Astrophysics

Sky, stars and the universe from six principles

Chandra Prakash

lunar-photon.github.io

Preface

Six principles

Astrophysics can look like a long list of formulas. Yet almost every result about stars, gas and the universe follows from six general principles, applied with ordinary classical and quantum physics. A formula that has been forgotten can always be rebuilt from them.

P1 **Conservation.** What is in a region changes only by what crosses its boundary and what acts on it.

P2 **Gravity of a sphere.** Inside a spherical body, only the mass inside a radius pulls.

P3 **Counting states.** Pressure, energy density and populations follow from counting quantum states and how full they are.

P4 **Balance of rates.** In a steady state gains match losses. In equilibrium each process matches its reverse.

P5 **Random walk.** Distance grows as the square root of the number of steps, so scattering things flow down their gradients.

P6 **Compare two scales.** A ratio of two times or two lengths decides which physics matters.

Each of them is stated and derived in Chapter 9.

Orders of magnitude

Quantity	Value
Solar mass, $M_\odot$	1.99×10^{30} kg
Solar radius, $R_\odot$	6.957×10^{5} km
Solar luminosity, $L_\odot$	3.83×10^{26} W
Astronomical Unit, 1 AU	1.496×10^{8} km
Light-year, 1 ly	9.461×10^{12} km
Parsec, 1 pc	3.086×10^{13} km = 3.26156 ly

Table 1: Units of mass, length, distance and power.

- **Solar System.** 8 planets, 9 dwarf planets, 468 known moons, about 1.5 million catalogued asteroids, and a vast reservoir of comets in the Kuiper belt (30–50 AU) and the Oort cloud (2000–200,000 AU, about 3 ly at its outer edge). The total mass is about $1.0014\,M_\odot$. The nearest star, Proxima Centauri, lies 4.24 ly away.
- **The Milky Way.** 10^{11} to 4×10^{11} stars, a total mass of about $1.15 \times 10^{12}\,M_\odot$, a central black hole of about $4.3 \times 10^{6}\,M_\odot$, and a radius of about 27 kpc. The Sun lies about 8 kpc from the centre, the frame used for the (U, V, W) velocities of Chapter 3.

- **The observable universe.** About 2×10^{12} galaxies within a diameter of about 93 Gly. The mean density of all its matter and energy is about $8.5 \times 10^{-27}\,\mathrm{kg\,m^{-3}}$, the weight of about five protons per cubic metre. Ordinary matter makes up 4.9% of it, about one proton in every four cubic metres. Dark matter makes up 26.8% and dark energy 68.3%.

The messengers of astrophysics

Information reaches us from the cosmos through four channels. They are electromagnetic waves, gravitational waves, neutrinos and cosmic rays. They are produced by different processes, travel and are absorbed differently, and so carry different information about the same source. Seeing one event through more than one channel is multi-messenger astronomy. The neutron-star merger GW170817 was detected both in gravitational waves and across the electromagnetic spectrum. Light still carries almost everything known about stars and galaxies, from its passage through the atmosphere and the interstellar medium to the physics of the bodies that emit it. Neutrinos carry away most of the energy of a collapsing stellar core (Chapter 13).

Contents

Preface **iii**

1 Time Systems and the Rotating Earth **1**
1.1 The Celestial Sphere . . . 1
1.2 Two Clocks: the Solar Day and the Sidereal Day . . . 2
1.3 Three Years: Tropical, Sidereal and Anomalistic . . . 4
1.4 Longitude and Local Solar Time . . . 6
1.5 Sidereal Time and the Hour Angle . . . 7
1.6 The Sun's Coordinates through the Year . . . 8
1.7 The Equation of Time and the Analemma . . . 10
1.8 Precession of the Equinoxes and Reference Epochs . . . 13
1.9 Nutation: Mean and Apparent Sidereal Time . . . 17
1.10 Modern Time Scales: UT1, TAI, UTC, TT and TDB . . . 18
1.11 Julian Date and Modified Julian Date . . . 19
1.12 Problem Set: Time Systems and Solar Motion . . . 21

2 Coordinate Systems on the Celestial Sphere **23**
2.1 The Equatorial Coordinate System . . . 23
2.2 Turning the Axes . . . 26
2.3 Horizontal Coordinates . . . 29
2.4 Rising, Setting, and Circumpolar Stars . . . 33
2.5 Altitude at Meridian Transit . . . 35
2.6 Angular Distance on the Sphere . . . 35
2.7 The Ecliptic Coordinate System . . . 37
2.8 The Galactic Coordinate System . . . 39
2.9 Atmospheric Refraction and Stellar Aberration . . . 40
2.10 Airmass and the Parallactic Angle . . . 43
2.11 Problem Set: Celestial Coordinate Transformations . . . 45

3 Astrometry: Distances, Motions, and the Space Velocity Vector **49**
3.1 Trigonometric Parallax and the Parsec . . . 49
3.2 Three-Dimensional Cartesian Position Vector . . . 52
3.3 Proper Motion and Transverse Velocity . . . 55
3.4 Radial Velocity and the Doppler Effect . . . 56
3.5 The Full 3D Space Velocity Vector . . . 58
3.6 Galactic Space Velocity Components (U, V, W) . . . 59
3.7 The Cosmic Distance Ladder . . . 60
3.8 Problem Set: Astrometry and Kinematics . . . 71

4 Stellar Photometry: Flux, Magnitudes, and Radii **75**

4.1 Flux and the Inverse-Square Law . . . 75
4.2 The Magnitude Scale . . . 76
4.3 Light from a Hot Surface . . . 79
4.4 Colours and the Dimming by Dust . . . 83
4.5 The Radii of Stars . . . 88
4.6 Problem Set: Photometry, Colours and Extinction . . . 91

5 Angular Resolution: Baselines, Diffraction, and Pixel Sampling 95
5.1 Windows on the Universe . . . 95
5.2 Interferometric Baselines . . . 97
5.3 Telescope Optics . . . 101
5.4 Diffraction . . . 107
5.5 Pixel Scale and Nyquist Sampling . . . 114
5.6 Atmospheric Seeing, PSF Profiles, and Adaptive Optics . . . 117
5.7 Problem Set: Interferometric Baselines, Optics, and Resolution . . . 122

6 The Two-Body Problem 129
6.1 Central Force Problem and Conservation Laws . . . 129
6.2 Coordinate Systems and Orbital Elements . . . 131
6.3 Planar Motion and Kepler's Second Law . . . 134
6.4 The Orbit Equation . . . 136
6.5 Keplerian Orbits . . . 137
6.6 Energy and Angular Momentum in Terms of a, e . . . 139
6.7 True and Eccentric Anomaly; Kepler's Equation . . . 141
6.8 Kepler's Third Law . . . 143
6.9 Kepler's Equation . . . 144
6.10 Keplerian Parametrization . . . 144
6.11 Hyperbolic-Trajectory Parameters . . . 145
6.12 Fourier Solution of Kepler's Equation . . . 146
6.13 The Laplace–Runge–Lenz Vector . . . 147
6.14 Perturbed Orbits . . . 149
6.15 Planar perturbations . . . 153
6.16 Orbit-averaged evolution . . . 158
6.17 Nearly circular orbits . . . 164
6.18 Rømer delay for a Keplerian binary . . . 166
6.19 Radial Velocities and Planets . . . 167
6.20 Other Ways to Find Planets . . . 170
6.21 Applied Orbital Mechanics . . . 174

7 Signal, Noise, and Detection in CCD Photometry 193
7.1 Counting Photoelectrons . . . 193
7.2 Shot Noise and the Poisson Distribution . . . 194
7.3 The Noise Budget and the Signal-to-Noise Equation . . . 199
7.4 Three Regimes of the Signal-to-Noise Equation . . . 202
7.5 From Signal-to-Noise to Magnitudes and Detections . . . 205
7.6 Noise Budgets Beyond the CCD . . . 213
7.7 Problem Set: Poisson Statistics and CCD Signal-to-Noise . . . 213

8 The Interstellar Medium 221
8.1 Composition, Density and Temperature . . . 221
8.2 How the Medium Acts on Light . . . 222
8.3 Spectral Lines and Their Widths . . . 231

8.4 Hydrogen in the ISM . . . 236
8.5 Interstellar Clouds and Nebulae . . . 239
8.6 Heating and Cooling of the ISM . . . 242
8.7 Problem Set: The Interstellar Medium . . . 243

9 Six Principles **249**
9.1 Conservation (P1) . . . 249
9.2 Gravity of a Sphere (P2) . . . 253
9.3 Counting States (P3) . . . 256
9.4 Balance of Rates (P4) . . . 259
9.5 Random Walk (P5) . . . 261
9.6 Compare Two Scales (P6) . . . 263
9.7 Problem Set: Six Principles . . . 264

10 Self-Gravitating Gas **267**
10.1 Hydrostatic Equilibrium . . . 267
10.2 The Virial Theorem . . . 269
10.3 Negative Heat Capacity . . . 272
10.4 The Jeans Criterion . . . 275
10.5 Free Fall and the Three Clocks . . . 277
10.6 Rotation and the End of Collapse . . . 279
10.7 From Protostar to Star . . . 281
10.8 Problem Set: Self-Gravitating Gas . . . 286

11 Radiation in Matter **289**
11.1 Specific Intensity . . . 289
11.2 The Transfer Equation . . . 290
11.3 Kirchhoff's Law and Local Thermodynamic Equilibrium . . . 293
11.4 Diffusion of Light . . . 294
11.5 The Eddington Luminosity . . . 297
11.6 Einstein Coefficients . . . 298
11.7 Spectral Classes . . . 299
11.8 The Grey Atmosphere . . . 305
11.9 Problem Set: Radiation in Matter . . . 307

12 The Equations of a Star **311**
12.1 The Unknowns . . . 311
12.2 The Four Equations . . . 312
12.3 Pressure, Opacity and Burning . . . 315
12.4 Boundary Conditions and How Stars Are Computed . . . 322
12.5 Polytropes and the Isothermal Sphere . . . 323
12.6 Scaling Laws and the HR Diagram . . . 326
12.7 Pulsating Stars . . . 330
12.8 Problem Set: The Equations of a Star . . . 337

13 Degenerate and Compact Stars **341**
13.1 Compactness . . . 341
13.2 White Dwarfs . . . 342
13.3 The Relativistic Structure Equations . . . 345
13.4 Computing a Neutron Star . . . 347
13.5 Black Holes . . . 349
13.6 Gravitational Waves from Compact Binaries . . . 349

13.7 Problem Set: Degenerate and Compact Stars . . . 357

14 Galaxies and the Motions of Stars **361**
14.1 Our Galaxy . . . 361
14.2 Rotation Curves and Dark Matter . . . 366
14.3 Stars as a Gas That Rarely Collides . . . 371
14.4 Weighing Galaxies and Clusters . . . 376
14.5 Normal Galaxies . . . 382
14.6 Telling Stars, Galaxies, Nebulae and Quasars Apart . . . 384
14.7 Groups, Clusters and the Cosmic Web . . . 390
14.8 Problem Set: Galaxies and the Motions of Stars . . . 394

15 Radiation from Hot and Fast Particles **401**
15.1 An Accelerated Charge Radiates . . . 401
15.2 Bremsstrahlung . . . 404
15.3 Synchrotron Radiation . . . 408
15.4 Compton and Inverse Compton Scattering . . . 412
15.5 Accretion and Active Galaxies . . . 416
15.6 Active Nuclei Seen From Different Angles . . . 420
15.7 Problem Set: Radiation from Hot and Fast Particles . . . 426

16 The Expanding Universe **431**
16.1 Hubble's Law and the Scale Factor . . . 431
16.2 The Friedmann Equations . . . 432
16.3 The Microwave Background and Recombination . . . 434
16.4 Freeze-Out and the First Three Minutes . . . 436
16.5 The Growth of Structure . . . 438
16.6 The Lyman-α Forest . . . 440
16.7 The Accelerating Expansion . . . 443
16.8 Problem Set: The Expanding Universe . . . 444

Chapter 1

Time Systems and the Rotating Earth

Pointing a telescope requires us to know both where an object is and when we will observe it. The Earth's eastward rotation makes the sky appear to turn westward, so the angle through which the sky has turned can serve as a clock. One turn of 360° takes 24 hours on that clock, giving the conversion

$$1^{\mathrm{h}} \equiv 15^\circ, \qquad 1^{\mathrm{m}} \equiv 15' = 0.25^\circ, \qquad 1^{\mathrm{s}} \equiv 15'' = 0.004167^\circ. \tag{1.1}$$

That simple conversion becomes less simple because our observatory moves. The Earth spins daily, orbits the Sun yearly, and its axis slowly sweeps out a cone over about 26 000 years. Each motion changes the reference direction against which we measure a turn. We will use those differences to define the time systems behind the sky coordinates of Chapter 2.

1.1 The Celestial Sphere

We can tell the direction of a star at once, but not its distance. So every star can be drawn as a point on a large sphere centred on us, of any radius we like. This is the **celestial sphere**. A planetarium dome is a good picture of it. The stars painted on the dome look right whatever its size, because only the directions matter.

Two features of the sphere are fixed by the Earth's spin alone, wherever the observer stands. The **celestial poles** are the two points where the Earth's spin axis, extended, meets the sphere. The **celestial equator** is the great circle where the plane of the Earth's equator meets the sphere.

The sky turns rigidly about the poles once per sidereal day (Section 1.2). The observer's **celestial meridian** is the great circle through both poles and the **zenith**, the point straight overhead. As the sky turns, every point of the celestial equator crosses the meridian once per turn. So naming the point of the equator that sits on the meridian now is the same as reading a clock. This one fact makes a sidereal clock possible.

A direction on the sphere is named by two angles, built like latitude and longitude on the Earth. The **declination** δ is the angle north (+) or south (−) of the celestial equator. The **right ascension** α is the angle measured east along the celestial equator from a chosen zero point. Stars are so far away that both angles hardly change in a lifetime, so a catalogue of (α, δ) stays useful for years.

The zero point is the **vernal equinox** $\hat{\gamma}$: the intersection of the celestial equator with the Sun's apparent yearly path where the Sun crosses from south to north. Around 20 March the Sun points in that direction. On later dates the Sun moves on, but the intersection remains the zero point of right ascension. Think of a runner passing a fixed mark on a track; the runner's position changes, while the mark still tells us where to start measuring. The mark itself drifts slowly against the stars because of precession (Section 1.8), so catalogue coordinates state an epoch. Section 1.6 follows the Sun away from this zero point through the year.

1.2 Two Clocks: the Solar Day and the Sidereal Day

We know that the Earth turns once a day. That is enough for daily life, where noon comes round every 24 hours. But "once" needs a reference. One turn relative to the Sun and one turn relative to the stars are not the same, because the Sun itself moves against the stars as the Earth goes round its orbit.

Look at the hands of a clock. The minute hand starts on top of the hour hand at noon. After 60 minutes it is back at the top, but the hour hand has moved on to the 1. The minute hand needs a little more than five extra minutes to catch it. It laps the hour hand every 65.45 minutes, not every 60. The observer's meridian is the minute hand. The direction to the Sun is the hour hand, which creeps east by about a degree a day as the Earth moves along its orbit. One turn relative to the stars is the **sidereal day**. One turn relative to the Sun, from one noon to the next, is the **solar day**. The solar day is the time the meridian takes to lap the Sun.

The two clock hands make the difference precise. The minute hand turns at $1/60$ of a turn per minute and the hour hand at $1/720$. Their relative rate is $1/60 - 1/720 = 11/720$ of a turn per minute, so the minute hand catches the hour hand after $720/11 = 65.45$ minutes. Earth's meridian similarly needs a little extra time to catch the Sun after completing a turn relative to the stars. The same lapping explains why the Moon takes 27.32 days to orbit relative to the stars but 29.53 days between full Moons: the Sun has moved in the meantime. For Earth's spin and orbit, the rate difference is small because a year contains about 366 sidereal turns. We now calculate how much longer the solar day is.

The two days. The **solar day** is the time between two transits of the Sun across the meridian, from one local noon to the next. Its average over the year, the **mean solar day**, is defined to be exactly

$$24^{\mathrm{h}} = 1440^{\mathrm{m}} = 86\,400\,\mathrm{s}.$$

The **sidereal day** is the time between two transits of the vernal equinox $\hat{\gamma}$. In mean solar time it lasts

$$T_{\mathrm{sid}} \simeq 23^{\mathrm{h}}56^{\mathrm{m}}04.0905^{\mathrm{s}} = 86\,164.0905\,\mathrm{s}, \tag{1.2}$$

which we derive below.

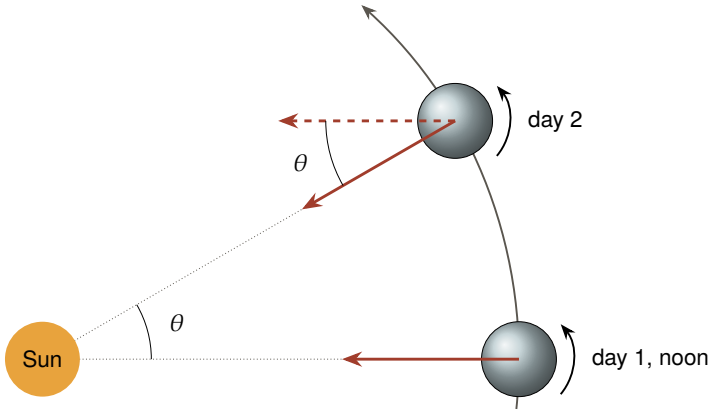


Figure 1.1: The Earth on two successive days, seen from above the north pole. Day 1 is at local noon. The red arrows show the direction of the observer's meridian. On day 2 the dashed arrow points the same way in space as on day 1, and the solid arrow points at the Sun. The angle θ moved along the orbit in one day is drawn much larger than its true value of about 1°.

A first estimate. The Earth goes round its orbit in the same sense as it spins, both anticlockwise seen from above the north pole (Figure 1.1). In one day it moves along its orbit by

$$\frac{360^\circ}{365.2422 \text{ days}} = 0.985647^\circ \text{ per day}.$$

After one turn relative to the stars, the meridian points the same way in space as the day before (dashed arrow). The Sun now lies 0.9856° further east, so the Earth must turn this extra angle before the meridian reaches it again (solid arrow). At 360° per 1440^{m} of solar time, that is 4^{m} per degree, and the extra angle takes

$$0.985647^\circ \times 4^{\mathrm{m}}/^\circ = 3.94259^{\mathrm{m}} = 3^{\mathrm{m}}56.555^{\mathrm{s}}.$$

This estimate is slightly off, because the extra angle is turned at the Earth's spin rate and not at the Sun's rate of 360° per solar day. The lapping calculation gets it right.

Lapping, as a rotation. Put the z-axis along the Earth's spin axis, and let

$$R_z(\varphi) = \begin{pmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.3}$$

turn a vector by the angle φ anticlockwise about it, seen from the north. Let $\omega_\oplus = 2\pi/T_{\text{sid}}$ be the Earth's spin rate against the equinox, and let $n = 2\pi/T_{\text{yr}}$ be the mean rate at which the Sun's direction turns east, where T_{yr} is the year measured against the equinox. This n is the mean motion of the Earth's orbit, the same n as in Eq. (6.50). The meridian direction turns as $R_z(\omega_\oplus t)$ and the mean Sun's direction as $R_z(nt)$. To see the meridian from the Sun, we turn everything back by nt,

$$\begin{aligned} R_z(-nt)\,R_z(\omega_\oplus t) &\overset{(1)}{=} \begin{pmatrix} \cos nt & \sin nt & 0 \\ -\sin nt & \cos nt & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\omega_\oplus t & -\sin\omega_\oplus t & 0 \\ \sin\omega_\oplus t & \cos\omega_\oplus t & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &\overset{(2)}{=} \begin{pmatrix} \cos(\omega_\oplus - n)t & -\sin(\omega_\oplus - n)t & 0 \\ \sin(\omega_\oplus - n)t & \cos(\omega_\oplus - n)t & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= R_z\big((\omega_\oplus - n)t\big). \end{aligned} \tag{1.4}$$

(1) Eq. (1.3) with $\varphi = -nt$ and $\varphi = \omega_\oplus t$.

(2) Matrix product, then $\cos x \cos y + \sin x \sin y = \cos(x - y)$ and $\sin x \cos y - \cos x \sin y = \sin(x - y)$.

Seen from the Sun, the meridian turns at $\omega_\oplus - n$, and so

$$\omega_\oplus - n = \frac{2\pi}{T_{\text{sol}}}, \tag{1.5}$$

where T_{sol} is the mean solar day. This is the clock-hand rule. The rates subtract because the two motions go the same way.

Days per year. Writing each rate as 2π over its period,

$$\frac{1}{T_{\text{sol}}} = \frac{1}{T_{\text{sid}}} - \frac{1}{T_{\text{yr}}}. \tag{1.6}$$

We multiply by T_{yr}. Then $T_{\text{yr}}/T_{\text{sol}}$ is the number of solar days in a year, 365.2422, and $T_{\text{yr}}/T_{\text{sid}}$ the number of sidereal days,

Result 1.1

$$\frac{T_{\mathrm{yr}}}{T_{\mathrm{sid}}} = \frac{T_{\mathrm{yr}}}{T_{\mathrm{sol}}} + 1 = 365.2422 + 1 = 366.2422 \textit{ sidereal days per year}. \tag{1.7}$$

In one year the Earth turns once more against the stars than against the Sun. The orbit itself supplies the extra turn.

The length of the sidereal day. Dividing the two counts,

$$\begin{aligned} T_{\mathrm{sid}} &\overset{(1)}{=} T_{\mathrm{sol}} \frac{365.2422}{366.2422} \\ &= 86\,400\,\mathrm{s} \times 0.99726957 = 86\,164.0905\,\mathrm{s}, \end{aligned} \tag{1.8}$$

(1) Eq. (1.7), the same year counted in the two kinds of day.

which is Eq. (1.2). The difference of the two days is

$$\boxed{T_{\mathrm{sol}} - T_{\mathrm{sid}} = \frac{T_{\mathrm{sol}}}{366.2422} = 235.909\,\mathrm{s} = 3^{\mathrm{m}}55.909^{\mathrm{s}}.}$$

It depends only on how many days there are in a year, not on the size or shape of the orbit. The first estimate, $3^{\mathrm{m}}56.555^{\mathrm{s}}$, is larger by the factor $366.2422/365.2422 = 1.0027379$. That number is not wrong. It is the amount by which a sidereal clock gains on a solar clock in one solar day, counted in sidereal seconds, and it returns in Eq. (1.45).

The year in Eq. (1.6) is the one measured against the equinox, so T_{sid} is the turn against the equinox. The equinox drifts slowly west (Section 1.8). Measured against the distant stars instead, with the sidereal year of 365.256363 days in place of 365.2422, the turn takes about 0.009 s longer. The difference matters only for the most precise work.

Two clocks on the same sky. Imagine fixing the equatorial coordinate grid to the stars for the duration of a night. A star keeps nearly the same right ascension and declination in that grid. The observer's meridian, attached to the rotating Earth, sweeps through it. A sidereal clock reads the right ascension currently crossing the meridian; a solar clock follows the Sun instead. We may choose an instant when both clock faces read 06:00. One mean solar day later, the solar clock again reads 06:00, but the sidereal clock reads about 06:03:56.6, by Eq. (1.45). At a fixed site, the two 24-hour readings come back into agreement about once a year, after the sidereal clock has gained a full turn. Their agreement is only a changing clock offset. At the same civil time on successive nights, the meridian points to progressively larger right ascensions. The stars' coordinates have not been redefined; the Earth-bound meridian has moved relative to their grid.

Why the night sky changes with the seasons. A sidereal clock gains about $3^{\mathrm{m}}56^{\mathrm{s}}$ on a solar clock each day. That adds up to about 2^{h} a month and a full 24^{h} in a year. A star that crosses the meridian at midnight on 1 October crosses it at 22:00 on 1 November, and at noon on 1 April, when it is lost in daylight. This is why each season has its own constellations.

1.3 Three Years: Tropical, Sidereal and Anomalistic

The lapping rule of Eq. (1.6) turned the two references for a day into two kinds of day. A year has the same problem. We count it as one trip of the Sun round the sky, but we must say where the trip starts and ends. The equinox, the stars and the point of closest approach to the Sun all give different answers, because two of them move.

Think of a lap of a running track where the finish line is moved a little each lap. If the line is moved back toward the runner, she crosses it before a full lap and her lap time is short. If it is moved forward, she must run past her last lap to reach it, and her lap time is long. The Sun's yearly trip has one fixed reference, the distant stars, and two moving ones. The equinox moves west, toward the oncoming Sun, by 50.29″ a year (Section 1.8). The perihelion, the point of the orbit closest to the Sun, moves east, ahead of the Sun, by about 11.6″ a year, pulled by the other planets. So the year to the equinox is shorter than the year to the stars, and the year to perihelion is longer.

In numbers, the Sun covers $360^\circ = 1\,296\,000''$ in one sidereal year of 365.256 days. Shortening the trip by 50.29″ saves

$$\frac{50.29''}{1\,296\,000''} \times 365.256 \text{ d} = 0.014174 \text{ d} = 20.41 \text{ min}.$$

The lapping rule is at work again. A reference that moves changes the rate at which the moving body reaches it, by adding or subtracting the reference's own rate. The three years are listed in Table 1.1.

Year	Measured from	Length (days)	Minus tropical
Tropical	equinox to equinox, the cycle of the seasons	365.242190	0
Sidereal	a fixed direction among the stars back to itself	365.256363	$+20^{\text{m}}24.5^{\text{s}}$
Anomalistic	perihelion to perihelion	365.259636	$+25^{\text{m}}07.3^{\text{s}}$

Table 1.1: The three astronomical years.

The tropical year. The **tropical year** runs from one passage of the Sun through the equinox $\hat{\gamma}$ to the next. It sets the seasons and the calendar. Let the Sun move at the mean rate 360° per sidereal year against the stars, while the equinox moves toward it at 50.29″ per year. The Sun reaches the equinox when it has gained a full turn on it,

$$\begin{aligned} \text{tropical year} &\overset{(1)}{=} \frac{360^\circ}{360^\circ/\text{sidereal year} + 50.29''/\text{year}} \\ &\overset{(2)}{=} \text{sidereal year} \times \left(1 - \frac{50.29''}{1\,296\,000''}\right) \\ &= 365.256363 \text{ d} - 0.014174 \text{ d} = 365.242189 \text{ d}. \end{aligned} \tag{1.9}$$

(1) The rates add, since the Sun and the equinox move toward each other (the lapping rule, Eq. (1.5), with the sign of n reversed).
(2) $1/(1+x) \simeq 1 - x$ for $x = 50.29''/1\,296\,000'' = 3.9 \times 10^{-5}$.

This agrees with the measured value in Table 1.1 to 10^{-6} days. The difference from the sidereal year is 20.41 minutes, which is $20^{\text{m}}24.5^{\text{s}}$.

The sidereal year. The **sidereal year** is one full turn of the Earth round the Sun against the distant stars, 365.256363 days. It is the true orbital period of Eq. (6.50).

The anomalistic year. The **anomalistic year** runs from one perihelion passage to the next. The perihelion moves ahead of the Earth at about 11.6″ per year, so the rates now subtract, and

the same steps give

$$\text{anomalistic year} - \text{sidereal year} \simeq \frac{11.6''}{1\,296\,000''} \times 365.256\ \text{d} = 0.00327\ \text{d} = 4.71\ \text{min}.$$

The table gives $365.259636 - 365.256363 = 0.003273$ days, or 4.71 minutes. The two years differ by the time the Earth needs to catch up with its own perihelion.

The Gregorian calendar. The Julian calendar took the year to be exactly 365.25 days, with a leap year every fourth year. That is longer than the tropical year by 0.0078 days, or 11.2 minutes, each year, which adds up to one day in 128 years. By 1582 the March equinox had drifted from 21 March, its date in the year 325, to 11 March. Pope Gregory XIII dropped ten days from the calendar and changed the rule. A year divisible by 4 is a leap year, except a century year not divisible by 400. So 1600 and 2000 were leap years, while 1700, 1800, 1900 and 2100 are not. Of every 400 years, 97 are leap years, and the mean year is

$$365 + \frac{97}{400} = 365.2425\ \text{days}.$$

This exceeds the tropical year by 0.00031 days a year, one day in about 3200 years.

1.4 Longitude and Local Solar Time

The mean solar day of Section 1.2 gives the length of a day. It does not say when noon is. Noon comes when the Sun crosses the local meridian, and the Sun reaches different meridians at different moments. A town to the east has its noon first.

Two friends talk on the phone, one in Kolkata and one in Mumbai. The friend in Kolkata sees the Sun set about an hour earlier. Their clocks show the same time, because India keeps one time across the country, but the Sun does not. The Earth turns east relative to the Sun at the rate $\omega_\oplus - n$ of Eq. (1.5), 360° per mean solar day,

$$\omega_\oplus - n = \frac{360^\circ}{24^\text{h}} = 15^\circ \text{ per hour} = 1^\circ \text{ per } 4^\text{m} = 1' \text{ per } 4^\text{s}.$$

So each degree of longitude moves the time of noon by four minutes. Kolkata (88.4°E) and Mumbai (72.9°E) are 15.5° apart, which is 62 minutes between their noons. Their sunsets differ by between about 55 and 70 minutes over the year, because the two cities are also at different latitudes.

Noon therefore sweeps west by one degree of longitude every four minutes. Civil time zones use a chosen reference meridian instead of changing the clock continuously with longitude. Comparing an observer's longitude with that meridian tells us when local noon occurs on the civil clock.

Two meridians. Let longitude λ be measured east from the Greenwich meridian. The Sun crosses the meridian at λ_1 earlier than the one at λ_2 by

$$\Delta t = \frac{\lambda_1 - \lambda_2}{15^\circ \text{ per hour}}. \tag{1.10}$$

The observer further east sees the Sun cross first.

Civil time. Daily life cannot run on a clock that differs from town to town. A country instead keeps the **standard time** of one chosen meridian, at longitude λ_{ref}. An observer at longitude λ_{obs} is east of it if $\lambda_{\text{obs}} > \lambda_{\text{ref}}$, and then has local noon before the standard clock shows 12:00. By Eq. (1.10),

Result 1.2

$$\textit{local mean solar time} = \textit{standard time} + \frac{\lambda_{\text{obs}} - \lambda_{\text{ref}}}{15^\circ \textit{ per hour}}, \tag{1.11}$$

for an observer on either side of the reference meridian. This is *mean* solar time, the time of a Sun that crosses each meridian at the same moment every day. The real Sun crosses up to 16 minutes earlier or later (Section 1.7).

Worked example 1.1: Local mean solar time at IISER Thiruvananthapuram

IISER TVM is at longitude 77.1382°E. Indian Standard Time is the mean time of the meridian at 82.5°E, which passes near Prayagraj. Then

$$\frac{\lambda_{\text{obs}} - \lambda_{\text{ref}}}{15^\circ \text{ per hour}} = \frac{77.1382^\circ - 82.5^\circ}{15^\circ \text{ per hour}} = -5.3618^\circ \times 4^{\text{m}}/^\circ = -21.4472^{\text{m}} = -21^{\text{m}}26.83^{\text{s}}.$$

When the clock reads 12:00:00 IST, the local mean solar time at IISER TVM is 11:38:33. Local mean noon comes at 12:21:27 IST.

1.5 Sidereal Time and the Hour Angle

Solar time follows the Sun, while the equatorial grid follows the stars. The daily difference just calculated means a fixed civil clock reading does not pick out the same part of that grid each night. To point a telescope, we need the angle of the Earth-bound meridian within the equatorial grid at the moment of observation.

Sit on a merry-go-round that faces a ring of shops. Whichever shop is straight ahead tells you how far the ride has turned. Read the shop name and you know the time on the ride's own clock. The sky is the ring of shops, and right ascension is written along its equator. The right ascension that sits on the meridian now is the reading of the sky's clock. This is sidereal time.

The zero of right ascension is the vernal equinox $\hat{\gamma}$, not a moment when a solar and a sidereal clock happen to agree. A star's catalogue coordinates refer to a specified coordinate epoch because the equinox and pole slowly precess (Section 1.8). Over one night we can treat that grid as fixed and let the meridian turn through it.

Sirius has right ascension $6^{\text{h}}45^{\text{m}}$. When local sidereal time reads $6^{\text{h}}45^{\text{m}}$, Sirius crosses the meridian. Two sidereal hours later the sky has carried it 30° west of that line. Thus the difference between sidereal time and a star's right ascension tells us how far the star has moved from the meridian; we call this its hour angle.

Local sidereal time. The **local sidereal time** is the right ascension on the observer's meridian,

$$\text{LST} \equiv \alpha_{\text{Meridian}}. \tag{1.12}$$

The **Greenwich sidereal time** GST is the same at the Greenwich meridian.

From Greenwich to any longitude. Use the sky's frame, with x toward $\hat{\gamma}$ and z toward the north celestial pole. The point of the celestial equator on the Greenwich meridian has right

ascension GST, so it lies along $R_z(\mathrm{GST})\,\hat{x}$. The observer's meridian is a further λ east, so its point of the equator is

$$R_z(\mathrm{GST})\,R_z(\lambda)\,\hat{x} \overset{(1)}{=} R_z(\mathrm{GST}+\lambda)\,\hat{x} \overset{(2)}{=} \begin{pmatrix}\cos(\mathrm{GST}+\lambda)\\ \sin(\mathrm{GST}+\lambda)\\ 0\end{pmatrix}. \tag{1.13}$$

(1) Two turns about the same axis add, as in Eq. (1.4).
(2) The first column of Eq. (1.3).

A point of the equator at angle α from $\hat{x}$ has right ascension α, so

Result 1.3

$$\mathrm{LST} = \mathrm{GST} + \lambda, \tag{1.14}$$

with λ turned into time by Eq. (1.1). Only the longitude enters. Observers on the same meridian, at any latitude, share one sidereal time.

The hour angle. A star has its own right ascension α, and the meridian has right ascension LST. Both are angles along the celestial equator, so their difference is how far the sky has turned since the star was on the meridian. This is the **hour angle**,

Result 1.4

$$H \equiv \mathrm{LST} - \alpha. \tag{1.15}$$

For the vernal equinox itself, right ascension is always 0^{h} in a grid measured from that equinox. Its *hour angle* is the changing angle from the observer's meridian to that zero direction, and is therefore the local sidereal time. It changes throughout the day as the Earth turns, and at a fixed civil time it changes from one day to the next as the Earth orbits the Sun. Over years, precession slowly moves the equinox relative to the background stars; that is why a catalogue states the epoch of its coordinate grid. It is a separate effect from the daily change of hour angle.

- $H = 0^{\mathrm{h}}$. The object is on the meridian, at its highest point (culmination, or transit).
- $H > 0$. The object has crossed the meridian and is moving toward the western horizon.
- $H < 0$. The object is in the eastern sky and has not yet reached the meridian.

For Sirius in the example above, LST $= 8^{\mathrm{h}}45^{\mathrm{m}}$ gives $H = +2^{\mathrm{h}}$. The hour angle grows at one sidereal hour per sidereal hour, for every star alike.

Solar time as an hour angle. The same definition works for the Sun. Its hour angle is $H_\odot = \mathrm{LST} - \alpha_\odot$, and it is zero at local noon. The time on a sundial, the **apparent solar time**, is this hour angle shifted so that noon reads 12:00,

$$\mathrm{AST} = H_\odot + 12^{\mathrm{h}}. \tag{1.16}$$

At Greenwich apparent noon $H_\odot = 0$ at the Greenwich meridian, so GST $= \alpha_\odot$ at that moment.

1.6 The Sun's Coordinates through the Year

A star keeps nearly the same (α, δ) for years, and the sidereal clock tells when it crosses the meridian. The Sun passes the vernal equinox around 20 March, then moves about a degree a

day along its yearly path. Its changing direction does not move the equinox; it changes the Sun's own coordinates. The Sun also climbs north and sinks south over the year, which is why noon is higher in summer.

Drive round a circular track whose surface is tilted, high on one side and low on the other. Seen from the middle, the car rises above a level line for half the lap and falls below it for the other half. At the two points where the track crosses the level line, the car is level. The track is the **ecliptic**, the Sun's yearly path among the stars. It is the plane of the Earth's orbit seen from the Earth. The level line is the celestial equator. The Earth's spin axis is tilted from the normal to its orbit by the **obliquity of the ecliptic**, $\epsilon \simeq 23.44°$, so the ecliptic and the equator are two great circles tilted by ϵ. They cross at two points, the equinoxes. The Sun's declination follows from one rotation by ϵ.

At the June solstice, the Sun's angle along the ecliptic from $\hat{\gamma}$, its **ecliptic longitude** $\lambda_\odot$, is 90°. The tilt then puts it as far north as it gets, at $\delta_\odot = +23.44°$. At $\lambda_\odot = 45°$ in early May, its declination is only 16.3°, as we will calculate below. To obtain the Sun's coordinates at any point in the year, we turn its position in the ecliptic plane through the tilt angle ϵ.

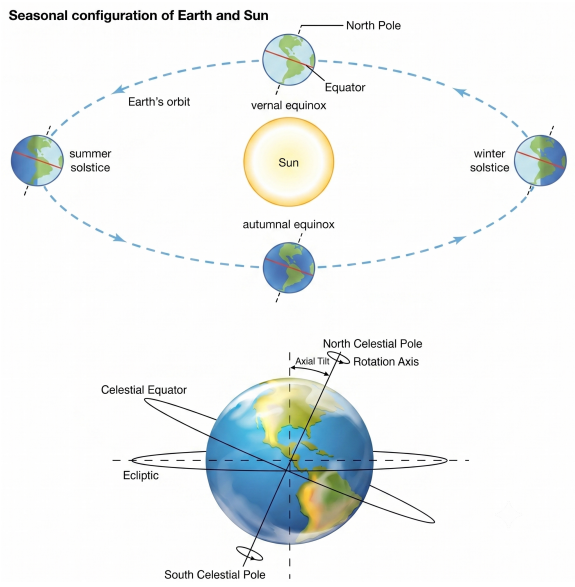


Figure 1.2: The Earth at the two equinoxes and the two solstices on its orbit (top), and the tilt of its axis between the ecliptic and the celestial equator (bottom).

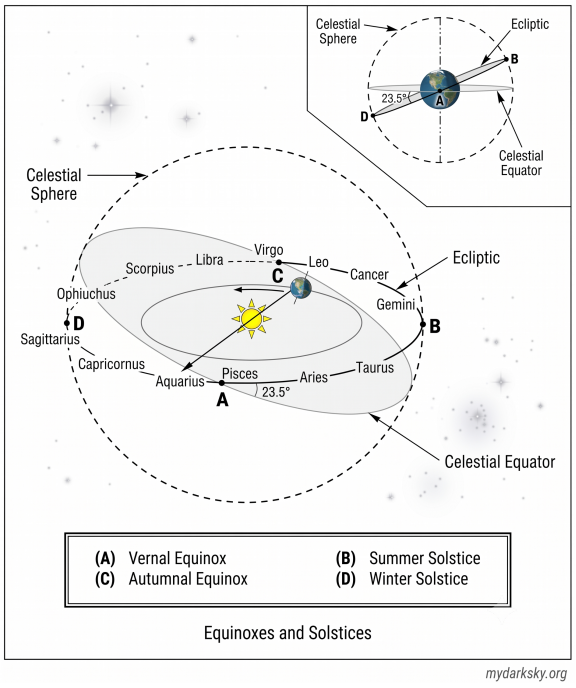


Figure 1.3: The ecliptic and the celestial equator on the celestial sphere, with the equinoxes (A, C), the solstices (B, D) and the constellations along the ecliptic.

One rotation about the line of equinoxes. Use the ecliptic frame, with x toward $\hat{\gamma}$ and z toward the north pole of the ecliptic. The Sun lies in the ecliptic plane, along $(\cos\lambda_\odot, \sin\lambda_\odot, 0)$. The equatorial frame shares the x-axis, since $\hat{\gamma}$ lies on both circles, and it is tilted about that axis by ϵ. In equatorial coordinates the north pole of the ecliptic is at $(0, -\sin\epsilon, \cos\epsilon)$, toward $\alpha = 18^{\mathrm{h}}$. The rotation that carries the ecliptic pole $(0, 0, 1)$ there is a turn by ϵ about x, so the

Sun's equatorial unit vector is

$$\begin{pmatrix} \cos\delta_\odot \cos\alpha_\odot \\ \cos\delta_\odot \sin\alpha_\odot \\ \sin\delta_\odot \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\epsilon & -\sin\epsilon \\ 0 & \sin\epsilon & \cos\epsilon \end{pmatrix} \begin{pmatrix} \cos\lambda_\odot \\ \sin\lambda_\odot \\ 0 \end{pmatrix} = \begin{pmatrix} \cos\lambda_\odot \\ \cos\epsilon\, \sin\lambda_\odot \\ \sin\epsilon\, \sin\lambda_\odot \end{pmatrix}. \tag{1.17}$$

(1) The left side is a unit vector at $(\alpha_\odot, \delta_\odot)$, written as in Eq. (2.2). The matrix turns by ϵ about the x-axis, and it sends $(0, 0, 1)$ to $(0, -\sin\epsilon, \cos\epsilon)$, the ecliptic pole.

