[2\alpha+\gamma-1+1-2\alpha\right]\frac{\mathrm{d}v}{\mathrm{d}\xi} \\ + \left[\alpha(\alpha-1)+\alpha(1-2\alpha)+(\beta\gamma z^\gamma)^2+\alpha^2-\nu^2\gamma^2\right]v = 0.\end{aligned} \tag{4}$$

The coefficient of $\frac{\mathrm{d}v}{\mathrm{d}\xi}$ simplifies to $\beta\gamma^2 z^\gamma = \gamma\xi$. The bracket in the zeroth-order term reduces to:

$$\alpha^2 - \alpha + \alpha - 2\alpha^2 + \alpha^2 + (\beta\gamma z^\gamma)^2 - \nu^2\gamma^2 = \gamma^2\xi^2 - \nu^2\gamma^2. \tag{5}$$

Noting that $(\beta\gamma)^2 z^{2\gamma} = \gamma^2\xi^2$, the entire equation becomes

$$\gamma^2\xi^2\frac{\mathrm{d}^2v}{\mathrm{d}\xi^2} + \gamma^2\xi\frac{\mathrm{d}v}{\mathrm{d}\xi} + \gamma^2(\xi^2-\nu^2)v = 0. \tag{6}$$

Dividing by γ^2 yields the standard Bessel equation:

$$\xi^2\frac{\mathrm{d}^2v}{\mathrm{d}\xi^2} + \xi\frac{\mathrm{d}v}{\mathrm{d}\xi} + (\xi^2-\nu^2)v = 0. \tag{7}$$

Hence, $v(\xi) = Z_\nu(\xi) = Z_\nu(\beta z^\gamma)$, and therefore $u(z) = z^\alpha Z_\nu(\beta z^\gamma)$, establishing Eq. (B.3.15).

For the spherical Bessel equation (B.3.19), set $\alpha = -1/2$, $\gamma = 1$, $\beta = 1$, and $\nu = \ell + 1/2$. Then:

$$\frac{1-2\alpha}{z} = \frac{1-2(-1/2)}{z} = \frac{2}{z}, \qquad \alpha^2 - \nu^2\gamma^2 = \frac{1}{4} - \left(\ell + \frac{1}{2}\right)^2 = -\ell(\ell+1). \tag{8}$$

The differential equation becomes

$$u'' + \frac{2}{z}u' + \left[1 - \frac{\ell(\ell+1)}{z^2}\right] u = 0, \tag{9}$$

which is precisely Eq. (B.3.19). Its regular solution is $u(z) = z^{-1/2}J_{\ell+1/2}(z) = \sqrt{2/\pi}j_\ell(z)$, matching definition (B.3.16).

B.4 Legendre Polynomials

B.4.1 Associated Legendre Polynomials

An associated Legendre polynomial, $P_\nu^\mu(z)$, is a solution of the differential equation

$$(1-z^2)\frac{\mathrm{d}^2u}{\mathrm{d}z^2} - 2z\frac{\mathrm{d}u}{\mathrm{d}z} + \left[\nu(\nu+1) - \frac{\mu^2}{1-z^2}\right] u = 0, \tag{B.4.1}$$

μ and ν being two complex numbers. These polynomials satisfy the recursion relation—(8.733) of Ref. [2]—

$$P_{\nu-1}^\mu - P_{\nu+1}^\mu = (2\nu+1)\sqrt{1-x^2}P_\nu^{\mu-1}. \tag{B.4.2}$$

B.4.2 Legendre Polynomials

Legendre polynomials are in fact a special case of associated Legendre polynomials for which $\mu = 0$ and $\nu = n \in \mathbb{N}$. These polynomials form an orthogonal basis for functions defined on the segment $[-1, 1]$, satisfying the orthogonality relation

$$\int_{-1}^{1} \left(n + \frac{1}{2}\right)^{1/2} P_n(x) \left(m + \frac{1}{2}\right)^{1/2} P_m(x)\,\mathrm{d}x = \delta_{mn}. \tag{B.4.3}$$

In particular, one has

$$P_0(x) = 1, \qquad P_1(x) = x, \qquad P_2(x) = \frac{1}{2}(3x^2-1), \qquad P_3(x) = \frac{x}{2}(5x^2-3).$$

Any given function of the variable $\cos\phi$ can be expanded as

$$f(\cos\phi) = \sum_n \frac{2n+1}{2} a_n P_n(\cos\phi). \tag{B.4.4}$$

The Legendre polynomials for $\ell = 3, 10, 30$ are plotted in Fig. B.3.

Legendre polynomials $P_n(x)$ can be generated from the Rodrigues formula:

$$P_n(x) = \frac{1}{2^n n!}\frac{\mathrm{d}^n\ {}^n}{\mathrm{d}x^n}. \tag{1}$$

Let $I_{mn} \equiv \int_{-1}^{1} P_m(x)P_n(x)\,\mathrm{d}x$, and assume without loss of generality that $m \le n$. Inte-

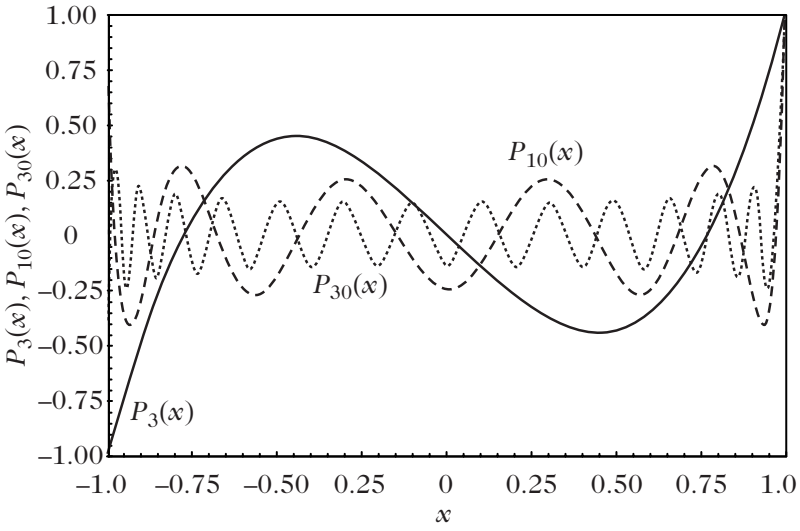


Figure B.3: Legendre polynomials for $\ell = 3, 10, 30$.

grating by parts n times:

$$I_{mn} = \frac{1}{2^n n!} \int_{-1}^{1} P_m(x) \frac{\mathrm{d}^n \ ^n}{\mathrm{d}x^n} \, \mathrm{d}x = \frac{(-1)^n}{2^n n!} \int_{-1}^{1} \frac{\mathrm{d}^n P_m \ ^n}{\mathrm{d}x^n} \, \mathrm{d}x \,, \tag{2}$$

since the boundary terms at $x = \pm 1$ vanish because $(x^2 - 1)^n = (x - 1)^n (x + 1)^n$ and its first $n - 1$ derivatives vanish at $x = \pm 1$.
If $m < n$, then $P_m(x)$ is a polynomial of degree $m < n$, so its n-th derivative identically vanishes: $\frac{\mathrm{d}^n P_m}{\mathrm{d}x^n} = 0$. Thus $I_{mn} = 0$ for $m \neq n$.
For $m = n$, the leading coefficient of $P_n(x)$ is $\frac{(2n)!}{2^n (n!)^2}$, so $\frac{\mathrm{d}^n P_n}{\mathrm{d}x^n} = \frac{(2n)!}{2^n n!}$. Then

$$I_{nn} = \frac{(2n)!}{2^{2n} (n!)^2} \int_{-1}^{1} (1 - x^2)^n \, \mathrm{d}x \,. \tag{3}$$

Setting $x = 2t - 1$, the integral maps to a Beta function:

$$\int_{-1}^{1} (1 - x^2)^n \, \mathrm{d}x = 2^{2n+1} \int_{0}^{1} t^n (1 - t)^n \, \mathrm{d}t = 2^{2n+1} B(n + 1, n + 1) = 2^{2n+1} \frac{(n!)^2}{(2n + 1)!}. \tag{4}$$

Substituting this back yields

$$I_{nn} = \frac{(2n)!}{2^{2n} (n!)^2} \cdot \frac{2^{2n+1} (n!)^2}{(2n + 1)!} = \frac{2}{2n + 1}. \tag{5}$$

Rewriting this as

$$\int_{-1}^{1} \left(\sqrt{n + \frac{1}{2}} P_n(x) \right) \left(\sqrt{m + \frac{1}{2}} P_m(x) \right) \mathrm{d}x = \delta_{mn}, \tag{6}$$

reproduces Eq. (B.4.3).

B.5 Fourier Transform and Eigenmodes of the Laplacian

B.5.1 Fourier Transform

In a D-dimensional maximally symmetric space, the Helmholtz equation is an eigenvalue equation of the Laplacian

$$\Delta Q_k = -(k^2 - K)Q_k, \tag{B.5.1}$$

where the wavevector k labels the modes and the sign of K depends on the spatial curvature; Δ is the Laplacian defined from the covariant derivative D_i associated to the spatial metric γ_{ij}. We also impose the normalization condition

$$\int Q_k(\boldsymbol{x})Q^*_{k'}(\boldsymbol{x})\sqrt{\gamma}\,\mathrm{d}^D x = (2\pi)^D\delta^{(D)}(\boldsymbol{k}-\boldsymbol{k}'). \tag{B.5.2}$$

Any function can be expanded in Fourier modes on the basis of the functions $Q_k(\boldsymbol{x})$ according to

$$f(\boldsymbol{x}) = \int \frac{\mathrm{d}^D k}{(2\pi)^{D/2}}\hat{f}(\boldsymbol{k})Q_k(\boldsymbol{x}), \tag{B.5.3}$$

where $\hat{f}(\boldsymbol{k})$ are the Fourier coefficients of the function $f(\boldsymbol{x})$. This relation is inverted using (B.5.2), yielding

$$\hat{f}(\boldsymbol{k}) = \int \frac{\sqrt{\gamma}\,\mathrm{d}^D x}{(2\pi)^{D/2}} f(\boldsymbol{x})Q^*_k(\boldsymbol{x}). \tag{B.5.4}$$

B.5.2 Power Spectrum

If $f(\boldsymbol{x})$ is a given homogeneous and isotropic stochastic field, one can define its correlation function by

$$\xi_f(r) = \langle f(\boldsymbol{x})f(\boldsymbol{x}+\boldsymbol{r})\rangle\,. \tag{B.5.5}$$

Its power spectrum is defined in Fourier space as

$$\left\langle \hat{f}(\boldsymbol{k})\hat{f}^*(\boldsymbol{k}')\right\rangle = \delta^{(D)}(\boldsymbol{k}-\boldsymbol{k}')P_f(k). \tag{B.5.6}$$

It turns out to be handy to define a spectrum with different dimensions, namely

$$\mathcal{P}_f(k) = \frac{2S_{[D-2]}}{(2\pi)^D}k^D P_f(k), \tag{B.5.7}$$

where $S_{[D-1]}$ is the surface of the $(D-1)$-dimensional sphere of unit radius,

$$S_{[D-1]} = \frac{2\pi^{D/2}}{\Gamma(D/2)}, \tag{B.5.8}$$

where we recover the particular value for the surface of the 1-sphere, i.e. a circle, $S_{[1]} = 2\pi$, and that of the 2-sphere, for which $S_{[2]} = 4\pi$.