The third row, and the ratio of the second row to the first, give

$$\boxed{\sin\delta_\odot = \sin\epsilon\, \sin\lambda_\odot, \qquad \tan\alpha_\odot = \cos\epsilon\, \tan\lambda_\odot.} \tag{1.18}$$

The declination swings between $-\epsilon$ and $+\epsilon$ and depends only on the tilt and the season. The right ascension keeps step with $\lambda_\odot$ at the four points where $\tan\lambda_\odot$ is 0 or infinite, and runs ahead of it or behind it in between. This mismatch is half of the equation of time (Section 1.7).

In early May, $\lambda_\odot = 45°$, and $\sin\delta_\odot = 0.3978 \times 0.7071 = 0.2813$, so $\delta_\odot = 16.3°$.

The four seasonal points. Setting $\lambda_\odot = 0°, 90°, 180°, 270°$ in Eq. (1.18) gives Table 1.2. The last column uses GST = $\alpha_\odot$ at Greenwich apparent noon, from Section 1.5.

Point	Date (about)	$\lambda_\odot$	$\alpha_\odot$	$\delta_\odot$	GST at Greenwich noon
March (vernal) equinox	20 March	$0°$	0^h	$0°$	0^h
June solstice	21 June	$90°$	6^h	$+23.44°$	6^h
September equinox	23 September	$180°$	12^h	$0°$	12^h
December solstice	21 December	$270°$	18^h	$-23.44°$	18^h

Table 1.2: The Sun's coordinates at the equinoxes and solstices, and the Greenwich sidereal time at Greenwich apparent noon.

1.7 The Equation of Time and the Analemma

A clock keeps mean solar time, in which every day is exactly 24 hours (Section 1.2).

A sundial does not agree with it. In mid-February the shadow of a sundial reaches noon about 14 minutes after the clock does, and in early November about 16 minutes before. The days measured by the real Sun are not all equal.

Two runners go round a track. One runs at a steady pace. The other runs faster on one side of the track and slower on the other, but finishes each lap together with the first. The gap between them grows and shrinks over the lap and never adds up.

The steady runner is the **mean Sun**, an imagined Sun that moves along the celestial equator at a uniform rate. The other is the real Sun. It moves along the ecliptic at a varying speed, because the Earth's orbit is an ellipse, and its motion along the tilted ecliptic counts unevenly in right ascension. The **equation of time** is the gap between the two, read as a time.

In numbers, the eccentricity of the Earth's orbit is $e = 0.0167$. Kepler's second law makes the Sun run ahead of the mean Sun by up to $2e = 0.0334$ radians, which is $1.91°$. At four minutes per degree, that is 7.66 minutes.

If we photograph the Sun at the same clock time every few days, its changing declination moves it up and down, while the equation of time moves it left and right. Together those motions trace the figure-eight analemma. We can calculate the clock difference by combining the changing orbital speed with the tilt of the ecliptic.

Definition. Apparent solar time is the Sun's hour angle, Eq. (1.16). **Mean solar time** MST is the hour angle of the mean Sun, shifted in the same way. The mean Sun's right ascension α_{mean} grows uniformly, and it equals the **mean longitude** $\bar{\lambda}_\odot$, the ecliptic longitude the Sun would have if it moved at the uniform rate n. The difference of the two times is

Result 1.5

$$\text{EoT} \equiv \text{AST} - \text{MST} = \alpha_{\text{mean}} - \alpha_\odot. \tag{1.19}$$

The last step follows from Eq. (1.15), since the two hour angles are LST $-\,\alpha_\odot$ and LST $-\,\alpha_{\text{mean}}$, read at the same LST.

Splitting into two parts. We add and subtract the Sun's true ecliptic longitude $\lambda_\odot$,

$$\begin{aligned} \text{EoT} &\overset{(1)}{=} \bar{\lambda}_\odot - \alpha_\odot \\ &= \underbrace{\left(\bar{\lambda}_\odot - \lambda_\odot\right)}_{\text{uneven speed}} + \underbrace{\left(\lambda_\odot - \alpha_\odot\right)}_{\text{tilt}}. \end{aligned} \tag{1.20}$$

(1) $\alpha_{\text{mean}} = \bar{\lambda}_\odot$, from the definition of the mean Sun.

The first part would vanish on a circular orbit. The second would vanish if the axis were not tilted.

Uneven speed, from Kepler's second law. The Sun's yearly path is the Earth's orbit seen from the other end, so it has the same eccentricity e, true anomaly v and mean anomaly ℓ as the Earth's orbit (Chapter 6). Both $\lambda_\odot$ and $\bar{\lambda}_\odot$ are measured from $\hat{\gamma}$, and they differ from v and ℓ by the same fixed angle, the longitude of perihelion. So $\bar{\lambda}_\odot - \lambda_\odot = \ell - v$. To find $v - \ell$ we write the rate of the true anomaly with angular momentum conservation (P1),

$$\begin{aligned} \frac{\mathrm{d}v}{\mathrm{d}t} &\overset{(1)}{=} \frac{L}{\mu r^2} \\ &\overset{(2)}{=} \frac{\sqrt{GMa(1-e^2)}\,(1+e\cos v)^2}{a^2(1-e^2)^2} \\ &\overset{(3)}{=} n\,\frac{(1+e\cos v)^2}{(1-e^2)^{3/2}} \\ &\overset{(4)}{\simeq} n\,(1+2e\cos v). \end{aligned} \tag{1.21}$$

(1) Eq. (6.22). The true anomaly differs from the polar angle ϕ by a constant, so they share a rate.
(2) The orbit $r = a(1-e^2)/(1+e\cos v)$ from Eq. (6.28), whose semi-latus rectum $L^2/(GM\mu^2)$ equals $a(1-e^2)$, so $L/\mu = \sqrt{GMa(1-e^2)}$.
(3) Kepler's third law, $GM = n^2a^3$, Eq. (6.50).
(4) Keep only the first order in e, since $e^2 = 2.8 \times 10^{-4}$.

The mean anomaly is $\ell = n(t - t_0)$, with t_0 the time of perihelion, where $v = \ell = 0$. Separating the variables and integrating from perihelion,

$$\begin{aligned} \ell &\overset{(1)}{=} \int_0^v \frac{dv'}{1 + 2e\cos v'} \\ &\overset{(2)}{\simeq} \int_0^v (1 - 2e\cos v')\, dv' \\ &= v - 2e\sin v. \end{aligned} \tag{1.22}$$

(1) $n\,dt = dv/(1 + 2e\cos v)$ from the last line of the previous chain.
(2) $1/(1 + x) \simeq 1 - x$ for small x, again to first order in e.

So $v - \ell \simeq 2e\sin v \simeq 2e\sin\ell$, where replacing v by ℓ in the small term changes it only at order e^2. The Sun runs ahead of the mean Sun from perihelion to aphelion and behind it in the other half year.

Tilt, from the rotation. Eq. (1.18) gives $\tan\alpha_\odot = \cos\epsilon\tan\lambda_\odot$. To find $\lambda_\odot - \alpha_\odot$ we take its tangent,

$$\begin{aligned} \tan(\lambda_\odot - \alpha_\odot) &\overset{(1)}{=} \frac{\tan\lambda_\odot - \tan\alpha_\odot}{1 + \tan\lambda_\odot\tan\alpha_\odot} \\ &\overset{(2)}{=} \frac{(1 - \cos\epsilon)\tan\lambda_\odot}{1 + \cos\epsilon\tan^2\lambda_\odot} \\ &\overset{(3)}{=} \frac{(1 - \cos\epsilon)\sin\lambda_\odot\cos\lambda_\odot}{\cos^2\lambda_\odot + \cos\epsilon\sin^2\lambda_\odot} \\ &\overset{(4)}{=} \frac{(1 - \cos\epsilon)\sin 2\lambda_\odot}{(1 + \cos\epsilon) + (1 - \cos\epsilon)\cos 2\lambda_\odot} \\ &\overset{(5)}{=} \frac{\tan^2(\epsilon/2)\,\sin 2\lambda_\odot}{1 + \tan^2(\epsilon/2)\cos 2\lambda_\odot}. \end{aligned} \tag{1.23}$$

(1) The tangent of a difference.
(2) $\tan\alpha_\odot = \cos\epsilon\tan\lambda_\odot$.
(3) Multiply the top and the bottom by $\cos^2\lambda_\odot$.
(4) Multiply the top and the bottom by 2, and use $2\sin\lambda\cos\lambda = \sin 2\lambda$, $2\cos^2\lambda = 1 + \cos 2\lambda$ and $2\sin^2\lambda = 1 - \cos 2\lambda$.
(5) Divide the top and the bottom by $1 + \cos\epsilon$, and use $(1 - \cos\epsilon)/(1 + \cos\epsilon) = \tan^2(\epsilon/2)$.

For $\epsilon = 23.44°$, $\tan^2(\epsilon/2) = 0.0430$ is small, and expanding the arctangent to first order gives

$$\lambda_\odot - \alpha_\odot \simeq \tan^2(\epsilon/2)\,\sin 2\lambda_\odot.$$

The next term, $-\frac{1}{2}\tan^4(\epsilon/2)\sin 4\lambda_\odot$, is at most 0.21 minutes. The tilt term vanishes at the equinoxes and at the solstices. Near an equinox the ecliptic crosses the equator at an angle, so only the fraction $\cos\epsilon$ of the Sun's motion counts in right ascension, and $\alpha_\odot$ falls behind. Near a solstice the ecliptic runs parallel to the equator but closer to the pole, where an hour of right ascension is a shorter arc, so $\alpha_\odot$ catches up.

The equation of time. With both parts in Eq. (1.20), and each radian turned into $4 \times 57.296 = 229.2$ minutes,

$$\begin{aligned} \mathrm{EoT} &\overset{(1)}{=} -(v - \ell) + (\lambda_\odot - \alpha_\odot) \\ &\overset{(2)}{\simeq} -2e\sin\ell + \tan^2(\epsilon/2)\,\sin 2\lambda_\odot \\ &\overset{(3)}{=} -7.66^{\mathrm{m}}\sin\ell + 9.86^{\mathrm{m}}\sin 2\lambda_\odot. \end{aligned} \tag{1.24}$$

(1) Eq. (1.20) with $\bar{\lambda}_\odot - \lambda_\odot = \ell - v$.
(2) The two results above, each to first order.
(3) $2e = 0.0334$ and $\tan^2(\epsilon/2) = 0.0430$ radians, times 229.2 minutes per radian.

$$\boxed{\mathrm{EoT} \simeq -7.66^{\mathrm{m}} \sin \ell + 9.86^{\mathrm{m}} \sin 2\lambda_\odot .} \tag{1.25}$$

The first term has a period of one year and is set by e alone. The second has a period of half a year and is set by ϵ alone. Neither depends on where the observer is.

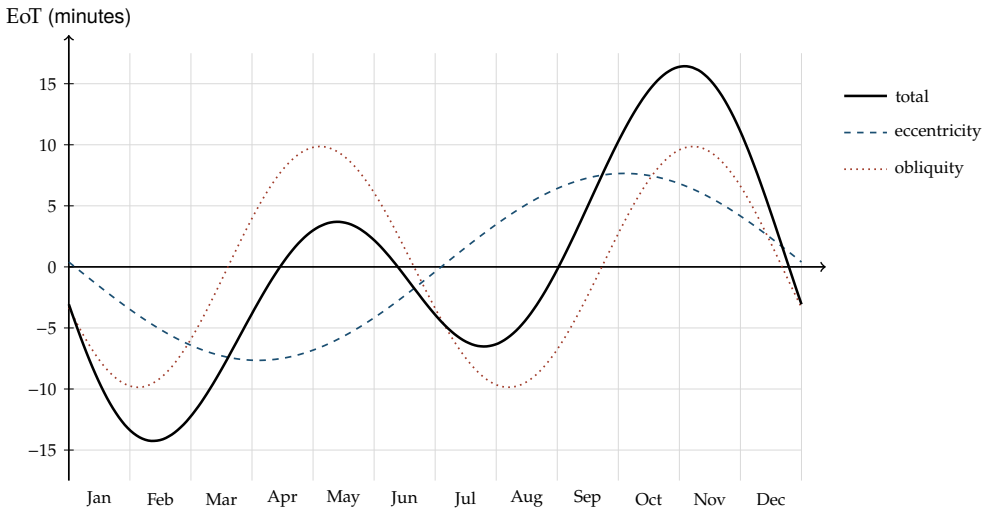


Figure 1.4: The equation of time over one year (solid), with the eccentricity term $-7.66^{\mathrm{m}} \sin \ell$ (dashed) and the obliquity term $9.86^{\mathrm{m}} \sin 2\lambda_\odot$ (dotted). The solid curve is computed from Eq. (1.19) without expanding in e or ϵ.

Figure 1.4 compares Eq. (1.25) with the exact gap. The two terms add up to within 0.3 minutes of it. The perihelion is in early January, so $\sin \ell$ peaks in early April, while $\sin 2\lambda_\odot$ peaks in early May and early November. The two terms add in November and in February. The equation of time ranges from -14.2 minutes in mid-February to $+16.4$ minutes in early November, with smaller turning points of $+3.7$ minutes in mid-May and -6.5 minutes in late July.

Worked example 1.2: Apparent noon at IISER TVM

Local mean noon at IISER TVM is at 12:21:27 IST (Section 1.4). The Sun crosses the meridian when AST $= 12^{\mathrm{h}}$, that is when MST $= 12^{\mathrm{h}} -$ EoT. On 3 November, with EoT $= +16.4^{\mathrm{m}} = +16^{\mathrm{m}}24^{\mathrm{s}}$, the Sun crosses at 12:05:03 IST. On 11 February, with EoT $= -14.2^{\mathrm{m}}$, it crosses at 12:35:39 IST.

The analemma. Plot the Sun's declination $\delta_\odot$ (Eq. (1.18)) against the equation of time for each day of the year. The curve is a figure eight, the **analemma**. Its height is $2\epsilon = 46.9°$. Its two loops are unequal because the eccentricity term adds to the obliquity term in one half of the year and works against it in the other.

1.8 Precession of the Equinoxes and Reference Epochs

So far the Earth's axis has pointed at a fixed place among the stars, and the equinox $\hat{\gamma}$ has been a fixed zero point.

Over centuries this fails. Hipparchus, comparing his star positions with ones made 150 years earlier, found that the equinox had moved west among the stars. The pole star today, Polaris,

was not near the pole in his time.

Spin a top and tilt it. Gravity pulls its centre down and tries to tip it over. It does not fall. Its axis sweeps slowly round a cone instead, sideways to the pull.

The Earth is a spinning top with a bulge at its equator. The Sun and the Moon pull on the bulge and try to turn the equator into the plane of the ecliptic. The axis responds by sweeping round a cone about the pole of the ecliptic. A torque at right angles to the spin turns the axis sideways, which is angular momentum conservation (P1).

A bicycle wheel makes a good gyroscope. Take a wheel of mass 2 kg and radius 0.3 m, spinning 10 times a second, with its axle held horizontal at one end 0.2 m from the wheel. Its angular momentum is $2 \times 0.3^2 \times 2\pi \times 10 = 11.3\,\mathrm{kg\,m^2\,s^{-1}}$, and the weight gives a torque of $2 \times 9.81 \times 0.2 = 3.92\,\mathrm{N\,m}$. The axle swings round at $3.92/11.3 = 0.35$ radians per second, once in 18 seconds.

The Earth's axis traces a circle of radius 23.44° on the sky in about 26 000 years. As its direction changes, the celestial equator and its intersection with the ecliptic change too. The equinox therefore drifts through the stars, so a catalogue position must state the date to which its coordinate grid refers.

The gyroscope rule. Let $\hat{p}$ be the unit vector along the spin axis and $\hat{k}$ the north pole of the ecliptic, at angle ϵ from $\hat{p}$. Suppose the angular momentum $\vec{L}$ turns steadily about $\hat{k}$ at rate ω_{prec}, with ϵ fixed. In the ecliptic frame, with z along $\hat{k}$, $\vec{L}(t) = R_z(-\omega_{\mathrm{prec}} t)\,\vec{L}(0)$, where the minus sign anticipates that the turn goes clockwise seen from the north (we find this below). Differentiating at $t = 0$,

$$\frac{\mathrm{d}\vec{L}}{\mathrm{d}t} \overset{(1)}{=} -\omega_{\mathrm{prec}} \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \vec{L} \overset{(2)}{=} -\omega_{\mathrm{prec}}\, \hat{k} \times \vec{L}. \tag{1.26}$$

(1) The derivative of Eq. (1.3) at $\varphi = 0$ is the matrix shown, and $\varphi = -\omega_{\mathrm{prec}} t$.

(2) The matrix times (L_x, L_y, L_z) is $(-L_y, L_x, 0) = \hat{k} \times \vec{L}$.

The size of $\hat{k} \times \vec{L}$ is $L \sin\epsilon$, so

$$\omega_{\mathrm{prec}} = \frac{|\dot{\vec{L}}|}{L \sin\epsilon}. \tag{1.27}$$

The rate change $\dot{\vec{L}}$ is the torque, $\sum \vec{x} \times \vec{F}$ over the Earth's mass (Eq. (6.99)). The torque must be perpendicular to $\vec{L}$ for L and ϵ to stay fixed.

The pull on the bulge. Let a body of mass M, the Sun or the Moon, be at distance a from the Earth's centre in the direction $\hat{s}$. A mass element dm of the Earth at $\vec{x}$ from the centre, with $|\vec{x}| \ll a$, feels the force $dm\, GM(a\hat{s} - \vec{x})/|a\hat{s} - \vec{x}|^3$. To first order in $|\vec{x}|/a$,

$$\begin{aligned} \frac{a\hat{s} - \vec{x}}{|a\hat{s} - \vec{x}|^3} &\overset{(1)}{=} \frac{a\hat{s} - \vec{x}}{a^3}\left(1 - \frac{2\hat{s}\cdot\vec{x}}{a} + \frac{x^2}{a^2}\right)^{-3/2} \\ &\overset{(2)}{\simeq} \frac{a\hat{s} - \vec{x}}{a^3}\left(1 + \frac{3\hat{s}\cdot\vec{x}}{a}\right) \\ &\overset{(3)}{\simeq} \frac{1}{a^2}\left[\hat{s} + \frac{3(\hat{s}\cdot\vec{x})\hat{s} - \vec{x}}{a}\right]. \end{aligned} \tag{1.28}$$

(1) $|a\hat{s} - \vec{x}|^2 = a^2 - 2a\,\hat{s}\cdot\vec{x} + x^2$.

(2) $(1+y)^{-3/2} \simeq 1 - \frac{3}{2}y$, dropping x^2/a^2.
(3) Multiply out and drop the term of order x^2/a^3.

The torque about the Earth's centre is

$$\begin{aligned}\dot{\vec{L}} &\overset{(1)}{=} \frac{GM}{a^2}\Big(\int \vec{x}\, dm\Big) \times \hat{s} + \frac{GM}{a^3}\int \big[3(\hat{s}\cdot\vec{x})\,\vec{x}\times\hat{s} - \vec{x}\times\vec{x}\big]dm \\ &\overset{(2)}{=} \frac{3GM}{a^3}\Big(\int \vec{x}\,\vec{x}^{\mathsf{T}} dm\; \hat{s}\Big)\times\hat{s}. \end{aligned} \tag{1.29}$$

(1) $\vec{x}\times$ the last line of the previous chain, times $GM\,dm$, summed over the Earth.
(2) $\int \vec{x}\,dm = 0$ about the centre of mass, $\vec{x}\times\vec{x} = 0$, and $(\hat{s}\cdot\vec{x})\,\vec{x} = \vec{x}\,\vec{x}^{\mathsf{T}}\hat{s}$.

The matrix $\int \vec{x}\,\vec{x}^{\mathsf{T}} dm$ holds the Earth's shape. In axes fixed in the Earth, with z along $\hat{p}$, it is diagonal, and by symmetry about the axis $\int x^2 dm = \int y^2 dm$. The moments of inertia about the polar axis and about an equatorial axis are $C = \int (x^2+y^2)\,dm = 2\int x^2 dm$ and $A = \int (y^2+z^2)\,dm = \int x^2 dm + \int z^2 dm$, so

$$\begin{aligned}\int \vec{x}\,\vec{x}^{\mathsf{T}} dm &= \begin{pmatrix} \int x^2 dm & 0 & 0 \\ 0 & \int x^2 dm & 0 \\ 0 & 0 & \int z^2 dm \end{pmatrix} \\ &\overset{(1)}{=} \Big(\int x^2 dm\Big)\,\mathbb{1} - (C-A)\,\hat{p}\,\hat{p}^{\mathsf{T}}. \end{aligned} \tag{1.30}$$

(1) $\int z^2 dm - \int x^2 dm = A - C$, and $\hat{p}\,\hat{p}^{\mathsf{T}}$ is the matrix with a single 1 in the zz place.

The term with $\mathbb{1}$ gives $\hat{s}\times\hat{s} = 0$. The flattening $C - A$ alone produces the torque,

$$\dot{\vec{L}} = \frac{3GM}{a^3}\,(C-A)\,(\hat{p}\cdot\hat{s})\,\big(\hat{s}\times\hat{p}\big). \tag{1.31}$$

A round Earth, $C = A$, feels no torque. Nor does an Earth with the Sun in its equatorial plane, $\hat{p}\cdot\hat{s} = 0$.

Averaged over the orbit. The Sun goes round the ecliptic in a year. That is long compared with a day and short compared with the precession period, so the axis hardly moves in a year and we may average Eq. (1.31) over $\hat{s}$ with $\hat{p}$ held fixed (P6). In the ecliptic frame, $\hat{s} = (\cos\theta, \sin\theta, 0)$ for a circular orbit, and

$$\begin{aligned}\big\langle (\hat{p}\cdot\hat{s})\,\hat{s}\times\hat{p}\big\rangle &\overset{(1)}{=} \big(\big\langle \hat{s}\,\hat{s}^{\mathsf{T}}\big\rangle\,\hat{p}\big)\times\hat{p} \\ &\overset{(2)}{=} \left[\begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix}\hat{p}\right]\times\hat{p} \\ &\overset{(3)}{=} \tfrac{1}{2}\big(\hat{p} - \cos\epsilon\,\hat{k}\big)\times\hat{p} \\ &\overset{(4)}{=} -\tfrac{1}{2}\cos\epsilon\;\hat{k}\times\hat{p}. \end{aligned} \tag{1.32}$$

(1) $(\hat{p}\cdot\hat{s})\,\hat{s} = \hat{s}\,\hat{s}^{\mathsf{T}}\hat{p}$, and $\hat{p}$ is constant during the average.
(2) $\big\langle\cos^2\theta\big\rangle = \big\langle\sin^2\theta\big\rangle = \frac{1}{2}$ and $\langle\cos\theta\sin\theta\rangle = 0$ over one turn, and $\hat{s}$ has no z part.
(3) The matrix removes the z part of $\hat{p}$, which is $(\hat{p}\cdot\hat{k})\hat{k} = \cos\epsilon\,\hat{k}$, and halves the rest.
(4) $\hat{p}\times\hat{p} = 0$.

With $\vec{L} = C\omega_\oplus \hat{p}$, which holds to the accuracy of $\omega_{\text{prec}}/\omega_\oplus \sim 10^{-7}$ (P6),

$$\begin{aligned} \dot{\vec{L}} &\overset{(1)}{=} -\frac{3}{2}\frac{GM}{a^3}(C-A)\cos\epsilon\ \hat{k}\times\hat{p} \\ &\overset{(2)}{=} -\left[\frac{3}{2}\frac{GM}{a^3}\frac{C-A}{C}\frac{\cos\epsilon}{\omega_\oplus}\right]\hat{k}\times\vec{L}. \end{aligned} \tag{1.33}$$

(1) Eq. (1.31), averaged with the result above.
(2) $\hat{p} = \vec{L}/(C\omega_\oplus)$.

This has the form $-\omega_{\text{prec}}\,\hat{k}\times\vec{L}$ of the gyroscope rule. The bracket is positive, so the axis turns clockwise seen from the north ecliptic pole, against the Earth's spin and orbit. The equinox, where the moving equator meets the ecliptic, therefore moves west. For the Sun, $GM_\odot/a^3 = n^2$ by Kepler's third law, Eq. (6.50), so

$$\boxed{\omega_{\text{prec},\odot} = \frac{3}{2}\frac{n^2}{\omega_\oplus}\frac{C-A}{C}\cos\epsilon.} \tag{1.34}$$

The rate grows with the flattening and falls with the spin. A faster top precesses more slowly.

Numbers for the Sun and the Moon. Satellite orbits give the Earth's flattening as $(C-A)/(M_\oplus R_\oplus^2) = 1.0826\times10^{-3}$, and the density profile from seismology gives $C = 0.3307\,M_\oplus R_\oplus^2$, so $(C-A)/C = 0.003274$. With $n/\omega_\oplus = T_{\text{sid}}/T_{\text{yr}} = 1/366.2422$ from Eq. (1.7),

$$\omega_{\text{prec},\odot} = \frac{3}{2}\times\frac{360^\circ}{\text{year}}\times\frac{1}{366.2422}\times 0.003274\times 0.9175 = 15.9'' \text{ per year}.$$

The Moon gives the same expression with $GM_\odot/a^3$ replaced by $GM_{\text{Moon}}/a_{\text{Moon}}^3$. That is larger by $(M_{\text{Moon}}/M_\odot)(1\,\text{AU}/a_{\text{Moon}})^3 = 2.18$, with $a_{\text{Moon}} = 384\,400$ km, so the Moon gives 34.7″ per year. The Moon's orbit is tilted by 5° to the ecliptic, which changes this by about one percent. The total, about 50.6″ per year, is within one percent of the measured value,

$$\boxed{\omega_{\text{prec}} \simeq 50.29'' \text{ per year} = 1^\circ \text{ in } 71.6 \text{ years},} \tag{1.35}$$

one turn in $1\,296\,000''/50.29'' = 25\,770$ years. The measured rate also includes a slow turn of the ecliptic itself, caused by the planets, of about 0.1″ per year.

Epochs. The grid of (α, δ) is tied to the moving equinox $\hat{\gamma}$, so a catalogue position must state the moment, or **epoch**, whose equator and equinox it uses (Section 2.1.1). The current standard is **J2000.0**, 2000 January 1, 12:00:00 TT.

Precessing a catalogue position. In the frame of the moving equinox, every star turns about the ecliptic pole at ω_{prec} per year, anticlockwise seen from the north, so its ecliptic longitude grows by 50.29″ per year. In equatorial coordinates the ecliptic pole is $\hat{k} = (0, -\sin\epsilon, \cos\epsilon)$ (Eq. (1.17)). The star's unit vector $\hat{r}$ then changes as

$$\begin{aligned} \frac{\mathrm{d}\hat{r}}{\mathrm{d}t} &\overset{(1)}{=} \omega_{\text{prec}}\,\hat{k}\times\hat{r} \\ &\overset{(2)}{=} \omega_{\text{prec}}\begin{pmatrix} 0 & -\cos\epsilon & -\sin\epsilon \\ \cos\epsilon & 0 & 0 \\ \sin\epsilon & 0 & 0 \end{pmatrix}\begin{pmatrix} \cos\delta\cos\alpha \\ \cos\delta\sin\alpha \\ \sin\delta \end{pmatrix}. \end{aligned} \tag{1.36}$$

(1) A turn about $\hat{k}$ at rate ω_{prec}, as in the gyroscope rule but anticlockwise.

(2) $\hat{k} \times \hat{r}$ written as a matrix acting on $\hat{r}$, with $\hat{k} = (0, -\sin\epsilon, \cos\epsilon)$.

The left side is $\dot{\hat{r}} = \dot{\alpha}\,\partial_\alpha \hat{r} + \dot{\delta}\,\partial_\delta \hat{r}$. Its third row is $\cos\delta\,\dot{\delta} = \omega_{\text{prec}} \sin\epsilon \cos\delta \cos\alpha$. Its second row is

$$\cos\delta\cos\alpha\,\dot{\alpha} - \sin\delta\sin\alpha\,\dot{\delta} = \omega_{\text{prec}} \cos\epsilon \cos\delta \cos\alpha.$$

Putting $\dot{\delta}$ from the third row into the second and dividing by $\cos\delta\cos\alpha$,

Result 1.6

$$\alpha(t) \simeq \alpha_0 + \left(\omega_{\text{prec}} \cos\epsilon + \omega_{\text{prec}} \sin\epsilon\, \sin\alpha_0 \tan\delta_0\right) \Delta t, \tag{1.37}$$

$$\delta(t) \simeq \delta_0 + \left(\omega_{\text{prec}} \sin\epsilon\, \cos\alpha_0\right) \Delta t, \tag{1.38}$$

for a star at (α_0, δ_0) at J2000.0 and Δt years later, as long as the change stays small. With $\omega_{\text{prec}} = 50.29''$ per year, $\omega_{\text{prec}} \cos\epsilon = 46.14''$ and $\omega_{\text{prec}} \sin\epsilon = 20.00''$ per year. The tabulated values, $46.12'' = 3.075^{\text{s}}$ and $20.04'' = 1.336^{\text{s}}$ per year, differ in the last digits because of the slow turn of the ecliptic. The first term moves every star equally in α. The second depends on where the star is, and it grows near the celestial poles, where $\tan\delta_0$ is large.

Worked example 1.3: Sirius from J2000.0 to late 2026

Sirius is at $\alpha_0 = 6^{\text{h}}45^{\text{m}}09^{\text{s}}$, $\delta_0 = -16°43'$. Its right ascension grows at $3.075^{\text{s}} + 1.336^{\text{s}} \sin 101.29° \tan(-16.72°) = 2.68^{\text{s}}$ per year, and its declination changes at $20.04'' \cos 101.29° = -3.92''$ per year. Over 26.8 years, α grows by 72^{s} and δ falls by $105''$. The star's own motion across the sky, its proper motion, is a separate correction (Chapter 3).

1.9 Nutation: Mean and Apparent Sidereal Time

The averaged torque of Section 1.8 turns the axis smoothly. The real torque does not stay at its average. It changes as the Sun and the Moon move north and south of the equator, and the Moon's orbit itself turns, once in 18.6 years. The axis wobbles about its smooth cone with small periodic terms, called **nutation**.

The largest term follows the 18.6-year turn of the Moon's orbit. It moves the equinox back and forth along the ecliptic by up to $17.2''$. This shift is the **nutation in longitude** $\Delta\psi$.

Mean and true equinox. The **mean equinox** moves with the smooth precession alone. The **true equinox** also includes nutation. Sidereal time measured from the mean equinox is **Greenwich mean sidereal time** (GMST). Measured from the true equinox it is **Greenwich apparent sidereal time** (GAST).

The equation of the equinoxes. Sidereal time is the hour angle of the equinox. A shift $\Delta\psi$ of the equinox along the ecliptic is a small turn about the ecliptic pole, which is what Eq. (1.37) describes with $\omega_{\text{prec}}\Delta t$ replaced by $\Delta\psi$. For a point on the equator, $\delta = 0$, so only the first term survives, and the shift along the equator is $\Delta\psi \cos\epsilon$. So

Result 1.7

$$\text{GAST} - \text{GMST} = \Delta\psi \cos\epsilon, \tag{1.39}$$

the **equation of the equinoxes**. Here ϵ is the true obliquity. For the largest term, $17.2'' \times 0.917 = 15.8''$, or 1.05^{s}. With the smaller terms the difference stays below 1.2^{s}. GMST is enough for pointing a telescope. GAST is needed for positions better than an arcsecond.

1.10 Modern Time Scales: UT1, TAI, UTC, TT and TDB

Every clock so far has been the turning Earth. The turning Earth is not a perfect clock. The day grows longer as the tides slow the Earth, and it changes by milliseconds with the winds and with the motion of the Earth's liquid core.

A flywheel with a little friction in its bearing loses speed slowly. A clock that counts its turns falls behind a quartz watch, by more each day. The tides raised by the Moon are dragged ahead of the Moon by the Earth's spin, and the Moon pulls back on them. Angular momentum (P1) flows from the Earth's spin into the Moon's orbit, and the Moon moves away by about 3.8 cm a year. Records of ancient eclipses show that the day has grown longer by about 1.7 ms per century on average. Atomic clocks keep a steady time, and the time kept by the Earth is corrected against them.

Even a 1 ms excess over 86 400 atomic seconds each day makes an Earth-rotation clock fall behind by 0.37 s in a year. Atomic time supplies a steadier rate, while Earth-rotation time keeps track of the actual orientation of our observatory. The time scales below serve different needs by using one of these rates or by specifying how they are kept together.

1. **UT1 (Universal Time).** The angle through which the Earth has actually turned, read as a time. It is measured with radio telescopes (VLBI) that time the arrival of signals from distant quasars. Because of tidal friction, winds and the core, UT1 does not run at a uniform rate.
2. **TAI (International Atomic Time).** A uniform time kept by a weighted average of several hundred atomic clocks around the world. Its second is the SI second, 9 192 631 770 periods of the radiation from the ground-state hyperfine transition of caesium-133.
3. **UTC (Coordinated Universal Time).** Civil time. UTC runs at the rate of TAI, but whole **leap seconds** are added at the end of June or December to keep $|\mathrm{UTC} - \mathrm{UT1}| < 0.9\,\mathrm{s}$. Since 2017,
$$\mathrm{TAI} - \mathrm{UTC} = 37\,\mathrm{s}.$$
In 2022 the General Conference on Weights and Measures decided to stop adding leap seconds by 2035.
4. **TT (Terrestrial Time).** The uniform time of a clock on the geoid, the surface of mean sea level, free of the Earth's irregular spin. It is defined to run with TAI,
$$\mathrm{TT} = \mathrm{TAI} + 32.184\,\mathrm{s}, \qquad \mathrm{TT} - \mathrm{UTC} = 37\,\mathrm{s} + 32.184\,\mathrm{s} = 69.184\,\mathrm{s}.$$
The 32.184 s keeps TT continuous with the older Ephemeris Time.
5. **TDB (Barycentric Dynamical Time).** The time coordinate for the motion of bodies in the Solar System, referred to its centre of mass, the barycentre. It is used for planetary ephemerides, spacecraft navigation and pulsar timing. It differs from TT because a clock on the Earth moves and sits in the Sun's gravity.

TT against TDB. A clock on the Earth ticks proper time. In a weak field, its rate against the barycentric time t is
$$\frac{\mathrm{d}\tau}{\mathrm{d}t} = 1 - \frac{1}{c^2}\left(\frac{v^2}{2} + \frac{GM_\odot}{r}\right),$$

with v the Earth's orbital speed and r its distance from the Sun. The Earth's own gravity and spin add a constant, which the definition of TT absorbs. The orbital terms are not constant, because the orbit is an ellipse. With the vis-viva relation $v^2 = GM_\odot(2/r - 1/a)$, which follows from Eq. (6.34) with $E = \frac{1}{2}\mu v^2 - GM\mu/r$,

$$\frac{v^2}{2} + \frac{GM_\odot}{r} = \frac{2GM_\odot}{r} - \frac{GM_\odot}{2a}.$$

The time average of $1/r$ over an orbit is $1/a$, by the same substitution used below. The average rate is a constant, which the definitions of the time scales take up. The part that changes over the year adds up to

$$\begin{aligned}
\mathrm{TDB} - \mathrm{TT} &\overset{(1)}{=} \frac{2GM_\odot}{c^2} \int \left(\frac{1}{r} - \frac{1}{a}\right) dt \\
&\overset{(2)}{=} \frac{2GM_\odot}{c^2} \int \left(\frac{1}{a(1 - e\cos u)} - \frac{1}{a}\right) \frac{(1 - e\cos u)}{n}\, du \\
&= \frac{2GM_\odot}{c^2\, na} \int e\cos u\, du \\
&\overset{(3)}{=} \frac{2e\sqrt{GM_\odot a}}{c^2} \sin u\,.
\end{aligned} \tag{1.40}$$

(1) The rate above, with its mean removed, integrated over time. The clock on the Earth runs slow when $1/r$ is above its mean, so TT falls behind TDB there.

(2) $r = a(1 - e\cos u)$, Eq. (6.42), and $dt = (1 - e\cos u)\, du/n$ from Eq. (6.56). Here u is the eccentric anomaly of the Earth's orbit. The same steps show that $\int dt/r$ over one orbit is $2\pi/(na)$, so $1/r$ averages to $1/a$.

(3) $GM_\odot/(na) = na^2 = \sqrt{GM_\odot a}$ by Eq. (6.50), and the constant of integration is part of the definition.

With $e = 0.0167$, $GM_\odot = 1.327 \times 10^{20}\ \mathrm{m^3\,s^{-2}}$ and $a = 1.496 \times 10^{11}\ \mathrm{m}$,

$$\boxed{\mathrm{TDB} \simeq \mathrm{TT} + 0.001657\,\mathrm{s}\ \sin u\,, \qquad u \simeq \ell + e\sin\ell\,.} \tag{1.41}$$

The second relation is Kepler's equation, Eq. (6.52), to first order in e, with ℓ the Earth's mean anomaly. The difference is 1.7 ms at most and depends only on the eccentricity and the size of the orbit. The Moon and the planets add terms of a few microseconds. A pulsar timed to a microsecond needs all of them.

1.11 Julian Date and Modified Julian Date

A calendar date is hard to subtract. Months have different lengths, leap years come and go, and the calendar itself was changed in 1582. Astronomers count days instead, with one number that never resets.

Julian Date. The **Julian Date** (JD) is the number of days since Greenwich mean noon on 1 January 4713 BC, in the Julian calendar. The day starts at noon, so that a night of observing has a single date.

Modified Julian Date. The **Modified Julian Date** (MJD) moves the start of the day to midnight and removes the large leading digits,

Result 1.8

$$\mathrm{MJD} \equiv \mathrm{JD} - 2\,400\,000.5. \tag{1.42}$$

MJD = 0.0 is midnight at the start of 17 November 1858.

Standard epochs.

- **J2000.0**, JD = 2 451 545.0, 2000 January 1, 12:00:00 TT.
- **B1950.0**, JD = 2 433 282.423.

From a calendar date. For year Y, month M ($1 \le M \le 12$) and day D, with the time of day as a fraction of D, the following rule (Meeus 1998) gives the Julian Date for a date in the Gregorian calendar. If M is 1 or 2, replace Y by $Y-1$ and M by $M+12$. Then

Result 1.9

$$\mathrm{JD} = \lfloor 365.25(Y+4716) \rfloor + \lfloor 30.6001(M+1) \rfloor + D + 2 - \lfloor Y/100 \rfloor + \lfloor Y/400 \rfloor - 1524.5. \tag{1.43}$$

Each piece has a job. Moving January and February to the end of the previous year puts the leap day at the very end of the year, so it never shifts the months that follow it. The first term counts the days in whole years, one in four of them a leap year. From March, the month lengths run 31, 30, 31, 30, 31 and then repeat, 153 days in five months, or 30.6 a month, so the second term counts the days in the whole months before M. The terms $2 - \lfloor Y/100 \rfloor + \lfloor Y/400 \rfloor$ remove the leap days that the Gregorian rule drops in century years. The constant 1524.5 sets the zero at noon in 4713 BC.

For J2000.0, $Y = 1999$, $M = 13$ and $D = 1.5$. Then $\lfloor 365.25 \times 6715 \rfloor = 2\,452\,653$, $\lfloor 30.6001 \times 14 \rfloor = 428$, and $2 - 19 + 4 = -13$, so

$$\mathrm{JD} = 2\,452\,653 + 428 + 1.5 - 13 - 1524.5 = 2\,451\,545.0,$$

as listed above.

1.11.1 Sidereal Time on Any Date

To point a telescope at a given moment we need LST, and by Eq. (1.14) that needs GMST. The IAU formula gives GMST at 0^{h} UT1 from the number of Julian centuries since J2000.0, $T = (\mathrm{JD} - 2\,451\,545.0)/36\,525$,

$$\mathrm{GMST}(0^{\mathrm{h}}\mathrm{UT1}) = 6^{\mathrm{h}}41^{\mathrm{m}}50.54841^{\mathrm{s}} + 8\,640\,184.812866^{\mathrm{s}}\,T + 0.093104^{\mathrm{s}}\,T^2 - 6.2\times10^{-6\,\mathrm{s}}\,T^3. \tag{1.44}$$

The linear term is the daily gain of the sidereal clock. A Julian century has 36 525 days, and $8\,640\,184.81^{\mathrm{s}}/36\,525 = 236.555^{\mathrm{s}}$ per day, which is the $3^{\mathrm{m}}56.555^{\mathrm{s}}$ found in Section 1.2. At a time t_{UT} later in the same day, the sidereal clock has run faster by the ratio of the two days,

$$\mathrm{GMST}(t_{\mathrm{UT}}) = \mathrm{GMST}(0^{\mathrm{h}}\mathrm{UT1}) + 1.00273790935 \times t_{\mathrm{UT}}, \tag{1.45}$$

where $1.00273790935 = T_{\mathrm{sol}}/T_{\mathrm{sid}} = 366.2422/365.2422$ by Eq. (1.7).