The correlation function stems from the power spectrum through

$$\xi_f(r) = \frac{2S_{[D-2]}}{(2\pi)^D}\int P_f(k)k^D\frac{\sin kr}{kr}\frac{\mathrm{d}k}{k} = \int \mathcal{P}_f(k)\frac{\sin kr}{kr}\frac{\mathrm{d}k}{k}, \tag{B.5.9}$$

after integrating out over the $(D-1)$ angular directions.

B.5.3 Cartesian Coordinates

To get the explicit form of the functions Q_k, let us restrict attention to the case $D = 3$. Depending on the curvature sign, the wave numbers take the values

$$K < 0 : \qquad k \in [0, \infty[\quad \text{and} \quad ik \in [0, \sqrt{-K}], \tag{B.5.10}$$

$$K = 0 : \qquad k \in [0, \infty[, \tag{B.5.11}$$

$$K > 0 : \qquad k = (\nu + 1)\sqrt{K}, \qquad \nu \in \mathbb{N}. \tag{B.5.12}$$

In Cartesian coordinates, only the functions for $K = 0$ take a simple form, namely

$$Q_{\boldsymbol{k}}(\boldsymbol{x}) = \exp(i\boldsymbol{k} \cdot \boldsymbol{x}), \tag{B.5.13}$$

and Eqs. (B.5.3)–(B.5.4) take the usual form of the Fourier transform. Note also that, still assuming flat space $K = 0$, we get the following useful expression

$$\delta^{(D)}(\boldsymbol{x} - \boldsymbol{x}') = \frac{1}{(2\pi)^D} \int \mathrm{d}^D \boldsymbol{k}\, \mathrm{e}^{i\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{x}')}, \tag{B.5.14}$$

for the Dirac distribution $\delta^{(D)}$ in D dimensions.

B.5.4 Spherical Coordinates

In a spherical coordinate system (χ, θ, ϕ), the eigenfunctions can be expanded as

$$Q_{k\ell m} = (2\pi)^{3/2} R_{k\ell}(\chi) Y_\ell^m(\theta, \phi), \tag{B.5.15}$$

with the normalization taking the form

$$\int Q_{k\ell m}(\chi, \theta, \phi) Q^*_{k'\ell'm'}(\chi, \theta, \phi) f_K^2(\chi)\, \mathrm{d}\chi\, \mathrm{d}\Omega = \frac{(2\pi)^3}{k^2} \delta(k - k') \delta_{\ell\ell'} \delta_{mm'}. \tag{B.5.16}$$

With the decomposition (B.5.15), the radial part of the Helmholtz equation becomes

$$\frac{1}{f_K^2(\chi)} \frac{\mathrm{d}}{\mathrm{d}\chi} \left[f_K^2(\chi) \frac{\mathrm{d}R_{k\ell}}{\mathrm{d}\chi} \right] + \left[(k^2 - K) - \frac{\ell(\ell + 1)}{f_K^2(\chi)} \right] R_{k\ell} = 0. \tag{B.5.17}$$

This second-order differential equation has two independent solutions, only one of which is regular at $\chi = 0$, so that the radial function is fully determined once the normalization is chosen. Solving this equation yields the following three solutions [4, 5]:

$$R_{k\ell}(\chi) = \sqrt{\frac{N_{k\ell}}{k f_K(\chi)}} P^{-1/2-\ell}_{-1/2+i\omega} \left(\cosh \sqrt{-K}\chi \right) \qquad (K < 0), \tag{B.5.18}$$

$$R_{k\ell}(\chi) = \sqrt{\frac{2}{\pi}} j_\ell(k\chi) \qquad (K = 0), \tag{B.5.19}$$

$$R_{k\ell}(\chi) = \sqrt{\frac{M_{k\ell}}{k f_K(\chi)}} P^{-1/2-\ell}_{-1/2+\nu} \left(\cos \sqrt{K}\chi \right) \qquad (K > 0). \tag{B.5.20}$$

The numerical normalization coefficients are given by

$$N_{k\ell} = \prod_{n=0}^{\ell} (\omega^2 + n^2), \qquad M_{k\ell} = \begin{cases} \prod_{n=0}^{\ell} \left[(\nu + 1)^2 - n^2 \right] & \text{for } \ell \leq \nu, \\ 0 & \text{for } \ell > \nu, \end{cases} \tag{B.5.21}$$

with

$$\omega = \frac{k}{\sqrt{-K}}, \qquad \nu = \frac{k}{\sqrt{K}} - 1. \tag{B.5.22}$$

Note that in the case $K < 0$, the normalization condition only holds for the modes for which k is real [6], which correspond to wavelengths smaller than the curvature radius of space. For the supercurvature modes, i.e. $ik \in [0, \sqrt{-K}]$, the radial function can be obtained by analytic continuation.

Any function can then be expanded into spherical Fourier modes as

$$f(\chi,\theta,\phi) = \frac{1}{(2\pi)^{3/2}} \sum_{\ell m} \begin{Bmatrix} \int_0^\infty k^2\,\mathrm{d}k \\ K^{3/2} \sum_{\nu=2}^{\infty} (\nu+1)^2 \end{Bmatrix} f_{\ell m}(k) Q_{k\ell m}(\chi,\theta,\phi), \tag{B.5.23}$$

the Fourier coefficients being given by

$$f_{\ell m}(k) = \frac{1}{(2\pi)^{3/2}} \int f(\chi,\theta,\phi) Q^*_{k\ell m}(\chi,\theta,\phi) f_K^2(\chi)\,\mathrm{d}\chi\,\mathrm{d}\Omega\,. \tag{B.5.24}$$

Using Eqs. (B.2.8) and (B.2.15), it can be shown that these Fourier coefficients are related by

$$f_{\ell m}(k) = (-i)^\ell \int \hat{f}(\boldsymbol{k}) Y_{\ell m}(\hat{\boldsymbol{k}})\,\mathrm{d}\Omega_{\boldsymbol{k}}\,. \tag{B.5.25}$$

In a 3-dimensional maximally symmetric space with metric $\mathrm{d}s^2 = \mathrm{d}\chi^2 + f_K^2(\chi)(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)$, the metric determinant is $\sqrt{\gamma} = f_K^2(\chi)\sin\theta$. The Laplace–Beltrami operator on a scalar function Q is

$$\Delta Q = \frac{1}{\sqrt{\gamma}}\partial_i\left(\sqrt{\gamma}\gamma^{ij}\partial_j Q\right) = \frac{1}{f_K^2(\chi)}\frac{\partial}{\partial\chi}\left[f_K^2(\chi)\frac{\partial Q}{\partial\chi}\right] + \frac{1}{f_K^2(\chi)}\Delta_{S^2} Q, \tag{1}$$

where $\Delta_{S^2} \equiv \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}$ is the standard Laplacian on the unit 2-sphere. Separating variables according to $Q_{k\ell m}(\chi,\theta,\phi) = (2\pi)^{3/2} R_{k\ell}(\chi) Y_\ell^m(\theta,\phi)$ and recalling the spherical harmonic eigenvalue equation (B.2.1):

$$\Delta_{S^2} Y_\ell^m = -\ell(\ell+1) Y_\ell^m, \tag{2}$$

the Helmholtz equation (B.5.1), $\Delta Q = -(k^2 - K)Q$, becomes

$$\frac{1}{f_K^2(\chi)}\frac{\mathrm{d}}{\mathrm{d}\chi}\left[f_K^2(\chi)\frac{\mathrm{d}R_{k\ell}}{\mathrm{d}\chi}\right] Y_\ell^m - \frac{\ell(\ell+1)}{f_K^2(\chi)} R_{k\ell} Y_\ell^m = -(k^2 - K) R_{k\ell} Y_\ell^m. \tag{3}$$

Factoring out Y_ℓ^m yields the radial Helmholtz equation:

$$\frac{1}{f_K^2(\chi)}\frac{\mathrm{d}}{\mathrm{d}\chi}\left[f_K^2(\chi)\frac{\mathrm{d}R_{k\ell}}{\mathrm{d}\chi}\right] + \left[(k^2 - K) - \frac{\ell(\ell+1)}{f_K^2(\chi)}\right] R_{k\ell} = 0, \tag{4}$$

which is Eq. (B.5.17).

In flat space $K = 0$, $f_0(\chi) = \chi$, and this reduces to

$$\frac{\mathrm{d}^2 R_{k\ell}}{\mathrm{d}\chi^2} + \frac{2}{\chi}\frac{\mathrm{d}R_{k\ell}}{\mathrm{d}\chi} + \left[k^2 - \frac{\ell(\ell+1)}{\chi^2}\right] R_{k\ell} = 0. \tag{5}$$

Setting $x = k\chi$, this matches the spherical Bessel equation (B.3.19). The regular solution at $\chi = 0$ is $R_{k\ell}(\chi) = C j_\ell(k\chi)$. Using the normalization condition (B.5.16) with (B.3.22) fixes $C = \sqrt{2/\pi}$, proving Eq. (B.5.19). In curved spaces ($K \neq 0$), the transformation $R_{k\ell}(\chi) = [f_K(\chi)]^{-1/2} u(\chi)$ maps (B.5.17) into the associated Legendre equation with variable $z = \cosh\left(\sqrt{-K}\chi\right)$ or $z = \cos\left(\sqrt{K}\chi\right)$, yielding solutions (B.5.18) and (B.5.20).

Appendix C

Useful Cosmological Quantities

This appendix gathers the background and perturbed geometric and matter tensors in Friedmann–Lemaître–Robertson–Walker (FLRW) cosmologies. Throughout this appendix, we use conformal time η unless explicitly stated otherwise, with prime denoting a derivative with respect to conformal time, $' \equiv \mathrm{d}/\mathrm{d}\eta$, and $\mathcal{H} \equiv a'/a = aH$, where $H \equiv \dot{a}/a$ is the usual Hubble expansion rate with dot denoting cosmic time derivative, $\cdot \equiv \mathrm{d}/\mathrm{d}t$. Spatial indices $i, j, k, \ldots$ are raised and lowered with the background 3-metric γ_{ij}, and D_i denotes the covariant derivative compatible with γ_{ij}.

C.1 Background Space

C.1.1 Geometry

In cosmic time t, the line element of a homogeneous and isotropic FLRW universe is

$$\mathrm{d}s^2 = -\,\mathrm{d}t^2 + a^2(t)\gamma_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j\,, \tag{C.1.1}$$

while in conformal time η, defined through $\mathrm{d}t = a(\eta)\,\mathrm{d}\eta$, the metric is conformally flat (or conformally static in curved spaces),

$$\mathrm{d}s^2 = a^2(\eta)\left[-\,\mathrm{d}\eta^2 + \gamma_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j\right], \tag{C.1.2}$$

with components $g_{00} = -a^2$, $g_{0i} = 0$, $g_{ij} = a^2\gamma_{ij}$, and inverse metric

$$g^{00} = -a^{-2}, \qquad g^{0i} = 0, \qquad g^{ij} = a^{-2}\gamma^{ij}. \tag{C.1.3}$$

The spatial metric can also take the standard form in spherical coordinates:

$$\gamma_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j = \frac{\mathrm{d}r^2}{1-Kr^2} + r^2\,\mathrm{d}\Omega^2 = \mathrm{d}\chi^2 + f_K^2(\chi)\,\mathrm{d}\Omega^2\,, \tag{C.1.4}$$

with $\mathrm{d}\Omega^2 \equiv \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$ and

$$f_K(\chi) = \begin{cases} \dfrac{1}{\sqrt{K}}\sin\!\left(\sqrt{K}\chi\right) & (K>0), \\ \chi & (K=0), \\ \dfrac{1}{\sqrt{-K}}\sinh\!\left(\sqrt{-K}\chi\right) & (K<0), \end{cases} \tag{C.1.5}$$

respectively for spatial sections with positive, zero, and negative curvature.