Worked example 1.4: What is on the meridian at IISER TVM?

Take 2026 October 16 at 19:00 IST, which is 13:30 UT. Eq. (1.43) with $Y = 2026$, $M = 10$, $D = 16.0$ gives $\lfloor 365.25 \times 6742 \rfloor = 2\,462\,515$, $\lfloor 30.6001 \times 11 \rfloor = 336$ and $2 - 20 + 5 = -13$,

so JD = 2 461 329.5 at 0^{h} UT. Then $T = 0.267885$, and Eq. (1.44), reduced modulo 24^{h}, gives GMST(0^{h}) = $1^{h}38^{m}06.6^{s}$. Adding $1.0027379 \times 13.5^{h} = 13^{h}32^{m}13.1^{s}$ gives GMST = $15^{h}10^{m}19.6^{s}$. The longitude 77.1382°E is $5^{h}08^{m}33.2^{s}$, so by Eq. (1.14) LST = $20^{h}18^{m}52.8^{s}$. The bright star Deneb, at $\alpha = 20^{h}41^{m}$, has $H = -22^{m}$ and will cross the meridian about 22 minutes later.

1.12 Problem Set: Time Systems and Solar Motion

1.12.1 Problem 1: Local Solar Time and Noon Zenith Angle at IISER TVM

Problem Statement: IISER TVM is at latitude 8.6833°N and longitude 77.1382°E. Indian Standard Time (IST) is the mean solar time of the meridian at 82.5°E, near Prayagraj.

(a) What is the local mean solar time at IISER TVM when it is 12:00 noon IST?

(b) What is the zenith angle of the Sun at noon on the June solstice at IISER TVM?

(c) At what IST does the Sun cross the meridian at IISER TVM on 3 November, when the equation of time is +16.4 minutes?

Essential Ideas:

- The Earth turns 15° per hour relative to the mean Sun, so local mean time differs from standard time by four minutes per degree of longitude, Eq. (1.11). Solar time follows the Sun, and sidereal time follows the stars. A sidereal clock gains about four minutes a day on a solar one (Figure 1.5).

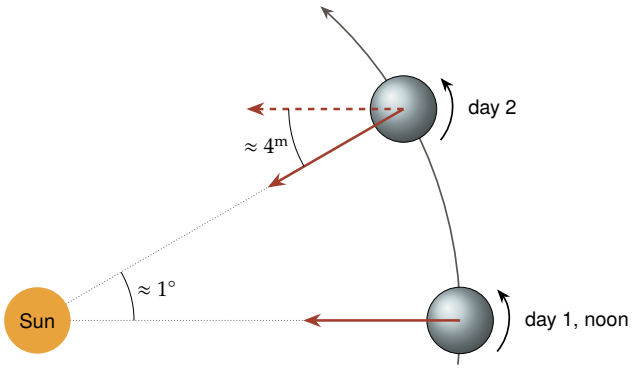


Figure 1.5: The Earth on two successive days, as in Figure 1.1, with the angle moved along the orbit in one day (≈ 1°) and the extra spin needed to face the Sun again (≈ 4^{m}).

- At noon the Sun is on the meridian, and its declination $\delta_{\odot}$ is the latitude of the point on the Earth where it is overhead. The Sun's rays are parallel, so the noon zenith angle is the angle at the Earth's centre between the observer and that point (Figure 1.6). On the June solstice $\delta_{\odot} = +\epsilon = +23.44°$ (Table 1.2).
- The Sun crosses the meridian when apparent solar time is 12^{h}, which is when mean solar time is 12^{h} − EoT, Eq. (1.19).

Detailed Solutions:

(a) **Local mean solar time**. IISER TVM is west of the IST meridian by

$$\lambda_{\text{ref}} - \lambda_{\text{obs}} = 82.5° - 77.1382° = 5.3618°.$$

At four minutes per degree this is

$$5.3618° \times 4^{m}/° = 21.4472^{m} = 21^{m}26.83^{s} \approx 21^{m}27^{s}.$$

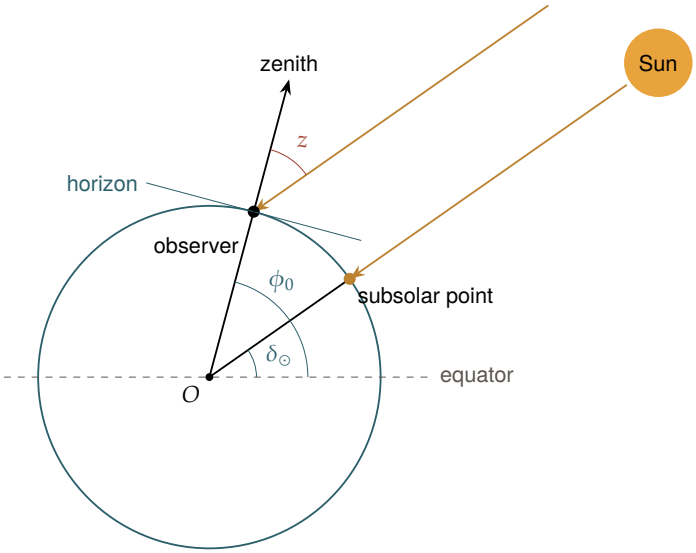


Figure 1.6: A cut through the Earth along the observer's meridian at local noon. The observer is at latitude ϕ_0 and the subsolar point at latitude $\delta_\odot$. Parallel rays from the Sun reach both, and z is the angle at the observer between the zenith and the Sun.

IISER TVM is west of the reference meridian, so its noon comes later and its local time lags IST,

$$\boxed{\text{local mean solar time} = 12{:}00{:}00 - 0{:}21{:}27 = 11{:}38{:}33.}$$

(b) **Zenith angle at solstice noon**. The radii to the observer and to the subsolar point make angles ϕ_0 and $\delta_\odot$ with the equator. The zenith at the observer lies along its radius, and the ray to the Sun is parallel to the radius through the subsolar point. So the angle between them at the observer equals the angle between the two radii at the centre, $|\phi_0 - \delta_\odot|$. (The same result follows from the full rotation between horizon and equator, Eq. (2.33).) With $\phi_0 = 8.6833^\circ$ and $\delta_\odot = 23.44^\circ$,

$$z = |\phi_0 - \delta_\odot| = |8.6833^\circ - 23.44^\circ| = 14.7567^\circ, \qquad \boxed{z \approx 14.76^\circ.}$$

The Sun is north of the zenith, since $\delta_\odot > \phi_0$. Its altitude above the northern horizon is $90^\circ - z = 75.24^\circ$.

(c) **Apparent noon on 3 November**. Mean noon at IISER TVM is at 12:21:27 IST, from part (a). The Sun crosses the meridian when mean solar time is $12^{\mathrm{h}} - 16.4^{\mathrm{m}}$, that is $16^{\mathrm{m}}24^{\mathrm{s}}$ before mean noon,

$$\boxed{12{:}21{:}27 - 0{:}16{:}24 = 12{:}05{:}03 \text{ IST}.}$$

A sundial at IISER TVM in early November reads noon only five minutes after the IST clock does, against twenty-one minutes on average. Anyone setting up a telescope or a sundial from a civil clock must allow for both the longitude and the equation of time.

Chapter 2

Coordinate Systems on the Celestial Sphere

Chapter 1 gave us the celestial sphere and a clock tied to its rotation. To point a telescope, we must locate a direction on that sphere with two angles. Those angles depend on the reference plane we choose: the horizon is natural for an observer, the Earth's equator for a star catalogue, the orbital plane for planets, and the Milky Way's disk for Galactic objects.

The star's direction is the same in every system; only its coordinates change. We will represent that direction by a unit vector and transform it by turning the axes. Once the geometry is settled, we will account for two shifts in the direction seen by a telescope: bending by the air and aberration from the Earth's motion.

2.1 The Equatorial Coordinate System

Chapter 1 introduced the celestial poles, the celestial equator and the vernal equinox $\hat{\gamma}$, and built a sidereal clock from them (Section 1.1, Eq. (1.12)). These define the equatorial grid in which a catalogue gives a star's right ascension α and declination δ. The grid does not rotate with the Earth over a night; the observer's meridian rotates through it. To compute angles and change frames, we first turn each catalogued direction into a vector.

Think of a town on a globe. Its latitude says how far north of the equator it lies. Its longitude says how far east of Greenwich it lies, counted along the equator. An arrow from the centre of the globe to the town is fixed by these two angles. Its height above the plane of the equator is the sine of the latitude. Its shadow on that plane has the cosine of the latitude as its length, and it points in the direction given by the longitude. The sky uses the same two angles. Declination plays the part of latitude, and right ascension plays the part of longitude, with $\hat{\gamma}$ in place of Greenwich. A direction on the sky is a unit vector whose three components are products of sines and cosines of these two angles.

For Vega, right ascension is $18^{\mathrm{h}}36^{\mathrm{m}}56^{\mathrm{s}} = 279.23^\circ$ and declination is $+38.78^\circ$. Its direction has a component $\sin 38.78^\circ = 0.626$ above the equatorial plane. The projection onto that plane has length $\cos 38.78^\circ = 0.779$ and points 279.23° around from $\hat{\gamma}$, giving components $0.779\cos 279.23^\circ = 0.125$ and $0.779\sin 279.23^\circ = -0.769$. This split between height and projection gives the unit vector used in every coordinate transformation below.

The two angles. The **declination** $\delta \in [-90^\circ, +90^\circ]$ is the angle of a direction north (+) or south (−) of the celestial equator. The great circle through the two poles and a star is the star's **hour circle**. The **right ascension** α is the angle along the celestial equator from $\hat{\gamma}$ eastward to the hour circle. It runs from 0^{h} to 24^{h}, with $1^{\mathrm{h}} = 15^\circ$ (Eq. (1.1)). Both are shown in Figure 2.2.

The unit vector. Put the z axis toward the north celestial pole and the x axis toward $\hat{\gamma}$. The y axis then points to $\alpha = 6^{\mathrm{h}}$, which makes the three axes right-handed. Split the direction $\hat{\mathbf{r}}$ of a star into a part along the pole and a part in the plane of the equator,

$$\begin{aligned}\hat{\mathbf{r}} &\overset{(1)}{=} \sin\delta\,\hat{\mathbf{z}} + \cos\delta\,\bigl(\cos\alpha\,\hat{\mathbf{x}} + \sin\alpha\,\hat{\mathbf{y}}\bigr)\\ &\overset{(2)}{=} \cos\delta\cos\alpha\,\hat{\mathbf{x}} + \cos\delta\sin\alpha\,\hat{\mathbf{y}} + \sin\delta\,\hat{\mathbf{z}}.\end{aligned} \tag{2.1}$$

(1) The angle between $\hat{\mathbf{r}}$ and the plane of the equator is δ. So the part along the pole is $\sin\delta$, and the shadow in the plane has length $\cos\delta$ and makes the angle α with $\hat{\mathbf{x}}$.
(2) Multiplying out the bracket.

We write the three components as a column, labelled by the frame they belong to,

$$\boxed{\hat{\mathbf{r}}_{\mathrm{eq}} = \begin{pmatrix}\cos\delta\cos\alpha\\ \cos\delta\sin\alpha\\ \sin\delta\end{pmatrix}.} \tag{2.2}$$

Its length is one for every (α, δ). It does not depend on the time or on where on the Earth the observer stands, because the grid turns with the stars.

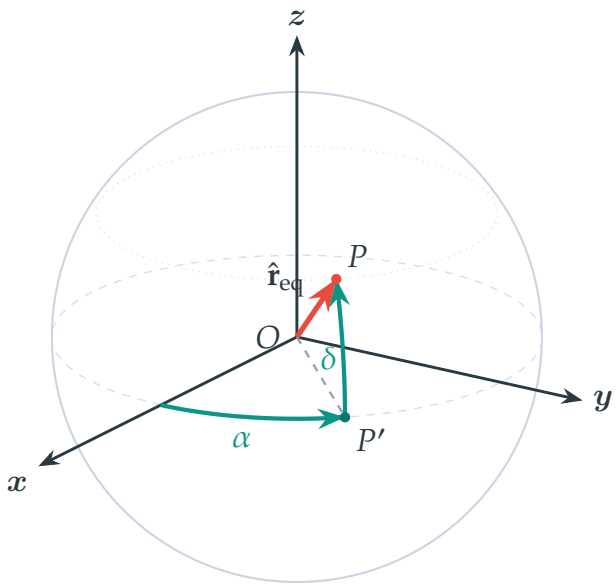


Figure 2.1: The unit vector $\hat{\mathbf{r}}_{\mathrm{eq}}$ toward a point P in the equatorial frame, with its right ascension α measured from the x axis and its declination δ measured up from the equatorial plane.

The values of (α, δ) in a catalogue change only slowly, through the star's own motion across the sky and through precession. The daily rising and setting of the stars comes from the observer's horizon turning under this fixed grid.

The grid of Figure 2.3 is the one that planetarium software draws on the real sky (Figure 2.4). In both, α grows clockwise around the pole for an observer facing north. The reason is that a star of larger α reaches the meridian later, so at any moment it lies further east, and east is on the right for someone facing north.

The local sidereal time and the hour angle of Chapter 1 are two more angles on the same sphere. Figure 2.5 marks $\hat{\gamma}$, the observer's meridian and a star S together. The right ascension now on the meridian is the local sidereal time, Eq. (1.12). In that figure the star is still east of the

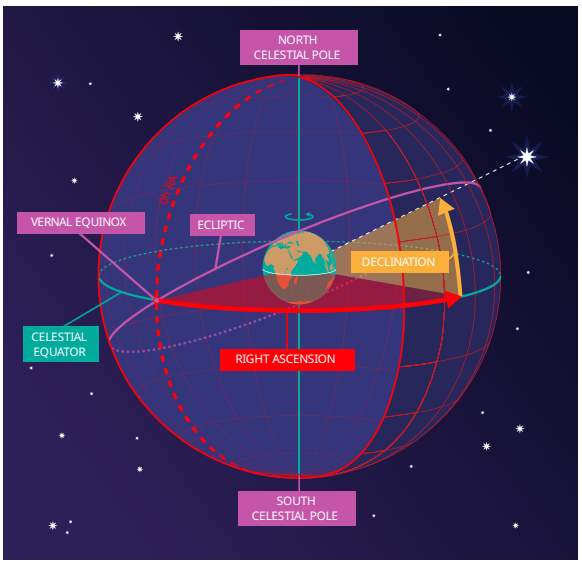


Figure 2.2: The celestial sphere around the Earth, with the north and south celestial poles, the celestial equator, the ecliptic, the vernal equinox, and the right ascension and declination of a star.

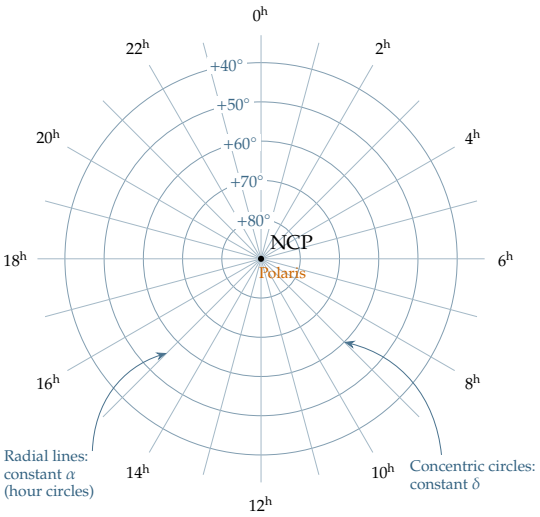


Figure 2.3: The equatorial grid near the north celestial pole, as seen by an observer facing north. Radial lines are hour circles 1^h apart. Circles are parallels of constant δ, $10°$ apart.

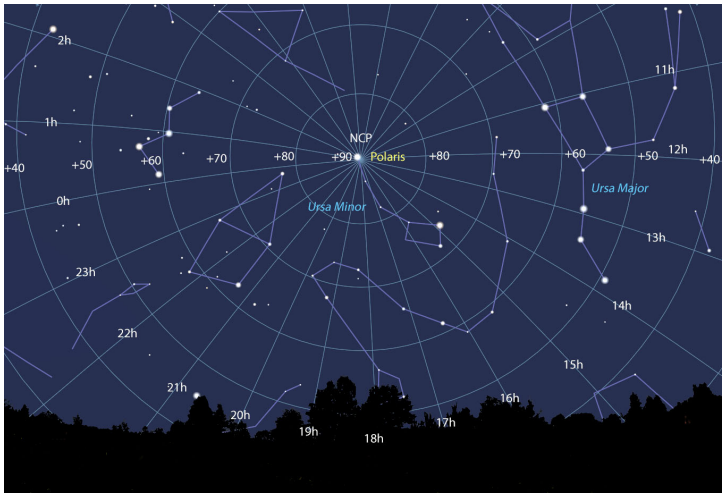


Figure 2.4: An equatorial grid drawn over the sky near the north celestial pole, with Polaris and the constellations Ursa Minor and Ursa Major.

meridian. As the Earth turns, the LST grows while the star's α stays fixed. Figure 2.6 shows the same sky later, after the meridian has passed the star. The arc from the meridian back to the star is then the hour angle H, defined in Section 2.3. In the earlier figure the same arc points the other way and has length $-H$.

2.1.1 Coordinate Epochs

The equinox $\hat{\gamma}$ that anchors the grid moves. The Earth's axis precesses, and $\hat{\gamma}$ slides west along the ecliptic by about 50.3″ per year, Eq. (1.35). A position (α, δ) therefore means nothing until it states the instant at which the grid was taken. That instant is the **epoch**. The standard today is J2000.0, and older catalogues used B1950.0. Some names carry the epoch. The pulsar PSR J1939+2134 is named after its J2000 position, $\alpha \approx 19^h39^m$ and $\delta \approx +21°34'$.

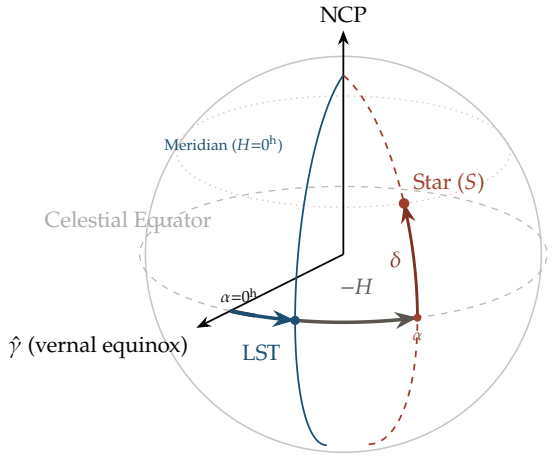


Figure 2.5: The vernal equinox $\hat{\gamma}$, the observer's meridian and a star S at (α, δ) east of the meridian, on one celestial sphere.

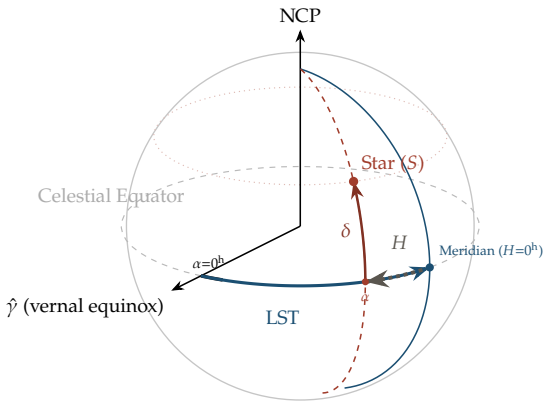


Figure 2.6: The same sky as Figure 2.5 at a later instant. $\hat{\gamma}$ and S are unchanged, and the meridian has moved past the star.

Moving a position from one epoch to another needs two corrections. One is the precession of the grid, Eqs. (1.37)–(1.38). The other is the star's own motion across the sky, its proper motion $(\mu_\alpha^*, \mu_\delta)$ of Chapter 3. The two are independent, and over a few decades they simply add. To restate a catalogue from B1950.0 to J2000.0 with the stars held fixed, only the precession is needed.

2.2 Turning the Axes

The equatorial frame gives each star a column of three numbers, Eq. (2.2). The other frames of the sky put their pole or their zero point somewhere else. We need the same star's components in those frames.

Think of a plan of a room drawn with the x axis along one wall. A lamp stands 3 m along that wall from the corner. A friend draws the same room with axes turned 30° anticlockwise. The lamp has not moved. Only the axes have. Its new coordinates are its distances measured along the friend's axes, and each of those is a dot product of the lamp's position with one of the friend's axes. A star's direction does not change when we change frame either. The idea is that the new components are dot products with the new axes, so the matrix of the change has the new axes as its rows.

The lamp in numbers. The friend's axes are $(\cos 30°, \sin 30°)$ and $(-\sin 30°, \cos 30°)$ in the old frame. The lamp at $(3, 0)$ has new coordinates $3\cos 30° = 2.60$ and $-3\sin 30° = -1.50$. Stacking the two axes as rows gives the same answer as one matrix product,

$$\begin{pmatrix} \cos 30° & \sin 30° \\ -\sin 30° & \cos 30° \end{pmatrix} \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} 2.60 \\ -1.50 \end{pmatrix}.$$

The lamp has stayed in place; its numbers change because the axes have turned. We will use this same rule for sky coordinates. It also explains why the inverse transformation is the transpose of the original rotation matrix.

The rule. Let a new frame have orthonormal axes $\hat{\mathbf{x}}', \hat{\mathbf{y}}', \hat{\mathbf{z}}'$, each written by its components in the old frame. The new components of the direction $\hat{\mathbf{r}}$ are

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} \hat{\mathbf{x}}' \cdot \hat{\mathbf{r}} \\ \hat{\mathbf{y}}' \cdot \hat{\mathbf{r}} \\ \hat{\mathbf{z}}' \cdot \hat{\mathbf{r}} \end{pmatrix} \overset{(2)}{=} \underbrace{\begin{pmatrix} — \, \hat{\mathbf{x}}'^T \, — \\ — \, \hat{\mathbf{y}}'^T \, — \\ — \, \hat{\mathbf{z}}'^T \, — \end{pmatrix}}_{\mathbf{M}} \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{2.3}$$

(1) In an orthonormal frame each component of a vector is its dot product with one axis.
(2) Each dot product is the row of the new axis's old components times the column of old components.

The entries of $\mathbf{M}\mathbf{M}^T$ are the dot products of the new axes with each other,

$$\mathbf{M}\mathbf{M}^T = \begin{pmatrix} \hat{\mathbf{x}}' \cdot \hat{\mathbf{x}}' & \hat{\mathbf{x}}' \cdot \hat{\mathbf{y}}' & \hat{\mathbf{x}}' \cdot \hat{\mathbf{z}}' \\ \hat{\mathbf{y}}' \cdot \hat{\mathbf{x}}' & \hat{\mathbf{y}}' \cdot \hat{\mathbf{y}}' & \hat{\mathbf{y}}' \cdot \hat{\mathbf{z}}' \\ \hat{\mathbf{z}}' \cdot \hat{\mathbf{x}}' & \hat{\mathbf{z}}' \cdot \hat{\mathbf{y}}' & \hat{\mathbf{z}}' \cdot \hat{\mathbf{z}}' \end{pmatrix} = \mathbb{1}, \qquad \text{so} \qquad \boxed{\mathbf{M}^{-1} = \mathbf{M}^T.} \tag{2.4}$$

Going back to the old frame multiplies by the transpose, whose columns are the new axes. The determinant of $\mathbf{M}$ is $+1$ when the new axes are right-handed, and then $\mathbf{M}$ is a rotation. It is -1 when they are left-handed. The horizontal frame of Section 2.3 is such a case.

Axes and components turn with the same matrix. The rows of $\mathbf{M}$ are the new axes written in the old ones, so the same matrix also gives the axes themselves. Write the old axes as $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$ (the $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$ above) and the new ones with primes. For a turn by the angle ϵ about the x axis, the turn that takes the equatorial frame to the ecliptic frame of Section 2.7,

$$\begin{pmatrix} \mathbf{e}'_x \\ \mathbf{e}'_y \\ \mathbf{e}'_z \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\epsilon & \sin\epsilon \\ 0 & -\sin\epsilon & \cos\epsilon \end{pmatrix} \begin{pmatrix} \mathbf{e}_x \\ \mathbf{e}_y \\ \mathbf{e}_z \end{pmatrix}.$$

This is a **passive** rotation matrix. It turns the axes and leaves the vector where it is. An **active** rotation would instead turn the vector by ϵ and keep the axes fixed, and its matrix is the transpose of this one. Every matrix in this chapter is passive, because changing frames moves the axes and never the stars. One thing looks odd at first. The axes and the components of one vector should change in opposite ways, so that the vector itself stays where it is, yet Eq. (2.3) turns the components with the same matrix. Indices settle it. Let $R_i{}^j$ be the entry in row i and column j of the matrix, so that $\mathbf{e}'_i = R_i{}^j \mathbf{e}_j$, and drop the summation sign over a repeated index, so that $V^j \mathbf{e}_j$ means $\sum_j V^j \mathbf{e}_j$. A vector $\mathbf{V}$ with components V^j in the old frame and V'^i in the new one then gives

$$\begin{aligned} & V^j \mathbf{e}_j \overset{(1)}{=} V'^i \mathbf{e}'_i \overset{(2)}{=} V'^i R_i{}^j \mathbf{e}_j \\ \Longrightarrow \quad & V^j \overset{(3)}{=} R_i{}^j V'^i \quad \Longrightarrow \quad V \overset{(4)}{=} \mathbf{R}^T V' \quad \Longrightarrow \quad V' \overset{(5)}{=} \left(\mathbf{R}^T\right)^{-1} V = \mathbf{R}\, V. \end{aligned} \tag{2.5}$$

(1) The vector $\mathbf{V}$ is the same in both frames. Only its components and its axes differ.
(2) $\mathbf{e}'_i = R_i{}^j \mathbf{e}_j$.
(3) The old axes are independent, so the coefficients of each $\mathbf{e}_j$ agree on the two sides.
(4) The sum runs over i, the first index of $\mathbf{R}$, its row. Summing over the row index is a product with the transpose. V and V' are the columns of components.
(5) $\mathbf{R}^{-1} = \mathbf{R}^T$ for a rotation, Eq. (2.4).

The axes go forward with $\mathbf{R}$ and the components go back with $\mathbf{R}^T$, the inverse law that keeps the vector fixed. The difference sits only in where the summed index stands. The sum over the row index brings in a transpose, and solving for the new components turns $\mathbf{R}^T$ back into $\mathbf{R}$, so the two laws are written with the same matrix.

Turning about the z axis. Turn the x and y axes by θ about z, anticlockwise as seen from $+z$. The new axes are $\hat{\mathbf{x}}' = \cos\theta\,\hat{\mathbf{x}} + \sin\theta\,\hat{\mathbf{y}}$, $\hat{\mathbf{y}}' = -\sin\theta\,\hat{\mathbf{x}} + \cos\theta\,\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}' = \hat{\mathbf{z}}$. Stacked as rows by Eq. (2.3),

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \underbrace{\begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{R}_z(\theta)} \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{2.6}$$

The same matrix from polar angles. A second route uses angles in the plane of the turn. Take the turn about the x axis, in the y–z plane (Figure 2.7). A point at distance r from the axis and at angle Θ from the old y axis has $(y, z) = (r\cos\Theta, r\sin\Theta)$. Measured from the new y' axis, turned by θ, its angle is $\Theta - \theta$. So

$$\begin{aligned} y' &= r\cos(\Theta-\theta) \overset{(1)}{=} r\cos\Theta\cos\theta + r\sin\Theta\sin\theta \overset{(2)}{=} y\cos\theta + z\sin\theta, \\ z' &= r\sin(\Theta-\theta) \overset{(1)}{=} r\sin\Theta\cos\theta - r\cos\Theta\sin\theta \overset{(2)}{=} -y\sin\theta + z\cos\theta. \end{aligned} \tag{2.7}$$

(1) The angle-difference identities.
(2) $r\cos\Theta = y$ and $r\sin\Theta = z$.

With $x' = x$ this is

$$\mathbf{R}_x(\theta) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}, \tag{2.8}$$

which is $\mathbf{R}_z(\theta)$ with the axes relabelled $x \to y \to z \to x$, as the rule of rows gives it.

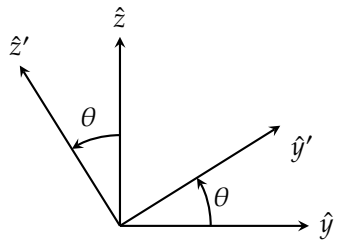


Figure 2.7: The y and z axes and the axes y' and z' turned by θ about the x axis.

What a turn about the pole does to a star. Apply $\mathbf{R}_z(\theta)$ to the equatorial column (2.2),

$$\begin{aligned} \mathbf{R}_z(\theta)\,\hat{\mathbf{r}}_{\text{eq}} &\overset{(1)}{=} \begin{pmatrix} \cos\delta\cos\alpha\cos\theta + \cos\delta\sin\alpha\sin\theta \\ -\cos\delta\cos\alpha\sin\theta + \cos\delta\sin\alpha\cos\theta \\ \sin\delta \end{pmatrix} \\ &\overset{(2)}{=} \begin{pmatrix} \cos\delta\cos(\alpha-\theta) \\ \cos\delta\sin(\alpha-\theta) \\ \sin\delta \end{pmatrix}. \end{aligned} \tag{2.9}$$

(1) Each row of Eq. (2.6) times the column.
(2) The angle-difference identities.

A turn by θ about the pole subtracts θ from every right ascension and leaves every declination alone.

2.3 Horizontal Coordinates

The equatorial frame gives a star an almost fixed direction over a night, Eq. (2.2). The horizon frame is attached to the observer and rotates with the Earth. A star's (α, δ) alone therefore cannot say whether it is above the horizon now. We need the observer's latitude and the current orientation of the meridian, supplied by local sidereal time. Their difference $H = \mathrm{LST} - \alpha$ is the star's hour angle: it changes as the Earth turns even while α and δ stay nearly fixed.

Think of pointing out an aeroplane to a friend. You say "thirty degrees up, toward the east-north-east". One angle is the height above the horizon. The other is the compass direction, counted round from north. Neither needs any knowledge of the sky, only of the ground. The answer changes from minute to minute because the plane moves. For a star it is the ground that moves. The **altitude** and **azimuth** use the zenith as their pole and turn with the Earth. The hour angle measures how far the sky has turned since the star crossed the meridian. The idea is that the horizontal frame is the equatorial frame turned about the pole to the meridian, and then tipped about the east–west line so that the zenith takes the place of the pole.

At IISER TVM, latitude is $\phi_0 = 8.68°\mathrm{N}$. The north celestial pole is therefore $8.68°$ above the northern horizon. A star on the celestial equator crosses the meridian $8.68°$ south of the zenith, at altitude $90° - 8.68° = 81.32°$. The transformation below extends this meridian geometry to any hour angle, so we can find where to point at any time.

The angles. The **altitude** $h \in [-90°, +90°]$ is the angle of a star above the horizon, $+90°$ at the zenith and $-90°$ at the nadir. The **zenith angle** is its distance from the zenith,

$$z = 90° - h. \tag{2.10}$$

The **azimuth** $A \in [0°, 360°)$ is the angle along the horizon from north ($0°$) through east ($90°$), south ($180°$) and west ($270°$). The **hour angle** of Eq. (1.15) is the angle along the celestial equator from the meridian westward to the star's hour circle,

$$H = \mathrm{LST} - \alpha, \tag{2.11}$$

with $H > 0$ after the star has crossed the meridian.

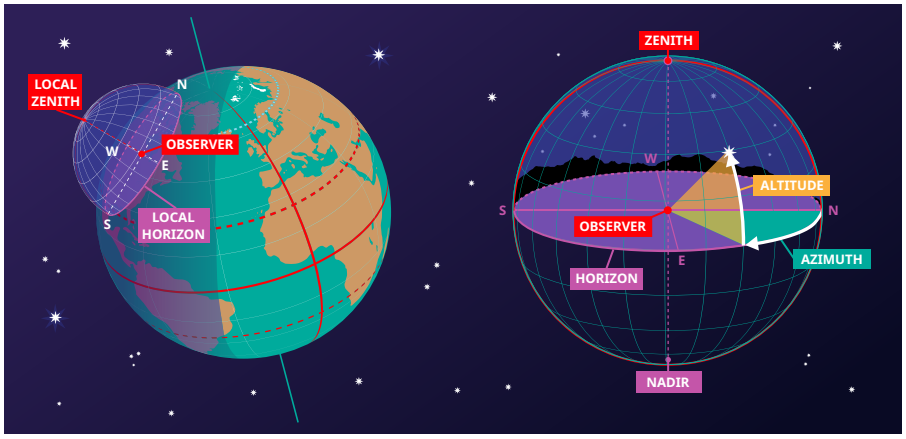


Figure 2.8: Left, an observer on the Earth with the local horizon and zenith. Right, the observer's sky with the horizon, zenith, nadir, the cardinal points, and the altitude and azimuth of a star.

2.3.1 The Meridian Frame

To compare the two systems we first turn the equatorial axes about the pole until the x axis lies in the observer's meridian. The right ascension on the meridian is LST, so the turn is $\mathbf{R}_z(\mathrm{LST})$,

and Eq. (2.9) gives

$$\hat{\mathbf{r}}_{\text{mer}} \overset{(1)}{=} \begin{pmatrix} \cos \text{LST} & \sin \text{LST} & 0 \\ -\sin \text{LST} & \cos \text{LST} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\delta \cos\alpha \\ \cos\delta \sin\alpha \\ \sin\delta \end{pmatrix} \tag{2.12}$$

$$\overset{(2)}{=} \begin{pmatrix} \cos\delta \cos(\alpha - \text{LST}) \\ \cos\delta \sin(\alpha - \text{LST}) \\ \sin\delta \end{pmatrix} \tag{2.13}$$

$$\overset{(3)}{=} \begin{pmatrix} \cos\delta \cos H \\ -\cos\delta \sin H \\ \sin\delta \end{pmatrix}. \tag{2.14}$$

(1) Eq. (2.6) with $\theta = \text{LST}$, acting on Eq. (2.2).
(2) Eq. (2.9).
(3) $\alpha - \text{LST} = -H$ by Eq. (2.11), with $\cos(-H) = \cos H$ and $\sin(-H) = -\sin H$.

In this frame the x axis points to where the meridian crosses the celestial equator, the y axis points to the east point of the horizon ($H = -90°$, $\delta = 0$), and the z axis still points to the pole. The right ascension and the sidereal time now enter only through their difference H.

2.3.2 The Local Axes

Next we need the observer's north, east and up directions in the meridian frame.

Up. The zenith lies on the meridian. Its declination is the latitude ϕ_0, because the angle between the celestial equator and the zenith is the angle between the Earth's equator and the local vertical. Setting $H = 0$ and $\delta = \phi_0$ in Eq. (2.14) gives the up direction $\hat{\mathbf{u}}$.

East. The east point of the horizon is 90° from the zenith and 90° from the pole, since it is on the horizon and both the zenith and the pole lie on the meridian. So it lies on the celestial equator, a quarter turn east of the meridian, at $\delta = 0$ and $H = -90°$.

North. North is the horizontal direction that completes the frame, $\hat{\mathbf{n}} = \hat{\mathbf{u}} \times \hat{\mathbf{e}}$. With the two vectors just found,

$$\hat{\mathbf{u}} = \begin{pmatrix} \cos\phi_0 \\ 0 \\ \sin\phi_0 \end{pmatrix}, \qquad \hat{\mathbf{e}} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \tag{2.15}$$

$$\hat{\mathbf{n}} = \hat{\mathbf{u}} \times \hat{\mathbf{e}} \overset{(1)}{=} \begin{pmatrix} 0 \cdot 0 - \sin\phi_0 \cdot 1 \\ \sin\phi_0 \cdot 0 - \cos\phi_0 \cdot 0 \\ \cos\phi_0 \cdot 1 - 0 \cdot 0 \end{pmatrix} = \begin{pmatrix} -\sin\phi_0 \\ 0 \\ \cos\phi_0 \end{pmatrix}. \tag{2.16}$$

(1) The components of a cross product, $(u_y e_z - u_z e_y,\ u_z e_x - u_x e_z,\ u_x e_y - u_y e_x)$.

A second route gives the same $\hat{\mathbf{n}}$. The north point is on the meridian, 90° beyond the zenith, past the pole. Putting $H = 0$ and $\delta = 90° + \phi_0$ into Eq. (2.14) gives $(\cos(90° + \phi_0), 0, \sin(90° + \phi_0)) = (-\sin\phi_0, 0, \cos\phi_0)$. For later reference the east vector is

$$\hat{\mathbf{e}} = (0, 1, 0)^T. \tag{2.17}$$

The triple $(\hat{\mathbf{n}}, \hat{\mathbf{e}}, \hat{\mathbf{u}})$ is left-handed, $\hat{\mathbf{n}} \times \hat{\mathbf{e}} = -\hat{\mathbf{u}}$, because the azimuth runs clockwise as seen from above.

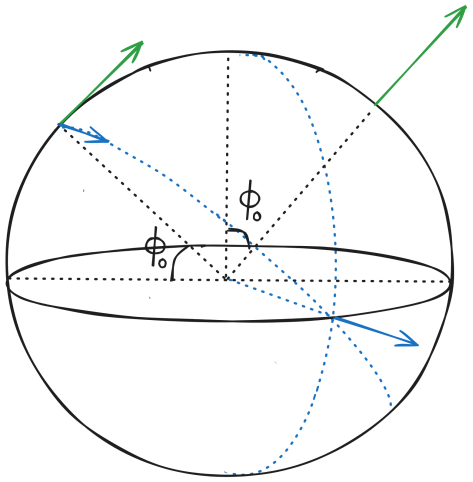


Figure 2.9: The local north (blue) and up (green) directions at an observer at latitude ϕ_0, drawn on the Earth.

2.3.3 The Transformation

The goal is (h, A) in terms of (δ, H). A star at altitude h and azimuth A has, by the same split as Eq. (2.2) with h for δ and A counted from north,

$$\hat{\mathbf{r}} = (\cos h \cos A)\,\hat{\mathbf{n}} + (\cos h \sin A)\,\hat{\mathbf{e}} + (\sin h)\,\hat{\mathbf{u}}. \tag{2.18}$$

In the meridian frame this is a matrix whose columns are the three local axes,

$$\hat{\mathbf{r}}_{\text{mer}} = \begin{pmatrix}\hat{\mathbf{n}} & \hat{\mathbf{e}} & \hat{\mathbf{u}}\end{pmatrix}\hat{\mathbf{r}}_{\text{hor}} = \underbrace{\begin{pmatrix}-\sin\phi_0 & 0 & \cos\phi_0\\ 0 & 1 & 0\\ \cos\phi_0 & 0 & \sin\phi_0\end{pmatrix}}_{\mathbf{M}(\phi_0)}\begin{pmatrix}\cos h\cos A\\ \cos h\sin A\\ \sin h\end{pmatrix}. \tag{2.19}$$

To solve for the horizontal column we invert $\mathbf{M}(\phi_0)$,

$$\begin{aligned}\begin{pmatrix}\cos h\cos A\\ \cos h\sin A\\ \sin h\end{pmatrix} &\overset{(1)}{=} \mathbf{M}^T(\phi_0)\,\hat{\mathbf{r}}_{\text{mer}}\\ &\overset{(2)}{=}\begin{pmatrix}-\sin\phi_0 & 0 & \cos\phi_0\\ 0 & 1 & 0\\ \cos\phi_0 & 0 & \sin\phi_0\end{pmatrix}\begin{pmatrix}\cos\delta\cos H\\ -\cos\delta\sin H\\ \sin\delta\end{pmatrix}\\ &\overset{(3)}{=}\begin{pmatrix}-\sin\phi_0\cos\delta\cos H + \cos\phi_0\sin\delta\\ -\cos\delta\sin H\\ \cos\phi_0\cos\delta\cos H + \sin\phi_0\sin\delta\end{pmatrix}.\end{aligned} \tag{2.20}$$

(1) The columns of $\mathbf{M}(\phi_0)$ are orthonormal, so $\mathbf{M}^{-1} = \mathbf{M}^T$, Eq. (2.4).
(2) $\mathbf{M}(\phi_0)$ is symmetric, so $\mathbf{M}^T = \mathbf{M}$. The column $\hat{\mathbf{r}}_{\text{mer}}$ is Eq. (2.14).
(3) Each row times the column.

Reading off the three rows,

Result 2.1

$$\sin h = \sin\phi_0 \sin\delta + \cos\phi_0 \cos\delta \cos H, \tag{2.21}$$
$$\cos h \sin A = -\cos\delta \sin H, \tag{2.22}$$
$$\cos h \cos A = -\sin\phi_0 \cos\delta \cos H + \cos\phi_0 \sin\delta. \tag{2.23}$$

The star enters only through δ, the time only through H, and the site only through ϕ_0. The site's longitude enters only through the LST inside H. The last row is also $\hat{\mathbf{u}} \cdot \hat{\mathbf{r}} = \cos z$, the dot product of the up vector (2.15) with Eq. (2.14).