The non-vanishing Christoffel symbols computed from the metric (C.1.2) are

$$\Gamma^0_{00} = \mathcal{H}, \qquad \Gamma^0_{ij} = \mathcal{H}\gamma_{ij}, \qquad \Gamma^j_{0i} = \mathcal{H}\delta^j_i, \qquad \Gamma^i_{jk} = \gamma^i_{jk}, \tag{C.1.6}$$

where γ^i_{jk} are the Christoffel symbols derived solely from the spatial metric γ_{ij}.

The non-vanishing components of the Riemann curvature tensor $R^\mu{}_{\nu\alpha\beta} \equiv \partial_\alpha \Gamma^\mu_{\nu\beta} - \partial_\beta \Gamma^\mu_{\nu\alpha} + \Gamma^\mu_{\sigma\alpha}\Gamma^\sigma_{\nu\beta} - \Gamma^\mu_{\sigma\beta}\Gamma^\sigma_{\nu\alpha}$ are given by

$$R^0{}_{i0j} = \mathcal{H}'\gamma_{ij}, \qquad R^i{}_{00j} = \mathcal{H}'\delta^i_j, \qquad R^i{}_{jml} = (\mathcal{H}^2 + K)\left(\delta^i_m\gamma_{jl} - \delta^i_l\gamma_{jm}\right), \tag{C.1.7}$$

leading upon contraction to the components of the Ricci tensor $R_{\mu\nu} \equiv R^\alpha{}_{\mu\alpha\nu}$:

$$R_{00} = -3\mathcal{H}', \qquad R_{ij} = \left(2\mathcal{H}^2 + \mathcal{H}' + 2K\right)\gamma_{ij}, \tag{C.1.8}$$

and eventually the spacetime Ricci scalar curvature $R \equiv g^{\mu\nu}R_{\mu\nu}$:

$$Ra^2 = 6\left(\mathcal{H}^2 + \mathcal{H}' + K\right). \tag{C.1.9}$$

Using these tensors, one obtains the non-vanishing components of the Einstein tensor $G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R$, namely

$$G_{00} = 3\left(\mathcal{H}^2 + K\right), \qquad G_{ij} = -\left(\mathcal{H}^2 + 2\mathcal{H}' + K\right)\gamma_{ij}. \tag{C.1.10}$$

The spatial scalar curvature of the 3-dimensional spatial hypersurfaces is

$${}^{(3)}R = 6K, \tag{C.1.11}$$

which is computed by contracting the spatial Riemann and Ricci tensors associated to the spatial metric γ_{ij}, namely

$${}^{(3)}R^i{}_{jkl} = K\left(\delta^i_k\gamma_{jl} - \delta^i_l\gamma_{jk}\right), \qquad {}^{(3)}R_{ij} = {}^{(3)}R^k{}_{ikj} = 2K\gamma_{ij}, \tag{C.1.12}$$

respectively.

C.1.2 Matter

The stress-energy tensor for a perfect fluid of energy density ρ and isotropic pressure P is

$$T_{\mu\nu} = \rho u_\mu u_\nu + P\left(g_{\mu\nu} + u_\mu u_\nu\right), \tag{C.1.13}$$

with normalized 4-velocity $u^\mu u_\mu = -1$, so that for a comoving cosmological fluid one can take

$$u^0 = \frac{1}{a}, \qquad u^i = 0, \qquad u_0 = -a, \qquad u_i = 0. \tag{C.1.14}$$

The non-vanishing components of this tensor are then given by

$$T_{00} = \rho a^2, \qquad T_{ij} = Pa^2\gamma_{ij}, \tag{C.1.15}$$

or equivalently, with one index raised,

$$T^0{}_0 = -\rho, \qquad T^i{}_j = P\delta^i_j. \tag{C.1.16}$$

The covariant energy-momentum conservation $\nabla_\mu T^\mu{}_\nu = 0$ for a fluid takes the form

$$\rho' + 3\mathcal{H}(1+w)\rho = 0, \tag{C.1.17}$$

where the equation-of-state parameter w is defined by

$$P = w\rho. \tag{C.1.18}$$

After some algebra, we obtain that

$$w' = -3\mathcal{H}(1+w)\left(c_{\rm s}^2 - w\right), \tag{C.1.19}$$

with the adiabatic speed of sound $c_{\rm s}$ defined by

$$c_{\rm s}^2 \equiv \frac{P'}{\rho'}. \tag{C.1.20}$$

Using the Einstein field equations $G_{\mu\nu} = \kappa T_{\mu\nu}$ (where $\kappa \equiv 8\pi G_{\rm N}$), it turns out to be extremely useful to express both the density and pressure as functions of geometric quantities as

$$\kappa\rho a^2 = 3\left(\mathcal{H}^2 + K\right), \qquad \kappa P a^2 = -\left(\mathcal{H}^2 + 2\mathcal{H}' + K\right), \qquad \kappa(\rho + 3P)a^2 = -6\mathcal{H}', \tag{C.1.21}$$

from where we obtain the following formula after some calculations:

$$\mathcal{H}' = -\frac{1}{2}\left(\mathcal{H}^2 + K\right)(1+3w). \tag{C.1.22}$$

For a homogeneous scalar field $\phi(\eta)$ minimally coupled to gravity with potential $V(\phi)$, we have

$$T_{00} = \frac{1}{2}\phi'^2 + a^2 V(\phi), \qquad T_{ij} = \left[\frac{1}{2}\phi'^2 - a^2 V(\phi)\right]\gamma_{ij}. \tag{C.1.23}$$

From the Christoffel symbols (C.1.6), the contraction $\Gamma^{\mu}_{\mu 0} = \Gamma^0_{00} + \Gamma^i_{i0} = \mathcal{H} + 3\mathcal{H} = 4\mathcal{H}$. The conservation equation $\nabla_\mu T^\mu{}_0 = 0$ reads

$$\partial_0 T^0{}_0 + \partial_i T^i{}_0 + \Gamma^0_{\mu 0} T^\mu{}_0 + \Gamma^i_{\mu i} T^\mu{}_0 - \Gamma^\mu_{00} T^0{}_\mu - \Gamma^\mu_{i0} T^i{}_\mu = 0. \tag{1}$$

Since $T^i{}_0 = 0$, $T^0{}_0 = -\rho$, and $T^i{}_j = P\delta^i_j$, this reduces to

$$-\rho' - 4\mathcal{H}\rho + \mathcal{H}\rho - 3\mathcal{H}P = -\rho' - 3\mathcal{H}(\rho + P) = 0, \tag{2}$$

yielding Eq. (C.1.17): $\rho' + 3\mathcal{H}(1+w)\rho = 0$.
Differentiating $w = P/\rho$ with respect to η:

$$w' = \frac{P'\rho - P\rho'}{\rho^2} = \frac{P'}{\rho'}\frac{\rho'}{\rho} - w\frac{\rho'}{\rho} = \frac{\rho'}{\rho}\left(c_{\rm s}^2 - w\right). \tag{3}$$

Substituting $\rho'/\rho = -3\mathcal{H}(1+w)$ from (C.1.17) immediately establishes Eq. (C.1.19):

$$w' = -3\mathcal{H}(1+w)\left(c_{\rm s}^2 - w\right). \tag{4}$$

To derive Eq. (C.1.22), combine the first and third Friedmann relations in (C.1.21):

$$\kappa(\rho + 3P)a^2 = \kappa\rho(1+3w)a^2 = 3\left(\mathcal{H}^2 + K\right)(1+3w). \tag{5}$$

Equating this to $-6\mathcal{H}'$ gives $-6\mathcal{H}' = 3(\mathcal{H}^2 + K)(1+3w)$, which simplifies to

$$\mathcal{H}' = -\frac{1}{2}\left(\mathcal{H}^2 + K\right)(1+3w), \tag{6}$$

proving Eq. (C.1.22).

C.2 Perturbed Quantities

C.2.1 Geometry

The perturbed Friedmann–Lemaître metric in conformal time is given by

$$\begin{aligned}
\delta g_{00} &= -2a^2 A, \\
\delta g_{0i} &= a^2 \left(D_i B + \overline{B}_i \right) \equiv a^2 \mathcal{B}_i, \\
\delta g_{ij} &= a^2 \left[2C\gamma_{ij} + 2D_i D_j E + 2D_{(i}\overline{E}_{j)} + 2\overline{E}_{ij} \right] \equiv a^2 h_{ij},
\end{aligned} \tag{C.2.1}$$

where we have explicitly separated out the scalar, vector, and tensor parts (SVT decomposition). Let us recall that D_i represents the covariant derivative with respect to the spatial background metric γ_{ij}, and that Latin indices $(i, j, \dots)$ are raised and lowered by means of γ_{ij}. The vectors $\overline{B}_i$ and $\overline{E}_i$ are divergenceless ($D_i\overline{B}^i = D_i\overline{E}^i = 0$), and so is the tensor $\overline{E}_{ij}$, which is also traceless ($D_i\overline{E}^{ij} = 0$ and $\overline{E}^i{}_i = 0$).

At the same linear order in perturbations, the inverse metric is

$$\begin{aligned}
\delta g^{00} &= \frac{2A}{a^2}, \\
\delta g^{0i} &= \frac{1}{a^2} \left(D^i B + \overline{B}^i \right), \\
\delta g^{ij} &= -\frac{1}{a^2} \left[2C\gamma^{ij} + 2D^i D^j E + 2D^{(i}\overline{E}^{j)} + 2\overline{E}^{ij} \right],
\end{aligned} \tag{C.2.2}$$

from which one derives the perturbed Christoffel symbols as

$$\begin{aligned}
\delta\Gamma^0_{00} &= A', \\
\delta\Gamma^i_{00} &= D^i \left(A + B' + \mathcal{H}B \right) + \overline{B}^{i\prime} + \mathcal{H}\overline{B}^i, \\
\delta\Gamma^0_{0i} &= D_i \left(A + \mathcal{H}B \right) + \mathcal{H}\overline{B}_i, \\
\delta\Gamma^0_{ij} &= -2\mathcal{H}A\gamma_{ij} + \mathcal{H}h_{ij} + \frac{1}{2}h'_{ij} - D_{(i}D_{j)}B - D_{(i}\overline{B}_{j)}, \\
\delta\Gamma^i_{0j} &= \frac{1}{2}h^{i\prime}{}_j + \gamma^{il}D_{[j}\overline{B}_{l]}, \\
\delta\Gamma^i_{jk} &= -\mathcal{H}\left(D^i B + \overline{B}^i \right)\gamma_{jk} + D_{(j}h^i{}_{k)} - \frac{1}{2}D^i h_{jk}.
\end{aligned} \tag{C.2.3}$$

The perturbed Christoffel symbols allow us to compute the perturbed Ricci tensor. This is

$$\begin{aligned}
\delta R_{00} &= 3\mathcal{H}A' + \Delta A + \Delta(B' + \mathcal{H}B) - \frac{1}{2}\left(h'' + \mathcal{H}h' \right), \\
\delta R_{0i} &= 2\mathcal{H}D_i A + \left(\mathcal{H}' + 2\mathcal{H}^2 \right)\overline{B}_i + D^k D_{[i}\overline{B}_{k]} + \frac{1}{2}\left(D_k h^k{}_i{}' - D_i h' \right), \\
\delta R_{ij} &= -\left[2\left(2\mathcal{H}^2 + \mathcal{H}' \right)A + \mathcal{H}A' + \mathcal{H}\Delta B \right]\gamma_{ij} - D_i D_j A - D_{(i}\left[\overline{B}'_{j)} + 2\mathcal{H}\overline{B}_{j)} \right] \\
&\quad + \left(2\mathcal{H}^2 + \mathcal{H}' \right)h_{ij} + \mathcal{H}h'_{ij} + \frac{1}{2}\mathcal{H}h'\gamma_{ij} + \frac{1}{2}h''_{ij} \\
&\quad + D_{(i|}D_k h^k{}_{|j)} - \frac{1}{2}\Delta h_{ij} - \frac{1}{2}D_i D_j h,
\end{aligned} \tag{C.2.4}$$

leading, upon contraction, to the curvature perturbation

$$a^2\delta R = h'' + 3\mathcal{H}h' - 4\Delta C - 2Kh - 2\Delta A - 12\left(\mathcal{H}' + \mathcal{H}^2 \right)A - 6\mathcal{H}A' - 2\Delta(B' + 3\mathcal{H}B), \tag{C.2.5}$$

where use has been made of the trace relation $h \equiv h^i{}_i = 6C + 2\Delta E$, and of the formula $\Delta h - D^{(i}D^{j)}h_{ij} = 4\Delta C$.