The azimuth. Dividing Eq. (2.22) by Eq. (2.23),

$$\tan A = \frac{-\cos\delta \sin H}{-\sin\phi_0 \cos\delta \cos H + \cos\phi_0 \sin\delta} \overset{(1)}{=} \frac{\sin H}{\sin\phi_0 \cos H - \cos\phi_0 \tan\delta}. \tag{2.24}$$

(1) Numerator and denominator divided by $-\cos\delta$.

That division flips the sign of both parts, so the second form cannot tell A from $A + 180°$. The quadrant comes from the signs of Eqs. (2.22) and (2.23) themselves. $\sin A$ has the sign of $\cos h \sin A$, and $\cos A$ the sign of $\cos h \cos A$, because $\cos h \geq 0$.

Going back. The product of $\mathbf{M}(\phi_0)$ with itself is

$$\mathbf{M}(\phi_0)\,\mathbf{M}(\phi_0) = \begin{pmatrix} \sin^2\phi_0 + \cos^2\phi_0 & 0 & -\sin\phi_0\cos\phi_0 + \cos\phi_0\sin\phi_0 \\ 0 & 1 & 0 \\ -\cos\phi_0\sin\phi_0 + \sin\phi_0\cos\phi_0 & 0 & \cos^2\phi_0 + \sin^2\phi_0 \end{pmatrix} = \mathbb{1}, \tag{2.25}$$

so the same matrix also takes horizontal to equatorial coordinates. Eq. (2.19) multiplied out reads

$$\boxed{\begin{aligned} \sin\delta &= \sin\phi_0 \sin h + \cos\phi_0 \cos h \cos A, \\ \cos\delta \cos H &= \cos\phi_0 \sin h - \sin\phi_0 \cos h \cos A, \\ -\cos\delta \sin H &= \cos h \sin A. \end{aligned}} \tag{2.26}$$

These are Eqs. (2.21)–(2.23) with the pairs (δ, H) and (h, A) exchanged. The reason is that $\mathbf{M}(\phi_0)$ sends the pole $(0, 0, 1)^T$ to the zenith $\hat{\mathbf{u}}$ and the zenith back to the pole, and leaves east alone. It is a mirror, not a rotation. Its determinant is $-\sin^2\phi_0 - \cos^2\phi_0 = -1$, as it must be for a left-handed set of axes. Tipping the pole onto the zenith about the east–west line would be a rotation. Counting the azimuth clockwise reverses one axis on top of that.

The astronomical triangle. The pole, the zenith and the star are the corners of a spherical triangle (Figure 2.10). Its sides are the co-latitude $90° - \phi_0$, the star's distance from the pole $90° - \delta$, and its zenith angle z. Its angle at the pole is H. For a star west of the meridian its angle at the zenith is $360° - A$. Eq. (2.21) is the cosine rule of spherical trigonometry for the side z. The matrix route gives all three relations at once and with their signs. The third angle, at the star, is the parallactic angle q of Section 2.10.2.

Worked example, the Sun at an equinox. At an equinox $\delta_\odot = 0$. Three hours after local noon the Sun's hour angle is $H = 3^{\mathrm{h}} = 45°$. At IISER TVM, $\phi_0 = 8.6833°$, Eqs. (2.21)–(2.23) give

$$\sin h = \cos(8.6833°)\cos 45° = 0.6990, \qquad h = 44.35°,$$
$$\cos h \sin A = -\sin 45° = -0.7071,$$
$$\cos h \cos A = -\sin(8.6833°)\cos 45° = -0.1068.$$

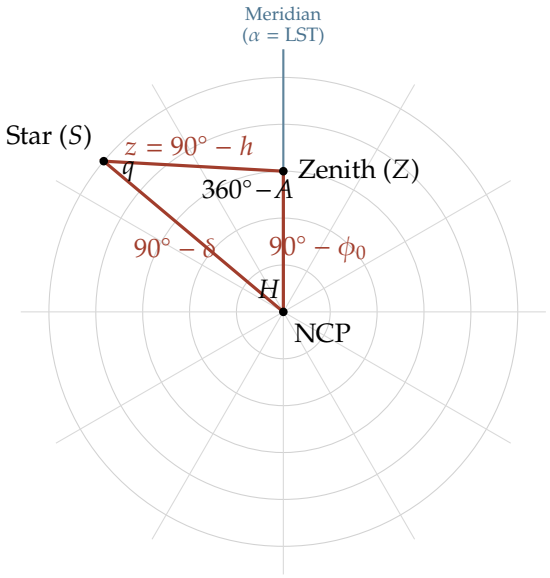


Figure 2.10: The pole–zenith–star triangle for a star west of the meridian, drawn over the equatorial grid near the pole as seen facing north. The sides are $90° - \phi_0$, $90° - \delta$ and z. The angles are H at the pole, $360° - A$ at the zenith and q at the star. The sides are drawn straight, though each is an arc of a great circle.

Both $\sin A$ and $\cos A$ are negative, so A is in the third quadrant,

$$A = 180° + \arctan(0.7071/0.1068) = 261.41°.$$

The Sun is 8.59° south of due west. At an equinox the Sun runs along the celestial equator, which crosses the horizon at the east and west points and passes south of the zenith at this site. So the Sun stays a little south of the east–west line all day.

2.4 Rising, Setting, and Circumpolar Stars

Eq. (2.21) gives the altitude of a star at any hour angle. To plan a night one wants the reverse, the times when a star is up at all.

Think of a big wheel at a fair whose lower part dips behind a fence. Seats near the hub may never go behind the fence. Seats on the rim spend part of each turn hidden. How much of each turn a seat is hidden depends on its distance from the hub and on how high the fence is compared with the hub. Each star moves once a sidereal day on a circle of constant δ around the pole. The horizon cuts that circle, and the height of the pole above the horizon is the latitude. The idea is to set $h = 0$ in Eq. (2.21) and solve for the hour angle.

At the Earth's equator, $\phi_0 = 0$, Eq. (2.21) becomes $\sin h = \cos\delta \cos H$. Every star reaches the horizon at $H = \pm 90° = \pm 6^{\mathrm{h}}$ and spends half a sidereal day above it. Away from the equator, the horizon cuts each star's daily circle differently (Figure 2.11). Setting $h = 0$ will tell us whether that circle crosses the horizon at all and, if it does, at which hour angles.

The rising and setting hour angle. A star rises or sets when its centre crosses the horizon, $h = 0$. In Eq. (2.21),

$$0 \overset{(1)}{=} \sin\phi_0 \sin\delta + \cos\phi_0 \cos\delta \cos H_0,$$
$$\cos H_0 \overset{(2)}{=} -\frac{\sin\phi_0 \sin\delta}{\cos\phi_0 \cos\delta}. \tag{2.27}$$

(1) Eq. (2.21) with $\sin h = 0$.
(2) Divided by $\cos\phi_0 \cos\delta$, which is positive away from the poles.

$$\boxed{\cos H_0 = -\tan\phi_0 \tan\delta.} \tag{2.28}$$

The equation has two roots. The star sets at $H = +H_0$ and rises at $H = -H_0$, with $H_0 = \arccos\bigl(-\tan\phi_0 \tan\delta\bigr)$ between $0°$ and $180°$. By Eq. (2.11) the sidereal times are

$$\mathrm{LST}_{\mathrm{rise}} = \alpha - H_0, \qquad \mathrm{LST}_{\mathrm{set}} = \alpha + H_0. \tag{2.29}$$

The star is up for $2H_0$ of sidereal time. A star north of the equator seen from a northern site has $\tan\phi_0 \tan\delta > 0$, so $H_0 > 90°$ and it is up for more than half the day. The formula ignores refraction, which lifts a star on the horizon by about half a degree (Section 2.9), so a star is seen to rise a little before $-H_0$.

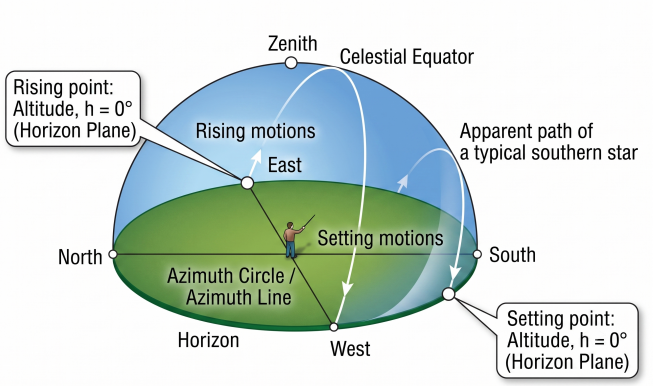


Figure 2.11: Paths of stars across an observer's sky, with a star rising in the east and setting in the west, and the shorter path of a southern star.

Stars that never set or never rise. A root exists only if $|\tan\phi_0 \tan\delta| \le 1$. Take a northern site, $\phi_0 > 0$. If $\tan\phi_0 \tan\delta > 1$ then $\cos H_0 < -1$, there is no crossing, and $h > 0$ all day. The star is **circumpolar**. The condition is

$$\tan\delta \overset{(1)}{>} \frac{1}{\tan\phi_0} \overset{(2)}{=} \tan\bigl(90° - \phi_0\bigr) \quad \Longleftrightarrow \quad \delta > 90° - \phi_0. \tag{2.30}$$

(1) $\tan\phi_0 \tan\delta > 1$ divided by $\tan\phi_0 > 0$.
(2) $1/\tan\phi_0 = \cot\phi_0 = \tan\bigl(90° - \phi_0\bigr)$. The last step uses that tan increases on $(-90°, 90°)$.

In the same way a star with $\tan\phi_0 \tan\delta < -1$, that is $\delta < -(90° - \phi_0)$, never rises. For a southern site the roles of the two poles swap.

A second route checks the condition. The star is lowest at lower transit, $H = 180°$, where Eq. (2.21) gives

$$\sin h_{\mathrm{low}} = \sin\phi_0 \sin\delta - \cos\phi_0 \cos\delta \overset{(1)}{=} -\cos\bigl(\phi_0 + \delta\bigr) \overset{(2)}{=} \sin\bigl(\phi_0 + \delta - 90°\bigr), \tag{2.31}$$

(1) The angle-sum identity for the cosine.
(2) $-\cos x = \sin(x - 90^\circ)$.

so $h_{\text{low}} = \phi_0 + \delta - 90^\circ$. This is above the horizon exactly when $\delta > 90^\circ - \phi_0$, as in Eq. (2.30).

From IISER TVM, $90^\circ - \phi_0 = 81.32^\circ$. Polaris, at $\delta = +89.26^\circ$, is circumpolar. It circles the pole between $h_{\text{low}} = 8.68^\circ + 89.26^\circ - 90^\circ = 7.95^\circ$ and 9.42°, low in the north all night.

2.5 Altitude at Meridian Transit

Eq. (2.21) gives the altitude at any hour angle. A star is highest when it crosses the meridian, at $H = 0$. That is the best moment to observe it, and for the Sun it is local noon. On the meridian the pole, the zenith and the star lie on one great circle. The zenith is the point of that circle at declination ϕ_0, so the zenith angle is just the difference of two declinations, as the distance between two towns on one meridian of the Earth is the difference of their latitudes. Chapter 1 used this for the noon Sun at IISER TVM in its problem set.

The highest altitude. Put $H = 0$ in Eq. (2.21),

$$\begin{aligned} \sin h_{\max} &\overset{(1)}{=} \sin\phi_0 \sin\delta + \cos\phi_0 \cos\delta \\ &\overset{(2)}{=} \cos\big(\phi_0 - \delta\big) \\ &\overset{(3)}{=} \sin\big(90^\circ - |\phi_0 - \delta|\big). \end{aligned} \tag{2.32}$$

(1) $\cos H = 1$.
(2) The angle-difference identity for the cosine.
(3) The cosine is even, so $\cos\big(\phi_0 - \delta\big) = \cos|\phi_0 - \delta|$, and $\cos x = \sin(90^\circ - x)$.

Both $h_{\max}$ and $90^\circ - |\phi_0 - \delta|$ lie between -90° and 90°, so they are equal, and by Eq. (2.10)

$$\boxed{z_{\min} = |\phi_0 - \delta|.} \tag{2.33}$$

The star crosses south of the zenith if $\delta < \phi_0$ and north of it if $\delta > \phi_0$. At local solar noon the Sun is on the meridian, so its zenith angle is $z_{\text{noon}} = |\phi_0 - \delta_\odot|$.

Worked example, the June solstice at IISER TVM. At the June solstice $\delta_\odot = +23.44^\circ$. With $\phi_0 = 8.6833^\circ\text{N}$,

$$z_{\text{noon}} = |8.6833^\circ - 23.44^\circ| = 14.76^\circ.$$

Since $\delta_\odot > \phi_0$, the noon Sun is 14.76° north of the zenith.

2.6 Angular Distance on the Sphere

Each star has a unit vector, Eq. (2.2). The angle between two stars is not found by subtracting their coordinates. Near the pole, two stars 12^{h} apart in right ascension can be only a few degrees apart on the sky.

Think of two towns at latitude 60°N on opposite sides of the globe, 180° apart in longitude. Walking west along the parallel covers $180^\circ \times \cos 60^\circ = 90^\circ$ of arc. Going straight over the pole covers only $30^\circ + 30^\circ = 60^\circ$. The shortest way is the great circle, and its angle is the angle between the two arrows from the centre of the Earth. Two stars are two such arrows. The idea is that the cosine of the angle between them is the dot product of their unit vectors.

For the two towns, $\delta_1 = \delta_2 = 60^\circ$ and $\alpha_1 - \alpha_2 = 180^\circ$. Their direction vectors give $\cos\theta = \sin^2 60^\circ + \cos^2 60^\circ \cos 180^\circ = 0.75 - 0.25 = 0.5$, hence $\theta = 60^\circ$: the shortest route goes over the pole. The same dot product gives the angular separation of two stars. We will derive it and then rewrite it in a form that remains accurate when the separation is small.

The cosine rule. For two unit vectors, $\cos\theta = \hat{\mathbf{r}}_1 \cdot \hat{\mathbf{r}}_2$. With Eq. (2.2),

$$\begin{aligned}\cos\theta &\overset{(1)}{=} (\cos\delta_1\cos\alpha_1)(\cos\delta_2\cos\alpha_2) + (\cos\delta_1\sin\alpha_1)(\cos\delta_2\sin\alpha_2) + \sin\delta_1\sin\delta_2 \\ &\overset{(2)}{=} \cos\delta_1\cos\delta_2\big(\cos\alpha_1\cos\alpha_2 + \sin\alpha_1\sin\alpha_2\big) + \sin\delta_1\sin\delta_2 \\ &\overset{(3)}{=} \sin\delta_1\sin\delta_2 + \cos\delta_1\cos\delta_2\cos(\alpha_1 - \alpha_2). \end{aligned} \tag{2.34}$$

(1) The dot product of the two columns (2.2), component by component.
(2) $\cos\delta_1\cos\delta_2$ taken out of the first two terms.
(3) The angle-difference identity for the cosine.

Result 2.2

$$\cos\theta = \hat{\mathbf{r}}_1 \cdot \hat{\mathbf{r}}_2 = \sin\delta_1\sin\delta_2 + \cos\delta_1\cos\delta_2\cos(\alpha_1 - \alpha_2). \tag{2.35}$$

The result depends on the right ascensions only through their difference, as it must, since a turn about the pole changes neither the angle between two stars nor that difference.

Small angles. For close pairs Eq. (2.35) loses accuracy. Then $\cos\theta \approx 1 - \theta^2/2$, and θ comes from the small difference of numbers close to one. A computer stores numbers with a relative error of about 10^{-16}, so $\theta^2/2$ is lost once θ is below about 1.5×10^{-8} rad, which is 3 milliarcseconds. The cure is to compute $1 - \cos\theta = 2\sin^2(\theta/2)$ directly,

$$\begin{aligned} 2\sin^2\frac{\theta}{2} &\overset{(1)}{=} 1 - \sin\delta_1\sin\delta_2 - \cos\delta_1\cos\delta_2\cos(\alpha_2 - \alpha_1) \\ &\overset{(2)}{=} 1 - \sin\delta_1\sin\delta_2 - \cos\delta_1\cos\delta_2 + 2\cos\delta_1\cos\delta_2\sin^2\frac{\alpha_2 - \alpha_1}{2} \\ &\overset{(3)}{=} 1 - \cos(\delta_2 - \delta_1) + 2\cos\delta_1\cos\delta_2\sin^2\frac{\alpha_2 - \alpha_1}{2} \\ &\overset{(4)}{=} 2\sin^2\frac{\delta_2 - \delta_1}{2} + 2\cos\delta_1\cos\delta_2\sin^2\frac{\alpha_2 - \alpha_1}{2}. \end{aligned} \tag{2.36}$$

(1) $1 - \cos\theta = 2\sin^2(\theta/2)$, with $\cos\theta$ from Eq. (2.35).
(2) $\cos x = 1 - 2\sin^2(x/2)$ for $x = \alpha_2 - \alpha_1$.
(3) The angle-difference identity, $\cos\delta_1\cos\delta_2 + \sin\delta_1\sin\delta_2 = \cos(\delta_2 - \delta_1)$.
(4) $1 - \cos x = 2\sin^2(x/2)$ again, for $x = \delta_2 - \delta_1$.

Dividing by two gives the **haversine formula**,

$$\sin^2\left(\frac{\theta}{2}\right) = \sin^2\left(\frac{\delta_2 - \delta_1}{2}\right) + \cos\delta_1\cos\delta_2\sin^2\left(\frac{\alpha_2 - \alpha_1}{2}\right). \tag{2.37}$$

Every term is now small when the stars are close, so nothing is lost.

The flat-sky limit. For small separations each sine equals its argument. Writing $\Delta\delta = \delta_2 - \delta_1$ and $\Delta\alpha = \alpha_2 - \alpha_1$, and keeping the lowest order,

$$\boxed{\theta^2 \approx \Delta\delta^2 + \cos^2\delta\,\Delta\alpha^2.} \tag{2.38}$$

A step in right ascension is shortened on the sky by $\cos\delta$, because the hour circles crowd together toward the pole. This is why the proper motion in right ascension carries the factor $\cos\delta$ in Eq. (3.19).

2.7 The Ecliptic Coordinate System

Section 2.2 gives the matrix for a turn about any axis. The Sun and the planets move close to the ecliptic, which is tilted to the celestial equator. In equatorial coordinates the Sun's declination swings between $-23.44°$ and $+23.44°$ every year. A frame whose equator is the ecliptic removes that swing.

Think of a model railway laid on a tilted table. Measured from the floor, the train rises and falls as it goes round. Measured from the table top, it stays at height zero. Choosing the table as the reference plane removes the rising and falling. The ecliptic is the table, tilted by the obliquity $\epsilon = 23.44°$, and the Sun is the train. In ecliptic coordinates its latitude is always zero. The idea is that the ecliptic frame is the equatorial frame turned by ϵ about the line to $\hat{\gamma}$, which lies in both planes.

At the June solstice the Sun is at ecliptic longitude $90°$, a quarter turn from $\hat{\gamma}$. There the tilt puts it at declination $\delta_\odot = +\epsilon = +23.44°$ and right ascension $\alpha = 6^{\mathrm{h}}$. Figure 2.12 shows the two great circles meeting at the equinox. We now express their tilt as a rotation matrix, which will let us convert any direction between equatorial and ecliptic coordinates.

The angles. The **ecliptic longitude** $\lambda \in [0°, 360°)$ is measured eastward along the ecliptic from $\hat{\gamma}$. The **ecliptic latitude** $\beta \in [-90°, +90°]$ is the angle north or south of the ecliptic (Figure 2.13). By the split of Eq. (2.2), a star has the ecliptic column

$$\hat{\mathbf{r}}_{\mathrm{ecl}} = (\cos\beta\cos\lambda,\ \cos\beta\sin\lambda,\ \sin\beta)^T.$$

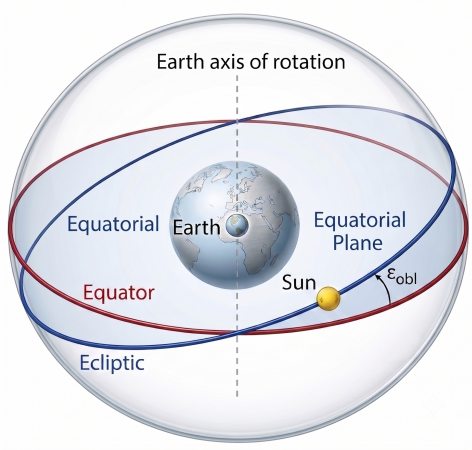


Figure 2.12: The celestial equator and the ecliptic around the Earth, crossing at the obliquity $\epsilon \simeq 23.44°$.

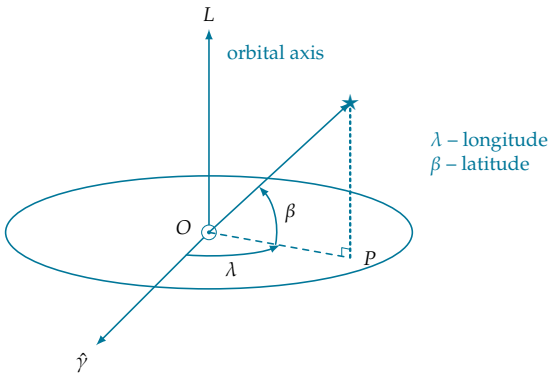


Figure 2.13: Ecliptic longitude λ, measured from $\hat{\gamma}$ in the ecliptic plane, and ecliptic latitude β, measured from that plane toward the orbital axis.

The ecliptic axes in equatorial components. The vernal equinox is on both circles (Section 1.6), so both frames share $\hat{\mathbf{x}}$. The June solstice point is on the ecliptic at $\alpha = 90°$, $\delta = \epsilon$, so its column (2.2) is $(0, \cos\epsilon, \sin\epsilon)^T$. The ecliptic pole is at right angles to both, so it is their cross product,

and the third axis completes the set,

$$\hat{\mathbf{z}}_{\text{ecl}} \overset{(1)}{=} \hat{\mathbf{x}} \times \begin{pmatrix} 0 \\ \cos\epsilon \\ \sin\epsilon \end{pmatrix} = \begin{pmatrix} 0\cdot\sin\epsilon - 0\cdot\cos\epsilon \\ 0\cdot 0 - 1\cdot\sin\epsilon \\ 1\cdot\cos\epsilon - 0\cdot 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -\sin\epsilon \\ \cos\epsilon \end{pmatrix},$$
$$\hat{\mathbf{y}}_{\text{ecl}} \overset{(2)}{=} \hat{\mathbf{z}}_{\text{ecl}} \times \hat{\mathbf{x}} = \begin{pmatrix} -\sin\epsilon\cdot 0 - \cos\epsilon\cdot 0 \\ \cos\epsilon\cdot 1 - 0\cdot 0 \\ 0\cdot 0 - (-\sin\epsilon)\cdot 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \cos\epsilon \\ \sin\epsilon \end{pmatrix}. \tag{2.39}$$

(1) The pole of the ecliptic is perpendicular to every direction in the ecliptic, here $\hat{\mathbf{x}}$ and the solstice. The order of the factors puts it on the north side, $\cos\epsilon > 0$.
(2) A right-handed set has $\hat{\mathbf{y}} = \hat{\mathbf{z}} \times \hat{\mathbf{x}}$.

The ecliptic y axis points to the solstice, $\lambda = 90°$, as it should. The ecliptic pole lies at $\alpha = 18^{\text{h}}$ and $\delta = 90° - \epsilon$.

The transformation. Stacking $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}_{\text{ecl}}$ and $\hat{\mathbf{z}}_{\text{ecl}}$ as rows, by Eq. (2.3), gives exactly $\mathbf{R}_x(\epsilon)$ of Eq. (2.8),

$$\begin{pmatrix} \cos\beta\cos\lambda \\ \cos\beta\sin\lambda \\ \sin\beta \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\epsilon & \sin\epsilon \\ 0 & -\sin\epsilon & \cos\epsilon \end{pmatrix} \begin{pmatrix} \cos\delta\cos\alpha \\ \cos\delta\sin\alpha \\ \sin\delta \end{pmatrix} \tag{2.40}$$
$$\overset{(2)}{=} \begin{pmatrix} \cos\delta\cos\alpha \\ \cos\delta\sin\alpha\cos\epsilon + \sin\delta\sin\epsilon \\ -\cos\delta\sin\alpha\sin\epsilon + \sin\delta\cos\epsilon \end{pmatrix}.$$

(1) Eq. (2.3) with the ecliptic axes as rows, acting on Eq. (2.2).
(2) Each row times the column.

Row by row,

Result 2.3

$$\sin\beta = \sin\delta\cos\epsilon - \cos\delta\sin\alpha\sin\epsilon, \tag{2.41}$$
$$\cos\beta\cos\lambda = \cos\delta\cos\alpha, \tag{2.42}$$
$$\cos\beta\sin\lambda = \cos\delta\sin\alpha\cos\epsilon + \sin\delta\sin\epsilon. \tag{2.43}$$

The longitude needs both of the last two rows. Their ratio alone, $\tan\lambda$, cannot tell λ from $\lambda + 180°$.

Going back, and the Sun's declination. The inverse is the transpose, Eq. (2.4), so $\hat{\mathbf{r}}_{\text{eq}} = \mathbf{R}_x^T(\epsilon)\,\hat{\mathbf{r}}_{\text{ecl}}$. For the Sun, $\beta = 0$ and $\hat{\mathbf{r}}_{\text{ecl}} = (\cos\lambda_\odot, \sin\lambda_\odot, 0)^T$, so

$$\begin{pmatrix} \cos\delta_\odot\cos\alpha_\odot \\ \cos\delta_\odot\sin\alpha_\odot \\ \sin\delta_\odot \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\epsilon & -\sin\epsilon \\ 0 & \sin\epsilon & \cos\epsilon \end{pmatrix} \begin{pmatrix} \cos\lambda_\odot \\ \sin\lambda_\odot \\ 0 \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} \cos\lambda_\odot \\ \cos\epsilon\sin\lambda_\odot \\ \sin\epsilon\sin\lambda_\odot \end{pmatrix}. \tag{2.44}$$

(1) The transpose of Eq. (2.8) with $\theta = \epsilon$.
(2) Each row times the column.

$$\boxed{\sin\delta_\odot = \sin\epsilon\,\sin\lambda_\odot.} \tag{2.45}$$

The Sun's declination follows from its longitude alone. It is zero at the equinoxes, $\lambda_\odot = 0°$ and $180°$, and reaches $\pm\epsilon$ at the solstices, which is the table of Section 1.6.

2.8 The Galactic Coordinate System

The ecliptic frame was one turn about one axis, because the two planes share the line to $\hat{\gamma}$. The disk of the Milky Way is tilted by about 63° to the celestial equator, and its zero point, the direction of the Galactic centre, does not lie on the line where the two planes cross. One turn is not enough.

Think of a camera on a tripod. Any way the camera can point is reached by three turns of the tripod head. The head pans about the vertical, tilts about a horizontal axis, and then the camera rolls about its own lens axis. The Galactic frame is reached from the equatorial frame by the same three turns. A pan about the celestial pole brings the x axis to the line where the Galactic equator crosses the celestial equator. A tilt about that line carries the z axis to the north Galactic pole. A roll about the new pole brings the x axis to the Galactic centre. The idea is that the Galactic frame is the equatorial frame turned by three known angles.

The three angles are a pan of 282.86°, a tilt of 62.87° and a roll of −32.93°. Their values follow from the Galactic pole and the longitude of one reference direction. Once combined into a matrix, these turns convert both stellar positions and stellar velocities into the Galactic frame used in Chapter 3.

The angles and the pole. The **Galactic latitude** $b \in [-90°, +90°]$ is the angle north (+) or south (−) of the Galactic plane. The **Galactic longitude** $l \in [0°, 360°)$ is measured along the Galactic equator from the direction of the Galactic centre, toward the direction of the Galaxy's rotation at $l = 90°$. The radio source Sagittarius A*, at the centre, lies within 0.1° of $l = b = 0$. By IAU convention the north Galactic pole and the longitude of the north celestial pole are fixed at the J2000.0 values

$$\alpha_{\mathrm{GP}} = 12^{\mathrm{h}}51^{\mathrm{m}}26.28^{\mathrm{s}} = 192.85948°, \qquad \delta_{\mathrm{GP}} = +27°07'41.7'' = +27.12825°,$$
$$l_{\mathrm{CP}} = 122.93192°. \tag{2.46}$$

Here l_{CP} is the Galactic longitude of the north celestial pole.

The three turns. The Galactic equator is 90° from its pole, so it crosses the celestial equator 90° from the pole's hour circle, at $\alpha = \alpha_{\mathrm{GP}} + 90°$. The angle between the two poles is $90° - \delta_{\mathrm{GP}}$. In the Galactic frame that crossing is 90° from the half circle of longitude l_{CP} that runs through the celestial pole, at $l = l_{\mathrm{CP}} - 90° = 32.93°$, so the last turn is by $-(l_{\mathrm{CP}} - 90°)$. With Eqs. (2.6) and (2.8),

$$\hat{\mathbf{r}}_{\mathrm{gal}} = \mathbf{R}_z(90° - l_{\mathrm{CP}})\, \mathbf{R}_x(90° - \delta_{\mathrm{GP}})\, \mathbf{R}_z(\alpha_{\mathrm{GP}} + 90°)\, \hat{\mathbf{r}}_{\mathrm{eq}}. \tag{2.47}$$

We multiply from the right. The first turn is Eq. (2.9) with $\theta = \alpha_{\mathrm{GP}} + 90°$,

$$\mathbf{R}_z(\alpha_{\mathrm{GP}} + 90°)\, \hat{\mathbf{r}}_{\mathrm{eq}} \overset{(1)}{=} \begin{pmatrix} \cos\delta\cos(\alpha - \alpha_{\mathrm{GP}} - 90°) \\ \cos\delta\sin(\alpha - \alpha_{\mathrm{GP}} - 90°) \\ \sin\delta \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} \cos\delta\sin(\alpha - \alpha_{\mathrm{GP}}) \\ -\cos\delta\cos(\alpha - \alpha_{\mathrm{GP}}) \\ \sin\delta \end{pmatrix}. \tag{2.48}$$

(1) Eq. (2.9).
(2) $\cos(x - 90°) = \sin x$ and $\sin(x - 90°) = -\cos x$.

The tilt uses $\cos(90° - \delta_{\mathrm{GP}}) = \sin\delta_{\mathrm{GP}}$ and $\sin(90° - \delta_{\mathrm{GP}}) = \cos\delta_{\mathrm{GP}}$,

$$\mathbf{R}_x(90° - \delta_{\mathrm{GP}}) \begin{pmatrix} \cos\delta\sin(\alpha - \alpha_{\mathrm{GP}}) \\ -\cos\delta\cos(\alpha - \alpha_{\mathrm{GP}}) \\ \sin\delta \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sin\delta_{\mathrm{GP}} & \cos\delta_{\mathrm{GP}} \\ 0 & -\cos\delta_{\mathrm{GP}} & \sin\delta_{\mathrm{GP}} \end{pmatrix} \begin{pmatrix} \cos\delta\sin(\alpha - \alpha_{\mathrm{GP}}) \\ -\cos\delta\cos(\alpha - \alpha_{\mathrm{GP}}) \\ \sin\delta \end{pmatrix}$$
$$\overset{(2)}{=} \begin{pmatrix} \cos\delta\sin(\alpha - \alpha_{\mathrm{GP}}) \\ \cos\delta_{\mathrm{GP}}\sin\delta - \sin\delta_{\mathrm{GP}}\cos\delta\cos(\alpha - \alpha_{\mathrm{GP}}) \\ \sin\delta_{\mathrm{GP}}\sin\delta + \cos\delta_{\mathrm{GP}}\cos\delta\cos(\alpha - \alpha_{\mathrm{GP}}) \end{pmatrix}. \tag{2.49}$$

(1) Eq. (2.8) with $\theta = 90^\circ - \delta_{GP}$.
(2) Each row times the column.

For the roll it is simpler to move it to the left side. By Eq. (2.4), $\mathbf{R}_z^T(\theta) = \mathbf{R}_z(-\theta)$, so the left side of Eq. (2.47) times $\mathbf{R}_z(l_{CP} - 90^\circ)$ is

$$\mathbf{R}_z(l_{CP} - 90^\circ)\begin{pmatrix}\cos b\cos l\\ \cos b\sin l\\ \sin b\end{pmatrix} \overset{(1)}{=} \begin{pmatrix}\cos b\cos(l - l_{CP} + 90^\circ)\\ \cos b\sin(l - l_{CP} + 90^\circ)\\ \sin b\end{pmatrix} \overset{(2)}{=} \begin{pmatrix}\cos b\sin(l_{CP} - l)\\ \cos b\cos(l_{CP} - l)\\ \sin b\end{pmatrix}. \tag{2.50}$$

(1) Eq. (2.9) with $\theta = l_{CP} - 90^\circ$.
(2) $\cos(x + 90^\circ) = -\sin x$ and $\sin(x + 90^\circ) = \cos x$, with $x = l - l_{CP}$.

Setting the last two columns equal, row by row,

Result 2.4

$$\sin b = \sin\delta_{GP}\sin\delta + \cos\delta_{GP}\cos\delta\cos(\alpha - \alpha_{GP}), \tag{2.51}$$
$$\cos b\sin(l_{CP} - l) = \cos\delta\sin(\alpha - \alpha_{GP}), \tag{2.52}$$
$$\cos b\cos(l_{CP} - l) = \cos\delta_{GP}\sin\delta - \sin\delta_{GP}\cos\delta\cos(\alpha - \alpha_{GP}). \tag{2.53}$$

The latitude depends only on the star's angle from the Galactic pole, since Eq. (2.51) is the cosine rule (2.35) between the star and the pole. The longitude l needs both of the other two rows to fix its quadrant.

The same formulas as for the horizon. Eqs. (2.51)–(2.53) are Eqs. (2.21)–(2.23) with the Galactic pole in place of the zenith. Replace ϕ_0 by δ_{GP}, the hour angle H by $\alpha_{GP} - \alpha$, the altitude by b and the azimuth A by $l_{CP} - l$. The zenith is a pole at declination ϕ_0 on the meridian, and the Galactic pole is a pole at declination δ_{GP} on the hour circle α_{GP}. The minus sign in $l_{CP} - l$ is there because the azimuth runs clockwise about its pole and l runs anticlockwise.

The numbers. Multiplying out Eq. (2.47) with the values above gives

$$\hat{\mathbf{r}}_{gal} = \begin{pmatrix}-0.054876 & -0.873437 & -0.483835\\ +0.494109 & -0.444830 & +0.746982\\ -0.867666 & -0.198076 & +0.455984\end{pmatrix}\hat{\mathbf{r}}_{eq}. \tag{2.54}$$

By the rule of rows (2.3), the first row is the direction of the Galactic centre in equatorial components, the second is the direction $l = 90^\circ$, and the third is the Galactic pole. The third row is $(\cos\delta_{GP}\cos\alpha_{GP}, \cos\delta_{GP}\sin\alpha_{GP}, \sin\delta_{GP})$, as Eq. (2.2) requires. The same matrix appears in Eq. (3.34). Two checks follow. The point $\alpha = 266.405^\circ$, $\delta = -28.936^\circ$ comes out at $l = 0.0002^\circ$, $b = 0.0001^\circ$, so it is the zero point of the frame. Sagittarius A*, at $\alpha = 17^h45^m40.04^s$, $\delta = -29^\circ00'28.1''$, comes out at $l = 359.944^\circ$, $b = -0.046^\circ$.

2.9 Atmospheric Refraction and Stellar Aberration

Every formula so far gives the true direction of a star from the centre of the Earth. The direction in which its light arrives at a telescope is not quite that. The air bends the light, and the motion of the Earth tilts it.

Think of a straw in a glass of water. It looks bent where it enters the water, because light changes direction when it crosses into a medium with a different refractive index. Now think

of running through rain that falls straight down. The drops seem to come at you from in front, and you tilt your umbrella forward. The faster you run, the more you tilt it. Starlight entering the air bends toward the vertical, so a star looks higher than it is. Starlight met by the moving Earth seems to come from a little ahead of the Earth's motion. The idea is that refraction lifts a star by $(n-1)\tan z$ and aberration tilts it forward by v/c.

With $n-1 = 2.82\times10^{-4}$ for air, refraction lifts a star at $z = 45^\circ$ by 2.82×10^{-4} rad $= 58''$. Earth's orbital speed is $v = 29.78\,\mathrm{km\,s^{-1}}$, so $v/c = 0.99\times10^{-4}$ gives an aberration angle of $20.5''$. Both shifts matter for accurate pointing, but they have different causes. We derive refraction from the bending of a ray through air and aberration from the observer's motion.

2.9.1 Atmospheric Refraction

The goal is the lift $\Delta h = h_{\text{apparent}} - h_{\text{true}}$, the same as the drop in zenith angle.

Snell's law through many layers. Let the air be a stack of flat layers with refractive indices $n_1, n_2, \ldots$ from the top down to the ground, where $n_k = n$, and let space above have index 1. A ray from a star at true zenith angle z enters at the top and reaches the ground at the apparent zenith angle z'. Snell's law at each boundary gives

$$\sin z \overset{(1)}{=} n_1 \sin z_1 \overset{(1)}{=} n_2 \sin z_2 \overset{(1)}{=} \cdots \overset{(1)}{=} n \sin z'. \tag{2.55}$$

(1) Snell's law at each flat boundary. The boundaries are parallel, so the angle leaving one layer is the angle arriving at the next.

All the layers in between cancel. For flat layers the bending depends only on the index at the ground, not on how n varies with height.

The lift. Put $z' = z - \Delta h$ and expand for small Δh,

$$\begin{aligned} \sin z &= n\sin(z-\Delta h) \\ &\overset{(1)}{=} n\sin z\cos\Delta h - n\cos z\sin\Delta h \\ &\overset{(2)}{=} n\sin z - n\,\Delta h\cos z, \end{aligned} \tag{2.56}$$

(1) The angle-difference identity for the sine.
(2) $\cos\Delta h = 1$ and $\sin\Delta h = \Delta h$ to first order in Δh.

so that

$$\Delta h \approx \frac{(n-1)\sin z}{n\cos z} \approx (n-1)\tan z, \tag{2.57}$$

using $n \approx 1$ in the denominator. The index of air is a measured property. For visible light at 10°C and 1010 hPa it is $n-1 = 2.82\times10^{-4}$. In arcseconds ($1\,\text{rad} = 206\,265''$),

$$\boxed{\Delta h \approx 58.2''\tan z = 58.2''\cot h \qquad (h > 15^\circ).} \tag{2.58}$$

The lift is zero at the zenith and grows toward the horizon. To first order it does not matter whether z is the true or the apparent zenith angle. Since $n-1$ is proportional to the density of the air, which goes as P/T, the coefficient $58.2''$ must be scaled by the local pressure and temperature.

When the flat layers fail (P6). The air thins with a scale height of 8.4 km, Eq. (10.4). Call it H_{air} here, to keep it apart from the hour angle. A ray at zenith angle z crosses this lowest scale height while covering a horizontal distance $d = H_{\text{air}} \tan z$. Over a horizontal distance d the curved ground of radius $R_\oplus$ falls away below a flat plane by

$$R_\oplus - \sqrt{R_\oplus^2 - d^2} \overset{(1)}{=} R_\oplus\Big(1 - \sqrt{1 - d^2/R_\oplus^2}\Big) \overset{(2)}{=} \frac{d^2}{2R_\oplus}, \tag{2.59}$$

(1) $R_\oplus$ taken out of the bracket.
(2) $\sqrt{1-x} \approx 1 - x/2$ for $x = d^2/R_\oplus^2 \ll 1$.

The flat picture holds while this drop is small compared with H_{air} itself. The ratio of the two lengths is

$$\frac{d^2/(2R_\oplus)}{H_{\text{air}}} = \frac{H_{\text{air}} \tan^2 z}{2R_\oplus}. \tag{2.60}$$

With $R_\oplus = 6371$ km, $2R_\oplus/H_{\text{air}} = 1517$. The ratio is 0.01 at $\tan z = 3.9$, that is $z = 75.6°$ or $h \approx 14°$, which is why Eq. (2.58) is quoted for $h > 15°$. The ratio reaches one at $\tan z = 39$, or $z = 88.5°$, and there the flat picture fails outright.