The perturbed Einstein tensor is then given by

$$\begin{aligned}
\delta G_{00} &= 6\mathcal{H}C' - 2\Delta C + 6K(A-C) - 2\mathcal{H}\Delta(B-E'),\\
\delta G_{0i} &= D_i\left[2\mathcal{H}A - 2C' + 2KE' - \left(2\mathcal{H}' + \mathcal{H}^2 + 3K\right)B\right]\\
&\quad + \frac{1}{2}\Delta\left(\overline{E}_i' - \overline{B}_i\right) + K\overline{E}_i' - \left(2\mathcal{H}' + \mathcal{H}^2 + 2K\right)\overline{B}_i,\\
\delta G_{ij} &= D_iD_j\left[E'' + 2\mathcal{H}E' - 2\left(2\mathcal{H}' + \mathcal{H}^2 + K\right)E - B' - 2\mathcal{H}B - C - A\right]\\
&\quad + \gamma_{ij}\left[-\Delta(E'' + 2\mathcal{H}E') + \Delta(B' + 2\mathcal{H}B) - 2C'' - 4\mathcal{H}C' + \Delta(C+A)\right.\\
&\qquad \left. + 2\left(2\mathcal{H}' + \mathcal{H}^2\right)(A-C) + 2\mathcal{H}A'\right]\\
&\quad + D_{(i}\left\{\left[\overline{E}_{j)}' - \overline{B}_{j)}\right]' + 2\mathcal{H}\left[\overline{E}_{j)}' - \overline{B}_{j)}\right] - 2\left(2\mathcal{H}' + \mathcal{H}^2 + K\right)\overline{E}_{j)}\right\}\\
&\quad + \overline{E}_{ij}'' + 2\mathcal{H}\overline{E}_{ij}' - 2\left(2\mathcal{H}' + \mathcal{H}^2\right)\overline{E}_{ij} - \Delta\overline{E}_{ij}.
\end{aligned} \tag{C.2.6}$$

We also have the mixed components:

$$\begin{aligned}
a^2\delta G^0{}_0 &= 2\left[3\mathcal{H}^2A - 3\mathcal{H}C' + \Delta\left(C + \mathcal{H}B - \mathcal{H}E'\right) + 3KC\right],\\
a^2\delta G^0{}_i &= -2D_i\left[\mathcal{H}A - C' - K(B-E')\right] - \frac{1}{2}[\Delta + 2K]\left(\overline{E}_i' - \overline{B}_i\right),\\
a^2\delta G^i{}_j &= D^iD_j\left[(E'-B)' + 2\mathcal{H}(E'-B) - (C+A)\right]\\
&\quad + \delta^i_j\left[-\Delta(E'-B)' - 2\mathcal{H}\Delta(E'-B) - 2C'' - 4\mathcal{H}C' + \Delta C + 2KC\right.\\
&\qquad \left. + 2\mathcal{H}A' + \Delta A + 2\left(2\mathcal{H}' + \mathcal{H}^2\right)A\right]\\
&\quad + \gamma^{ik}D_{(k}\left\{\left[\overline{E}_{j)}' - \overline{B}_{j)}\right]' + 2\mathcal{H}\left[\overline{E}_{j)}' - \overline{B}_{j)}\right]\right\}\\
&\quad + \overline{E}'^i{}_j + 2\mathcal{H}\overline{E}^i{}_j - \Delta\overline{E}^i{}_j + 2K\overline{E}^i{}_j,
\end{aligned} \tag{C.2.7}$$

whereas the perturbed curvature of the spatial sections is

$$\delta^{(3)}R = -\frac{4}{a^2}(\Delta + 3K)C, \tag{C.2.8}$$

and the extrinsic curvature tensor for the hypersurfaces of constant conformal time is

$$K^i{}_j = -\frac{1}{2}\left[\mathcal{H}(1+\mathcal{K})\delta^i_j + \left(D^iD_j - \frac{1}{3}\delta^i_j\Delta\right)\sigma - 2D^{(i}\overline{\sigma}_{j)} - \overline{\sigma}^i{}_j\right], \tag{C.2.9}$$

with

$$\mathcal{K} = -A + \frac{1}{\mathcal{H}}\left(C' + \frac{1}{3}\Delta\sigma\right), \qquad \sigma = E' - B, \qquad \overline{\sigma}^i = \overline{\Phi}^i, \qquad \overline{\sigma}^i{}_j = \overline{E}^i{}_j. \tag{C.2.10}$$

The gauge-invariant quantities associated to the geometry are summarized in Table C.1.

C.2.2 Matter

The components of the stress-energy tensor for a perfect fluid, to first order, are

$$\begin{aligned}
\delta T_{00} &= \rho a^2(\delta + 2A),\\
\delta T_{0i} &= -\rho a^2\left[(1+w)\left(D_iv + \overline{v}_i\right) + D_iB + \overline{B}_i\right],\\
\delta T_{ij} &= Pa^2\left(h_{ij} + \frac{\delta P}{P}\gamma_{ij} + \pi_{ij}\right),
\end{aligned} \tag{C.2.11}$$

Table C.1: Gauge-invariant quantities associated to cosmological geometry.

Sector	**Quantity**	**Definition**
Scalar	Ψ	$= -C - \mathcal{H}(B - E') = -C + \mathcal{H}\sigma$
	Φ	$= A + \mathcal{H}(B - E') + (B - E')' = A - \mathcal{H}\sigma + \sigma'$
	$\mathcal{R}$	$= C - \mathcal{H}\dfrac{\delta\phi}{\phi'}$
	ζ	$= \Phi + \dfrac{2}{3}\dfrac{\Phi' + \mathcal{H}\Phi}{\mathcal{H}(1+w)}$
	ζ_{BST}	$= -C + \dfrac{\delta\rho}{3(\rho + P)}$
Vector	$\overline{\Phi}_i$	$= \overline{E}_i' - \overline{B}_i = \overline{\sigma}_i$
Tensor	$\overline{E}_{ij}$	directly gauge-invariant

or equivalently

$$\begin{aligned} \delta T^0{}_0 &= -\rho\delta, \\ \delta T^0{}_i &= \rho(1+w)\left(D_i v + \overline{v}_i + D_i B + \overline{B}_i\right), \\ \delta T^i{}_j &= \delta P \delta^i_j + P\pi^i{}_j, \end{aligned} \tag{C.2.12}$$

where the anisotropic stress tensor can be SVT-decomposed as

$$\pi_{ij} = \left(D_i D_j - \frac{1}{3}\gamma_{ij}\Delta\right)\pi + D_{(i}\overline{\pi}_{j)} + \overline{\pi}_{ij}. \tag{C.2.13}$$

Similarly, the scalar pressure perturbation is expanded into adiabatic and entropic parts as

$$\delta P = c_{\mathrm{s}}^2 \delta\rho + P\Gamma, \tag{C.2.14}$$

from which one deduces that

$$\delta w = \left(c_{\mathrm{s}}^2 - w\right)\delta + \Gamma w. \tag{C.2.15}$$

In the special case of a scalar field $\phi(\eta, \boldsymbol{x}) = \phi_0(\eta) + \delta\phi(\eta, \boldsymbol{x})$, we obtain

$$\begin{aligned} \delta T_{00} &= \phi'\delta\phi' + 2a^2 V(\phi) A + a^2 \frac{\mathrm{d}V}{\mathrm{d}\phi}\delta\phi, \\ \delta T_{0i} &= D_i\left[\phi'\delta\phi + \left(\frac{1}{2}\phi'^2 - a^2 V\right)B\right] + \left(\frac{1}{2}\phi'^2 - a^2 V\right)\overline{B}_i, \\ \delta T_{ij} &= \left(\phi'\delta\phi' - \phi'^2 A - a^2\frac{\mathrm{d}V}{\mathrm{d}\phi}\delta\phi\right)\gamma_{ij} + \left(\frac{1}{2}\phi'^2 - a^2 V\right)h_{ij}, \end{aligned} \tag{C.2.16}$$

and therefore

$$\begin{aligned} a^2\delta T^0{}_0 &= -\phi'\delta\phi' - a^2\frac{\mathrm{d}V}{\mathrm{d}\phi}\delta\phi + A\phi'^2, \\ a^2\delta T^0{}_i &= -D_i\left(\phi'\delta\phi\right), \\ a^2\delta T^i{}_j &= \left(\phi'\delta\phi' - \phi'^2 A - a^2\frac{\mathrm{d}V}{\mathrm{d}\phi}\delta\phi\right)\delta^i_j. \end{aligned} \tag{C.2.17}$$

All the gauge-invariant quantities appearing in the description of matter are summarized in Table C.2.

Table C.2: Gauge-invariant quantities appearing in cosmological matter and scalar field descriptions.