Near the horizon. Close to the horizon an empirical fit is used instead. Bennett (1982) fitted the refraction as a function of the apparent altitude, $\Delta h[\text{arcmin}] \approx 1/\tan\big(h_{\text{app}} + 7.31°/(h_{\text{app}} + 4.4°)\big)$. Sæmundsson (1986) turned it round to use the true altitude h, and added the scaling of $n-1$ with pressure P and temperature T,

$$\Delta h[\text{arcmin}] \approx \frac{1.02}{\tan\left(h + \dfrac{10.3°}{h + 5.11°}\right)} \times \left(\frac{P}{1010\,\text{hPa}}\right)\left(\frac{283}{273 + T[°\text{C}]}\right). \tag{2.61}$$

A star whose true position is on the horizon is lifted by 29′. A star that is seen exactly on the horizon is really 34.5′ below it, by Bennett's fit at $h_{\text{app}} = 0$. That is more than the Sun's angular diameter of about 32′, so when the setting Sun appears to touch the horizon it is already geometrically below it. At $h = 45°$ Eq. (2.61) gives 1.01′ = 61″, close to the 58″ of Eq. (2.58).

2.9.2 Stellar Aberration

The goal is the shift of a star's apparent direction caused by the observer's velocity $\vec{v}$.

The general shift. Light from a star in the direction $\hat{\mathbf{r}}$ travels with velocity $-c\,\hat{\mathbf{r}}$ in the frame of the Sun. Since $v \ll c$, Galilean addition of velocities is enough. In the Earth's frame the light moves with velocity $-c\,\hat{\mathbf{r}} - \vec{v}$, so it seems to come from the direction of $c\,\hat{\mathbf{r}} + \vec{v}$. The apparent direction is

$$\begin{aligned} \hat{\mathbf{r}}' = \frac{c\,\hat{\mathbf{r}} + \vec{v}}{|c\,\hat{\mathbf{r}} + \vec{v}|} &\overset{(1)}{=} \frac{\hat{\mathbf{r}} + \vec{v}/c}{\sqrt{1 + 2\,\hat{\mathbf{r}}\cdot\vec{v}/c + v^2/c^2}} \\ &\overset{(2)}{=} \Big(\hat{\mathbf{r}} + \frac{\vec{v}}{c}\Big)\Big(1 - \frac{\hat{\mathbf{r}}\cdot\vec{v}}{c}\Big) \overset{(3)}{=} \hat{\mathbf{r}} + \frac{\vec{v} - (\hat{\mathbf{r}}\cdot\vec{v})\,\hat{\mathbf{r}}}{c}. \end{aligned} \tag{2.62}$$

(1) Divided above and below by c, and $|\hat{\mathbf{r}} + \vec{v}/c|^2$ written out with $\hat{\mathbf{r}}\cdot\hat{\mathbf{r}} = 1$.
(2) $(1+x)^{-1/2} \approx 1 - x/2$, keeping first order in v/c.
(3) Multiplied out, dropping the term of order v^2/c^2.

$$\boxed{\hat{\mathbf{r}}' - \hat{\mathbf{r}} = \frac{\vec{v}_\perp}{c},} \tag{2.63}$$

where $\vec{v}_\perp = \vec{v} - (\hat{\mathbf{r}} \cdot \vec{v})\,\hat{\mathbf{r}}$ is the part of the velocity across the line of sight. The star moves toward the direction of motion by the angle $(v/c)\sin\psi$, where ψ is the angle between $\hat{\mathbf{r}}$ and $\vec{v}$. A velocity along the line of sight gives no shift. The relativistic formula differs from this only by a fraction of order $v^2/c^2 \approx 10^{-8}$.

The constant of aberration. The Earth's orbital speed is $v = 2\pi \times 1\,\text{AU}$ per sidereal year, $29.78\,\text{km s}^{-1}$. The largest shift, for $\psi = 90°$, is

$$\boxed{\kappa = \frac{v}{c} = \frac{29.78}{299\,792} \times 206\,265'' = 20.49''.} \tag{2.64}$$

The yearly ellipse. To see the path of a star over a year, work in ecliptic coordinates and take the Earth's orbit as a circle. The Earth is opposite the Sun, so with the Sun at ecliptic longitude $\lambda_\odot$ the Earth is at $-1\,\text{AU}\,(\cos\lambda_\odot, \sin\lambda_\odot, 0)$. Its velocity is the time derivative of that position, with $\lambda_\odot$ growing at 2π per year,

$$\vec{v} = v\,(\sin\lambda_\odot,\ -\cos\lambda_\odot,\ 0)^T. \tag{2.65}$$

The directions of increasing λ and β at the star are Eq. (3.11) with $\alpha \to \lambda$ and $\delta \to \beta$, namely $\hat{\boldsymbol{\lambda}} = (-\sin\lambda, \cos\lambda, 0)^T$ and $\hat{\boldsymbol{\beta}} = (-\sin\beta\cos\lambda, -\sin\beta\sin\lambda, \cos\beta)^T$. Both are perpendicular to $\hat{\mathbf{r}}$, so the components of the shift (2.63) along them are $\vec{v}\cdot\hat{\boldsymbol{\lambda}}/c$ and $\vec{v}\cdot\hat{\boldsymbol{\beta}}/c$,

$$\begin{aligned}
\cos\beta\,\Delta\lambda &= \frac{\vec{v}\cdot\hat{\boldsymbol{\lambda}}}{c} \overset{(1)}{=} -\kappa\,(\sin\lambda_\odot\sin\lambda + \cos\lambda_\odot\cos\lambda) \overset{(2)}{=} -\kappa\cos(\lambda_\odot - \lambda),\\
\Delta\beta &= \frac{\vec{v}\cdot\hat{\boldsymbol{\beta}}}{c} \overset{(1)}{=} -\kappa\sin\beta\,(\sin\lambda_\odot\cos\lambda - \cos\lambda_\odot\sin\lambda) \overset{(2)}{=} -\kappa\sin\beta\sin(\lambda_\odot - \lambda).
\end{aligned} \tag{2.66}$$

(1) The dot products of the columns, with $v/c = \kappa$.
(2) The angle-difference identities.

As $\lambda_\odot$ runs through the year,

$$\left(\frac{\cos\beta\,\Delta\lambda}{\kappa}\right)^2 + \left(\frac{\Delta\beta}{\kappa\sin\beta}\right)^2 = 1. \tag{2.67}$$

Every star traces an ellipse with semi-axes $20.5''$ along the ecliptic and $20.5''\sin\beta$ across it. At the ecliptic pole it is a circle, and on the ecliptic it closes to a line.

The parallax ellipse of Chapter 3 comes from the same projection with the Earth's position in place of its velocity. A star at distance d is shifted by $-(\text{Earth's position})_\perp/d$, which gives $\cos\beta\,\Delta\lambda = p\sin(\lambda_\odot - \lambda)$ and $\Delta\beta = -p\sin\beta\cos(\lambda_\odot - \lambda)$ with $p = 1\,\text{AU}/d$. The two ellipses have the same shape and the same one-year period, and they are a quarter of a year out of step. They differ in size. The parallax of even the nearest star is below $0.8''$ and shrinks with distance. The aberration ellipse is $20.5''$ for every star at every distance.

2.10 Airmass and the Parallactic Angle

The zenith angle of a star is known from Eq. (2.21), and refraction from Eq. (2.57). Two practical questions remain. Photometry needs to know how much air the light has crossed. Spectroscopy needs to know in which direction on the sky the air spreads the colours.

Think of crossing a road 10 m wide. Walking straight across covers 10 m. Walking at 60° to the straight line covers $10/\cos 60° = 20$ m. Now think of a prism. It bends blue light more than red, so white light fans out into colours along the direction of bending. Light from a star at zenith angle z crosses the air on a slant. Air also bends blue light a little more than red, so every star is drawn out into a short spectrum pointing at the zenith. The idea is that the path through flat layers grows as $1/\cos z$, and the parallactic angle gives the direction of the zenith as seen at the star.

At $z = 60°$, a ray crosses twice as much air as one arriving from the zenith because its path through each flat layer is longer by $1/\cos z$. Its optical depth and extinction therefore grow with the same factor. We call this relative column of air the airmass. Air also refracts different colours by different amounts; the parallactic angle tells us the direction of that colour spread on the sky and how to align a spectrograph slit with it.

2.10.1 Airmass

Let the air be flat layers, with density $\rho(y)$ at height y. A ray at zenith angle z covers a path ds while it rises dy, and the slanted path is the long side of a right triangle whose side along the vertical is dy, so $ds = dy/\cos z$. The **airmass** X is the mass of air along the slanted path divided by that along the vertical,

$$X \equiv \frac{\int \rho \, ds}{\int \rho \, dy} \overset{(1)}{=} \frac{\frac{1}{\cos z}\int \rho(y)\, dy}{\int \rho(y)\, dy} \overset{(2)}{=} \frac{1}{\cos z}. \tag{2.68}$$

(1) $ds = dy/\cos z$, with the same z in every flat layer. Refraction changes z by less than 10^{-3} rad above $h = 15°$, Eq. (2.57).
(2) The integrals cancel, whatever the profile $\rho(y)$.

$$\boxed{X \approx \sec z = \frac{1}{\sin h}.} \tag{2.69}$$

The airmass depends only on the zenith angle, not on how the density falls with height. It diverges at the horizon, while the real path stays finite because the ground curves away. By Eq. (2.60) the curvature matters at the one per cent level from about $z = 75°$. There the empirical fit of Kasten and Young (1989) is used,

$$X = \frac{1}{\cos z + 0.50572\,(96.07995° - z)^{-1.6364}}, \tag{2.70}$$

with z in degrees. It gives 1.994 at $z = 60°$, where $\sec z = 2.000$, and 4.071 at $z = 76°$, where $\sec z = 4.134$, a difference of 1.5%. At the horizon it gives $X = 37.9$ instead of infinity.

2.10.2 The Parallactic Angle

The **parallactic angle** q is the angle at the star between the direction toward the celestial pole and the direction toward the zenith. It is the third angle of the astronomical triangle of Figure 2.10. We measure it from the pole direction through east, which is how position angles on the sky are measured.

The two directions at the star. Work in the meridian frame of Eq. (2.14), with the pole $\hat{\mathbf{p}} = (0, 0, 1)^T$ and the zenith $\hat{\mathbf{u}}$ of Eq. (2.15). Seen from the star, the great circle toward the pole

leaves along the part of $\hat{\mathbf{p}}$ perpendicular to $\hat{\mathbf{r}}$, and the great circle toward the zenith leaves along the part of $\hat{\mathbf{u}}$ perpendicular to $\hat{\mathbf{r}}$,

$$\vec{t}_P = \hat{\mathbf{p}} - \sin\delta\,\hat{\mathbf{r}}, \qquad \vec{t}_Z = \hat{\mathbf{u}} - \sin h\,\hat{\mathbf{r}}, \tag{2.71}$$

since $\hat{\mathbf{p}}\cdot\hat{\mathbf{r}} = \sin\delta$ and $\hat{\mathbf{u}}\cdot\hat{\mathbf{r}} = \sin h$. Their lengths are $|\vec{t}_P|^2 = 1 - 2\sin^2\delta + \sin^2\delta = \cos^2\delta$ and, in the same way, $|\vec{t}_Z| = \cos h$.

The cosine. The angle between the two directions has

$$\begin{aligned}\cos q\,\cos\delta\cos h = \vec{t}_P\cdot\vec{t}_Z &\overset{(1)}{=} \hat{\mathbf{p}}\cdot\hat{\mathbf{u}} - \sin h\,\hat{\mathbf{p}}\cdot\hat{\mathbf{r}} - \sin\delta\,\hat{\mathbf{r}}\cdot\hat{\mathbf{u}} + \sin\delta\sin h\,\hat{\mathbf{r}}\cdot\hat{\mathbf{r}}\\ &\overset{(2)}{=} \sin\phi_0 - \sin h\sin\delta - \sin\delta\sin h + \sin\delta\sin h\\ &= \sin\phi_0 - \sin h\sin\delta.\end{aligned} \tag{2.72}$$

(1) The dot product multiplied out.
(2) $\hat{\mathbf{p}}\cdot\hat{\mathbf{u}} = \sin\phi_0$ from Eq. (2.15), $\hat{\mathbf{p}}\cdot\hat{\mathbf{r}} = \sin\delta$, $\hat{\mathbf{r}}\cdot\hat{\mathbf{u}} = \sin h$ and $\hat{\mathbf{r}}\cdot\hat{\mathbf{r}} = 1$.

The sine. The direction a quarter turn east of the pole direction at the star is $\vec{t}_P\times\hat{\mathbf{r}}/\cos\delta$. (On the meridian at the equator, $\vec{t}_P = \hat{\mathbf{p}}$ and $\hat{\mathbf{r}} = \hat{\mathbf{x}}$, and $\hat{\mathbf{p}}\times\hat{\mathbf{x}} = \hat{\mathbf{y}}$ points east.) So

$$\begin{aligned}\sin q\,\cos\delta\cos h = \vec{t}_Z\cdot(\vec{t}_P\times\hat{\mathbf{r}}) &\overset{(1)}{=} \hat{\mathbf{u}}\cdot(\hat{\mathbf{p}}\times\hat{\mathbf{r}})\\ &\overset{(2)}{=} \begin{pmatrix}\cos\phi_0\\0\\\sin\phi_0\end{pmatrix}\cdot\begin{pmatrix}0\cdot\sin\delta - 1\cdot(-\cos\delta\sin H)\\1\cdot\cos\delta\cos H - 0\cdot\sin\delta\\0\end{pmatrix}\\ &\overset{(3)}{=} \cos\phi_0\cos\delta\sin H.\end{aligned} \tag{2.73}$$

(1) The parts of $\vec{t}_P$ and $\vec{t}_Z$ along $\hat{\mathbf{r}}$ drop out, since a triple product with $\hat{\mathbf{r}}$ twice is zero.
(2) $\hat{\mathbf{p}} = (0,0,1)^T$, $\hat{\mathbf{r}}$ from Eq. (2.14) and $\hat{\mathbf{u}}$ from Eq. (2.15).
(3) The dot product.

Dividing both results by $\cos\delta\cos h$,

$$\boxed{\sin q = \frac{\sin H\cos\phi_0}{\cos h}, \qquad \cos q = \frac{\sin\phi_0 - \sin h\sin\delta}{\cos h\cos\delta}.} \tag{2.74}$$

The angle q is zero on the meridian, where the zenith lies straight toward the pole, and has the sign of H. A star in the west has its zenith direction turned toward the east of north. These are the sine and cosine rules of spherical trigonometry for the astronomical triangle.

Why it matters for a slit. The index of air is larger for blue light than for red. Between 400 nm and 700 nm, $n-1$ changes by 7.0×10^{-6} at 15°C and 1013 hPa. By Eq. (2.57), the blue image of a star is lifted more than the red one by $7.0\times10^{-6}\tan z$ radians, or $1.43''\tan z$. At $z = 60°$ that is 2.5″. A spectrograph slit is often about 1″ wide, so a slit set across this spread loses the blue or the red light. Set along the vertical through the star, at position angle q, it takes in all colours.

2.11 Problem Set: Celestial Coordinate Transformations

2.11.1 Problem 1: Horizontal ↔ Equatorial Transformations and Sirius at IISER TVM

Problem Statement: Two coordinate systems turn with the Earth. One is the **horizontal system**, altitude h and azimuth A. The other is the turning **equatorial system**, declination δ and hour angle H.

(a) Show how the two systems are related.

(b) Sirius is at $\alpha = 06^{\mathrm{h}}45^{\mathrm{m}}08.917^{\mathrm{s}}$, $\delta = -16°42'58.02''$. Find its hour angle, altitude and azimuth at LST $= 22^{\mathrm{h}}00^{\mathrm{m}}00^{\mathrm{s}}$ at IISER TVM ($\phi_0 = 8.6833°$N).

(c) Find the sidereal times at which Sirius rises and sets at IISER TVM.

Essential Ideas:

- The local sidereal time is the right ascension on the meridian, Eq. (1.12), and the hour angle is $H = \mathrm{LST} - \alpha$, Eq. (2.11).
- In a frame turned to the meridian the star is, by Eq. (2.14),

$$\hat{\mathbf{r}}_{\mathrm{mer}} = \begin{pmatrix} \cos\delta\cos(\alpha - \mathrm{LST}) \\ \cos\delta\sin(\alpha - \mathrm{LST}) \\ \sin\delta \end{pmatrix} = \begin{pmatrix} \cos\delta\cos H \\ -\cos\delta\sin H \\ \sin\delta \end{pmatrix}. \tag{2.75}$$

  The matrix $\mathbf{M}(\phi_0)$ whose columns are north, east and up, Eq. (2.19), turns it into $(\cos h\cos A, \cos h\sin A, \sin h)^T$.
- Rising and setting happen at $h = 0$, so $\cos H_0 = -\tan\phi_0\tan\delta$, Eq. (2.28).

Detailed Solutions:

(a) **The transformation**. The derivation is in Section 2.3. Since $\mathbf{M}(\phi_0)$ is orthogonal and symmetric, $(\cos h\cos A, \cos h\sin A, \sin h)^T = \mathbf{M}(\phi_0)\,\hat{\mathbf{r}}_{\mathrm{mer}}$, Eq. (2.20), which gives

$$\boxed{\begin{aligned} \sin h &= \cos\phi_0\cos\delta\cos H + \sin\phi_0\sin\delta, \\ \cos h\sin A &= -\cos\delta\sin H, \\ \cos h\cos A &= -\sin\phi_0\cos\delta\cos H + \cos\phi_0\sin\delta. \end{aligned}} \tag{2.76}$$

The azimuth is fixed by the last two rows together. The same matrix takes (h, A) back to (δ, H), Eq. (2.26), because $\mathbf{M}(\phi_0)^2 = \mathbb{1}$.

(b) **Hour angle, altitude and azimuth**. In degrees, $\delta = -16.7161°$ and $\alpha = 15\,(6 + 45/60 + 8.917/3600)° = 101.2872°$. The hour angle is

$$H = \mathrm{LST} - \alpha = 22^{\mathrm{h}}00^{\mathrm{m}}00^{\mathrm{s}} - 06^{\mathrm{h}}45^{\mathrm{m}}08.917^{\mathrm{s}} = 15^{\mathrm{h}}14^{\mathrm{m}}51.08^{\mathrm{s}} = 228.7128°.$$

With $\sin\phi_0 = 0.15097$, $\cos\phi_0 = 0.98854$, $\sin\delta = -0.28763$, $\cos\delta = 0.95774$, $\sin H = -0.75141$ and $\cos H = -0.65983$,

$$\begin{aligned} \sin h &= 0.15097\times(-0.28763) + 0.98854\times 0.95774\times(-0.65983) = -0.6681, \\ \cos h\sin A &= -0.95774\times(-0.75141) = +0.7197, \\ \cos h\cos A &= -0.15097\times 0.95774\times(-0.65983) + 0.98854\times(-0.28763) = -0.1889. \end{aligned}$$

So $h = \arcsin(-0.6681)$. The sine of A is positive and its cosine negative, so A is in the second quadrant, $A = 180° - \arctan(0.7197/0.1889)$.

$$\boxed{H = 228.71°\ (15^{\mathrm{h}}14^{\mathrm{m}}51^{\mathrm{s}}), \qquad h = -41.92°, \qquad A = 104.71°.}$$

Sirius is 42° below the horizon, toward the east-southeast. Its hour angle, $228.71° = -131.29°$, says it is still east of the meridian and has yet to rise. The ratio form (2.24) gives $\arctan(-3.809) = -75.29°$, which is 180° away from the right answer, so the quadrant must come from the two rows separately.

(c) **Rising and setting**. With $\tan\phi_0 = 0.15272$ and $\tan\delta = -0.30032$, Eq. (2.28) gives

$$\cos H_0 = -0.15272 \times (-0.30032) = 0.04587, \qquad H_0 = 87.371^\circ = 5^{\mathrm{h}}49^{\mathrm{m}}29^{\mathrm{s}}.$$

Then $\mathrm{LST}_{\mathrm{rise}} = \alpha - H_0$ and $\mathrm{LST}_{\mathrm{set}} = \alpha + H_0$,

$$\boxed{\mathrm{LST}_{\mathrm{rise}} = 00^{\mathrm{h}}55^{\mathrm{m}}40^{\mathrm{s}}, \qquad \mathrm{LST}_{\mathrm{set}} = 12^{\mathrm{h}}34^{\mathrm{m}}38^{\mathrm{s}}.}$$

Sirius is up for $2H_0 = 11^{\mathrm{h}}39^{\mathrm{m}}$ of sidereal time, a little less than half the day, because it is south of the equator and the site is north of it. At LST $= 22^{\mathrm{h}}$ it is indeed below the horizon, as part (b) found.

2.11.2 Problem 2: Angular Distance on the Sphere and Ecliptic Coordinates of Sirius and Regulus

Problem Statement: Two stars are at equatorial coordinates (α_1, δ_1) and (α_2, δ_2). What is the angle between them on the celestial sphere? Regulus is at $\alpha = 10^{\mathrm{h}}08^{\mathrm{m}}22.311^{\mathrm{s}}$, $\delta = +11^\circ 58' 01.95''$.

(a) Find the angle between Sirius (Problem 2.11.1) and Regulus.

(b) Find the ecliptic coordinates of Sirius and of Regulus. Use $\epsilon = 23.4393^\circ$.

Essential Ideas:

- Each star is a unit vector, Eq. (2.2), and the angle between two unit vectors has $\cos\theta = \hat{\mathbf{r}}_1 \cdot \hat{\mathbf{r}}_2$, Eq. (2.35).
- The equatorial and ecliptic frames share the x axis toward $\hat{\gamma}$. The ecliptic column is $\mathbf{R}_x(\epsilon)$ times the equatorial column, Eq. (2.40), with the turn of Eq. (2.8) and Figure 2.7.
- The longitude needs both $\cos\beta\cos\lambda$ and $\cos\beta\sin\lambda$ to fix its quadrant.

Detailed Solutions:

(a) **Angular distance**. In degrees, Sirius is at $(\alpha_1, \delta_1) = (101.2872^\circ, -16.7161^\circ)$ from Problem 2.11.1. For Regulus,

$$(\alpha_2, \delta_2) = \left(15\left(10 + \tfrac{8}{60} + \tfrac{22.311}{3600}\right)^\circ,\ 11 + \tfrac{58}{60} + \tfrac{1.95}{3600}\right) = (152.0930^\circ,\ 11.9672^\circ).$$

The two columns (2.2) are

$$\hat{\mathbf{r}}_1 = \begin{pmatrix} -0.18746 \\ 0.93922 \\ -0.28763 \end{pmatrix}, \qquad \hat{\mathbf{r}}_2 = \begin{pmatrix} -0.86450 \\ 0.45787 \\ 0.20735 \end{pmatrix},$$

and their dot product, Eq. (2.35), is

$$\cos\theta = \sin\delta_1\sin\delta_2 + \cos\delta_1\cos\delta_2\cos(\alpha_1 - \alpha_2) = 0.5325,$$

$$\boxed{\theta = 57.83^\circ.}$$

(b) **Ecliptic coordinates**. With $\cos\epsilon = 0.91748$ and $\sin\epsilon = 0.39778$, Eq. (2.40) gives for Sirius

$$\hat{\mathbf{r}}_{\mathrm{ecl}} = \begin{pmatrix} -0.18746 \\ 0.93922 \times 0.91748 + (-0.28763) \times 0.39778 \\ -0.93922 \times 0.39778 + (-0.28763) \times 0.91748 \end{pmatrix} = \begin{pmatrix} -0.18746 \\ 0.74730 \\ -0.63749 \end{pmatrix}.$$

So $\beta = \arcsin(-0.63749)$. For λ, $\cos\lambda < 0$ and $\sin\lambda > 0$, so λ is in the second quadrant, $\lambda = 180^\circ - \arctan(0.74730/0.18746)$. For Regulus,

$$\hat{\mathbf{r}}_{\text{ecl}} = \begin{pmatrix} -0.86450 \\ 0.45787 \times 0.91748 + 0.20735 \times 0.39778 \\ -0.45787 \times 0.39778 + 0.20735 \times 0.91748 \end{pmatrix} = \begin{pmatrix} -0.86450 \\ 0.50256 \\ 0.00811 \end{pmatrix},$$

again with λ in the second quadrant.

$$\boxed{\text{Sirius } (\lambda, \beta) = (104.08^\circ, -39.61^\circ), \qquad \text{Regulus } (\lambda, \beta) = (149.83^\circ, +0.46^\circ).}$$

Taking arctan of the ratio alone gives -75.92° for Sirius and -30.17° for Regulus, each 180° from the true longitude. Regulus lies less than half a degree from the ecliptic. The Moon's orbit is inclined about 5° to the ecliptic, so the Moon often passes in front of Regulus, while Sirius, 40° south of the ecliptic, is never covered by the Moon or a planet.

Chapter 3

Astrometry: Distances, Motions, and the Space Velocity Vector

Chapter 2 gives each star a direction on the sky, the two angles (α, δ). To place the star in space we also need its distance, and to follow it in time we need its velocity. That makes six numbers, three for the position $\vec{r}$ and three for the velocity $\vec{v}$.

The direction supplies two position coordinates. A star's small yearly shift, its **parallax**, supplies the distance and completes its position. Its slow drift across the sky, the **proper motion**, gives two components of velocity once the distance is known. A Doppler shift gives the third component, toward or away from us. These measurements therefore build a star's position and velocity one component at a time, letting us map both the structure and motion of the Galaxy.

3.1 Trigonometric Parallax and the Parsec

Chapter 2 gives each star a direction, the unit vector $\hat{\mathbf{r}}_{\mathrm{eq}}$ of Eq. (2.2). A direction alone does not say how far away the star is. A faint star close by and a bright star far away can sit at the same point of the sky.

Hold a thumb up at arm's length and look at it with the left eye, then with the right. The thumb jumps against the far wall, and the wall does not move. The eyes are about 6.5 cm apart and the thumb is about 65 cm away, so the jump is $0.065/0.65 = 0.1$ radian, about 6°. Bring the thumb closer and the jump grows. The jump and the gap between the eyes together give the distance to the thumb. The Earth's orbit gives the astronomer two eyes. In January and in July the Earth is on opposite sides of the Sun, 2 AU apart. A nearby star then jumps against the far background stars, which hardly move at all. The idea is that a distance is a baseline divided by an angle.

For stars, the triangle is extremely narrow. One second of arc is only 1/206 265 of a radian, so a star 206 265 AU away shifts by 1″ across a 1 AU baseline. Even Sirius, one of the nearest stars, has a parallax of only 0.38″. Figure 3.1 shows the Earth at opposite points in its orbit and the star's two lines of sight. Their angular separation is $2p$, while p is the angle subtended by one orbital radius. Because $1\ \mathrm{AU}/d$ is so small, the triangle gives the simple distance rule below.

The triangle. Take the moment when the line from the Sun to the Earth is at right angles to the line from the Sun to the star. The triangle Sun–Earth–star then has its right angle at the Sun (Figure 3.1, left). The side from the Sun to the Earth is $1\ \mathrm{AU} = 1.495\,978\,707 \times 10^{11}$ m, a length first fixed in metres by radar echoes (Section 3.7.1), the side from the Sun to the star is

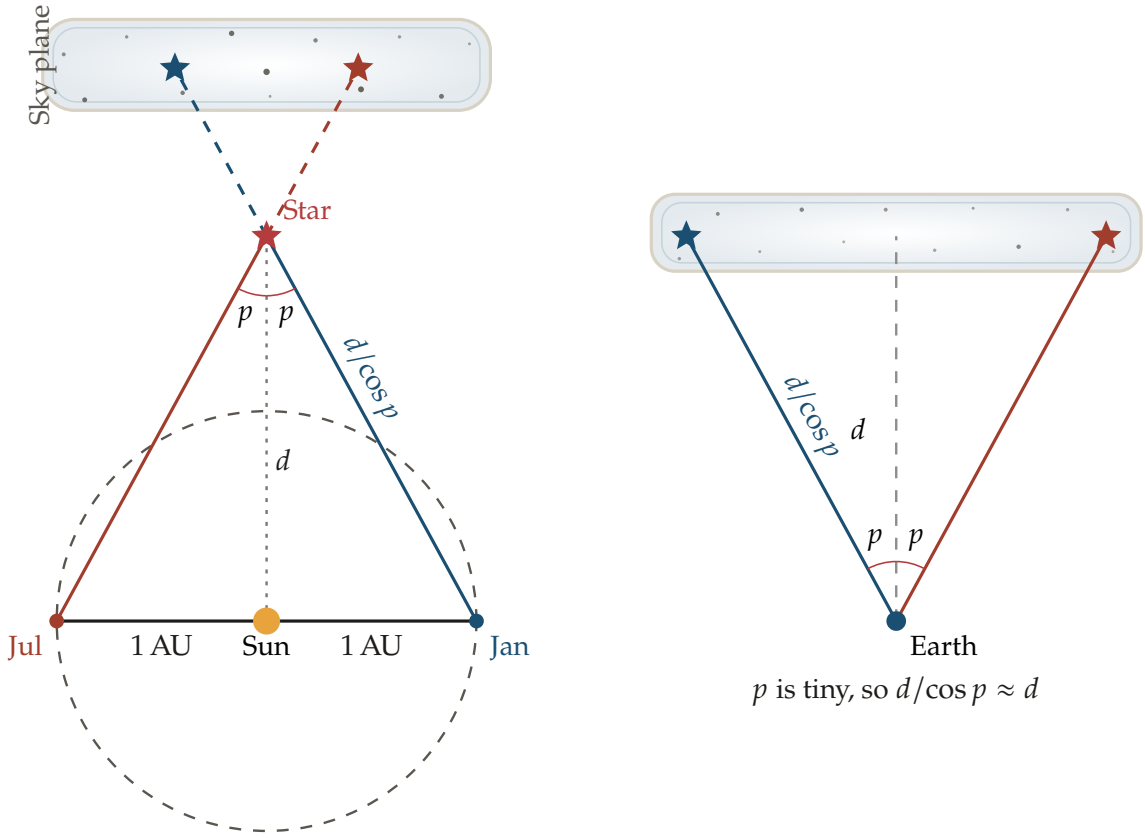


Figure 3.1: Left, the Sun, the Earth in January and in July, a nearby star at distance d from the Sun and $d/\cos p$ from the Earth, and the star's two apparent positions on the far sky. Right, the two lines of sight of the left panel, each moved parallel to itself so that their tails meet at one point, the Earth, instead of their heads meeting at the star, with the dashed line of length d and the line of sight of length $d/\cos p$. The baseline is 1 AU and the angle at the star is the parallax p. The angles are drawn far larger than real ones.

the distance d, and the angle at the star is the **parallax** p. So $\tan p = 1\,\text{AU}/d$, and

$$d = \frac{1\,\text{AU}}{\tan p} \overset{(1)}{=} 1\,\text{AU}\left(\frac{1}{p} - \frac{p}{3} - \ldots\right) \overset{(2)}{=} \frac{1\,\text{AU}}{p}\left(1 - \frac{p^2}{3} - \ldots\right). \tag{3.1}$$

(1) The series $\cot p = 1/p - p/3 - p^3/45 - \ldots$, with p in radians.
(2) Taking $1/p$ out of the bracket.

The distance d is measured from the Sun. From the Earth the star is the hypotenuse away, $d/\cos p$, which is longer by a fraction of about $p^2/2$. For $p = 1''$ that fraction is 1.2×10^{-11} and the correction $p^2/3$ above is 7.8×10^{-12}, and every star has $p < 1''$. So d is the distance from the Earth as well, and, with p in radians,

$$p = \frac{1\,\text{AU}}{d}. \tag{3.2}$$

The same p is the largest shift of the star from the place where it would appear if seen from the Sun. Over a year the star traces a small ellipse whose semi-major axis is p, Eq. (3.17).

The parsec. The **parsec** is the distance at which 1 AU subtends one second of arc. Since $1'' = \pi/648\,000$ radian, Eq. (3.2) gives

$$1\,\mathrm{pc} = \frac{1\,\mathrm{AU}}{\pi/648\,000} = 206\,264.806\,\mathrm{AU} \overset{(1)}{=} 3.085\,677\,58 \times 10^{16}\,\mathrm{m} \overset{(2)}{=} 3.261\,56\,\mathrm{ly}. \tag{3.3}$$

(1) $1\,\mathrm{AU} = 1.495\,978\,707 \times 10^{11}$ m.
(2) $1\,\mathrm{ly} = 9.460\,730\,47 \times 10^{15}$ m, the distance light travels in a Julian year of 365.25 days.

The IAU has defined the parsec in exactly this way, as $648\,000/\pi$ astronomical units, since 2015. For a parallax p'' measured in seconds of arc, the angle in radians is $p''\pi/648\,000$, and Eq. (3.2) becomes

$$d \overset{(1)}{=} \frac{1\,\mathrm{AU}}{p''\,\pi/648\,000} \overset{(2)}{=} \frac{1}{p''}\,\frac{1\,\mathrm{AU}}{\pi/648\,000} \overset{(3)}{=} \frac{1}{p''}\,\mathrm{pc}. \tag{3.4}$$

(1) Eq. (3.2) with the angle converted to radians.
(2) Taking the number $1/p''$ out.
(3) The definition of the parsec above.

Result 3.1

$$d\,[\mathrm{pc}] = \frac{1}{p\,[\mathrm{arcsec}]}. \tag{3.5}$$

A star with half the parallax is twice as far away. The rule uses nothing about the star itself, only the size of the Earth's orbit.

Worked example 3.1: Distance to Sirius

The *Hipparcos* catalogue gives Sirius a parallax of $p = 379.21\,\mathrm{mas} = 0.37921''$. By Eq. (3.5),

$$d = \frac{1}{0.37921}\,\mathrm{pc} = 2.637\,\mathrm{pc} = 8.601\,\mathrm{ly}.$$

Worked example 3.2: Parallax as Measured from Mars

From Mars, the baseline would be the radius of the orbit of Mars, 1.524 AU. The distance d belongs to the star and does not change. By Eq. (3.2) the parallax is proportional to the baseline,

$$\frac{p_{\mathrm{Mars}}}{p_{\mathrm{Earth}}} = \frac{1.524\,\mathrm{AU}/d}{1\,\mathrm{AU}/d} = 1.524, \qquad p_{\mathrm{Mars}} = 1.524 \times 0.37921'' = 0.578''.$$

A longer baseline gives a larger angle, which is easier to measure. The price is time. Mars goes round the Sun in 1.88 years, so the star takes 1.88 years to trace its full ellipse, against one year from the Earth.

From a measured parallax to a distance. A parallax is measured with an error, and the goal is to see what that error does to $d = 1/p$. Write the measured value as $p_{\mathrm{obs}} = p + \varepsilon$, where p is the

true parallax and the error ε has mean $\langle \varepsilon \rangle = 0$ and spread $\langle \varepsilon^2 \rangle = \sigma_p^2$. On average the inverse is

$$\left\langle \frac{1}{p_{\text{obs}}} \right\rangle = \left\langle \frac{1}{p}\,\frac{1}{1+\varepsilon/p} \right\rangle \overset{(1)}{=} \frac{1}{p}\left\langle 1 - \frac{\varepsilon}{p} + \frac{\varepsilon^2}{p^2} - \dots \right\rangle \overset{(2)}{=} \frac{1}{p}\left(1 + \frac{\sigma_p^2}{p^2} + \dots\right). \tag{3.6}$$

(1) The geometric series $1/(1+x) = 1 - x + x^2 - \dots$, valid while $|\varepsilon| < p$.
(2) $\langle \varepsilon \rangle = 0$ and $\langle \varepsilon^2 \rangle = \sigma_p^2$.

Even with errors that scatter evenly up and down, $1/p_{\text{obs}}$ overestimates the distance on average, by the fraction σ_p^2/p^2. At $\sigma_p/p = 0.15$ that is 2%. The series needs σ_p/p small (P6). At $\sigma_p/p = 0.5$, 2.3% of the measured parallaxes come out negative, and $1/p_{\text{obs}}$ means nothing for them.

A second bias comes from where the stars are. A shell of radius d and thickness $\mathrm{d}d$ has volume $4\pi d^2\,\mathrm{d}d$, so in a region of even density there are more stars far away than near. Among all the stars measured at one p_{obs}, more have a smaller true parallax and were scattered up than have a larger one and were scattered down. In terms of the parallax, the number of stars is

$$\mathrm{d}N \propto d^2\,\mathrm{d}d \overset{(1)}{=} \frac{1}{p^2}\,\frac{\mathrm{d}p}{p^2} = \frac{\mathrm{d}p}{p^4}. \tag{3.7}$$

(1) $d = 1/p$, so $|\mathrm{d}d| = \mathrm{d}p/p^2$.

By Bayes' rule, the chance of a true parallax p given the measured one is this count times the chance of the error, $p^{-4}\exp\left[-(p_{\text{obs}}-p)^2/2\sigma_p^2\right]$. Its peak is where the derivative of its logarithm vanishes,

$$0 \overset{(1)}{=} -\frac{4}{p} + \frac{p_{\text{obs}} - p}{\sigma_p^2} \quad\Longrightarrow\quad p^2 - p_{\text{obs}}\,p + 4\sigma_p^2 \overset{(2)}{=} 0$$

$$\Longrightarrow\quad p_{\text{peak}} \overset{(3)}{=} \frac{p_{\text{obs}}}{2}\left(1 + \sqrt{1 - \frac{16\sigma_p^2}{p_{\text{obs}}^2}}\right). \tag{3.8}$$

(1) The derivative of $-4\ln p - (p_{\text{obs}} - p)^2/2\sigma_p^2$ with respect to p.
(2) Multiplied by $-p\,\sigma_p^2$.
(3) The larger root of the quadratic. The smaller root is a minimum, since the chance grows without limit as $p \to 0$.

At $\sigma_p/p_{\text{obs}} = 0.15$ the bracket is 1.8, so $p_{\text{peak}} = 0.9\,p_{\text{obs}}$ and the most likely distance is 11% beyond $1/p_{\text{obs}}$. For $\sigma_p/p_{\text{obs}} > 1/4$ the square root is imaginary and there is no peak at all. The measurement alone then cannot fix the distance, and the answer depends on what is assumed about where the stars are. This is the **Lutz–Kelker bias**. The *Gaia* catalogues handle it by assuming a density of stars that falls off exponentially with distance, which keeps the chance finite as $p \to 0$.

3.2 Three-Dimensional Cartesian Position Vector

The parallax gives the distance d, and Chapter 2 gives the direction $\hat{\mathbf{r}}_{\text{eq}}$. Together they place the star in space. To describe how the star moves we need more. We need three directions at the star itself, because a motion is seen as a part along the line of sight and a part across the sky.

On the Earth every town has its own up, north and east. They differ from town to town, but in each town they are at right angles to each other. A pilot reports a speed as so much north and so much east, at the place where the plane is. A point on the sky has the same three directions. Up is outward along the line of sight, north is the way of increasing declination, and east is the way of increasing right ascension. The idea is that the three local directions, written in equatorial components, are the rows of a rotation matrix.

At $\alpha = 0$, $\delta = 0$, these three local directions coincide with the equatorial x, y and z axes. Elsewhere they turn with the star's position. We will first use the line-of-sight direction to write the position vector, then construct the other two directions to describe its motion across the sky.

The position. The star is at distance d along $\hat{\mathbf{r}}_{\text{eq}}$, Eq. (2.2), so its position relative to the Sun is

$$\boxed{\vec{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = d\,\hat{\mathbf{r}}_{\text{eq}} = \begin{pmatrix} d\cos\delta\cos\alpha \\ d\cos\delta\sin\alpha \\ d\sin\delta \end{pmatrix}.} \tag{3.9}$$

For Sirius, at $\alpha = 101.287°$, $\delta = -16.716°$ and $d = 2.637\,\text{pc}$, this is

$$\vec{r} = (-0.494,\ 2.477,\ -0.758)^T\ \text{pc}.$$

The local directions. The direction of increasing δ at the star is the rate of change of $\hat{\mathbf{r}}$ with δ, and the same for α. Differentiating Eq. (2.2),

$$\frac{\partial\hat{\mathbf{r}}}{\partial\delta} = \begin{pmatrix} -\sin\delta\cos\alpha \\ -\sin\delta\sin\alpha \\ \cos\delta \end{pmatrix}, \qquad \frac{\partial\hat{\mathbf{r}}}{\partial\alpha} = \begin{pmatrix} -\cos\delta\sin\alpha \\ \cos\delta\cos\alpha \\ 0 \end{pmatrix} = \cos\delta\begin{pmatrix} -\sin\alpha \\ \cos\alpha \\ 0 \end{pmatrix}. \tag{3.10}$$

The first has length $\sqrt{\sin^2\delta + \cos^2\delta} = 1$. The second has length $\cos\delta$, because a circle of constant declination has radius $\cos\delta$ on the unit sphere and shrinks to a point at the poles. Dividing it by $\cos\delta$ gives a unit vector. So the three directions at the star are

$$\hat{\mathbf{r}} = \begin{pmatrix} \cos\delta\cos\alpha \\ \cos\delta\sin\alpha \\ \sin\delta \end{pmatrix}, \qquad \hat{\boldsymbol{\alpha}} = \frac{1}{\cos\delta}\frac{\partial\hat{\mathbf{r}}}{\partial\alpha} = \begin{pmatrix} -\sin\alpha \\ \cos\alpha \\ 0 \end{pmatrix}, \qquad \hat{\boldsymbol{\delta}} = \frac{\partial\hat{\mathbf{r}}}{\partial\delta} = \begin{pmatrix} -\sin\delta\cos\alpha \\ -\sin\delta\sin\alpha \\ \cos\delta \end{pmatrix}. \tag{3.11}$$

They are at right angles to each other at every point,

$$\begin{aligned} \hat{\mathbf{r}}\cdot\hat{\boldsymbol{\alpha}} &= -\cos\delta\cos\alpha\sin\alpha + \cos\delta\sin\alpha\cos\alpha = 0, \\ \hat{\mathbf{r}}\cdot\hat{\boldsymbol{\delta}} &= -\sin\delta\cos\delta\,(\cos^2\alpha + \sin^2\alpha) + \sin\delta\cos\delta = 0, \\ \hat{\boldsymbol{\alpha}}\cdot\hat{\boldsymbol{\delta}} &= \sin\delta\sin\alpha\cos\alpha - \sin\delta\sin\alpha\cos\alpha = 0, \end{aligned} \tag{3.12}$$

and $\hat{\mathbf{r}}\times\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\delta}}$, so in the order $(\hat{\mathbf{r}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}})$ they are right-handed, like (x, y, z).