Sector	Quantity	Definition
Scalar	δ^{N}	$= \delta + \dfrac{\rho'}{\rho}(B - E')$
	δ^{F}	$= \delta - \rho' \dfrac{C}{\mathcal{H}}$
	δ^{C}	$= \delta + \dfrac{\rho'}{\rho}(v + B)$
	V	$= v + E'$
	π	scalar anisotropic stress
	Γ	intrinsic entropy perturbation
Vector	$\overline{V}_i$	$= \overline{v}_i + \overline{B}_i$
	$\overline{\pi}_i$	vector anisotropic stress
Tensor	$\overline{\pi}_{ij}$	tensor anisotropic stress
Scalar field	χ	$= \delta\phi + \phi'(B - E')$
	Q	$= \delta\phi - \dfrac{\phi' C}{\mathcal{H}} = -\phi' \dfrac{\mathcal{R}}{\mathcal{H}}$

Consider the metric perturbation $g_{\mu\nu} = \bar{g}_{\mu\nu} + \delta g_{\mu\nu}$ with $\delta g_{\mu\nu}$ given by (C.2.1). The variation of the Christoffel symbols to first order is given by the standard Palatini identity:

$$\delta\Gamma^{\alpha}_{\mu\nu} = \frac{1}{2} g^{\alpha\beta} \left(\nabla_\mu \delta g_{\nu\beta} + \nabla_\nu \delta g_{\mu\beta} - \nabla_\beta \delta g_{\mu\nu} \right). \tag{1}$$

Evaluating the $(0,0,0)$ component:

$$\delta\Gamma^{0}_{00} = \frac{1}{2} g^{00} \left(\partial_0 \delta g_{00} + \partial_0 \delta g_{00} - \partial_0 \delta g_{00} \right) - \Gamma^{0}_{00} \frac{\delta g_{00}}{g_{00}} = \frac{1}{2(-a^2)} \left(-2a^2 A \right)' - \mathcal{H}(2A) = A'. \tag{2}$$

Similarly, evaluating the $(0,i,j)$ component:

$$\delta\Gamma^{0}_{ij} = \frac{1}{2} g^{00} \left(\partial_i \delta g_{j0} + \partial_j \delta g_{i0} - \partial_0 \delta g_{ij} \right) + \Gamma^{\alpha}_{ij} \delta g^{0}_{\alpha}. \tag{3}$$

Substituting $\delta g_{0i} = a^2 \mathcal{B}_i$ and $\delta g_{ij} = a^2 h_{ij}$ along with background connections yields

$$\delta\Gamma^{0}_{ij} = -2\mathcal{H} A \gamma_{ij} + \mathcal{H} h_{ij} + \frac{1}{2} h'_{ij} - D_{(i} D_{j)} B - D_{(i} \overline{B}_{j)}, \tag{4}$$

matching the fourth line of (C.2.3). The remaining components follow by straightforward substitution of $g^{\mu\nu}$ and covariant spatial derivatives.
To derive the spatial scalar curvature perturbation (C.2.8), let $\tilde{\gamma}_{ij} = \gamma_{ij} + 2C\gamma_{ij} + 2D_i D_j E$. Under an infinitesimal conformal transformation of the 3-metric $\tilde{\gamma}_{ij} = \mathrm{e}^{2C}(\gamma_{ij} + 2D_i D_j E)$, the spatial scalar curvature transforms as

$$^{(3)}\tilde{R} = \mathrm{e}^{-2C} \left[{}^{(3)}R - 4\Delta C - 2\gamma^{ij} D_i C D_j C \right]. \tag{5}$$

Expanding to first order in C and noting that the purely longitudinal gauge part D_iD_jE does not alter intrinsic 3-curvature at linear order:

$$ {}^{(3)}\tilde{R} = (1-2C)\,{}^{(3)}R - 4\Delta C = {}^{(3)}R - 2C(6K) - 4\Delta C = {}^{(3)}R - 4(\Delta + 3K)C. \tag{6} $$

Since the 4-dimensional spatial sections have induced metric $q_{ij} = a^2\tilde{\gamma}_{ij}$, the physical 3-curvature is scaled by a^{-2}:

$$ \delta^{(3)}R = -\frac{4}{a^2}(\Delta + 3K)C, \tag{7} $$

which is precisely Eq. (C.2.8).

Appendix D

Monte Carlo Methods in CMB Lensing Reconstruction

D.1 Introduction and Observational Motivation

In observational cosmology, theoretical derivations frequently isolate formal expectation values and cumulant decompositions that cannot be measured directly from a single observed sky. In real data, the observer possesses only a single realization of the cosmos, wherein cosmological signals, instrumental noise, and higher-order Wick contractions are inextricably entangled. End-to-end Monte Carlo simulations provide an indispensable mathematical vehicle to bridge this divide: by constructing, pairing, and filtering controlled synthetic realizations, one can isolate and evaluate individual statistical quantities that are otherwise impossible to disentangle directly from observational data alone.

Gravitational lensing reconstruction of the Cosmic Microwave Background (CMB) provides a quintessential instance of this principle. In Section 8.1.6 and Section 8.1.6, we established the theoretical foundations of CMB lensing reconstruction under idealized flat-sky and full-sky assumptions. In that analytical treatment, the four-point correlation function (trispectrum) of the lensed CMB decomposes into the primary deflection power spectrum $C_L^{\phi\phi}$, a Gaussian disconnected noise floor $N_L^{(0)}$, and a secondary mode-coupling bias $N_L^{(1)}$. In an idealized, statistically isotropic universe with homogeneous detector noise, both $N_L^{(0)}$ and $N_L^{(1)}$ can be computed via two-dimensional semi-analytic integrals over the unlensed or lensed CMB power spectra.

In realistic cosmological experiments, however, several observational complications break translational and rotational invariance across the celestial sphere:

1. **Survey geometry and sky masks:** Ground-based and space-borne CMB telescopes observe through finite apertures and must mask out galactic synchrotron and dust emission, bright point sources, and dusty star-forming galaxies. The observed sky is multiplied by a spatial window function $W(\hat{\boldsymbol{n}})$, which convolves Fourier multipoles and breaks statistical isotropy.

2. **Inhomogeneous and anisotropic noise:** Due to scan patterns, cross-linking strategies, and atmospheric fluctuations, detector noise varies strongly across pixels and exhibits spatial correlations ($\mathbf{N}_{pp'} \neq \sigma_0^2 \delta_{pp'}$).

3. **Map filtering and transfer functions:** Time-ordered data processing, ground-pickup filtering, and beam deconvolution introduce anisotropic filter transfer functions that cannot be diagonalized in harmonic space.

Under these non-ideal conditions, evaluating the disconnected noise floor $N_L^{(0)}$ and the mode-coupling bias $N_L^{(1)}$ via idealized isotropic formulas leads to severe systematic errors. Modern cosmological pipelines—including those of the *Planck* satellite, the Atacama Cosmology Telescope (ACT), the South Pole Telescope (SPT), and the Simons Observatory—therefore evaluate these reconstruction biases directly through end-to-end Monte Carlo simulation realizations.

The purpose of this appendix is to derive the exact mathematical foundation of this Monte Carlo realization framework from the conditional Wick expansion [Eq. (8.1.283)]. We demonstrate that the same-sky disconnected Gaussian noise floor $N_L^{(0)}$ can be evaluated term-by-term through cross-paired independent simulation realizations, establishing the master observational identity:

$$\boxed{C_L^{\hat{\phi}_{\alpha\beta}\hat{\phi}_{\alpha\beta}} = C_L^{q_{ab}^{\rm ind} q_{ab}^{\rm ind}} + C_L^{q_{ab}^{\rm ind} q_{ba}^{\rm ind}} + C_L^{\phi\phi} + N_L^{(1),\alpha\beta} + \mathcal{O}((C^{\phi\phi})^2).} \tag{D.1.1}$$

D.2 The Same-Sky Quadratic Estimator and its Trispectrum

Let $\tilde{X}_\alpha(\boldsymbol{\ell})$ denote the observed lensed CMB Fourier multipoles in polarization channels $\alpha, \beta \in \{T, E, B\}$. On the observed sky, the normalized quadratic estimator for the lensing deflection potential Fourier mode $\boldsymbol{L}$ is defined explicitly by:

$$\hat{\phi}_{\alpha\beta}(\boldsymbol{L}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L}) \tilde{X}_\alpha(\boldsymbol{\ell}_1) \tilde{X}_\beta(\boldsymbol{L} - \boldsymbol{\ell}_1) - \langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L}) \rangle_{\rm MF}, \tag{D.2.1}$$

where $W_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L}) \equiv N_L^{\alpha\beta} g^{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L} - \boldsymbol{\ell}_1)$ incorporates the optimal minimum-variance filter and normalization factor, and $\langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L}) \rangle_{\rm MF}$ is the mean-field bias sourced by sky masks and anisotropic noise.

To streamline the algebra, we adopt the condensed multi-index notation where each index denotes both the polarization channel and the corresponding wavevector:

$$i \equiv (\alpha_i, \boldsymbol{\ell}_i). \tag{D.2.2}$$

In this notation, the four legs of the auto-covariance $\langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L}) \hat{\phi}^*_{\alpha\beta}(\boldsymbol{L}') \rangle$ correspond to:

$$1 \equiv (\alpha, \boldsymbol{\ell}_1), \qquad 2 \equiv (\beta, \boldsymbol{L} - \boldsymbol{\ell}_1), \qquad 3 \equiv (\alpha, \boldsymbol{\ell}_3), \qquad 4 \equiv (\beta, \boldsymbol{L}' - \boldsymbol{\ell}_3). \tag{D.2.3}$$

Writing out the auto-covariance of the mean-field-subtracted quadratic estimator explicitly gives:

$$\begin{aligned} \langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L}) \hat{\phi}^*_{\alpha\beta}(\boldsymbol{L}') \rangle &= \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} \frac{\mathrm{d}^2\boldsymbol{\ell}_3}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L}) W^*_{\alpha\beta}(\boldsymbol{\ell}_3, \boldsymbol{L}') \\ &\quad \times \langle \tilde{X}_\alpha(\boldsymbol{\ell}_1) \tilde{X}_\beta(\boldsymbol{L} - \boldsymbol{\ell}_1) \tilde{X}^*_\alpha(\boldsymbol{\ell}_3) \tilde{X}^*_\beta(\boldsymbol{L}' - \boldsymbol{\ell}_3) \rangle_c \\ &= \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} \frac{\mathrm{d}^2\boldsymbol{\ell}_3}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L}) W^*_{\alpha\beta}(\boldsymbol{\ell}_3, \boldsymbol{L}') \langle \tilde{X}_1 \tilde{X}_2 \tilde{X}^*_3 \tilde{X}^*_4 \rangle_c, \end{aligned} \tag{D.2.4}$$

where $\langle \cdots \rangle_c$ denotes the connected cumulant with respect to the mean field.

Conditioned on a fixed realization of the intervening gravitational deflection field ϕ, the primary CMB fluctuations remain Gaussian random fields. Wick's theorem [Eq. (8.1.283)] states that the conditional four-point expectation value decomposes into three distinct pairings of conditional two-point functions:

$$\begin{aligned} \langle \tilde{X}_1 \tilde{X}_2 \tilde{X}^*_3 \tilde{X}^*_4 \mid \phi \rangle = &\underbrace{\langle \tilde{X}_1 \tilde{X}_2 \mid \phi \rangle \langle \tilde{X}^*_3 \tilde{X}^*_4 \mid \phi \rangle}_{\rm I} \\ &+ \underbrace{\langle \tilde{X}_1 \tilde{X}^*_3 \mid \phi \rangle \langle \tilde{X}_2 \tilde{X}^*_4 \mid \phi \rangle}_{\rm II} + \underbrace{\langle \tilde{X}_1 \tilde{X}^*_4 \mid \phi \rangle \langle \tilde{X}_2 \tilde{X}^*_3 \mid \phi \rangle}_{\rm III}. \end{aligned} \tag{D.2.5}$$

Averaging over the cosmological ensemble of all intervening lens realizations, $\langle\cdots\rangle = \langle\langle\cdots \mid \phi\rangle\rangle_\phi$, yields the full unconditional four-point function:

$$\langle \tilde{X}_1\tilde{X}_2\tilde{X}_3^*\tilde{X}_4^*\rangle = \Big\langle \langle \tilde{X}_1\tilde{X}_2 \mid \phi\rangle\langle \tilde{X}_3^*\tilde{X}_4^* \mid \phi\rangle\Big\rangle_\phi + \Big\langle \langle \tilde{X}_1\tilde{X}_3^* \mid \phi\rangle\langle \tilde{X}_2\tilde{X}_4^* \mid \phi\rangle\Big\rangle_\phi + \Big\langle \langle \tilde{X}_1\tilde{X}_4^* \mid \phi\rangle\langle \tilde{X}_2\tilde{X}_3^* \mid \phi\rangle\Big\rangle_\phi . \tag{D.2.6}$$