The matrix of the local axes. By the rule of rows, Eq. (2.3), the matrix whose rows are the three local directions turns equatorial components into local ones. We call it $\mathbf{Q}$,

$$\mathbf{Q}(\alpha,\delta) = \begin{pmatrix} \cos\delta\cos\alpha & \cos\delta\sin\alpha & \sin\delta \\ -\sin\alpha & \cos\alpha & 0 \\ -\sin\delta\cos\alpha & -\sin\delta\sin\alpha & \cos\delta \end{pmatrix}. \tag{3.13}$$

Its rows are orthonormal and right-handed, so $\mathbf{Q}$ is a rotation and $\mathbf{Q}^{-1} = \mathbf{Q}^T$ by Eq. (2.4). The star itself has local components $\mathbf{Q}\,\hat{\mathbf{r}}_{\text{eq}} = (1, 0, 0)^T$, straight up.

The same matrix comes out of two turns, as a second check. A pan by α about the pole, $\mathbf{R}_z(\alpha)$ of Eq. (2.6), brings the star to $(\cos\delta, 0, \sin\delta)^T$ by Eq. (2.9). A tilt by δ about the new y axis then brings it to the first axis. By the rule of rows the tilt has the rows $(\cos\delta, 0, \sin\delta)$, $(0, 1, 0)$ and $(-\sin\delta, 0, \cos\delta)$, so

$$\begin{pmatrix} \cos\delta & 0 & \sin\delta \\ 0 & 1 & 0 \\ -\sin\delta & 0 & \cos\delta \end{pmatrix} \mathbf{R}_z(\alpha) \overset{(1)}{=} \begin{pmatrix} \cos\delta & 0 & \sin\delta \\ 0 & 1 & 0 \\ -\sin\delta & 0 & \cos\delta \end{pmatrix} \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
$$\overset{(2)}{=} \begin{pmatrix} \cos\delta\cos\alpha & \cos\delta\sin\alpha & \sin\delta \\ -\sin\alpha & \cos\alpha & 0 \\ -\sin\delta\cos\alpha & -\sin\delta\sin\alpha & \cos\delta \end{pmatrix} = \mathbf{Q}(\alpha, \delta). \tag{3.14}$$

(1) Eq. (2.6) with $\theta = \alpha$.
(2) Each row of the tilt times each column of $\mathbf{R}_z(\alpha)$.

The yearly parallax ellipse. The goal is the path a nearby star traces on the sky over a year, as seen from the moving Earth. We work in ecliptic coordinates (λ, β), Eq. (2.40), and take the Earth's orbit as a circle of radius 1 AU. With the Sun at ecliptic longitude $\lambda_\odot$, the Earth is on the opposite side, at $\vec{r}_\oplus = -1\,\text{AU}\,(\cos\lambda_\odot, \sin\lambda_\odot, 0)^T$ from the Sun. The star is at $d\,\hat{\mathbf{r}}$ from the Sun, so from the Earth it is at $d\,\hat{\mathbf{r}} - \vec{r}_\oplus$. Its apparent direction is

$$\hat{\mathbf{r}}' = \frac{d\,\hat{\mathbf{r}} - \vec{r}_\oplus}{|d\,\hat{\mathbf{r}} - \vec{r}_\oplus|} \overset{(1)}{=} \hat{\mathbf{r}} - \frac{\vec{r}_\oplus - (\hat{\mathbf{r}}\cdot\vec{r}_\oplus)\,\hat{\mathbf{r}}}{d}. \tag{3.15}$$

(1) The first-order expansion of Eq. (2.63), with $-\vec{r}_\oplus/d$ in place of $\vec{v}/c$. It holds since $1\,\text{AU}/d \ll 1$ (P6).

The star moves away from the Earth's position, by the part of $\vec{r}_\oplus/d$ across the line of sight. The local directions in the ecliptic frame are Eq. (3.11) with $\alpha \to \lambda$ and $\delta \to \beta$, so $\hat{\boldsymbol{\lambda}} = (-\sin\lambda, \cos\lambda, 0)^T$ and $\hat{\boldsymbol{\beta}} = (-\sin\beta\cos\lambda, -\sin\beta\sin\lambda, \cos\beta)^T$. Both are perpendicular to $\hat{\mathbf{r}}$, so the shifts along them are $-\vec{r}_\oplus\cdot\hat{\boldsymbol{\lambda}}/d$ and $-\vec{r}_\oplus\cdot\hat{\boldsymbol{\beta}}/d$. With $p = 1\,\text{AU}/d$ from Eq. (3.2),

$$\cos\beta\,\Delta\lambda \overset{(1)}{=} p\,(-\cos\lambda_\odot\sin\lambda + \sin\lambda_\odot\cos\lambda) \overset{(2)}{=} p\sin(\lambda_\odot - \lambda),$$
$$\Delta\beta \overset{(1)}{=} -p\sin\beta\,(\cos\lambda_\odot\cos\lambda + \sin\lambda_\odot\sin\lambda) \overset{(2)}{=} -p\sin\beta\cos(\lambda_\odot - \lambda). \tag{3.16}$$

(1) The dot products of $(\cos\lambda_\odot, \sin\lambda_\odot, 0)$ with $\hat{\boldsymbol{\lambda}}$ and with $\hat{\boldsymbol{\beta}}$.
(2) The angle-difference identities.

A step $\Delta\lambda$ in longitude is an arc $\cos\beta\,\Delta\lambda$ on the sky, for the same reason as the $\cos\delta$ in Eq. (3.10). As $\lambda_\odot$ runs through the year,

$$\boxed{\left(\frac{\cos\beta\,\Delta\lambda}{p}\right)^2 + \left(\frac{\Delta\beta}{p\sin\beta}\right)^2 = 1.} \tag{3.17}$$

Every star traces an ellipse with semi-axes p along the ecliptic and $p\sin\beta$ across it. Near the ecliptic pole it is a circle of radius p. On the ecliptic it closes to a line of half-length p. The semi-major axis is p for every star, which is why p is the number a catalogue lists. The aberration ellipse of Chapter 2 has the same shape and a quarter of a year's offset, but it is 20.5″ for every star, so the two are told apart by size and by timing.

3.3 Proper Motion and Transverse Velocity

The parallax and the direction give the star's position $\vec{r} = d\,\hat{\mathbf{r}}$, Eq. (3.9). Stars move, so $\vec{r}$ changes. A catalogue position slowly goes out of date, as the change of epoch in Section 2.1.1 already showed. We need to read a velocity from the change.

Stand by a straight road and watch two cars, both at 30 m/s. The one 300 m away crosses your view at 0.1 radian, about 6°, per second. The one 3 km away crosses at only 0.6° per second. The angle swept per second is the sideways speed divided by the distance. A star moving across the line of sight drifts across the sky in the same way. The drift is its **proper motion**. The idea is that the sideways speed is the proper motion times the distance.

The parsec makes the conversion easy to remember. A star 1 pc away that moves sideways at 4.74 km/s crosses 1 AU in a year and shifts by 1″ on the sky. A more distant star with the same sideways speed shifts less. We will resolve that angular drift into the directions of increasing right ascension and declination, then multiply by distance to recover its transverse speed.

How the direction changes. As the star moves, its α and δ change with time, and so does $\hat{\mathbf{r}}(\alpha, \delta)$. By the chain rule,

$$\frac{\mathrm{d}\hat{\mathbf{r}}}{\mathrm{d}t} = \frac{\partial \hat{\mathbf{r}}}{\partial \alpha}\frac{\mathrm{d}\alpha}{\mathrm{d}t} + \frac{\partial \hat{\mathbf{r}}}{\partial \delta}\frac{\mathrm{d}\delta}{\mathrm{d}t} \overset{(1)}{=} \underbrace{\cos\delta\,\frac{\mathrm{d}\alpha}{\mathrm{d}t}}_{\mu_\alpha^*}\,\hat{\boldsymbol{\alpha}} + \underbrace{\frac{\mathrm{d}\delta}{\mathrm{d}t}}_{\mu_\delta}\,\hat{\boldsymbol{\delta}}. \tag{3.18}$$

(1) The partial derivatives (3.10), written with the unit vectors (3.11).

The two components are the **proper motion in right ascension** $\mu_\alpha^* = \cos\delta\,\dot{\alpha}$ and the **proper motion in declination** $\mu_\delta = \dot{\delta}$. Both are true angles on the sky per unit time, usually given in mas/yr. The rate $\dot{\alpha}$ itself is not, because a step in α is an arc of only $\cos\delta$ times its size, as at Eq. (3.10). The two directions are at right angles, so the total proper motion is

$$\mu = \left|\frac{\mathrm{d}\hat{\mathbf{r}}}{\mathrm{d}t}\right| = \sqrt{(\mu_\alpha^*)^2 + \mu_\delta^2} = \sqrt{(\cos\delta\,\dot{\alpha})^2 + \mu_\delta^2}. \tag{3.19}$$

The velocity of the star. The goal is the star's velocity $\vec{v} = \dot{\vec{r}}$ in terms of what is measured. Differentiating $\vec{r} = d\,\hat{\mathbf{r}}$,

$$\vec{v} \overset{(1)}{=} \dot{d}\,\hat{\mathbf{r}} + d\,\frac{\mathrm{d}\hat{\mathbf{r}}}{\mathrm{d}t} \overset{(2)}{=} \underbrace{\dot{d}}_{v_r}\,\hat{\mathbf{r}} + \underbrace{d\,\mu_\alpha^*}_{v_\alpha}\,\hat{\boldsymbol{\alpha}} + \underbrace{d\,\mu_\delta}_{v_\delta}\,\hat{\boldsymbol{\delta}}. \tag{3.20}$$

(1) The product rule on $\vec{r} = d\,\hat{\mathbf{r}}$.
(2) Eq. (3.18).

The **radial velocity** $v_r = \dot{d}$ is the rate at which the distance grows. It does not change the direction, so astrometry cannot see it, and it is measured from the Doppler shift (Section 3.4). The other two components lie across the line of sight. Their size is the **transverse velocity**

$$v_t = \sqrt{v_\alpha^2 + v_\delta^2} = d\,\mu, \tag{3.21}$$

with μ in radians per unit time.

The constant 4.74. Catalogues give μ in seconds of arc per year and d through p in seconds of arc, and we want v_t in km/s.

$$v_t \overset{(1)}{=} \frac{1\,\mathrm{AU}}{p}\,\mu \overset{(2)}{=} \frac{\mu\,[''/\mathrm{yr}]}{p\,['']}\,\frac{1\,\mathrm{AU}}{1\,\mathrm{yr}} \overset{(3)}{=} \frac{\mu\,[''/\mathrm{yr}]}{p\,['']}\,\frac{1.495\,978\,707\times10^{8}\,\mathrm{km}}{3.155\,76\times10^{7}\,\mathrm{s}}$$
$$= 4.740\,47\,\frac{\mu\,[''/\mathrm{yr}]}{p\,['']}\ \mathrm{km/s}. \tag{3.22}$$

(1) Eq. (3.21) with $d = 1\,\mathrm{AU}/p$ from Eq. (3.2).
(2) μ/p is a ratio of two angles, the same in any angular unit, so μ in ″/yr and p in ″ can be used directly.
(3) 1 AU in km, and the Julian year of 365.25 days in seconds.

With $1/p'' = d$ in parsecs from Eq. (3.5),

Result 3.2

$$v_t\,[\mathrm{km/s}] = 4.74047\,\mu\,[\mathrm{arcsec/yr}]\,d\,[\mathrm{pc}] = 4.74047\,\frac{\mu\,[\mathrm{arcsec/yr}]}{p\,[\mathrm{arcsec}]}. \tag{3.23}$$

The number 4.74047 is just 1 AU per year in km/s. The transverse velocity depends only on the ratio of the two angles, the yearly drift over the yearly wobble. The same holds for each component, $v_\alpha = 4.74047\,\mu_\alpha^*/p$ and $v_\delta = 4.74047\,\mu_\delta/p$.

Worked example 3.3: Transverse Velocity of Sirius

Sirius has $\mu_\alpha^* = -546.01\,\mathrm{mas/yr}$ and $\mu_\delta = -1223.07\,\mathrm{mas/yr}$. By Eq. (3.19),

$$\mu = \sqrt{0.54601^2 + 1.22307^2}\ ''/\mathrm{yr} = 1.3394''/\mathrm{yr}.$$

With $d = 2.637\,\mathrm{pc}$, Eq. (3.23) gives

$$v_t = 4.74047 \times 1.3394 \times 2.637\,\mathrm{km/s} = 16.74\,\mathrm{km/s}.$$

The components are $v_\alpha = -6.83\,\mathrm{km/s}$ and $v_\delta = -15.29\,\mathrm{km/s}$, so Sirius drifts mainly south.

3.4 Radial Velocity and the Doppler Effect

Proper motion and parallax give the two components of the velocity across the line of sight, Eq. (3.20). The third component, $v_r = \dot{d}$, changes the distance but not the direction. A star moving straight toward us shows no drift on the sky at all, and its parallax changes far too slowly to measure.

Picture a friend who walks away from you at 1 m/s and throws you a ball every second, each at 10 m/s. Each ball starts 1 m further away than the one before, so it needs an extra 0.1 s to arrive. The balls reach you every 1.1 seconds, not every second. Walking toward you, the friend would send them every 0.9 seconds. A star sends wave crests instead of balls. If it moves away, each crest starts a little further off, and the crests arrive further apart. The light is shifted to longer wavelengths. The idea is that the fractional stretch of the wavelength is v_r/c.

Sirius approaches at 5.5 km/s, so its laboratory Hα wavelength of 656.46 nm is shortened by $656.46 \times 5.5/299\,792\,\mathrm{nm} = 0.012\,\mathrm{nm}$. The shift is small because stellar speeds are typically only $v_r/c \sim 10^{-5}$. We will derive the relation between this fractional wavelength shift and radial speed, then account for the Earth's own motion when measuring it.

The stretched period. Take a source that emits crests one period $T_0 = \lambda_0/c$ apart in its own frame, and moves straight away from the observer at $v_r \ll c$.

$$T_{\text{obs}} \overset{(1)}{=} T_0 + \frac{v_r T_0}{c} = T_0\left(1 + \frac{v_r}{c}\right),$$
$$z \equiv \frac{\lambda_{\text{obs}} - \lambda_0}{\lambda_0} \overset{(2)}{=} \frac{T_{\text{obs}} - T_0}{T_0} \overset{(3)}{=} \frac{v_r}{c}. \tag{3.24}$$

(1) Between two crests the source moves a further $v_r T_0$ away, so the second crest travels that far more and arrives $v_r T_0/c$ later.
(2) $\lambda = cT$ for both wavelengths.
(3) The first line.

Result 3.3

$$z = \frac{\lambda_{\text{obs}} - \lambda_0}{\lambda_0} = \frac{v_r}{c} \qquad (v_r \ll c). \tag{3.25}$$

A positive v_r is a recession and a shift to the red. A negative v_r is an approach and a shift to the blue. The shift fixes v_r alone, and the motion across the line of sight leaves it unchanged at this order.

The exact formula. For speeds near c, such as relativistic jets, the clock of the moving source runs slow. In the observer's frame its period is $T_0/\sqrt{1 - v^2/c^2}$, where v is its full speed. For motion straight along the line of sight, $v = v_r$, and

$$1 + z = \frac{T_{\text{obs}}}{T_0} \overset{(1)}{=} \frac{1 + v_r/c}{\sqrt{1 - v_r^2/c^2}} \overset{(2)}{=} \frac{1 + v_r/c}{\sqrt{(1 + v_r/c)(1 - v_r/c)}} = \sqrt{\frac{1 + v_r/c}{1 - v_r/c}}. \tag{3.26}$$

(1) The stretch of the first line of the classical chain, applied to the dilated period.
(2) $1 - v_r^2/c^2 = (1 + v_r/c)(1 - v_r/c)$.

Squaring and solving for v_r/c,

$$1 + z = \sqrt{\frac{1 + v_r/c}{1 - v_r/c}} \quad \Longrightarrow \quad \frac{v_r}{c} = \frac{(1+z)^2 - 1}{(1+z)^2 + 1}. \tag{3.27}$$

For small v_r/c the root is $1 + v_r/c + \frac{1}{2}(v_r/c)^2 + \ldots$, so Eq. (3.25) is its first term. For Sirius the second term is 1.7×10^{-10}, far below what a spectrograph can measure. A source moving across the line of sight also shows a shift, $z = 1/\sqrt{1 - v^2/c^2} - 1 \approx v^2/2c^2$, from the slow clock alone. The redshift of distant galaxies is a different effect. It comes from the expansion of space, and Eq. (3.27) does not apply to it.

The Earth's own motion. A spectrograph on the Earth moves with it, at about 29.78 km/s round the Sun. That is $2\pi \times 4.74$ km/s, since the Earth covers 2π AU a year. The shift measures how fast the distance from the observer to the star grows. With the star at $\vec{r}_*$ and the Earth at $\vec{r}_\oplus$,

$$\frac{\mathrm{d}}{\mathrm{d}t}|\vec{r}_* - \vec{r}_\oplus| \overset{(1)}{=} \frac{(\vec{r}_* - \vec{r}_\oplus)\cdot(\vec{v}_* - \vec{v}_\oplus)}{|\vec{r}_* - \vec{r}_\oplus|} \overset{(2)}{=} (\vec{v}_* - \vec{v}_\oplus)\cdot\hat{\mathbf{r}} = v_r - \vec{v}_\oplus\cdot\hat{\mathbf{r}}. \tag{3.28}$$

(1) $\mathrm{d}|\vec{a}|/\mathrm{d}t = \vec{a}\cdot\dot{\vec{a}}/|\vec{a}|$.
(2) The star is so far away that $(\vec{r}_* - \vec{r}_\oplus)/|\vec{r}_* - \vec{r}_\oplus|$ is $\hat{\mathbf{r}}$ to first order in 1 AU/d.

So the velocity relative to the Sun is the measured one plus $\vec{v}_\oplus \cdot \hat{\mathbf{r}}$. With the Earth's velocity of Chapter 2, this term swings once a year between $\pm 29.78 \cos\beta$ km/s, where β is the star's ecliptic latitude. It is larger than many stellar radial velocities, so every catalogue value has it removed. The Earth's spin adds up to 0.47 km/s at the equator, which is removed in the same way.

3.5 The Full 3D Space Velocity Vector

Eq. (3.20) gives the star's velocity as three components $(v_r, v_\alpha, v_\delta)$ along its local axes. Those axes differ from star to star. To compare two stars, or to add the Sun's own motion, we need all velocities in one fixed frame.

A ship's log gives the current as so much ahead and so much to starboard. To draw it on a chart the navigator turns those two numbers through the ship's heading into north and east. The current itself has not changed, only the axes it is written on. A star's velocity is written on its own local axes, and the fixed frame is the equatorial one. The idea is that the local components turn into equatorial ones with the transpose of $\mathbf{Q}$, Eq. (3.13).

At $\alpha = \delta = 0$ the local and equatorial axes coincide, so $\mathbf{Q}$ is the identity and $(v_x, v_y, v_z) = (v_r, v_\alpha, v_\delta)$. For any other direction we turn the same velocity vector into the fixed equatorial axes with the matrix already built for the star's local frame.

Turning to the equatorial frame. Stacking the three local directions of Eq. (3.11) as columns,

$$\begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} \overset{(1)}{=} v_r\,\hat{\mathbf{r}} + v_\alpha\,\hat{\boldsymbol{\alpha}} + v_\delta\,\hat{\boldsymbol{\delta}} \overset{(2)}{=} \begin{pmatrix} | & | & | \\ \hat{\mathbf{r}} & \hat{\boldsymbol{\alpha}} & \hat{\boldsymbol{\delta}} \\ | & | & | \end{pmatrix} \begin{pmatrix} v_r \\ v_\alpha \\ v_\delta \end{pmatrix} \overset{(3)}{=} \mathbf{Q}^T(\alpha,\delta) \begin{pmatrix} v_r \\ v_\alpha \\ v_\delta \end{pmatrix}. \tag{3.29}$$

(1) Eq. (3.20), with each unit vector in equatorial components.
(2) A matrix times a column is the sum of its columns weighted by the entries.
(3) The rows of $\mathbf{Q}$, Eq. (3.13), are $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\delta}}$, so its columns are those of $\mathbf{Q}^T$.

$$\boxed{\begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = \begin{pmatrix} \cos\delta\cos\alpha & -\sin\alpha & -\sin\delta\cos\alpha \\ \cos\delta\sin\alpha & \cos\alpha & -\sin\delta\sin\alpha \\ \sin\delta & 0 & \cos\delta \end{pmatrix} \begin{pmatrix} v_r \\ v_\alpha \\ v_\delta \end{pmatrix}.} \tag{3.30}$$

Going back costs only a transpose, $(v_r, v_\alpha, v_\delta)^T = \mathbf{Q}\,(v_x, v_y, v_z)^T$, by Eq. (2.4).

The speed. A rotation keeps lengths, so the speed is the same in both sets of components,

$$v_x^2 + v_y^2 + v_z^2 = \begin{pmatrix} v_r & v_\alpha & v_\delta \end{pmatrix} \mathbf{Q}\,\mathbf{Q}^T \begin{pmatrix} v_r \\ v_\alpha \\ v_\delta \end{pmatrix} \overset{(1)}{=} v_r^2 + v_\alpha^2 + v_\delta^2 \overset{(2)}{=} v_r^2 + v_t^2. \tag{3.31}$$

(1) $\mathbf{Q}\mathbf{Q}^T = \mathbb{1}$, Eq. (2.4).
(2) Eq. (3.21).

$$v_{\text{total}} = \sqrt{v_r^2 + v_t^2} = \sqrt{v_x^2 + v_y^2 + v_z^2}. \tag{3.32}$$

Worked example 3.4: Space Velocity of Sirius

Sirius is at $\alpha = 101.287°$, $\delta = -16.716°$, with $v_r = -5.50$ km/s and, from the previous

example, $v_\alpha = -6.83\,\mathrm{km/s}$, $v_\delta = -15.29\,\mathrm{km/s}$. Eq. (3.30) gives

$$\begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = \begin{pmatrix} -0.18746 & -0.98066 & -0.05630 \\ 0.93922 & -0.19573 & 0.28207 \\ -0.28763 & 0 & 0.95774 \end{pmatrix} \begin{pmatrix} -5.50 \\ -6.83 \\ -15.29 \end{pmatrix} \mathrm{km/s} = \begin{pmatrix} 8.59 \\ -8.14 \\ -13.06 \end{pmatrix} \mathrm{km/s}.$$

The speed is $\sqrt{8.59^2 + 8.14^2 + 13.06^2} = 17.62\,\mathrm{km/s}$, the same as $\sqrt{5.50^2 + 16.74^2}$ from Eq. (3.32).

3.6 Galactic Space Velocity Components (U, V, W)

Eq. (3.30) gives every star's velocity in one fixed frame, the equatorial one. That frame is set by the Earth's spin axis, which means nothing to the Galaxy. Stars orbit the Galactic centre in the Galactic disk, so their motions are simplest on Galactic axes. The velocities are also measured relative to the Sun, which itself moves.

A wind gauge on a moving train reads the wind relative to the train. To get the wind over the ground we first turn the reading to compass directions and then add the train's own velocity. The Sun is the train. We turn (v_x, v_y, v_z) to Galactic axes with the matrix of Chapter 2, and then add the Sun's own motion relative to its neighbours. The idea is that a velocity turns with the same matrix as a position, and that velocities relative to moving frames add.

For Sirius the heliocentric components are 12.96 km/s toward the Galactic centre, 1.89 km/s against Galactic rotation and 11.79 km/s toward the south Galactic pole. After accounting for the Sun's motion, its components relative to nearby stars are 24.1, 10.3 and −4.5 km/s on the same axes. We will make those two steps explicit: rotate the measured vector and then change the velocity reference frame.

The axes. Galactic velocities are given in a right-handed Cartesian frame centred on the Sun,

- U toward the Galactic centre ($l = 0°$, $b = 0°$),
- V along the direction of Galactic rotation ($l = 90°$, $b = 0°$),
- W toward the north Galactic pole ($b = +90°$).

Some authors take U positive toward the anticentre, which makes the frame left-handed. We do not.

Why a velocity turns like a position. Chapter 2 turned positions with a fixed matrix, $\hat{\mathbf{r}}_{\mathrm{gal}} = \mathbf{T}\,\hat{\mathbf{r}}_{\mathrm{eq}}$ with $\mathbf{T}$ the product of three turns, Eq. (2.47). Multiplying by d gives the same for the full position $\vec{r}$. The Galactic pole and centre are fixed by convention at their J2000.0 values, so $\mathbf{T}$ does not change with time, and

$$\begin{pmatrix} U \\ V \\ W \end{pmatrix} = \frac{\mathrm{d}}{\mathrm{d}t}\,\vec{r}_{\mathrm{gal}} \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{T}\,\vec{r}_{\mathrm{eq}}\right) \overset{(2)}{=} \mathbf{T}\,\frac{\mathrm{d}\vec{r}_{\mathrm{eq}}}{\mathrm{d}t} = \mathbf{T}\begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix}. \tag{3.33}$$

(1) Eq. (2.47), multiplied by d.
(2) $\mathbf{T}$ is constant.

The two frames share their origin at the Sun, so no extra velocity enters here.

The rotation. The three turns of Eq. (2.47), multiplied out as in Eq. (2.54), give

$$\begin{pmatrix} U \\ V \\ W \end{pmatrix} = \mathbf{R}_z(90^\circ - l_{\mathrm{CP}})\,\mathbf{R}_x(90^\circ - \delta_{\mathrm{GP}})\,\mathbf{R}_z(\alpha_{\mathrm{GP}} + 90^\circ) \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix}$$
$$\overset{(1)}{=} \begin{pmatrix} -0.054876 & -0.873437 & -0.483835 \\ +0.494109 & -0.444830 & +0.746982 \\ -0.867666 & -0.198076 & +0.455984 \end{pmatrix} \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix}. \tag{3.34}$$

(1) The IAU values of α_{GP}, δ_{GP} and l_{CP}, as in Eq. (2.54).

By the rule of rows, the rows of $\mathbf{T}$ are the directions of U, V and W in equatorial components. Putting in Eq. (3.30) for the equatorial components and Eq. (3.23) for v_α and v_δ takes us from the measured numbers to (U, V, W) in one line,

Result 3.4

$$\begin{pmatrix} U \\ V \\ W \end{pmatrix} = \mathbf{T}\,\mathbf{Q}^T(\alpha, \delta) \begin{pmatrix} v_r \\ 4.74047\,\mu_\alpha^*/p \\ 4.74047\,\mu_\delta/p \end{pmatrix}, \tag{3.35}$$

in km/s for v_r in km/s, μ in ″/yr and p in ″. All six measured numbers $(\alpha, \delta, p, \mu_\alpha^*, \mu_\delta, v_r)$ enter, and nothing else. For Sirius, $\mathbf{T}$ times the column $(8.59, -8.14, -13.06)$ km/s of the last example gives $(U, V, W) = (12.96,\ -1.89,\ -11.79)$ km/s.

The local standard of rest. The **local standard of rest** (LSR) is a frame that moves on a circular orbit round the Galactic centre at the Sun's distance from it, $R_0 \approx 8.2$ kpc, with circular speed $v_0 \approx 220$–240 km/s. The Sun moves relative to it. The measured value (Schönrich et al. 2010) is

$$(U_\odot, V_\odot, W_\odot) = (11.10^{+0.69}_{-0.75},\ 12.24^{+0.47}_{-0.47},\ 7.25^{+0.37}_{-0.36})\ \mathrm{km/s}. \tag{3.36}$$

The velocity of the star relative to the LSR is its velocity relative to the Sun plus that of the Sun relative to the LSR. Both are written on the same Galactic axes, so the components add,

$$(U_{\mathrm{LSR}}, V_{\mathrm{LSR}}, W_{\mathrm{LSR}}) = (U + U_\odot,\ V + V_\odot,\ W + W_\odot). \tag{3.37}$$

For Sirius this is $(24.06,\ 10.35,\ -4.54)$ km/s. Relative to the LSR, the velocity of a disk star is the small departure of its Galactic orbit from a circle.

3.7 The Cosmic Distance Ladder

Parallax gives a distance from geometry alone, Eq. (3.5). Its angle shrinks as $1/d$, while the error of the measurement does not. Beyond some distance the parallax is lost in the error. At the other end, the parallax itself is only as good as the length of its baseline, 1 AU, in metres.

To find how far away a lighthouse is at night, we cannot pace the distance. But if we have measured how bright such a lamp is from close by, its faintness tells us how far it is. Astronomers do the same with stars whose true brightness can be known from some other property. The nearby ones are calibrated by parallax, and they then give distances far beyond its reach. The idea is that each kind of distance calibrates the next, like the rungs of a ladder.

A distance is good to 10% when $\sigma_p/p = 0.1$. Since $d = 1/p$, the fractional errors of d and p are equal, and with $p = 1/d$ in seconds of arc and parsecs,

$$\frac{\sigma_d}{d} \overset{(1)}{=} \frac{\sigma_p}{p} \overset{(2)}{=} \sigma_p\, d \quad \Longrightarrow \quad d_{10\%} = \frac{0.1}{\sigma_p}. \tag{3.38}$$

(1) $|\mathrm{d}d/d| = |\mathrm{d}p/p|$ for $d = 1/p$.
(2) Eq. (3.5).

Hipparcos measured $\sigma_p \approx 1$ mas for bright stars, which gives 10% distances to 100 pc. *Gaia* reaches $\sigma_p \approx 0.02$ mas for bright stars, which gives 5 kpc, still well short of the far side of the Galaxy.

The picture is a ladder whose bottom rung stands on the Earth's orbit and whose top reaches the distant galaxies (Figure 3.2). Each bar is the range of one method, and each overlaps the one below it, where it is calibrated.

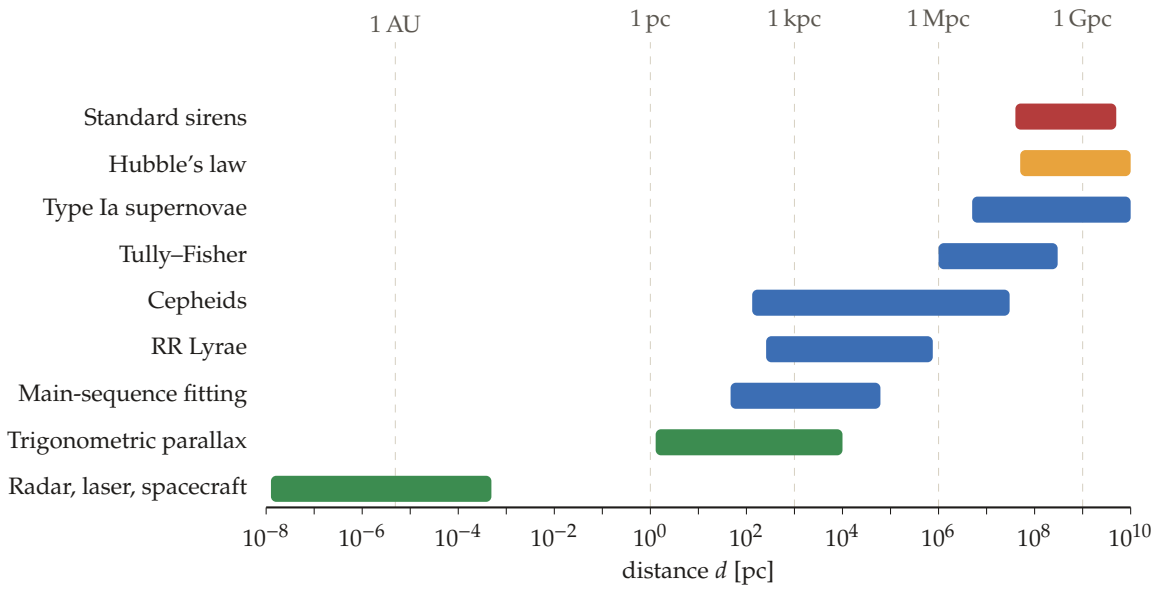


Figure 3.2: The rungs of the distance ladder. Each bar spans the distances over which one method is used, on a logarithmic axis in parsecs, with 1 AU, 1 pc, 1 kpc, 1 Mpc and 1 Gpc marked by dashed lines. Green bars are geometric methods, blue bars use a known luminosity, the orange bar uses the redshift, and the red bar uses gravitational waves. The ranges are approximate.

Each luminosity-based rung uses the inverse-square law (Eq. (4.2)) to turn measured flux into distance. Its calibration comes from objects whose distances are already known on a lower rung. We will follow that chain from the AU to stellar and galactic distances, then see how uncertainty in one calibration carries upward.

3.7.1 The Bottom Rung, the Astronomical Unit in Metres

Every parallax distance is 1 AU divided by an angle, Eq. (3.2). So every rung of the ladder rests on the length of the AU in metres.

Sizes from periods. Kepler's third law, Eq. (6.50), is $n^2 a^3 = GM$, with $n = 2\pi/P$ for the period P and M the total mass of the pair. For Venus and the Earth, both going round the Sun,

$$\frac{a_{\text{Venus}}^3}{a_\oplus^3} \overset{(1)}{=} \frac{G(M_\odot + m_{\text{Venus}})\, P_{\text{Venus}}^2/4\pi^2}{G(M_\odot + m_\oplus)\, P_\oplus^2/4\pi^2} = \frac{P_{\text{Venus}}^2}{P_\oplus^2}\,\frac{1 + m_{\text{Venus}}/M_\odot}{1 + m_\oplus/M_\odot}$$
$$\overset{(2)}{=} \frac{P_{\text{Venus}}^2}{P_\oplus^2} \quad\Longrightarrow\quad \frac{a_{\text{Venus}}}{a_\oplus} = \left(\frac{P_{\text{Venus}}}{P_\oplus}\right)^{2/3} \overset{(3)}{=} \left(\frac{224.701}{365.256}\right)^{2/3} = 0.7233. \tag{3.39}$$

(1) Eq. (6.50) solved for $a^3 = GMP^2/4\pi^2$, for each planet.

(2) The planets are light, $m_\oplus/M_\odot = 3\times10^{-6}$ and less for Venus, so the last factor differs from one by a few parts in a million (P6).
(3) The sidereal periods of Venus and the Earth in days. The ratio is 0.61519.

The ratio holds no length. Kepler's laws give the shape and the relative size of every planetary orbit, in AU, and nothing in them knows the metre. One distance in the Solar System measured in metres fixes all the others. In the eighteenth century that distance came from the transits of Venus across the Sun in 1761 and 1769. Observers far apart on the Earth saw Venus cross the Sun along slightly different lines, and the Earth's size, known in metres, served as the baseline. The AU came out to a few per cent.

The echo. Today the distance is measured with a radio pulse. The pulse leaves the Earth, bounces off Venus and comes back a time Δt later. It covers the distance D out and back at the speed c, so

$$\Delta t = \frac{2D}{c} \quad\Longrightarrow\quad D = \frac{c\,\Delta t}{2}. \tag{3.40}$$

The planets move at tens of km/s, a ten-thousandth of c, so D hardly changes while the pulse is on its way (P6).

From the echo to the AU. To compare the echo with Kepler's ratio we need D in AU. Take both orbits as circles in one plane, centred on the Sun, with the Earth at angle $n_\oplus t$ and Venus at $n_{\text{Venus}}t$. Then

$$\begin{aligned} D^2 &= \left|\vec r_{\text{Venus}} - \vec r_\oplus\right|^2 = a_{\text{Venus}}^2 + a_\oplus^2 - 2\,\vec r_{\text{Venus}}\cdot\vec r_\oplus \\ &\overset{(1)}{=} a_{\text{Venus}}^2 + a_\oplus^2 - 2a_{\text{Venus}}a_\oplus \begin{pmatrix}\cos n_\oplus t & \sin n_\oplus t & 0\end{pmatrix}\begin{pmatrix}\cos n_{\text{Venus}}t\\ \sin n_{\text{Venus}}t\\ 0\end{pmatrix} \\ &\overset{(2)}{=} a_{\text{Venus}}^2 + a_\oplus^2 - 2a_{\text{Venus}}a_\oplus\cos\left[(n_{\text{Venus}} - n_\oplus)t\right]. \end{aligned} \tag{3.41}$$

(1) $\vec r = a\,(\cos nt, \sin nt, 0)^T$ for each planet, and the dot product written as a row times a column.
(2) $\cos A\cos B + \sin A\sin B = \cos(A-B)$.

The distance swings between $a_\oplus - a_{\text{Venus}}$, when Venus passes between the Earth and the Sun, and $a_\oplus + a_{\text{Venus}}$, when it is behind the Sun. At the near point $D = (1-0.7233)\,a_\oplus = 0.2767$ AU by Eq. (3.39). The near point gives the strongest echo, and there

$$1\text{ AU} = a_\oplus \overset{(1)}{=} \frac{D}{1 - a_{\text{Venus}}/a_\oplus} \overset{(2)}{=} \frac{c\,\Delta t}{2\left[1 - (P_{\text{Venus}}/P_\oplus)^{2/3}\right]}. \tag{3.42}$$

(1) Eq. (3.41) at cos = 1, solved for $a_\oplus$.
(2) Eqs. (3.40) and (3.39).

$$\boxed{1\text{ AU} = \frac{c\,\Delta t}{2\times 0.2767}} \tag{3.43}$$

A time and two periods give the AU. The time carries the metres through c, and the periods carry the geometry through Kepler's third law. The real orbits are slightly eccentric and tilted, so in practice the echo is fitted with the full Kepler orbits of Chapter 6, but the logic is the same.

Worked example 3.5: The AU from a Venus echo

Suppose the echo from Venus at its near point returns after $\Delta t = 276.1\,\mathrm{s}$, an illustrative value. By Eq. (3.40),

$$D = \frac{299\,792.458\,\mathrm{km/s} \times 276.1\,\mathrm{s}}{2} = 4.139 \times 10^{7}\,\mathrm{km},$$

and by Eq. (3.43)

$$1\,\mathrm{AU} = \frac{4.139 \times 10^{7}\,\mathrm{km}}{0.2767} = 1.496 \times 10^{8}\,\mathrm{km}.$$

The radio pulse needs four and a half minutes for the round trip.

Ranging today. Radar echoes from Mercury, Venus, Mars and some asteroids fix their distances in metres. The Moon is ranged with laser pulses. They are sent from the Earth to the corner reflectors left on the Moon by the Apollo crews and the Lunokhod rovers, and the round trip of 2.56 s is timed to about 10^{-10} s, so the distance is known to about a centimetre. The outer planets are ranged by timing radio signals to and from spacecraft. With so many distances known in metres, the AU is no longer measured. Since 2012 it has been defined as exactly 149 597 870 700 m, and the parsec is defined from it. The ranging now fixes $GM_{\odot}$ through Eq. (6.50) instead.

3.7.2 Spectroscopic Parallax and Main-Sequence Fitting

The second rung uses the HR diagram, Figure 12.8. The spectral type of a star (Section 11.7) gives its temperature. The widths of its lines give its surface gravity, and so tell a dwarf from a giant of the same temperature. That is the luminosity class. Together they fix the star's place on the diagram, and so its absolute magnitude M. The apparent magnitude m, corrected for dust with Eq. (4.44), then gives the distance. Despite the name, no angle is measured. Inverting the distance modulus, Eq. (4.11),

$$\log_{10}\frac{d}{10\,\mathrm{pc}} \overset{(1)}{=} \frac{m-M}{5}$$
$$\Longrightarrow \quad d \overset{(2)}{=} 10\,\mathrm{pc} \times 10^{(m-M)/5}. \tag{3.44}$$

(1) Eq. (4.11) divided by 5.
(2) Ten raised to the power of each side.

How an error in M becomes an error in d. The main sequence is a band, not a line, so M read from the diagram is uncertain by some amount ΔM. Suppose the value used is too bright by ΔM, that is, $M - \Delta M$ in place of the true M. Then by Eq. (3.44)

$$\frac{d_{\mathrm{used}}}{d} = \frac{10\,\mathrm{pc} \times 10^{(m-M+\Delta M)/5}}{10\,\mathrm{pc} \times 10^{(m-M)/5}} = 10^{\Delta M/5} \overset{(1)}{=} 10^{0.1} = 1.259. \tag{3.45}$$

(1) $\Delta M = 0.5$ magnitudes.

A magnitude half a unit too bright puts the star 26% too far away. Half a unit too faint gives $10^{-0.1} = 0.794$, 21% too near. The two are not equal, because d grows exponentially with M. For a small spread σ_M the first-order rule follows from the logarithm,

$$\ln d \overset{(1)}{=} \ln(10\,\mathrm{pc}) + \frac{\ln 10}{5}(m-M) \quad \Longrightarrow \quad \frac{\sigma_d}{d} \overset{(2)}{=} \frac{\ln 10}{5}\,\sigma_M = 0.461\,\sigma_M. \tag{3.46}$$

(1) The natural logarithm of Eq. (3.44), with $\ln 10^x = x \ln 10$.
(2) $\mathrm{d}(\ln d) = \mathrm{d}d/d$, and only M is uncertain.

For $\sigma_M = 0.5$ this gives 23%, between the exact +26% and −21%. The spread of the band has several causes. Stars with fewer metals sit lower at a given colour. A star brightens slowly as it ages on the main sequence. Two equal stars too close to separate send twice the light of one, so they look $2.5 \log_{10} 2 = 0.75$ magnitudes brighter, and by Eq. (3.44) they are placed $10^{-0.15} = 1/\sqrt{2}$ of their true distance away. A single star's distance from its spectrum is typically uncertain by 10 to 50%.

Main-sequence fitting. A star cluster does much better. Its stars are all at nearly the same distance, since the cluster is a few parsecs across and far more than that away (P6). Every star then has the same $m - M$, so on a plot of m_V against the colour $(B-V)_0$ the whole cluster is the calibrated main sequence moved up or down by one amount. First each colour is corrected for reddening and each magnitude for $A_V = 3.1\,E(B-V)$, Eq. (4.44). Then the cluster sequence is slid vertically until it lies on the main sequence of nearby stars with parallaxes. The slide is the distance modulus. Each of N stars gives its own estimate with spread σ_M, and their average has the spread $\sigma_M/\sqrt{N}$, the same $\sqrt{N}$ as in the random walk (P5). For 100 stars and $\sigma_M = 0.5$ that is 0.05 magnitudes, a 2.3% distance by Eq. (3.46). Averaging does not remove an offset shared by all the stars. The calibrating stars must have the same metal content as the cluster, and the fit uses only the lower main sequence, below the turn-off where the most massive stars of the cluster have already left it. The same idea, one distance for a whole group, made the next rung possible.

3.7.3 Standard Candles, Cepheids, RR Lyrae and Miras

A **standard candle** is a kind of object whose luminosity L follows from something measured without knowing the distance, such as a period or the shape of a light curve. The measured flux F then gives the distance through Eq. (4.2),

$$F = \frac{L}{4\pi d^2} \quad \Longrightarrow \quad d = \sqrt{\frac{L}{4\pi F}}, \tag{3.47}$$

or, in magnitudes, Eq. (3.44). A good standard candle needs four things. Its luminosity must follow from a property read off the light alone. The relation must be tight. It must be bright, to be seen far away. And some members must be close enough to be calibrated on a lower rung.

Cepheids. Classical Cepheids are young, metal-rich (rich in elements other than hydrogen and helium, which astronomers call metals) stars of the disk, **Population I** like the Sun, as opposed to the old, metal-poor (with fewer such elements) **Population II** stars of the halo and the globular clusters. They have 4 to 10 solar masses and have left the main sequence (the long stage of burning hydrogen in the core) to become yellow giants and supergiants, swollen to tens or hundreds of times the radius of the Sun. Their outer layers swell and shrink in a regular radial pulsation (the whole star breathes in and out and stays round), with periods from about a day to about 50 days. The visual brightness changes by a few tenths of a magnitude up to about 2 magnitudes. The prototype, δ Cephei, repeats every $P = 5.37$ days. The pulsation is driven by a layer of partly ionised helium whose opacity rises when it is squeezed, the κ mechanism worked out in Section 12.7 (Figure 12.10).

Why the period tracks the luminosity follows from P6. A star's pulsation, like any motion driven by its own gravity, takes a time of order $(G\bar{\rho})^{-1/2}$, where $\bar{\rho}$ is the mean density (Section 9.6). Cepheids lie in a narrow range of surface temperature, the instability strip of the HR diagram (the band in which the helium layer of Section 12.7 can drive a pulsation). At a fixed temperature

a more luminous star has a larger radius, by Eq. (4.26), and so a lower mean density and a longer period. Brighter Cepheids pulse more slowly. Section 12.7 works this out into the scaling of Eq. (12.57).

The period–luminosity law. Henrietta Leavitt found the law in 1912 among the variable stars of the Small Magellanic Cloud. The cloud is far away compared with its own depth, so all its stars are at nearly the same distance (P6), the same reasoning as for a cluster. Differences of apparent magnitude among them are then differences of absolute magnitude, and Leavitt saw that the brighter variables have the longer periods, with the magnitude falling on a straight line in $\log_{10} P$. To turn the line into absolute magnitudes, one distance was needed. Today it comes from the parallaxes of nearby Cepheids in our Galaxy. One calibration in the visual band, from Hubble Space Telescope parallaxes (Benedict et al. 2007), is approximately

$$\boxed{M_V \approx -2.43 \left(\log_{10} \frac{P}{1\,\text{day}} - 1 \right) - 4.05.} \tag{3.48}$$

A Cepheid of 10 days has $M_V = -4.05$. Each factor of ten in period is 2.43 magnitudes, or a factor $10^{0.4 \times 2.43} = 9.4$ in luminosity. For δ Cephei, $\log_{10} 5.37 = 0.730$ and $M_V = -3.39$. At a given period the stars still spread by a few tenths of a magnitude, because the instability strip has a width. Dust adds its own error, so modern work uses the near infrared (wavelengths of about 1 to 2.5 μm, just beyond red light), where both are smaller. Population II Cepheids, the W Virginis stars, follow a separate law about 1.5 magnitudes fainter. Mixing the two kinds once made every distance to a galaxy too small by about a factor of two, until Walter Baade separated them in 1952.

Watching a Cepheid. The first step is to find the period from the data. Figure 3.3 follows one Cepheid of the Large Magellanic Cloud, OGLE-LMC-CEP-1252, which the OGLE survey (the Optical Gravitational Lensing Experiment, a long-running watch of the Magellanic Clouds and the centre of the Galaxy) observed 783 times in the I band over 11.2 years (Soszyński et al. 2008). Panel (a) shows the points of the first three seasons. They look like noise, but they are not. The period is about ten days, so the panel packs more than a hundred cycles side by side, and the star was seen only on scattered nights, so no single cycle can be followed by eye. The gaps are the seasons in which the Magellanic Clouds are too close to the Sun to observe. In the first part of the first season, shaded, from about day 455 to day 520, the star was seen almost every night, and there the cycle is plain. The star brightens quickly, fades slowly, and comes back to its maximum about every ten days.

Cutting and stacking. To test a trial period, cut the record into pieces one trial period long and lay the pieces on top of one another. If the cut length is the period, every piece holds the same part of the cycle, so the pieces repeat one another and the stack is one clean curve. Panel (b) cuts this nightly part at the dashed lines every 9.9935 days, and the inset lays the blue piece over the red one. They match. Panel (c) cuts the same data every 8.0 days. Now each piece starts at a different part of the cycle, the pieces do not repeat one another, and the two in the inset do not match. Stacking every piece of a wrong cut gives a scatter of points, not a curve. Written as a formula, the stack puts each point at its place in the cycle,

$$\text{phase} = \text{frac}\left(\frac{t - t_0}{P} \right), \tag{3.49}$$

where t_0 is a time of maximum light and frac keeps the part after the decimal point. This is called folding the light curve.

Why a long record helps. A cut that is only a little wrong shifts each piece by only a little, but the shift adds up from piece to piece. The whole series spans about 400 cycles. If the trial period is only 0.14% too short, 9.98 days, each year's pieces land about a twentieth of a cycle later than the year before. Over the eleven years the copies of the curve slide half a cycle apart and blur into a broad band, panel (e), coloured from the first year (blue) to the eleventh (red). At the right period, points taken years apart fall on one curve, panel (f). So a long record measures the period far better than any short part of it. A period wrong by δP puts cycle number N late by $N\,\delta P$, and a drift of about a twentieth of a cycle, half a day, already blurs the stack. Spread over 400 cycles, that pins the period to about $0.5/400 \approx 0.001$ days, against a few tenths of a day from one season by eye. To find the right period without guessing, we try many trial periods P and ask how well a single sine wave of that period fits the points. This is the Lomb–Scargle periodogram, which works for unevenly spaced times, and panel (d) shows the quality of the fit against P. It has a sharp peak at $P = 9.9935$ days, which agrees with the catalogue value 9.9933 days to 0.002 %. The small peaks beside it are echoes of the yearly gaps, and the one near 1.1 days is an echo of observing only at night, once per day. In panel (f) the rise is steep and the fall is slow, which is the usual shape of a Cepheid. The brightness changes by 0.47 magnitudes from the faintest to the brightest. The mean magnitude is found by averaging the *flux*, $\langle 10^{-0.4m} \rangle$, and turning the average back into a magnitude, since flux is what adds. It is $I = 13.72$, and $V = 14.55$ from the V series.

What the period–luminosity figure shows. Figure 3.4 plots the mean V and I magnitudes of the 1849 fundamental-mode Cepheids (those that pulsate in the simplest way, the whole star moving in and out together) of the same catalogue against $\log_{10} P$. All of them lie in the Large Magellanic Cloud, which is far away compared with its own depth, so they are at nearly the same distance (P6). The apparent magnitudes are then the absolute magnitudes plus one shared number, and they fall on a straight line in $\log_{10} P$, just as Leavitt found. The lines are least-squares fits $m = a + b \log_{10} P$, with points more than 3 standard deviations off dropped and the fit repeated until none are left. They give

$$\boxed{b_V = -2.76, \quad \sigma_V = 0.22, \qquad b_I = -2.96, \quad \sigma_I = 0.15,} \tag{3.50}$$

in magnitudes per decade of period, from 1747 stars in V after the clip, and about the same number in I. The scatter σ is the spread of the points about the line. It is smaller in I. Dust dims blue light more than red, and it is not the same in front of every star, so it adds scatter to V. The instability strip also has a width in temperature, and a star on its hot edge is brighter in V than one on its cool edge at the same period. That matters less in the infrared, where the light is less sensitive to temperature. The infrared line is also steeper. This is why the infrared is used for distances.

The slope $b_V = -2.76$ of this sample is not the -2.43 of Eq. (3.48). Slopes of different samples differ, with the Galaxy and the Large Magellanic Cloud having different metal content (different shares of elements heavier than helium) and different ranges of period. The distance below uses only the height of the lines at $\log_{10} P = 1$, where the two slopes matter least.

How a Cepheid gives a distance. Three steps turn the light curve into a distance. First, watch the star and find its period (Figure 3.3). Second, the period gives the absolute magnitude through Eq. (3.48), which was calibrated on the nearby Cepheids with parallaxes. Third, the mean apparent magnitude minus the absolute magnitude, corrected for dust, is the distance modulus, and the distance follows from Eq. (3.44). For the Large Magellanic Cloud we use the height of the fitted V line at $\log_{10} P = 1$, which Eq. (3.48) says belongs to $M_V = -4.05$, and the dust correction $A_V = 3.1\,E(B-V)$ of Eq. (4.44) with $E(B-V) = 0.08$, a typical value for the

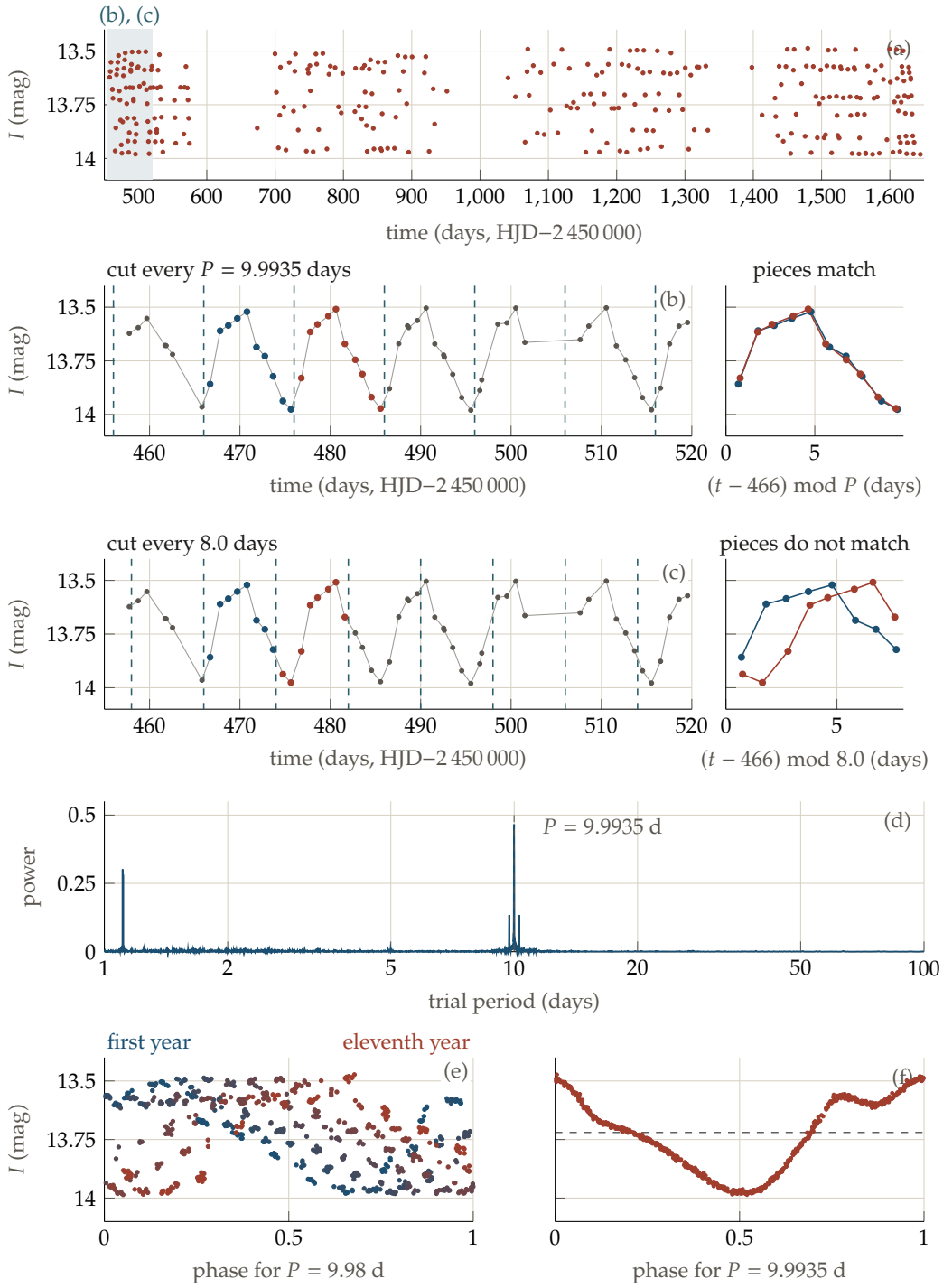


Figure 3.3: The *I*-band light curve of the Cepheid OGLE-LMC-CEP-1252 (OGLE-III). (a) The observed magnitudes over the first three seasons, with the first part, observed nearly every night, shaded. (b) That part with dashed cuts every 9.9935 days, and in the inset the blue and the red piece laid over one another. (c) The same part cut every 8.0 days, with two pieces laid over one another in the inset. (d) The Lomb–Scargle power against trial period, with the period of the highest peak marked by a dashed line. (e) All points folded at a period of 9.98 days, 0.14% too short, coloured from the first year (blue) to the eleventh (red). (f) All points folded at the period of the peak, with the mean magnitude as a dashed line. Brighter is up.

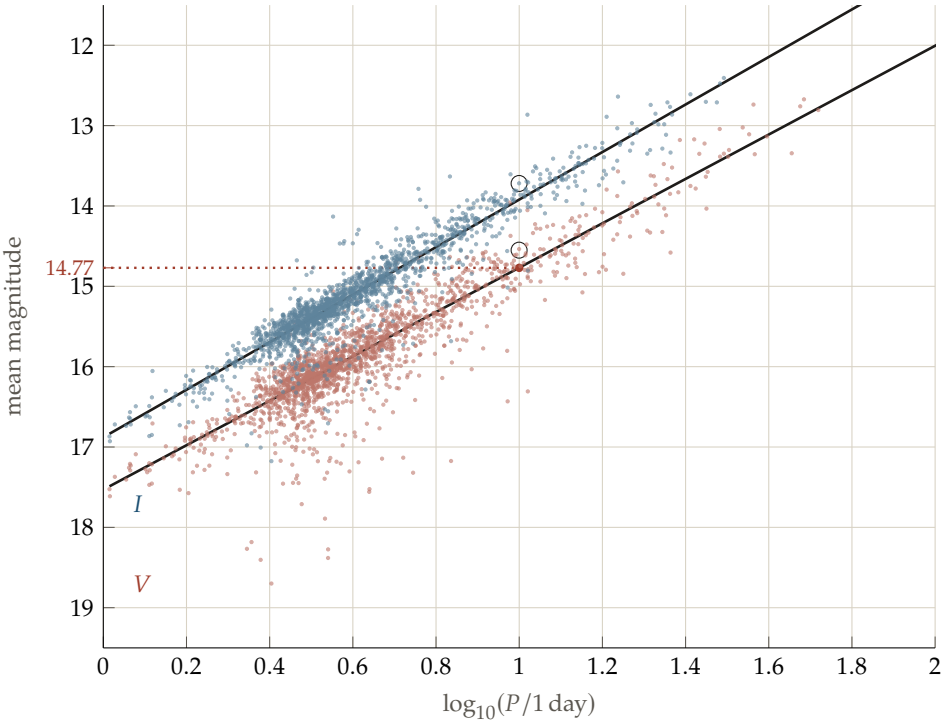


Figure 3.4: Mean V and I magnitudes of the fundamental-mode classical Cepheids of the Large Magellanic Cloud against the logarithm of the period (OGLE-III catalogue). V is red and I is blue. The black lines are the least-squares fits after a 3σ clip. The rings mark the star of Figure 3.3. The dotted line marks the height of the V fit at $\log_{10} P = 1$. Brighter is up.

cloud.

$$\begin{aligned}(m-M)_0 &\overset{(1)}{=} m_V(\log_{10} P = 1) - M_V(\log_{10} P = 1) - A_V \\ &\overset{(2)}{=} 14.77 - (-4.05) - 3.1 \times 0.08 \\ &\overset{(3)}{=} 18.82 - 0.25 = 18.57, \\ d &\overset{(4)}{=} 10\,\mathrm{pc} \times 10^{18.57/5} = 51.8\,\mathrm{kpc}. \end{aligned} \tag{3.51}$$

(1) The corrected distance modulus, Eq. (4.44), with the value of each term at a period of 10 days.
(2) The fitted V line gives $m_V = 14.77$ at $\log_{10} P = 1$ (the dotted line in Figure 3.4), and Eq. (3.48) gives $M_V = -4.05$. Dust gives $A_V = 3.1\,E(B-V)$.
(3) Arithmetic. The uncorrected modulus is 18.82 and the dust takes 0.25.
(4) The distance modulus inverted, Eq. (3.44).

The accepted distance to the Large Magellanic Cloud is about 50 kpc, a modulus of about 18.48 (quoted from memory). Our 51.8 kpc is close, and the difference is of the size of the uncertainty in the calibration of Eq. (3.48) and in the dust. The distance does not depend on which Cepheid we pick, since each one gives the same modulus once its period has fixed its absolute magnitude.

Worked example 3.6: A Cepheid in a distant galaxy

A Cepheid has $P = 30$ days and mean apparent magnitude $m_V = 25.0$. We ignore dust.

$$\begin{aligned} M_V &\overset{(1)}{=} -2.43\,(\log_{10} 30 - 1) - 4.05 = -2.43 \times 0.4771 - 4.05 = -5.21, \\ m_V - M_V &= 25.0 + 5.21 = 30.21, \\ d &\overset{(2)}{=} 10\,\text{pc} \times 10^{30.21/5} = 10^{7.042}\,\text{pc} = 1.10 \times 10^7\,\text{pc} = 11.0\,\text{Mpc}. \end{aligned}$$

(1) The period–luminosity law, Eq. (3.48).
(2) Eq. (3.44).

A spread of 0.2 magnitudes in M_V is a 9% spread in d by Eq. (3.46). Ten Cepheids in the same galaxy bring it to 3%.

RR Lyrae stars and Miras. RR Lyrae stars are Population II stars on the horizontal branch, old stars of under a solar mass that burn helium in their cores. The branch is named for where these stars sit on the HR diagram, a nearly level line at about the same luminosity. They pulsate by the same mechanism as Cepheids, with periods under a day, and all have about the same luminosity, $M_V \approx +0.6$. There is no period to fit, only the mean magnitude. They are 4.65 magnitudes, or $10^{0.4\times4.65} = 72$ times, fainter than a 10-day Cepheid, so by Eq. (3.47) they are seen $\sqrt{72} = 8.5$ times less far. At $m_V = 25$ they give $d = 10\,\text{pc} \times 10^{(25-0.6)/5} = 760\,\text{kpc}$, which spans the Local Group. They are found where Cepheids are not, in globular clusters and the halo. **Mira variables** are cool red giants near the end of their lives. They pulsate with periods of about 100 to 1000 days and large amplitudes in visible light. Most of their light comes out in the infrared, and there they also follow a tight period–luminosity law.

3.7.4 The Upper Rungs and a Rung That Needs No Ladder

Type Ia supernovae. A type Ia supernova is a white dwarf that explodes, and why each explosion is nearly alike is explained in Chapter 13. Here only its use matters. At its peak it reaches $M_B \approx -19.3$, with a spread of about 0.15 magnitudes once each peak is corrected by the width of its light curve. No type Ia supernova is near enough for a parallax. The peak is calibrated in galaxies that host both a supernova and Cepheids, so the supernova rung stands on the Cepheid rung. Its gain is its reach. A telescope detects any source down to some limiting flux F_{lim}. By Eq. (3.47) the farthest distance at which a candle is seen is

$$d_{\text{max}} = \sqrt{\frac{L}{4\pi F_{\text{lim}}}} \quad \Longrightarrow \quad \frac{d_{\text{max,SN}}}{d_{\text{max,Cep}}} \overset{(1)}{=} \sqrt{\frac{L_{\text{SN}}}{L_{\text{Cep}}}} \overset{(2)}{=} 10^{(M_{\text{Cep}} - M_{\text{SN}})/5}. \tag{3.52}$$

(1) The same F_{lim} for both, so it cancels.
(2) A difference of absolute magnitudes is a ratio of luminosities, $L_{\text{SN}}/L_{\text{Cep}} = 10^{0.4(M_{\text{Cep}} - M_{\text{SN}})}$, by Eq. (4.5), and the square root halves the exponent.

Against a 10-day Cepheid, $M_V = -4.05$, the gap is 15.25 magnitudes. That is $10^{6.1} = 1.3 \times 10^6$ in luminosity and $10^{3.05} = 1100$ in distance. Against the brightest Cepheids, of 50 days, Eq. (3.48) gives $M_V = -5.75$. The gap is then 13.55 magnitudes, 2.6×10^5 in luminosity and 510 in distance. These compare a B peak with V magnitudes, which changes the gap by a few tenths at most. A type Ia supernova is 10^5 to 10^6 times brighter than a Cepheid and is seen 500 to 1000 times farther. Cepheids reach about 30 Mpc, and the supernovae reach thousands of megaparsecs, far enough to show that the expansion of the universe is speeding up (Section 16.7).

The Tully–Fisher relation. The rotation speed v of a spiral galaxy fixes its luminosity, $L \propto v^4$, Eq. (14.44). The speed comes from the width of the 21 cm line, which needs no distance. The constant in front is fixed with spirals whose distances are known from their Cepheids, and is checked against galaxies that have hosted type Ia supernovae. No rare event is needed, so the relation gives distances to many thousands of spirals, out to a few hundred megaparsecs.

Hubble's law. Far enough out, a galaxy recedes at a speed proportional to its distance, $v = H_0 d$, Eq. (16.2). Once H_0 is known, a redshift alone gives the distance. The ladder measures H_0 itself. It finds d from type Ia supernovae in galaxies far enough away that their own motions, a few hundred km/s, are small next to the expansion speed. At 50 Mpc the expansion speed is already about 3500 km/s.

Standard sirens. A pair of neutron stars or black holes spiralling together sends out gravitational waves (Chapter 13). The frequency and how fast it rises fix the masses, and the masses fix how strong the waves are at the source. The measured strength falls off as $1/d$, so the two together give the distance. Only general relativity enters, with no calibration on any lower rung. Such a merger is called a **standard siren**. The redshift must still come from the host galaxy, and together they give H_0 with no ladder at all.

3.7.5 How Errors Climb the Ladder

Each rung multiplies the distance scale of the rung below by a calibration factor. A distance at the top is therefore a product $d_{\text{top}} = d_{\text{bottom}}\, f_1 f_2 \cdots f_k$, where d_{bottom} is the bottom distance and each factor f_i carries a small fractional error ε_i with $\langle \varepsilon_i \rangle = 0$ and spread $\langle \varepsilon_i^2 \rangle = \sigma_i^2$. Taking the logarithm turns the product into a sum, and to first order $\ln f_i$ changes by ε_i. The squared error of the top distance is

$$\left\langle \left(\frac{\delta d_{\text{top}}}{d_{\text{top}}} \right)^2 \right\rangle \overset{(1)}{=} \left\langle \Big(\sum_{i=1}^{k} \varepsilon_i \Big)^2 \right\rangle = \sum_{i=1}^{k} \langle \varepsilon_i^2 \rangle + \sum_{i \neq j} \langle \varepsilon_i \varepsilon_j \rangle$$
$$\overset{(2)}{=} \sum_{i=1}^{k} \sigma_i^2 . \tag{3.53}$$

(1) $\ln d_{\text{top}} = \ln d_{\text{bottom}} + \sum_i \ln f_i$, and $\delta(\ln d) = \delta d / d$.
(2) Independent errors, so $\langle \varepsilon_i \varepsilon_j \rangle = \langle \varepsilon_i \rangle \langle \varepsilon_j \rangle = 0$ for $i \neq j$.

This is the random walk of P5. For example, three rungs with 2% each give $\sqrt{3} \times 2\% = 3.5\%$, not 6%. The cross terms vanish only for independent errors. A mistake shared by all the rungs, such as a wrong correction for dust or for metal content used on every Cepheid, does not average out, and the errors then add in a straight line. Such shared errors are what the ladder must guard against.

The Hubble tension. The ladder of parallaxes, Cepheids and type Ia supernovae gives $H_0 \approx 73.0 \pm 1.0\,\text{km s}^{-1}\,\text{Mpc}^{-1}$. The cosmic microwave background, read with the standard model of the expanding universe, gives $67.4 \pm 0.5\,\text{km s}^{-1}\,\text{Mpc}^{-1}$ (Chapter 16). The two differ by 5.6, which is 8.3%, and by about five times their combined error, $\sqrt{1.0^2 + 0.5^2} = 1.1$. At a fixed redshift Eq. (16.2) gives $d = v/H_0$, so the ladder's distances would have to be 8.3% too small to explain it. In magnitudes that is

$$5 \log_{10} \frac{73.0}{67.4} = 0.17, \tag{3.54}$$

an offset of 0.17 magnitudes shared by every distance, for instance in the zero point of Eq. (3.48). Independent errors averaged over many stars are far smaller than this, so the question is whether some shared error has been missed or whether the standard model of the expanding universe is incomplete. That is why each rung is checked by a second route, with *Gaia* parallaxes of Cepheids, other candles in the same galaxies, and standard sirens that skip the ladder entirely.

3.8 Problem Set: Stellar Astrometry, Kinematics, and Space Velocities

3.8.1 Problem 1: Position, Space Velocity and Doppler Shift of Regulus

Problem Statement: Regulus (α Leonis), the brightest star in Leo, lies at $\alpha = 152.093°$, $\delta = +11.967°$. Its measured parallax is $p = 41.13\,\mathrm{mas}$, its radial velocity $v_r = 4.39\,\mathrm{km/s}$, and its proper motions are $\mu_\alpha^* = -248.73\,\mathrm{mas/yr}$ and $\mu_\delta = +5.59\,\mathrm{mas/yr}$.

(a) What is its distance d in parsecs?

(b) Find its position vector $\vec{r}$ in equatorial components.

(c) Find its velocity $\vec{v}$ relative to the Sun, in equatorial components, and its speed $|\vec{v}|$.

(d) At what wavelength is its Hα line observed ($\lambda_0 = 656.46\,\mathrm{nm}$ in vacuum)?

(e) Find its Galactic velocity (U, V, W) and its velocity relative to the local standard of rest.

Essential Ideas:

- Parallax gives the distance, $d\,[\mathrm{pc}] = 1/p\,['']$, Eq. (3.5).
- Proper motion gives the velocity across the line of sight, $v_\alpha = 4.74047\,\mu_\alpha^* d$ and $v_\delta = 4.74047\,\mu_\delta d$ in km/s, Eq. (3.23). The number 4.74047 is 1 AU per Julian year in km/s.
- The local components turn into equatorial ones with $\mathbf{Q}^T(\alpha, \delta)$, Eq. (3.30), and into Galactic ones with $\mathbf{T}$ after that, Eq. (3.34).
- For $v_r \ll c$ the Doppler shift is $(\lambda_{\mathrm{obs}} - \lambda_0)/\lambda_0 = v_r/c$, Eq. (3.25).

Detailed Solutions:

(a) **Distance**. With $p = 41.13\,\mathrm{mas} = 0.04113''$,

$$\boxed{d = \frac{1}{0.04113}\,\mathrm{pc} = 24.31\,\mathrm{pc}.}$$

(b) **Position vector**. With $\cos\delta = 0.97827$, $\sin\delta = 0.20735$, $\cos\alpha = -0.88371$ and $\sin\alpha = 0.46804$, Eq. (3.9) gives

$$\boxed{\vec{r} = 24.31\,\mathrm{pc}\begin{pmatrix} 0.97827 \times (-0.88371) \\ 0.97827 \times 0.46804 \\ 0.20735 \end{pmatrix} = \begin{pmatrix} -21.02 \\ 11.13 \\ 5.04 \end{pmatrix}\mathrm{pc}.}$$

(c) **Velocity and speed**. The proper motions give the local components

$$v_\alpha = 4.74047 \times (-0.24873) \times 24.31\,\mathrm{km/s} = -28.67\,\mathrm{km/s},$$
$$v_\delta = 4.74047 \times 0.00559 \times 24.31\,\mathrm{km/s} = +0.64\,\mathrm{km/s}.$$

Eq. (3.30) turns them into equatorial components,

$$\vec{v} = \begin{pmatrix} -0.86450 & -0.46804 & 0.18324 \\ 0.45787 & -0.88371 & -0.09705 \\ 0.20735 & 0 & 0.97827 \end{pmatrix}\begin{pmatrix} 4.39 \\ -28.67 \\ 0.64 \end{pmatrix}\mathrm{km/s} = \begin{pmatrix} 9.74 \\ 27.28 \\ 1.54 \end{pmatrix}\mathrm{km/s}.$$

By Eq. (3.32) the speed can be found from either set of components,

$$\boxed{|\vec{v}| = \sqrt{4.39^2 + 28.67^2 + 0.64^2}\,\text{km/s} = 29.01\,\text{km/s}.}$$

(d) **Observed wavelength of** Hα. The star recedes at $v_r = 4.39\,\text{km/s}$, so by Eq. (3.25)

$$\lambda_{\text{obs}} - \lambda_0 = \lambda_0 \frac{v_r}{c} = 656.46\,\text{nm} \times \frac{4.39}{299\,792.458} = 0.0096\,\text{nm},$$

$$\boxed{\lambda_{\text{obs}} = 656.4696\,\text{nm}.}$$

The shift is one part in 68 000 of the wavelength.

(e) **Galactic velocity**. Eq. (3.34) acting on the column of part (c) gives

$$\boxed{(U, V, W) = (-25.11,\ -6.17,\ -13.15)\,\text{km/s},}$$

with the same length, 29.01 km/s, as a rotation must keep. Adding the Sun's motion relative to the LSR,

$$\boxed{(U_{\text{LSR}}, V_{\text{LSR}}, W_{\text{LSR}}) = (-14.01,\ 6.07,\ -5.90)\,\text{km/s}.}$$

Most of the speed relative to the Sun is the U component, directed away from the Galactic centre. Relative to the LSR, Regulus moves outward from the centre at 14 km/s, a little ahead of circular rotation, and slowly toward the south side of the disk.

3.8.2 Problem 2: Closest Approach of a Fast Nearby Star

Problem Statement: A nearby red dwarf has parallax $p = 0.547''$, total proper motion $\mu = 10.39''/\text{yr}$ and radial velocity $v_r = -110\,\text{km/s}$. These are close to the values for Barnard's star. Treat its motion relative to the Sun as a straight line.

(a) Find its distance and its transverse velocity.

(b) Find its speed and the angle between its velocity and the line of sight.

(c) How close will it come to the Sun, and when?

(d) What will its proper motion be at closest approach?

Essential Ideas:

- With no force acting, momentum is conserved (P1) and the star moves on a straight line, $\vec{r}(t) = \vec{r}_0 + \vec{v}\,t$.
- The velocity splits into v_r along the line of sight and $v_t = 4.74047\,\mu/p$ across it, Eq. (3.23), and $|\vec{v}|^2 = v_r^2 + v_t^2$, Eq. (3.32).
- The closest approach is where $|\vec{r}(t)|^2$ has its minimum.

Detailed Solutions:

(a) **Distance and transverse velocity**. By Eqs. (3.5) and (3.23),

$$\boxed{d = \frac{1}{0.547}\,\text{pc} = 1.828\,\text{pc}, \qquad v_t = 4.74047 \times \frac{10.39}{0.547}\,\text{km/s} = 90.0\,\text{km/s}.}$$

(b) **Speed and direction**. By Eq. (3.32),

$$|\vec{v}| = \sqrt{110^2 + 90.0^2}\,\text{km/s} = 142.2\,\text{km/s}.$$

The angle ψ between $\vec{v}$ and $\hat{\mathbf{r}}$ has $\cos\psi = v_r/|\vec{v}| = -0.774$,

$$\boxed{|\vec{v}| = 142.2\,\text{km/s}, \qquad \psi = 140.7^\circ.}$$

The star is coming toward us, at 39.3° from the line that joins it to the Sun.

(c) **Closest approach**. Write $\vec{r}_0 = d\,\hat{\mathbf{r}}$ now. The squared distance and its minimum are

$$|\vec{r}(t)|^2 \overset{(1)}{=} d^2 + 2d\,(\hat{\mathbf{r}}\cdot\vec{v})\,t + |\vec{v}|^2 t^2 \overset{(2)}{=} d^2 + 2d\,v_r\,t + |\vec{v}|^2 t^2,$$

$$t_{\min} \overset{(3)}{=} -\frac{d\,v_r}{|\vec{v}|^2}, \qquad d^2_{\min} \overset{(4)}{=} d^2 - \frac{d^2 v_r^2}{|\vec{v}|^2} = d^2\,\frac{v_t^2}{|\vec{v}|^2}.$$

(1) $|\vec{r}_0 + \vec{v}\,t|^2$ multiplied out.
(2) $\hat{\mathbf{r}}\cdot\vec{v} = v_r$, Eq. (3.20).
(3) The derivative with respect to t set to zero.
(4) $t_{\min}$ put back, then $|\vec{v}|^2 - v_r^2 = v_t^2$.

So $d_{\min} = d\,v_t/|\vec{v}| = 1.828 \times 90.0/142.2$ pc. With $d = 1.828 \times 3.0857 \times 10^{13}$ km $= 5.64 \times 10^{13}$ km, $t_{\min} = 5.64 \times 10^{13} \times 110/142.2^2$ s $= 3.07 \times 10^{11}$ s,

$$\boxed{d_{\min} = 1.16\,\text{pc}, \qquad t_{\min} = 9.7 \times 10^3 \text{ years from now.}}$$

(d) **Proper motion at closest approach**. There the whole velocity is across the line of sight, so $v_t = |\vec{v}|$ and, by Eq. (3.23),

$$\boxed{\mu = \frac{|\vec{v}|}{4.74047\,d_{\min}} = \frac{142.2}{4.74047 \times 1.158}\,''/\text{yr} = 25.9''/\text{yr}.}$$

A second route uses the angular momentum about the Sun. With no force it is conserved (P1), and per unit mass it is $|\vec{r}\times\vec{v}| = d\,v_t = \mu\,d^2$ by Eq. (3.21). So $\mu\,d^2$ is the same now and at closest approach, and $\mu = 10.39\times(1.828/1.158)^2\,''/\text{yr} = 25.9''/\text{yr}$ again. The same law gives the present rate of growth. Differentiating $\mu\,d^2 = \text{const}$ gives $\dot{\mu} = -2\mu\,\dot{d}/d = -2\mu\,v_r/d$, which is $0.0013''/\text{yr}$ per year. The velocity is constant, yet the proper motion grows, because the distance shrinks. This growth is the **secular acceleration**, and for Barnard's star it shifts the position by about half a second of arc in thirty years.

3.8.3 Problem 3: A Cepheid Distance Through Dust, and a Supernova Calibrated by It

Problem Statement: A Cepheid in a nearby spiral galaxy has period $P = 20.0$ days and mean apparent magnitude $m_V = 24.00$. The dust in front of it gives the colour excess $E(B-V) = 0.10$. The same galaxy later hosts a type Ia supernova, seen through the same dust, whose peak apparent magnitude is $m_V = 9.50$.

(a) What is the absolute magnitude of the Cepheid?

(b) What distance follows if the dust is ignored?

(c) What is the distance with the dust taken into account?

(d) What is the peak absolute magnitude of the supernova?

Essential Ideas:

- The period gives the absolute magnitude through the period–luminosity law, Eq. (3.48).
- Dust dims the star by $A_V = 3.1\,E(B-V)$, and the corrected distance modulus is $(m_V - M_V)_0 = m_V - M_V - A_V$, Eq. (4.44). Then $d = 10\,\text{pc} \times 10^{(m_V - M_V)_0/5}$, Eq. (3.44).

- The galaxy is far away compared with its own size, so the Cepheid and the supernova have the same distance modulus (P6).

Detailed Solutions:

(a) **Absolute magnitude**. With $\log_{10} 20.0 = 1.3010$, Eq. (3.48) gives

$$\boxed{M_V = -2.43 \times 0.3010 - 4.05 = -4.78.}$$

(b) **Distance without dust**. The distance modulus is $m_V - M_V = 24.00 + 4.78 = 28.78$, so

$$\boxed{d = 10\,\mathrm{pc} \times 10^{28.78/5} = 10^{6.756}\,\mathrm{pc} = 5.70\,\mathrm{Mpc}.}$$

(c) **Distance with dust**. The dust takes $A_V = 3.1 \times 0.10 = 0.31$ magnitudes, so $(m_V - M_V)_0 = 28.78 - 0.31 = 28.47$ and

$$\boxed{d = 10\,\mathrm{pc} \times 10^{28.47/5} = 10^{6.694}\,\mathrm{pc} = 4.94\,\mathrm{Mpc}.}$$

Ignoring the dust puts the galaxy $10^{0.31/5} = 1.15$ times too far away.

(d) **Peak of the supernova**. The supernova has the same corrected modulus and the same dimming,

$$\boxed{M_V = m_V - A_V - (m_V - M_V)_0 = 9.50 - 0.31 - 28.47 = -19.28.}$$

This is how the peak of type Ia supernovae is calibrated. Any error in the Cepheid distance passes unchanged into every supernova distance built on it.

Chapter 4

Stellar Photometry: Flux, Magnitudes, and Radii

Parallax gives a star's distance, but does not tell us what the star itself is like. Photometry measures the light that reaches us. Once we know the distance, the received light gives the star's total power. The way that light is divided among colours gives its surface temperature; power and temperature together then reveal a radius even when the star looks like a point.

On its way to us the light spreads out with distance, and dust in space dims it and makes it redder. Each effect is undone by one formula, and astronomers record the result on a logarithmic scale of magnitudes.

4.1 Flux and the Inverse-Square Law

Chapter 3 gave the distance d of a nearby star from its parallax p, with $d\,[\mathrm{pc}] = 1/p\,['']$ (Eq. (3.5)). A distance alone says nothing about the star itself. A star can look bright because it is powerful or because it is close. To separate the two we need a rule for how light thins out with distance.