D.3 Separation of Disconnected and Connected Four-Point Statistics

For any two random variables A and B, the expectation value of their product decomposes into the product of their means plus their covariance:

$$\langle AB\rangle = \langle A\rangle\langle B\rangle + \big[\langle AB\rangle - \langle A\rangle\langle B\rangle\big]. \tag{D.3.1}$$

We now apply this identity to each of the three conditional covariance products in Eq. (D.2.6). Recalling that $\langle\langle \tilde{X}_i\tilde{X}_j^* \mid \phi\rangle\rangle_\phi = \langle \tilde{X}_i\tilde{X}_j^*\rangle$ is the unconditional two-point covariance of the lensed CMB (since $\langle\phi\rangle_\phi = 0$), the unconditional four-point function partitions cleanly into a disconnected Gaussian piece and a connected non-Gaussian cumulant:

$$\langle \tilde{X}_1\tilde{X}_2\tilde{X}_3^*\tilde{X}_4^*\rangle = \langle \tilde{X}_1\tilde{X}_2\tilde{X}_3^*\tilde{X}_4^*\rangle_{\rm disc} + \langle \tilde{X}_1\tilde{X}_2\tilde{X}_3^*\tilde{X}_4^*\rangle_{\rm conn}. \tag{D.3.2}$$

The disconnected four-point function collects the products of unconditional two-point functions:

$$\langle \tilde{X}_1\tilde{X}_2\tilde{X}_3^*\tilde{X}_4^*\rangle_{\rm disc} = \langle \tilde{X}_1\tilde{X}_2\rangle\langle \tilde{X}_3^*\tilde{X}_4^*\rangle + \langle \tilde{X}_1\tilde{X}_3^*\rangle\langle \tilde{X}_2\tilde{X}_4^*\rangle + \langle \tilde{X}_1\tilde{X}_4^*\rangle\langle \tilde{X}_2\tilde{X}_3^*\rangle. \tag{D.3.3}$$

The physical roles of these three disconnected terms are distinct:

- **Mean-field term:** The first term $\langle \tilde{X}_1\tilde{X}_2\rangle\langle \tilde{X}_3^*\tilde{X}_4^*\rangle$ represents the square of the mean field $\langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L})\rangle_{\rm MF}$. On a statistically isotropic sky, it vanishes for all $\boldsymbol{L} \neq \boldsymbol{0}$. On a cut sky, it is explicitly subtracted at the estimator level [Eq. (D.2.1)] and does not contribute to the connected covariance.
- **Direct and cross disconnected contractions:** The remaining two terms connect legs across the two quadratic estimators:

$$\text{Direct contraction (from pairing II):}\quad \langle \tilde{X}_1\tilde{X}_3^*\rangle\langle \tilde{X}_2\tilde{X}_4^*\rangle, \tag{D.3.4}$$

$$\text{Cross contraction (from pairing III):}\quad \langle \tilde{X}_1\tilde{X}_4^*\rangle\langle \tilde{X}_2\tilde{X}_3^*\rangle. \tag{D.3.5}$$

Substituting the direct and cross contractions into Eq. (D.2.4) defines the disconnected reconstruction noise floor:

$$\langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L})\hat{\phi}^*_{\alpha\beta}(\boldsymbol{L}')\rangle_{\rm disc} = \int \frac{{\rm d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\frac{{\rm d}^2\boldsymbol{\ell}_3}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1,\boldsymbol{L})W^*_{\alpha\beta}(\boldsymbol{\ell}_3,\boldsymbol{L}')\Big[\langle \tilde{X}_1\tilde{X}_3^*\rangle\langle \tilde{X}_2\tilde{X}_4^*\rangle + \langle \tilde{X}_1\tilde{X}_4^*\rangle\langle \tilde{X}_2\tilde{X}_3^*\rangle\Big]. \tag{D.3.6}$$

By spatial translation invariance, this disconnected covariance is proportional to $(2\pi)^2\delta^{(2)}(\boldsymbol{L}-\boldsymbol{L}')$, isolating the Gaussian reconstruction noise floor $N_L^{(0),\alpha\beta}$ [Eq. (8.1.287)]:

$$\langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L})\hat{\phi}^*_{\alpha\beta}(\boldsymbol{L}')\rangle_{\rm disc} = (2\pi)^2\delta^{(2)}(\boldsymbol{L}-\boldsymbol{L}')N_L^{(0),\alpha\beta}. \tag{D.3.7}$$

D.4 Realization Independence and the Mixed-Sky Formulation

We now address the central conceptual question: *Why can the same-sky disconnected noise floor $N_L^{(0)}$ be evaluated using distinct, independent simulation realizations?*

The answer lies in the factorization structure of Eq. (D.3.6): the disconnected noise $N_L^{(0)}$ is entirely determined by products of unconditional two-point functions $\langle \tilde{X}_i \tilde{X}_j^* \rangle$. In statistical cosmology, the two-point function is an ensemble expectation value characteristic of the underlying lensed cosmological field. It does not depend on the particular phase realization of the observer's sky.

Consequently, if we draw two statistically independent realizations a and b from the same underlying lensed CMB ensemble:

$$\tilde{X}^{(a)}, \qquad \tilde{X}^{(b)}, \tag{D.4.1}$$

each realization reproduces the exact unconditional two-point covariance of the observed data:

$$\langle \tilde{X}_i^{(a)} \tilde{X}_j^{(a)*} \rangle = \langle \tilde{X}_i \tilde{X}_j^* \rangle, \qquad \langle \tilde{X}_i^{(b)} \tilde{X}_j^{(b)*} \rangle = \langle \tilde{X}_i \tilde{X}_j^* \rangle, \tag{D.4.2}$$

while their inter-realization cross-covariance vanishes identically:

$$\langle \tilde{X}_i^{(a)} \tilde{X}_j^{(b)*} \rangle = 0. \tag{D.4.3}$$

D.4.1 Cross-Paired Estimator Construction with Realization Annotations

To exploit this factorization, we construct two cross-paired quadratic estimators where the two legs are evaluated on different realizations. Explicitly:

$$q_{ab}(\boldsymbol{L}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L})\, \underbrace{\tilde{X}_\alpha^{(a)}(\boldsymbol{\ell}_1)}_{\text{drawn from realization } (a)}\, \overbrace{\tilde{X}_\beta^{(b)}(\boldsymbol{L}-\boldsymbol{\ell}_1)}^{\text{drawn from realization } (b)}. \tag{D.4.4}$$

In condensed multi-index notation, this is written concisely as:

$$q_{ab}(\boldsymbol{L}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L}) \tilde{X}_1^{(a)} \tilde{X}_2^{(b)}. \tag{D.4.5}$$

Similarly, swapping the realization assignment between legs yields the reversed estimator:

$$q_{ba}(\boldsymbol{L}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L})\, \underbrace{\tilde{X}_\alpha^{(b)}(\boldsymbol{\ell}_1)}_{\text{drawn from realization } (b)}\, \overbrace{\tilde{X}_\beta^{(a)}(\boldsymbol{L}-\boldsymbol{\ell}_1)}^{\text{drawn from realization } (a)}. \tag{D.4.6}$$

In condensed notation, $q_{ba}(\boldsymbol{L}) = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L}) \tilde{X}_1^{(b)} \tilde{X}_2^{(a)}$.

Core Physical Principle: Exact Vanishing of Cosmological Signal

Because realizations a and b are statistically independent draws with independent primary CMB perturbations and independent lensing potentials ($\phi_a \neq \phi_b$), there is no common gravitational deflection between the two maps. The ensemble expectation value of both cross-paired estimators vanishes identically:

$$\langle q_{ab}(\boldsymbol{L}) \rangle = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1, \boldsymbol{L}) \langle \tilde{X}_1^{(a)} \tilde{X}_2^{(b)} \rangle = 0, \qquad \langle q_{ba}(\boldsymbol{L}) \rangle = 0. \tag{1}$$

Crucially, the true cosmological lensing potential spectrum $C_L^{\phi\phi}$ is completely unaffected

and drops out of the expectation value. The power spectra of q_{ab} and q_{ba} are purely the sample variance of these zero-mean estimators, which isolates the Gaussian disconnected noise floor $N_L^{(0)}$ with zero leakage from $C_L^{\phi\phi}$.

D.4.2 The Mixed Auto-Spectrum: Direct Disconnected Contraction

The auto-covariance of $q_{ab}(\boldsymbol{L})$ is:

$$\langle q_{ab}(\boldsymbol{L}) q_{ab}^*(\boldsymbol{L}')\rangle = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\frac{\mathrm{d}^2\boldsymbol{\ell}_3}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1,\boldsymbol{L}) W^*_{\alpha\beta}(\boldsymbol{\ell}_3,\boldsymbol{L}')\langle \tilde{X}_1^{(a)}\tilde{X}_2^{(b)}\tilde{X}_3^{(a)*}\tilde{X}_4^{(b)*}\rangle. \tag{D.4.7}$$

Because skies a and b are completely independent draws, the expectation value factorizes across realizations:

$$\langle \tilde{X}_1^{(a)}\tilde{X}_2^{(b)}\tilde{X}_3^{(a)*}\tilde{X}_4^{(b)*}\rangle = \langle \tilde{X}_1^{(a)}\tilde{X}_3^{(a)*}\rangle\langle \tilde{X}_2^{(b)}\tilde{X}_4^{(b)*}\rangle = \langle \tilde{X}_1\tilde{X}_3^*\rangle\langle \tilde{X}_2\tilde{X}_4^*\rangle. \tag{D.4.8}$$

All other pairings vanish identically because $\langle \tilde{X}^{(a)}\tilde{X}^{(b)*}\rangle = 0$. Substituting Eq. (D.4.8) into the covariance integral yields:

$$\langle q_{ab}(\boldsymbol{L}) q_{ab}^*(\boldsymbol{L}')\rangle_{\mathrm{ind}} = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\frac{\mathrm{d}^2\boldsymbol{\ell}_3}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1,\boldsymbol{L}) W^*_{\alpha\beta}(\boldsymbol{\ell}_3,\boldsymbol{L}')\langle \tilde{X}_1\tilde{X}_3^*\rangle\langle \tilde{X}_2\tilde{X}_4^*\rangle. \tag{D.4.9}$$

Comparing Eq. (D.4.9) with Eq. (D.3.6), we see that the auto-spectrum of q_{ab} reproduces the first (direct) disconnected contraction of the same sky with mathematical exactness:

$$\langle q_{ab}(\boldsymbol{L}) q_{ab}^*(\boldsymbol{L}')\rangle_{\mathrm{ind}} \longleftrightarrow \text{direct contraction } \langle \tilde{X}_1\tilde{X}_3^*\rangle\langle \tilde{X}_2\tilde{X}_4^*\rangle. \tag{D.4.10}$$

D.4.3 The Mixed Cross-Spectrum: Cross Disconnected Contraction

Next, evaluate the cross-covariance between $q_{ab}(\boldsymbol{L})$ and $q_{ba}(\boldsymbol{L})$:

$$\langle q_{ab}(\boldsymbol{L}) q_{ba}^*(\boldsymbol{L}')\rangle = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\frac{\mathrm{d}^2\boldsymbol{\ell}_3}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1,\boldsymbol{L}) W^*_{\alpha\beta}(\boldsymbol{\ell}_3,\boldsymbol{L}')\langle \tilde{X}_1^{(a)}\tilde{X}_2^{(b)}\tilde{X}_3^{(b)*}\tilde{X}_4^{(a)*}\rangle. \tag{D.4.11}$$

In this cross-pairing, realization a enters legs 1 and 4, while realization b enters legs 2 and 3. By statistical independence:

$$\langle \tilde{X}_1^{(a)}\tilde{X}_2^{(b)}\tilde{X}_3^{(b)*}\tilde{X}_4^{(a)*}\rangle = \langle \tilde{X}_1^{(a)}\tilde{X}_4^{(a)*}\rangle\langle \tilde{X}_2^{(b)}\tilde{X}_3^{(b)*}\rangle = \langle \tilde{X}_1\tilde{X}_4^*\rangle\langle \tilde{X}_2\tilde{X}_3^*\rangle. \tag{D.4.12}$$

Substituting Eq. (D.4.12) into the covariance integral gives:

$$\langle q_{ab}(\boldsymbol{L}) q_{ba}^*(\boldsymbol{L}')\rangle_{\mathrm{ind}} = \int \frac{\mathrm{d}^2\boldsymbol{\ell}_1}{(2\pi)^2}\frac{\mathrm{d}^2\boldsymbol{\ell}_3}{(2\pi)^2} W_{\alpha\beta}(\boldsymbol{\ell}_1,\boldsymbol{L}) W^*_{\alpha\beta}(\boldsymbol{\ell}_3,\boldsymbol{L}')\langle \tilde{X}_1\tilde{X}_4^*\rangle\langle \tilde{X}_2\tilde{X}_3^*\rangle. \tag{D.4.13}$$

Comparing Eq. (D.4.13) with Eq. (D.3.6), the cross-spectrum between q_{ab} and q_{ba} reproduces the second (cross) disconnected contraction of the same sky with mathematical exactness:

$$\langle q_{ab}(\boldsymbol{L}) q_{ba}^*(\boldsymbol{L}')\rangle_{\mathrm{ind}} \longleftrightarrow \text{cross contraction } \langle \tilde{X}_1\tilde{X}_4^*\rangle\langle \tilde{X}_2\tilde{X}_3^*\rangle. \tag{D.4.14}$$

D.4.4 Complete Exact Recovery of $N_L^{(0)}$ from Independent Skies

Adding Eq. (D.4.9) and Eq. (D.4.13) yields the complete same-sky disconnected covariance:

$$\langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L})\hat{\phi}^*_{\alpha\beta}(\boldsymbol{L}')\rangle_{\mathrm{disc}} = \langle q_{ab}(\boldsymbol{L}) q_{ab}^*(\boldsymbol{L}')\rangle_{\mathrm{ind}} + \langle q_{ab}(\boldsymbol{L}) q_{ba}^*(\boldsymbol{L}')\rangle_{\mathrm{ind}}. \tag{D.4.15}$$

Converting from covariances to angular power spectra via $\langle A(\boldsymbol{L})B^*(\boldsymbol{L}')\rangle = (2\pi)^2\delta^{(2)}(\boldsymbol{L}-\boldsymbol{L}')C_L^{AB}$, we arrive at the foundational Monte Carlo result:

$$\boxed{N_L^{(0),\alpha\beta} = C_L^{q_{ab}^{\mathrm{ind}} q_{ab}^{\mathrm{ind}}} + C_L^{q_{ab}^{\mathrm{ind}} q_{ba}^{\mathrm{ind}}}.} \tag{D.4.16}$$

Theorem: Exactness of the Monte Carlo $N_L^{(0)}$ Construction
Equation (D.4.16) is an *exact algebraic identity* for the disconnected Gaussian reconstruction noise floor at the ensemble level, subject only to the condition that the independent simulation realizations reproduce the unconditional two-point covariance of the lensed CMB [Eq. (D.4.2)].
Furthermore, it holds universally for all quadratic estimators, without requiring symmetry under leg exchange:

1. **Symmetric auto-estimators ($q_{ab} = q_{ba}$):** For auto-estimators such as TT or EE, swapping leg indices leaves the weighting kernel invariant, $W_{\alpha\alpha}(\boldsymbol{\ell}_1, \boldsymbol{L}) = W_{\alpha\alpha}(\boldsymbol{L} - \boldsymbol{\ell}_1, \boldsymbol{L})$. In that case, $C_L^{q_{ab}^{\rm ind} q_{ba}^{\rm ind}} = C_L^{q_{ab}^{\rm ind} q_{ab}^{\rm ind}}$, and the noise floor simplifies to the familiar factor-of-two formula:
$$N_L^{(0),\rm sym} = 2C_L^{q_{ab}^{\rm ind} q_{ab}^{\rm ind}}. \tag{1}$$
2. **General asymmetric cross-estimators ($q_{ab} \neq q_{ba}$):** For general cross-estimators such as unsymmetrized EB (where $f_{EB}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) \neq f_{BE}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_1)$), the auto-spectrum and cross-spectrum evaluate the direct and cross contractions with different filter weights. Retaining both terms in Eq. (D.4.16) is essential to correctly evaluate $N_L^{(0)}$.

D.5 Algorithmic Pipeline Implementation: Paired Simulation Execution

In practical CMB experiments, the ensemble expectation values are evaluated over a finite bank of $N_{\rm sim}$ mock simulations generated through the complete instrument pipeline. We now describe the standard algorithmic architecture used to compute $N_L^{(0)}$ and $N_L^{(1)}$ in $N_{\rm sim}/2$ paired realization steps.

D.5.1 Partitioning into $N_{\rm sim}/2$ Disjoint Realization Pairs

Suppose an experiment produces a library of $N_{\rm sim}$ independent mock sky realizations (typically $N_{\rm sim} \sim 400$–1000). While one could theoretically form all $N_{\rm sim}(N_{\rm sim} - 1)$ off-diagonal pairs, this is computationally prohibitive and introduces intricate statistical correlations between simulation pairs.

Instead, modern cosmological pipelines partition the $N_{\rm sim}$ simulations into exactly $N_{\rm sim}/2$ disjoint realization pairs (a, b). The entire pipeline executes in $N_{\rm sim}/2$ sequential or parallel steps, drawing two independent realizations at a time: realization (a) and realization (b).

D.5.2 Step-by-Step Paired Accumulation

For each disjoint pair (a, b), the pipeline performs three straightforward operations:

1. **Pairwise estimator evaluation:** Filter realizations a and b through the quadratic estimator weights to form the forward and reversed estimators $q_{ab}(\boldsymbol{L})$ and $q_{ba}(\boldsymbol{L})$ [Eqs. (D.4.4) and (D.4.6)].
2. **Pairwise power spectra:** Compute the auto- and cross-power spectra $C_L^{q_{ab}q_{ab}}$ and $C_L^{q_{ab}q_{ba}}$.
3. **Pairwise noise floor:** Sum the two spectra to form the pair's unbiased estimate of the disconnected noise floor, $C_L^{q_{ab}q_{ab}} + C_L^{q_{ab}q_{ba}}$.

Because realizations a and b are independent draws with distinct deflection fields ($\phi_a \neq \phi_b$), the true cosmological lensing spectrum $C_L^{\phi\phi}$ is completely unaffected and drops out identically from every pair ($\langle q_{ab} \rangle = 0$).

Averaging over all $N_{\rm sim}/2$ pairs yields the final Monte Carlo estimate of the Gaussian noise floor:

$$\boxed{N_L^{(0),\alpha\beta} = \frac{1}{N_{\rm sim}/2} \sum_{\text{pairs } (a,b)} \left[C_L^{q_{ab}^{\rm ind} q_{ab}^{\rm ind}} + C_L^{q_{ab}^{\rm ind} q_{ba}^{\rm ind}} \right].} \tag{D.5.1}$$

D.5.3 Statistical Convergence and Variance Scaling

The $N_{\rm sim}/2$ paired realization workflow provides three fundamental mathematical guarantees:

1. **Unbiased cosmological signal removal:** Because $\langle q_{ab} \rangle = 0$ for every disjoint pair, the true sky power spectrum $C_L^{\phi\phi}$ cannot leak into the estimate. The subtraction leaves the cosmological lensing signal strictly unperturbed.

2. **Mutually independent samples:** Because the pairs share no common simulations, each paired estimate is statistically independent. The Monte Carlo variance of the accumulated noise floor decays strictly inversely with the number of pairs:

$$\mathrm{Var}(N_L^{(0)}) \propto \frac{2}{N_{\rm sim}}. \tag{D.5.2}$$

   With $N_{\rm sim} = 400$, the fractional uncertainty in the noise floor is suppressed by $\sqrt{200} \approx 14.1$, ensuring that simulation noise contributes negligibly ($< 0.1\%$) to the final cosmological error budget.

3. **Optimal throughput:** Processing $N_{\rm sim}/2$ pairs scales linearly with simulation ensemble size $\mathcal{O}(N_{\rm sim})$, avoiding the $\mathcal{O}(N_{\rm sim}^2)$ bottleneck of all-to-all cross-pairing.

D.6 Master Decomposition of the Observed Power Spectrum

We now assemble the complete auto-power spectrum of the observed deflection estimator $\hat{\phi}_{\alpha\beta}$. The total covariance [Eq. (D.2.4)] is the sum of its disconnected Gaussian floor and its connected trispectrum cumulant:

$$\langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L}) \hat{\phi}^*_{\alpha\beta}(\boldsymbol{L}') \rangle = \langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L}) \hat{\phi}^*_{\alpha\beta}(\boldsymbol{L}') \rangle_{\rm disc} + \langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L}) \hat{\phi}^*_{\alpha\beta}(\boldsymbol{L}') \rangle_{\rm conn}. \tag{D.6.1}$$

As derived in Section 8.1.6, the connected cumulant decomposes into the true deflection potential power spectrum $C_L^{\phi\phi}$ (from s-channel single-lens exchange) and the secondary mode-coupling bias $N_L^{(1),\alpha\beta}$ (from t- and u-channel exchanges):

$$\langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L}) \hat{\phi}^*_{\alpha\beta}(\boldsymbol{L}') \rangle_{\rm conn} = (2\pi)^2 \delta^{(2)}(\boldsymbol{L} - \boldsymbol{L}') \left[C_L^{\phi\phi} + N_L^{(1),\alpha\beta} + \mathcal{O}((C^{\phi\phi})^2) \right]. \tag{D.6.2}$$

Substituting the independent-sky representation of the disconnected covariance [Eq. (D.4.15)] into Eq. (D.6.1) yields:

$$\begin{aligned} \langle \hat{\phi}_{\alpha\beta}(\boldsymbol{L}) \hat{\phi}^*_{\alpha\beta}(\boldsymbol{L}') \rangle = {} & \langle q_{ab}(\boldsymbol{L}) q^*_{ab}(\boldsymbol{L}') \rangle_{\rm ind} + \langle q_{ab}(\boldsymbol{L}) q^*_{ba}(\boldsymbol{L}') \rangle_{\rm ind} \\ & + (2\pi)^2 \delta^{(2)}(\boldsymbol{L} - \boldsymbol{L}') \left[C_L^{\phi\phi} + N_L^{(1),\alpha\beta} + \mathcal{O}((C^{\phi\phi})^2) \right]. \end{aligned} \tag{D.6.3}$$

Converting every term into its corresponding angular power spectrum establishes the master observational relation:

$$\boxed{C_L^{\hat{\phi}_{\alpha\beta}\hat{\phi}_{\alpha\beta}} = C_L^{q_{ab}^{\rm ind} q_{ab}^{\rm ind}} + C_L^{q_{ab}^{\rm ind} q_{ba}^{\rm ind}} + C_L^{\phi\phi} + N_L^{(1),\alpha\beta} + \mathcal{O}((C^{\phi\phi})^2).} \tag{D.6.4}$$

Equation (D.6.4) is the central result governing all practical CMB lensing measurements. It demonstrates that subtracting the independent-sky simulation spectra directly removes the Gaussian disconnected sample variance from the observed power spectrum.

D.7 Shared Lensing Potentials and Isolation of the $N_L^{(1)}$ Bias

While independent simulations isolate the Gaussian noise floor $N_L^{(0)}$, isolating the secondary mode-coupling bias $N_L^{(1)}$ requires constructing simulations that retain the connected t- and u-channel contractions while extinguishing the primary s-channel signal.

To achieve this, realization pairs (a, b) are lensed by the *same* intervening deflection field ϕ ($\phi_a = \phi_b = \phi$):

$$\tilde{X}^{(a)} = X^{(a)}(\hat{\boldsymbol{n}} + \boldsymbol{\nabla}\phi), \qquad \tilde{X}^{(b)} = X^{(b)}(\hat{\boldsymbol{n}} + \boldsymbol{\nabla}\phi). \tag{D.7.1}$$

Because the primordial CMB draws $X^{(a)}$ and $X^{(b)}$ are statistically independent, the conditional cross-covariance remains zero, $\langle \tilde{X}^{(a)} \tilde{X}^{(b)*} \mid \phi \rangle = 0$. Consequently, the mixed quadratic estimator $q_{ab}^{\rm shared}$ still possesses strictly zero mean:

$$\langle q_{ab}^{\rm shared}(\boldsymbol{L}) \rangle = 0. \tag{D.7.2}$$

This ensures that the primary cosmological signal $C_L^{\phi\phi}$ (contraction I) remains identically zero.

However, because realizations a and b are sheared by the identical deflection field ϕ, the ensemble average over lens realizations does not factorize. The mode-coupling response in realization a correlates with that in realization b through $\langle \phi(\boldsymbol{K})\phi^*(\boldsymbol{K}') \rangle \propto C_K^{\phi\phi}$.

The pipeline runs this shared-lens evaluation across the exact same $N_{\rm sim}/2$ pairs, accumulating:

$$\frac{1}{N_{\rm sim}/2} \sum_{\text{pairs } (a,b)} \left[C_L^{q_{ab}^{\rm shared} q_{ab}^{\rm shared}} + C_L^{q_{ab}^{\rm shared} q_{ba}^{\rm shared}} \right] = N_L^{(0),\alpha\beta} + N_L^{(1),\alpha\beta}. \tag{D.7.3}$$

Subtracting the independent-simulation result [Eq. (D.5.1)] from the shared-simulation result [Eq. (D.7.3)] cancels the zero-mean disconnected Gaussian noise floor $N_L^{(0)}$ completely, cleanly isolating the mode-coupling bias:

$$\boxed{N_L^{(1),\alpha\beta} = \frac{1}{N_{\rm sim}/2} \sum_{\text{pairs } (a,b)} \left[C_L^{q_{ab}^{\rm shared} q_{ab}^{\rm shared}} + C_L^{q_{ab}^{\rm shared} q_{ba}^{\rm shared}} \right] - N_L^{(0),\alpha\beta}.} \tag{D.7.4}$$

For symmetric estimators ($q_{ab} = q_{ba}$), this simplifies to:

$$N_L^{(1)} = \frac{2}{N_{\rm sim}/2} \sum_{\text{pairs } (a,b)} C_L^{q_{ab}^{\rm shared} q_{ab}^{\rm shared}} - N_L^{(0)}. \tag{D.7.5}$$

The realization pairing hierarchy, retained Wick contractions, and isolated bias components are summarized in Table D.1 and illustrated diagrammatically in Fig. D.1.

D.8 Practical Considerations and Methodological Nuances

We conclude this appendix with an overview of practical subtleties that arise when applying this framework to observational data:

Realization Pair	Primary CMB	Deflection Field	Mean $\langle q \rangle$	Isolated Power Spectrum
Observed Data	Real data	True sky ϕ	$\phi(\boldsymbol{L})$	$C_L^{\phi\phi} + N_L^{(0)} + N_L^{(1)} + \mathcal{O}\big((C^{\phi\phi})^2\big)$
Independent Sims	$a \neq b$	$\phi_a \neq \phi_b$	0	$N_L^{(0)} = C_L^{q_{ab}q_{ab}} + C_L^{q_{ab}q_{ba}}$
Shared Sims	$a \neq b$	$\phi_a = \phi_b = \phi$	0	$N_L^{(0)} + N_L^{(1)} = C_L^{q_{ab}^{\rm shared}} + C_L^{q_{ba}^{\rm shared}}$
Shared − Independent	$a \neq b$	Combined	0	$N_L^{(1)}$

Table D.1: Systematic realization cross-pairing scheme for isolating reconstruction noise floors and mode-coupling biases in CMB lensing analyses.

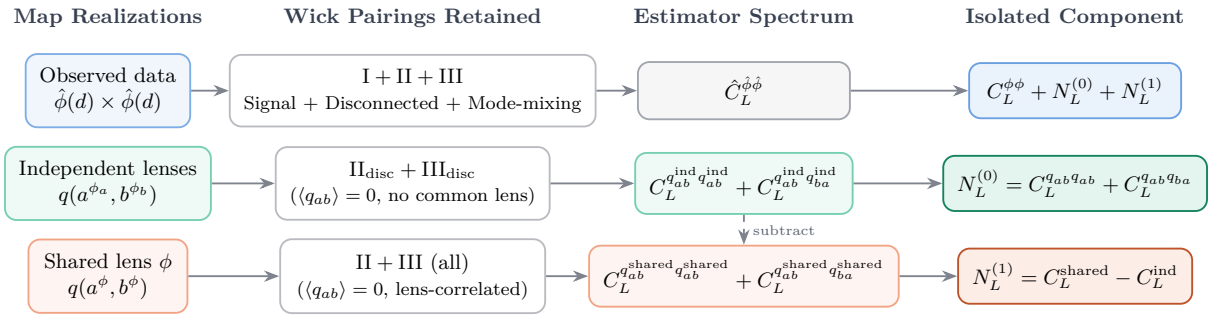


Figure D.1: Isolation of reconstruction noise floors and mode-coupling biases using Monte Carlo simulation realizations. Cross-pairing independent CMB realizations a and b renders the quadratic estimator zero-mean ($\langle q_{ab} \rangle = 0$), eliminating the primary signal (contraction I). Lensing with independent potentials $\phi_a \neq \phi_b$ ensures no common lensing occurs between the maps, causing the ensemble average to factorize; the resulting power spectrum is purely the sample variance of this zero-mean estimator, isolating the disconnected Gaussian noise floor $N_L^{(0)} = C_L^{q_{ab}q_{ab}} + C_L^{q_{ab}q_{ba}}$. Lensing with a shared potential ϕ preserves the lens-induced mode-coupling correlation between legs; subtracting the independent from the shared spectrum cancels the disconnected contribution and isolates the $N_L^{(1)}$ bias.

D.8.1 The Fatal Pitfall of Unlensed Simulations ($\phi = 0$) for Polarization

A common initial intuition in CMB data analysis is to attempt evaluating $N_L^{(0)}$ by simulating unlensed Gaussian CMB skies ($\phi = 0$). While this intuition works reasonably well for temperature fluctuations (where primordial T-modes dominate over lensing perturbations), it fails catastrophically for polarization B-mode reconstruction.

Primordial scalar density perturbations produce strictly curl-free E-modes; in standard ΛCDM cosmology, primordial tensor B-modes are negligible on small scales ($B^{\rm prim} \approx 0$). Consequently, in an unlensed simulation ($\phi = 0$), the simulated B-mode map contains only instrumental detector noise ($B = n^B$). Evaluating $N_L^{(0)}$ for the minimum-variance EB estimator on unlensed simulations would omit the cosmic variance of the lensed B-modes ($C_\ell^{BB,\rm lens} \sim 5\,\mu\mathrm{K} \cdot \mathrm{arcmin}$), which enters the denominator of the estimator as effective noise:

$$N_L^{(0),EB} \propto C_{\ell_1}^{EE,\rm tot} C_{\ell_2}^{BB,\rm tot} = (C_{\ell_1}^{EE} + N_{\ell_1}^{E})(C_{\ell_2}^{BB,\rm lens} + N_{\ell_2}^{B}). \tag{D.8.1}$$

Setting $\phi = 0$ deletes $C_\ell^{BB,\rm lens}$, predicting an unphysically small noise floor and leading to massive under-subtraction of $N_L^{(0)}$ in the observed data. Therefore, independent simulations *must* be lensed by independent deflection potentials ϕ_a, ϕ_b drawn from a fiducial cosmology, or

generated from the observed lensed power spectra, to maintain the correct two-point covariance across all multipoles.

D.8.2 Realization-Dependent Mean-Field Subtraction

On an incomplete sky, the quadratic estimator suffers from an additional realization-dependent noise bias (often denoted RD-$N_L^{(0)}$). When a finite ensemble of $N_{\rm sim}$ simulations is used to estimate the mean field $\bar{\phi}_{\rm MF} = \frac{1}{N_{\rm sim}} \sum_s \hat{\phi}(s)$, sample variance in the estimated mean field leaks into the power spectrum. Advanced CMB pipelines (Namikawa et al. 2013; Planck Collaboration 2018) evaluate a realization-dependent disconnected bias that cross-correlates the observed data map with simulation realizations:

$$N_L^{(0),\rm RD} = 2C_L^{q(d,s)q(d,s)} - C_L^{q(s,s')q(s,s')}, \tag{D.8.2}$$

where d denotes the real data and s, s' denote distinct simulation draws. This realization-dependent construction has the remarkable property that it reduces the covariance between power spectrum multipoles and insulates the reconstruction against small mismatches between the true sky power spectrum and the simulation fiducial model.

D.8.3 Comparison: Monte Carlo Simulation vs. Analytical Fixed-Point Iteration

Finally, it is instructive to contrast the Monte Carlo approach with the analytical fixed-point iteration developed in Section 8.1.6:

- **Analytical fixed-point iteration:** Operates entirely on 1D power spectra via the mode-coupling response matrix $\mathbf{K}$ [Eq. (8.1.302)]. It requires no large suites of simulated maps and converges in 3–4 matrix multiplications, making it computationally trivial. However, it relies on statistical isotropy and cannot easily account for irregular sky masks or anisotropic pixel noise.
- **Monte Carlo realization debiasing:** Operates directly at the 2D map level on simulated skies that have passed through the complete instrument pipeline (telescope pointing, beam smoothing, map making, and masking). It is computationally intensive (requiring hundreds of full-sky spherical harmonic transforms), but it provides unmatched fidelity against complex survey systematic effects.

Modern cosmological analyses therefore utilize both in tandem: the analytical response matrix $\mathbf{K}$ is used to explore parameter space and cosmological dependencies rapidly, while the Monte Carlo cross-pairing framework provides the final, debiased cosmological measurement.