Stand under a street lamp at night with a book, then walk away until you are twice as far. The page is much dimmer. The lamp sends out the same light every second as before. At twice the distance that light is spread over a sphere with four times the area, so the page gets a quarter of it. A star is such a lamp, and the Earth holds the page. The power of a star is shared out over a sphere centred on it, so the light per square metre falls as one over the square of the distance.

Suppose the lamp radiates 100 W equally in all directions. At 2 m its light crosses a sphere of area $4\pi(2\,\mathrm{m})^2 = 50.3\,\mathrm{m}^2$, giving $100/50.3 = 1.99\,\mathrm{W\,m^{-2}}$. At 4 m the area is $201\,\mathrm{m}^2$ and the flux is $0.497\,\mathrm{W\,m^{-2}}$, a quarter as much. The total power has not changed; it is spread over four times the area. We will use that conservation of energy to connect a star's luminosity, its measured flux and its distance.

Luminosity and flux. The **luminosity** L of a star is the total power it sends out, over all wavelengths, in watts. The **flux** F at a point is the power that crosses one square metre facing the star, in $\mathrm{W\,m^{-2}}$. A flux summed over all wavelengths is called **bolometric**.

Spreading over a sphere. To find the flux at distance d, we draw a sphere of radius d about the star and count the power that crosses it (P1). The star is steady, and nothing between it and

the sphere absorbs light.

$$\begin{aligned} L &\overset{(1)}{=} \oint F \, dA \\ &\overset{(2)}{=} F(d) \oint dA \\ &\overset{(3)}{=} 4\pi d^2 \, F(d). \end{aligned} \tag{4.1}$$

(1) In a steady state the energy that leaves the star each second crosses the sphere in the same second, because none is stored or absorbed on the way (P1).
(2) The star sends out light equally in all directions, so F is the same at every point of the sphere.
(3) The area of a sphere of radius d.

Result 4.1

$$F = \frac{L}{4\pi d^2} \quad \left[\mathrm{W\,m^{-2}}\right]. \tag{4.2}$$

The flux depends on the star and its distance only through L/d^2. A measured flux and a parallax distance therefore give the luminosity, $L = 4\pi d^2 F$.

The flux at the surface. The sphere can be drawn anywhere outside the star, including at its surface, at radius R. There the flux is $F_{\mathrm{surf}} = L/(4\pi R^2)$. Dividing Eq. (4.2) by it,

$$\boxed{F(d) = F_{\mathrm{surf}} \left(\frac{R}{d}\right)^2}. \tag{4.3}$$

The light thins out by the ratio of the two areas. This form returns in Section 4.5, where it gives the angular size of a star from its flux.

Worked example 4.1: The solar constant

The Sun has $L_\odot = 3.828 \times 10^{26}$ W. The Earth is at 1 AU $= 1.495\,978\,707 \times 10^{11}$ m from it, and Eq. (4.2) gives

$$F_\odot = \frac{3.828 \times 10^{26}\ \mathrm{W}}{4\pi \left(1.495\,978\,707 \times 10^{11}\ \mathrm{m}\right)^2} = 1361\ \mathrm{W\,m^{-2}}.$$

This is the **solar constant**, the bolometric flux of the Sun above the Earth's air. Satellites measure it, and the Sun's luminosity is found by running the same formula backwards.

4.2 The Magnitude Scale

We can now give the light of a star as a flux in $\mathrm{W\,m^{-2}}$. Fluxes of stars cover an enormous range. The Sun gives $1361\ \mathrm{W\,m^{-2}}$, and the faintest stars the eye can see give about 10^{-13} of that. A flux is also hard to measure in watts. A detector and the air above it change from night to night, so what an astronomer measures well is the ratio of a star's flux to that of a standard star seen through the same instrument.

Our ears face the same problem. They judge loudness roughly by ratios, not by differences. Going from one violin to ten sounds like one step, and going from ten to a hundred sounds like another step of the same size. Sound engineers therefore count in decibels, ten times the

logarithm of a ratio of powers. Ten decibels is a factor of 10 in power, and twenty decibels a factor of 100, whatever the level we start from. The eye also judges brightness roughly by ratios (the Weber–Fechner law). In the 2nd century BCE Hipparchus sorted the stars seen by the naked eye into six classes, from first magnitude for the brightest to sixth for the faintest. In 1856 Norman Pogson noticed that a first-magnitude star gives about 100 times the light of a sixth-magnitude one, and made this exact. A magnitude is a logarithm of a ratio of two fluxes, with five magnitudes set equal to a factor of 100.

Pogson's rule makes a difference of 5 magnitudes equal to a factor of 100 in flux. One magnitude therefore means a factor of $100^{1/5} = 2.512$, and 2.5 magnitudes a factor of 10. Brighter stars have smaller magnitude numbers, retaining Hipparchus's original ranking. A magnitude is always a comparison with a reference flux. Once the reference is fixed, differences of magnitudes can be used to compare stars at the same distance or to infer a distance from a known luminosity.

4.2.1 Apparent Magnitude and Pogson's Formula

Fixing the constant. A magnitude falls as the flux rises and changes by equal steps for equal ratios of flux. So it is a logarithm of the flux with a negative factor, $m = -k \log_{10} F + \text{constant}$. For two stars the constant cancels,

$$m_1 - m_2 = -k \log_{10} \frac{F_1}{F_2}.$$

Pogson's rule sets $m_1 - m_2 = 5$ when star 1 is fainter by a factor of 100,

$$\begin{aligned} 5 &\overset{(1)}{=} -k \log_{10} \frac{1}{100} \\ &= 2k, \end{aligned} \tag{4.4}$$

(1) Pogson's rule with $F_1/F_2 = 1/100$.

so $k = 5/2$. The **apparent magnitudes** of two stars with fluxes F_1 and F_2 then differ by

Result 4.2

$$m_1 - m_2 = -2.5 \log_{10} \frac{F_1}{F_2}. \tag{4.5}$$

Turned round, a difference of magnitudes gives a ratio of fluxes,

$$\frac{F_1}{F_2} = 10^{-0.4(m_1 - m_2)}, \tag{4.6}$$

and one magnitude is a factor of $10^{0.4} = 2.512$.

The zero point. One reference of known flux fixes the whole scale. Let star 2 be a reference that has magnitude zero, with flux F_{ref}. Eq. (4.5) becomes

$$m = -2.5 \log_{10} \frac{F}{F_{\text{ref}}} = -2.5 \log_{10} F + 2.5 \log_{10} F_{\text{ref}}. \tag{4.7}$$

The last term is a constant, different for each filter. In the visible bands F_{ref} is close to the flux of the star Vega, which has $V = 0.03$. For bolometric magnitudes the International Astronomical Union fixed $F_{\text{ref}} = 2.518 \times 10^{-8}\,\text{W m}^{-2}$ in 2015. The Sun then has the apparent bolometric magnitude

$$m_{\text{bol},\odot} = -2.5 \log_{10} \frac{1361}{2.518 \times 10^{-8}} = -26.83.$$

4.2.2 Absolute Magnitude and the Distance Modulus

An apparent magnitude mixes the star's power with its distance, just as the flux does. To compare stars we put them all at one distance. The **absolute magnitude** M of a star is the apparent magnitude it would have at a distance of 10 pc.

Absolute magnitude and luminosity. At one common distance the fluxes of two stars are in the ratio of their luminosities.

$$\begin{aligned} M_1 - M_2 &\overset{(1)}{=} -2.5 \log_{10} \frac{F_1(10\,\mathrm{pc})}{F_2(10\,\mathrm{pc})} \\ &\overset{(2)}{=} -2.5 \log_{10} \frac{L_1/\left(4\pi(10\,\mathrm{pc})^2\right)}{L_2/\left(4\pi(10\,\mathrm{pc})^2\right)} \\ &= -2.5 \log_{10} \frac{L_1}{L_2}. \end{aligned} \tag{4.8}$$

(1) Eq. (4.5) for the two stars placed at 10 pc.
(2) Eq. (4.2) for each star.

Taking the Sun as star 2,

$$\boxed{M - M_\odot = -2.5 \log_{10} \frac{L}{L_\odot}.} \tag{4.9}$$

For bolometric magnitudes, a star that gives the zero-point flux F_{ref} at 10 pc has the luminosity

$$4\pi(10\,\mathrm{pc})^2 F_{\mathrm{ref}} = 3.0128 \times 10^{28}\,\mathrm{W}.$$

Using this star as star 2 instead, the Sun has

$$M_{\mathrm{bol},\odot} = -2.5 \log_{10} \frac{3.828 \times 10^{26}}{3.0128 \times 10^{28}} = +4.74.$$

The distance modulus. Now compare one star with itself, at its true distance d and at 10 pc.

$$\begin{aligned} m - M &\overset{(1)}{=} -2.5 \log_{10} \frac{F(d)}{F(10\,\mathrm{pc})} \\ &\overset{(2)}{=} -2.5 \log_{10} \frac{L/(4\pi d^2)}{L/\left(4\pi(10\,\mathrm{pc})^2\right)} \\ &= -2.5 \log_{10} \left(\frac{10\,\mathrm{pc}}{d}\right)^2 \\ &\overset{(3)}{=} 5 \log_{10} \frac{d}{10\,\mathrm{pc}}. \end{aligned} \tag{4.10}$$

(1) Eq. (4.5), with m the magnitude at d and M the magnitude at 10 pc.
(2) Eq. (4.2) at both distances.
(3) $\log_{10} x^2 = 2 \log_{10} x$ and $\log_{10}(1/x) = -\log_{10} x$.

Result 4.3

$$m - M = 5 \log_{10} \frac{d}{10\,\mathrm{pc}} = 5 \log_{10} d\,[\mathrm{pc}] - 5. \tag{4.11}$$

The luminosity has cancelled. Without dust, $m - M$ depends on the distance alone, and every 5 magnitudes is a factor of 10 in distance. The difference $m - M$ is called the **distance modulus**. With the parallax distance of Eq. (3.5), $\log_{10} d\,[\mathrm{pc}] = -\log_{10} p\,['']$, and

$$m - M = -5 \log_{10} p\,[''] - 5. \tag{4.12}$$

A star with $p = 0.1''$ is at 10 pc and has $m = M$.

As a check, the Sun is at 1 AU $= 4.848 \times 10^{-6}$ pc, so Eq. (4.11) gives $m - M = -31.57$ for it. Then $m_{\mathrm{bol},\odot} = 4.74 - 31.57 = -26.83$, the value found above from the zero point.

Filter and bolometric magnitudes. A magnitude through a filter counts only part of the light. The difference $m_{\mathrm{bol}} - m_V$ is the **bolometric correction**. It depends on the temperature, because a hot star sends much of its light out in the ultraviolet and a cool star in the infrared, where the V filter does not look. Eqs. (4.5)–(4.12) hold in any one band and for bolometric magnitudes alike.

4.3 Light from a Hot Surface

We can now find the luminosity of a star from its flux and its distance. The luminosity does not say whether the star is small and hot or large and cool. For that we need to know how much light a hot surface gives off, and in which colours.

Watch the coil of an electric stove as it heats. At first it gives warmth but no glow. Then it glows dull red, then orange. A blacksmith's iron, hotter still, turns yellow and then almost white. Two things change together. The colour moves from red toward blue, and the heat felt on the face grows very fast. Stars do the same. In Orion, Betelgeuse is orange-red and Rigel is blue-white, because Rigel's surface is hotter. A dense hot body that absorbs all the light falling on it, a **blackbody**, gives off a spectrum fixed by its temperature alone, and the surface of a star is close to one.

At 1000 K, a stove coil emits most strongly at 2.9 μm in the infrared, leaving only the red tail visible. The Sun's 5772 K surface peaks at 502 nm and radiates 6.29×10^7 W per square metre. That is 1110 times the coil's power per unit area, even though the Sun is only 5.77 times hotter: the power grows as the fourth power of temperature, 5.772^4. We will obtain both this total power and the wavelength of the peak from the blackbody spectrum.

4.3.1 The Planck Function

The **specific intensity** of light is the power carried along a ray, per unit area, per unit solid angle of direction and per unit frequency. How it is defined in full and how it changes inside matter is the subject of Chapter 11. For a blackbody at temperature T the specific intensity is the **Planck function**, derived by counting states in Eq. (9.31),

$$B_\nu(T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{h\nu/k_BT} - 1} \quad \left[\mathrm{W\,m^{-2}\,sr^{-1}\,Hz^{-1}}\right]. \tag{4.13}$$

Here h is Planck's constant, c the speed of light and k_B Boltzmann's constant. B_ν depends on T and ν only, and it is the same in every direction. The photon energy in units of k_BT,

$$x = \frac{h\nu}{k_BT} = \frac{hc}{\lambda k_BT},$$

appears in every formula below.

Per unit wavelength. Filters are described by wavelength, so we also need the intensity per unit wavelength, B_λ. The power in a narrow band of the spectrum cannot depend on whether we label the band by wavelength or by frequency, so $B_\lambda\,|d\lambda| = B_\nu\,|d\nu|$.

$$\begin{aligned} B_\lambda &\overset{(1)}{=} B_\nu \left|\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right| \\ &\overset{(2)}{=} B_\nu \,\frac{c}{\lambda^2} \\ &\overset{(3)}{=} \frac{2hc^3}{c^2\lambda^3}\,\frac{c}{\lambda^2}\,\frac{1}{e^{hc/\lambda k_B T}-1}. \end{aligned} \tag{4.14}$$

(1) Equal power in the same band of the spectrum.
(2) $\nu = c/\lambda$, so $d\nu/d\lambda = -c/\lambda^2$.
(3) Eq. (4.13) with $\nu = c/\lambda$.

Collecting the powers of c and λ,

$$B_\lambda(T) = \frac{2hc^2}{\lambda^5}\,\frac{1}{e^{hc/\lambda k_B T}-1} \qquad \left[\mathrm{W\,m^{-2}\,sr^{-1}\,m^{-1}}\right]. \tag{4.15}$$

The factor c/λ^2 stretches the curve differently from B_ν, and this moves the peak, as we see next.

4.3.2 Wien's Displacement Law

To find where B_λ peaks, we write it through x. Since $x = hc/(\lambda k_B T)$ falls steadily as λ grows, a peak in λ is a peak in x.

$$\begin{aligned} B_\lambda &\overset{(1)}{=} 2hc^2\left(\frac{k_B T}{hc}\right)^5 \frac{x^5}{e^x-1} \\ &= \frac{2k_B^5 T^5}{h^4 c^3}\,\frac{x^5}{e^x-1}. \end{aligned} \tag{4.16}$$

(1) Eq. (4.15) with $1/\lambda = k_B T\,x/(hc)$.

At fixed T only the factor $x^5/(e^x-1)$ changes with wavelength. Its derivative vanishes at the peak,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x^5}{e^x-1}\right) &\overset{(1)}{=} \frac{5x^4(e^x-1)-x^5e^x}{(e^x-1)^2} = 0, \\ 5(e^x-1) &\overset{(2)}{=} xe^x, \\ x &\overset{(3)}{=} 5\left(1-e^{-x}\right). \end{aligned} \tag{4.17}$$

(1) The quotient rule.
(2) The numerator vanishes, $5x^4(e^x-1) = x^5e^x$. Divide by x^4.
(3) Divide by e^x.

This has no solution in closed form. Starting from $x = 5$ and putting each value back into the right-hand side gives 4.96631, then 4.96516, then 4.96512, and the values settle at $x = 4.965114$. The peak wavelength $\lambda_{\max}$ is then

$$\begin{aligned} \lambda_{\max} T &\overset{(1)}{=} \frac{hc}{4.965114\,k_B} \\ &= \frac{(6.626\times10^{-34})(2.998\times10^{8})}{4.965114\times1.381\times10^{-23}}\ \mathrm{m\,K}. \end{aligned} \tag{4.18}$$

(1) The definition of x at the peak.

Result 4.4

$$\lambda_{\max} T = 2.898 \times 10^{-3}\,\mathrm{m\,K} = 2898\,\mu\mathrm{m\,K}. \tag{4.19}$$

The wavelength of the peak is inversely proportional to the temperature and depends on nothing else. The Sun, at 5772 K, peaks at 502 nm.

Per unit frequency the peak lies elsewhere. In terms of x, $B_\nu = (2k_B^3T^3/h^2c^2)\,x^3/(e^x-1)$. The same steps with the power 3 in place of 5 give $x = 3(1-e^{-x})$, so $x = 2.821$. For the Sun this is $\nu_{\max} = 2.821\,k_BT/h = 339\,\mathrm{THz}$, whose wavelength $c/\nu_{\max} = 883\,\mathrm{nm}$ is far from 502 nm. Neither peak is wrong. A curve per unit wavelength and a curve per unit frequency put different widths of spectrum under each point, so "the colour at which a star is brightest" depends on which curve is meant. The total power, found next, does not.

4.3.3 The Stefan–Boltzmann Law

The flux leaving a blackbody surface is the power per square metre summed over all outward directions and all frequencies. This is the same counting of energy through a surface as in Section 4.1 (P1), now done at the surface itself.

Summing over directions. Take a small patch of surface. A ray leaving it at the angle ψ to the outward normal sees the patch foreshortened by $\cos\psi$. Per unit area of the patch, it carries the power $B_\nu \cos\psi\, d\Omega\, d\nu$ into the solid angle $d\Omega = \sin\psi\, d\psi\, d\phi$, with ϕ the angle around the normal. Only outward rays count, $0 \le \psi \le \pi/2$.

$$\begin{aligned} F_{\mathrm{surf}} &\overset{(1)}{=} \int_0^\infty d\nu \int_0^{2\pi} d\phi \int_0^{\pi/2} B_\nu \cos\psi \sin\psi\, d\psi \\ &\overset{(2)}{=} \int_0^\infty B_\nu\, d\nu \times 2\pi \times \left[\frac{\sin^2\psi}{2}\right]_0^{\pi/2} \\ &= \pi \int_0^\infty B_\nu\, d\nu. \end{aligned} \tag{4.20}$$

(1) The power per unit area, summed over the outward half of all directions and over frequency.

(2) B_ν is the same in every direction, so it leaves the angle integrals. The ϕ integral is 2π, and $\int \cos\psi \sin\psi\, d\psi = \frac{1}{2}\sin^2\psi$.

The factor is π, not the 2π of a half-sphere of directions, because rays that leave at a slant cross the surface at a slant. The same steps at one frequency give πB_ν per unit frequency, and πB_λ per unit wavelength.

Summing over frequency. The substitution $\nu = k_BTx/h$ takes all the temperature out of the integral.

$$\begin{aligned} \int_0^\infty B_\nu\, d\nu &\overset{(1)}{=} \frac{2h}{c^2} \int_0^\infty \frac{\nu^3\, d\nu}{e^{h\nu/k_BT} - 1} \\ &\overset{(2)}{=} \frac{2h}{c^2} \left(\frac{k_BT}{h}\right)^4 \int_0^\infty \frac{x^3\, dx}{e^x - 1} \\ &\overset{(3)}{=} \frac{2k_B^4T^4}{h^3c^2}\,\frac{\pi^4}{15}. \end{aligned} \tag{4.21}$$

(1) Eq. (4.13).
(2) $\nu = k_B T x/h$ and $d\nu = (k_B T/h)\,dx$, so $\nu^3 d\nu = (k_B T/h)^4 x^3 dx$. The limits stay 0 and ∞.
(3) The integral is $\pi^4/15$, Eq. (4.22) below.

The integral. Multiplying the top and bottom by e^{-x} turns the integrand into a geometric series.

$$\begin{aligned}\int_0^\infty \frac{x^3\,dx}{e^x-1} &\overset{(1)}{=} \int_0^\infty \frac{x^3 e^{-x}}{1-e^{-x}}\,dx \\ &\overset{(2)}{=} \sum_{n=1}^{\infty} \int_0^\infty x^3 e^{-nx}\,dx \\ &\overset{(3)}{=} \sum_{n=1}^{\infty} \frac{1}{n^4} \int_0^\infty y^3 e^{-y}\,dy \\ &\overset{(4)}{=} 6 \sum_{n=1}^{\infty} \frac{1}{n^4} \\ &\overset{(5)}{=} \frac{6\pi^4}{90} = \frac{\pi^4}{15}.\end{aligned} \tag{4.22}$$

(1) Multiply the top and the bottom by e^{-x}.
(2) $e^{-x}/(1-e^{-x}) = \sum_{n\geq 1} e^{-nx}$, a geometric series with ratio $e^{-x} < 1$ for $x > 0$. All terms are positive, so the sum and the integral may be swapped.
(3) The substitution $y = nx$.
(4) Three integrations by parts give $\int_0^\infty y^3 e^{-y} dy = 3! = 6$.
(5) $\sum_{n\geq 1} 1/n^4 = \pi^4/90$, shown next.

The sum of $1/n^4$. On $-\pi \leq x \leq \pi$ the function x^2 has the Fourier series $x^2 = \pi^2/3 + \sum_{n\geq 1} a_n \cos nx$, with $a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} x^2 \cos nx\,dx = 4(-1)^n/n^2$ from two integrations by parts. Parseval's theorem says the mean square of a function is the sum of the squares of its Fourier terms, $\frac{1}{\pi}\int_{-\pi}^{\pi} f^2\,dx = 2(\pi^2/3)^2 + \sum_n a_n^2$ for this series.

$$\begin{aligned}\frac{1}{\pi}\int_{-\pi}^{\pi} x^4\,dx &\overset{(1)}{=} 2\left(\frac{\pi^2}{3}\right)^2 + \sum_{n=1}^{\infty}\frac{16}{n^4}, \\ \frac{2\pi^4}{5} &\overset{(2)}{=} \frac{2\pi^4}{9} + 16\sum_{n=1}^{\infty}\frac{1}{n^4}, \\ \sum_{n=1}^{\infty}\frac{1}{n^4} &= \frac{1}{16}\left(\frac{2\pi^4}{5} - \frac{2\pi^4}{9}\right) = \frac{1}{16}\,\frac{8\pi^4}{45} = \frac{\pi^4}{90}.\end{aligned} \tag{4.23}$$

(1) Parseval's theorem with $f = x^2$ and $a_n^2 = 16/n^4$.
(2) $\int_{-\pi}^{\pi} x^4 dx = 2\pi^5/5$.

The law. Putting the frequency sum into Eq. (4.20),

$$F_{\text{surf}} = \pi\,\frac{2k_B^4 T^4}{h^3 c^2}\,\frac{\pi^4}{15} = \sigma_{\text{SB}} T^4, \tag{4.24}$$

where the **Stefan–Boltzmann constant** collects all the other factors,

$$\boxed{\sigma_{\text{SB}} = \frac{2\pi^5 k_B^4}{15 c^2 h^3} = 5.670 \times 10^{-8}\,\text{W m}^{-2}\,\text{K}^{-4}.} \tag{4.25}$$

The power per square metre grows as the fourth power of the temperature, so doubling T multiplies it by 16. It depends on nothing about the surface except its temperature, because the light of a blackbody is set by T alone.

4.3.4 Effective Temperature

A real star is not a perfect blackbody. Its spectrum has absorption lines, and its light comes from layers at different temperatures (Section 11.8). We still give it one temperature, by asking which blackbody of the same size gives off the same power. The **effective temperature** T_{eff} of a star of luminosity L and radius R is the temperature of a blackbody sphere of radius R that gives off L. The sphere has area $4\pi R^2$, and each square metre gives off $\sigma_{\text{SB}} T_{\text{eff}}^4$ by Eq. (4.24), so

Result 4.5

$$L = 4\pi R^2 \, \sigma_{\text{SB}} T_{\text{eff}}^4. \tag{4.26}$$

For the Sun, with $L_\odot = 3.828 \times 10^{26}\,\text{W}$ and $R_\odot = 6.957 \times 10^{8}\,\text{m}$,

$$T_{\text{eff},\odot} = \left(\frac{L_\odot}{4\pi R_\odot^2 \, \sigma_{\text{SB}}} \right)^{1/4} = 5772\,\text{K},$$

the value used throughout.

4.4 Colours and the Dimming by Dust

A magnitude through one filter measures one flux, and the Planck function says how the flux of a hot surface is shared among the colours. A temperature from Wien's law needs the whole spectrum, which takes a long time to measure for a faint star. Also, the light of distant stars passes through dust, which dims it. A dimmed star looks further away than it is, and the distance modulus gives the wrong distance.

The setting Sun shows both problems at once. Its light crosses a long path of air near the horizon. The air removes blue light from the beam more than red, so the Sun looks fainter and redder than at noon. Someone who did not know about the air might think the Sun had cooled. Comparing its brightness through a blue glass and a red glass shows the change, and knowing how the air treats each colour undoes it. Astronomers measure each star through a blue filter B and a yellow-green filter V. The difference of the two magnitudes, the colour, measures the temperature without needing the distance, because distance dims both filters alike. Dust between the stars removes blue light faster than yellow, like the air at sunset, so a dimmed star also looks redder. A colour and a magnitude, together with the spectral type of the star (Section 11.7), separate temperature, dust and distance, which one magnitude alone cannot do.

Temperature and dust can both change the measured colour, but in different ways. Blackbodies at 5000 K and 10 000 K differ by 0.71 magnitudes in $B - V$ (Eq. (4.30)). Dust that dims a star by one magnitude in V also makes it 0.32 magnitudes redder in $B - V$ (Eq. (4.42)). We will first use colour as a thermometer, then describe how extinction changes that colour and corrects the distance inferred from brightness.

4.4.1 Passbands and the Colour Index

Astronomical detectors measure flux through standard filters. The Johnson–Cousins set has passbands centred near U (365 nm), B (440 nm), V (550 nm), R (640 nm) and I (790 nm), shown as shaded strips in Figure 4.1.

A **colour index** is the difference of two magnitudes of the same star, such as $B - V$. Applying Eq. (4.7) in each band,

$$\begin{aligned} B - V &\overset{(1)}{=} -2.5 \log_{10} \frac{F_B}{F_{B,\mathrm{ref}}} + 2.5 \log_{10} \frac{F_V}{F_{V,\mathrm{ref}}} \\ &\overset{(2)}{=} -2.5 \log_{10} \frac{F_B}{F_V} + 2.5 \log_{10} \frac{F_{B,\mathrm{ref}}}{F_{V,\mathrm{ref}}}. \end{aligned} \tag{4.27}$$

(1) Eq. (4.7) for the B and V fluxes, each with its own zero point.
(2) Collecting the logarithms.

The last term is a constant, fixed so that Vega and other stars of type A0 V have $B - V \approx 0$. The distance and the radius of the star cancel from F_B/F_V, because by Eq. (4.3) each flux is its surface value times the same factor $(R/d)^2$.

Colour as a thermometer. For a blackbody, F_B/F_V depends on the temperature alone. We treat each filter as a single wavelength, λ_B and λ_V. The flux at the Earth per unit wavelength is then πB_λ (Eq. (4.20) at one wavelength) times $(R/d)^2$.

$$\begin{aligned} \frac{F_B}{F_V} &\overset{(1)}{=} \frac{\pi B_\lambda(\lambda_B, T)\,(R/d)^2}{\pi B_\lambda(\lambda_V, T)\,(R/d)^2} \\ &\overset{(2)}{=} \left(\frac{\lambda_V}{\lambda_B}\right)^5 \frac{e^{hc/\lambda_V k_B T} - 1}{e^{hc/\lambda_B k_B T} - 1} \\ &\overset{(3)}{\approx} \left(\frac{\lambda_V}{\lambda_B}\right)^5 \exp\left[-\frac{hc}{k_B T}\left(\frac{1}{\lambda_B} - \frac{1}{\lambda_V}\right)\right]. \end{aligned} \tag{4.28}$$

(1) Eq. (4.3) with the surface flux πB_λ at each wavelength.
(2) Eq. (4.15). The factors π, $2hc^2$ and $(R/d)^2$ cancel.
(3) Drop the -1 in each bracket, since $e^x \gg 1$. At 440 nm and 10^4 K, $x = 3.3$, and the -1 changes the ratio by a few per cent. For cooler stars it matters even less.

Putting this into Eq. (4.27),

$$\begin{aligned} B - V &\overset{(1)}{=} \frac{2.5}{\ln 10} \frac{hc}{k_B}\left(\frac{1}{\lambda_B} - \frac{1}{\lambda_V}\right)\frac{1}{T} - 12.5 \log_{10} \frac{\lambda_V}{\lambda_B} + 2.5 \log_{10} \frac{F_{B,\mathrm{ref}}}{F_{V,\mathrm{ref}}} \\ &\overset{(2)}{=} \frac{7100\,\mathrm{K}}{T} + \text{constant}. \end{aligned} \tag{4.29}$$

(1) $\log_{10} e^{-y} = -y/\ln 10$ and $\log_{10} a^5 = 5 \log_{10} a$.
(2) $hc/k_B = 1.4388 \times 10^{-2}$ m K, $\lambda_B = 440$ nm and $\lambda_V = 550$ nm. The constant collects the terms without T.

$$\boxed{B - V \approx \frac{7100\,\mathrm{K}}{T} + \text{constant}.} \tag{4.30}$$

Hotter stars have smaller $B - V$, and the hottest have negative values, bluer than Vega. Cool stars have large positive values. The colour depends on the temperature, not on the distance or the size of the star. Real filters are broad and real spectra have lines, so in practice $B - V$ is turned into T_{eff} by a relation measured on stars of known temperature. Eq. (4.30) explains its shape.

4.4.2 Extinction from Optical Depth

Interstellar dust grains are mostly smaller than a micrometre. Each grain removes light from a beam, either by absorbing it or by scattering it out of the beam. The two together are called **extinction**. The area of beam that one grain removes is its **extinction cross-section** $\sigma_{\mathrm{ext}}(\lambda)$.

The dimming law. Follow a beam through dust that has n_d grains per unit volume, with s the distance along the beam. In a thin slab of thickness ds, one square metre of beam meets $n_d\,ds$ grains, which together block the fraction $n_d\sigma_{\text{ext}}\,ds$ of it. The grains only remove light from the beam and add none (P1 for the energy in the beam), so

$$dF_\lambda = -n_d\,\sigma_{\text{ext}}\,F_\lambda\,ds. \tag{4.31}$$

Let $F_{\lambda,0}$ be the flux that would arrive with no dust. Dividing by F_λ and integrating along the whole path through the dust,

$$\begin{aligned}\ln\frac{F_\lambda}{F_{\lambda,0}} &\overset{(1)}{=} -\int n_d\,\sigma_{\text{ext}}\,ds \\ &\overset{(2)}{=} -\sigma_{\text{ext}}(\lambda)\,N_d.\end{aligned} \tag{4.32}$$

(1) $\int dF_\lambda/F_\lambda = \ln F_\lambda$, from $F_{\lambda,0}$ where the beam enters the dust to F_λ where it leaves.
(2) All grains are taken to be alike, so σ_{ext} is the same along the path. The **column density** $N_d = \int n_d\,ds$ is the number of grains in a tube of cross-section $1\,\text{m}^2$ along the line of sight.

The **optical depth** is

$$\tau_\lambda = \int n_d\,\sigma_{\text{ext}}\,ds = N_d\,\sigma_{\text{ext}}(\lambda), \tag{4.33}$$

and the flux falls exponentially with it,

$$\boxed{F_\lambda = F_{\lambda,0}\,e^{-\tau_\lambda}.} \tag{4.34}$$

The optical depth counts the grain cross-sections that lie along the line of sight. The mean distance a photon travels between grains is $1/(n_d\sigma_{\text{ext}})$, so τ_λ is the length of the path divided by this distance, a comparison of two lengths (P6). Light passes almost freely when $\tau_\lambda \ll 1$ and hardly at all when $\tau_\lambda \gg 1$.

Extinction in magnitudes. The **extinction** A_λ is the number of magnitudes by which dust dims a star, $A_\lambda = m_\lambda - m_{\lambda,0}$, where $m_{\lambda,0}$ is the magnitude the star would have with no dust.

$$\begin{aligned}A_\lambda &\overset{(1)}{=} -2.5\log_{10}\frac{F_\lambda}{F_{\lambda,0}} \\ &\overset{(2)}{=} -2.5\log_{10} e^{-\tau_\lambda} \\ &\overset{(3)}{=} \frac{2.5}{\ln 10}\,\tau_\lambda.\end{aligned} \tag{4.35}$$

(1) Eq. (4.5) between the dimmed flux and the flux with no dust.
(2) Eq. (4.34).
(3) $\log_{10} e^{-y} = -y/\ln 10$.

Result 4.6

$$A_\lambda = 1.086\,\tau_\lambda = 1.086\,N_d\,\sigma_{\text{ext}}(\lambda). \tag{4.36}$$

One magnitude of extinction is close to one optical depth. The extinction is proportional to the column of dust, so it adds up along the line of sight. The magnitude without dust is $m_\lambda - A_\lambda$, and Eq. (4.11) for it becomes

$$m_\lambda = M_\lambda + 5\log_{10} d\,[\text{pc}] - 5 + A_\lambda. \tag{4.37}$$

4.4.3 Reddening and the Colour Excess

Why dust reddens. The grains have radii a near 0.1 μm. How they treat light is decided by the ratio of the grain's circumference to the wavelength, $2\pi a/\lambda$ (P6). At 550 nm this ratio is 1.1, so visible light is in the Mie regime, between the λ^{-4} law of particles much smaller than the wavelength and the flat law of particles much larger (Section 8.2.1). Measured across the visible, the extinction by real dust, with its mix of grain sizes, falls roughly as $1/\lambda$ (Figure 4.1). At 2.2 μm the ratio is 0.29, and the extinction is only about a tenth of that in V. Dust clouds that are dark in visible light can therefore be seen through in the infrared.

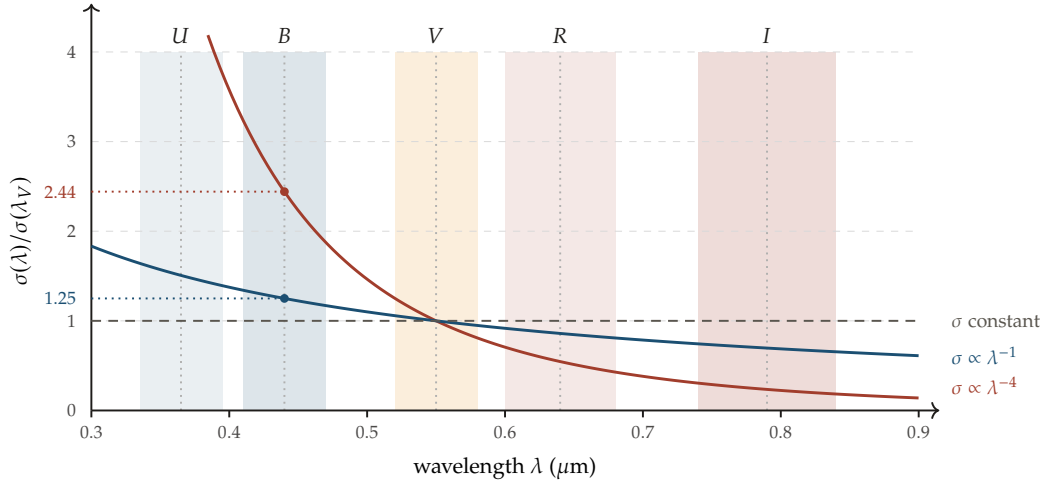


Figure 4.1: Cross-sections that vary as λ^{-4}, as λ^{-1} and not at all, each divided by its value at $\lambda_V = 0.55\,\mu$m, against wavelength from 0.3 to 0.9 μm. The shaded strips mark the U, B, V, R and I passbands, with dotted lines at their central wavelengths. The dots mark the heights of the λ^{-1} and λ^{-4} curves at the B line, carried to the vertical axis by dotted lines.

The colour excess. The measured colour of a star behind dust is

$$\begin{aligned}(B-V)_{\text{obs}} &\overset{(1)}{=} m_B - m_V \\ &\overset{(2)}{=} (m_{B,0} + A_B) - (m_{V,0} + A_V) \\ &\overset{(3)}{=} (B-V)_0 + (A_B - A_V).\end{aligned} \tag{4.38}$$

(1) The definition of the colour index.
(2) $A_\lambda = m_\lambda - m_{\lambda,0}$ in each band.
(3) $(B-V)_0 = m_{B,0} - m_{V,0}$ is the **intrinsic colour**, the colour with no dust.

The **colour excess** is $E(B-V) \equiv A_B - A_V$. The symbol $E(B-V)$ is one name, read as "the excess in the colour $B-V$", and not E multiplied by $(B-V)$. It is also written E_{B-V}, which says the same thing. It is how much redder the colour $B-V$ looks than it would without dust. By Eq. (4.36),

$$E(B-V) = 1.086\, N_d\, [\sigma_{\text{ext}}(\lambda_B) - \sigma_{\text{ext}}(\lambda_V)] > 0,$$

since σ_{ext} is larger at the shorter wavelength. So

$$(B-V)_{\text{obs}} = (B-V)_0 + E(B-V), \tag{4.39}$$

and the intrinsic colour is recovered as

$$\boxed{(B-V)_0 = (B-V)_{\text{obs}} - E(B-V).} \tag{4.40}$$

In practice it is used the other way round. The intrinsic colour of a star is known from its spectral type, so its measured colour gives $E(B-V)$.

From reddening to dimming. The distance needs the dimming A_V, but what we measure is the reddening $E(B-V)$. The reddening that comes with each magnitude of dimming is

$$\frac{E(B-V)}{A_V} \overset{(1)}{=} \frac{A_B - A_V}{A_V} \overset{(2)}{=} \frac{1.086\, N_d\, [\sigma_{\text{ext}}(\lambda_B) - \sigma_{\text{ext}}(\lambda_V)]}{1.086\, N_d\, \sigma_{\text{ext}}(\lambda_V)} = \frac{\sigma_{\text{ext}}(\lambda_B)}{\sigma_{\text{ext}}(\lambda_V)} - 1$$
$$\overset{(3)}{=} \frac{\lambda_V}{\lambda_B} - 1 = \frac{550}{440} - 1 = 0.25. \tag{4.41}$$

(1) The definition of the colour excess.
(2) Eq. (4.36) in the B and V bands.
(3) An extinction cross-section proportional to $1/\lambda$.

Each curve of Figure 4.1 is divided by its value at λ_V, so its height at the B line is $\sigma_{\text{ext}}(\lambda_B)/\sigma_{\text{ext}}(\lambda_V)$, and its height there minus one is the excess per magnitude of dimming. The $1/\lambda$ curve stands at 1.25, which gives $A_V/E(B-V) = 1/0.25 = 4.0$. The λ^{-4} curve of particles much smaller than the wavelength stands at $(550/440)^4 = 2.44$, so such particles give 1.44 magnitudes of excess for each magnitude of dimming, and this strong reddening is why the sky is blue and the setting Sun is red. Grains much larger than the wavelength have a flat curve at height 1. They dim every colour equally and give no excess, which is why clouds are white. The column of dust cancels in the ratio, so it depends on the kind of grains and not on how many there are. Measured along many lines of sight in the diffuse Milky Way, the ratio is

$$\boxed{A_V = 3.1\, E(B-V).} \tag{4.42}$$

The ratio 3.1 is often called R_V. It is smaller than the 4.0 of a pure $1/\lambda$ law, so between B and V the extinction falls a little faster than $1/\lambda$. A law $\lambda^{-\beta}$ with $(550/440)^\beta = 1 + 1/3.1$ fits it, which gives $\beta = 1.25$.

What sets the step. Going from V to B shortens the wavelength by the factor $550/440 = 1.25$, so for a particle of any size it multiplies $2\pi a/\lambda$ by 1.25. The step $\sigma(B)/\sigma(V)$ is how much the cross-section grows over that change, and it depends on where the particle sits on the curve of Figure 4.2. A particle much smaller than the wavelength feels the same field across its whole body, so all its electrons are shaken in step and radiate together. Its cross-section grows as $(2\pi a/\lambda)^4$ (Eq. (8.10)), and the step is $1.25^4 = 2.44$. When $2\pi a/\lambda$ nears one, the field changes across the grain and its parts radiate out of step, so the growth slows. The cross-section peaks at about 4.2 times the grain's area near $2\pi a/\lambda = 4.0$ and then settles at twice the area for every colour (Section 8.2.2). On the way the step falls from 2.44 through 1.25, near $2\pi a/\lambda = 2.80$, to 1. Just past the peak the cross-section falls as $2\pi a/\lambda$ grows, and grains of that size block red light a little more than blue. A single grain of radius 0.1 μm is at $2\pi a/\lambda = 1.14$ in V and gives a step of 1.95. Real dust is a mix of sizes and of materials that also absorb light, and the measured step of 1.32 is the average over that mix. Over the short range from V to B, a cross-section that rises as $\lambda^{-\beta}$ gives the step 1.25^β, which is how the $\beta = 1.25$ above was found.

The distance corrected for dust. Eq. (4.37) in the V band gives the distance.

$$5 \log_{10} d\,[\text{pc}] - 5 \overset{(1)}{=} m_V - M_V - A_V$$
$$\overset{(2)}{=} m_V - M_V - 3.1\, E(B-V). \tag{4.43}$$

(1) Eq. (4.37) solved for the distance term.

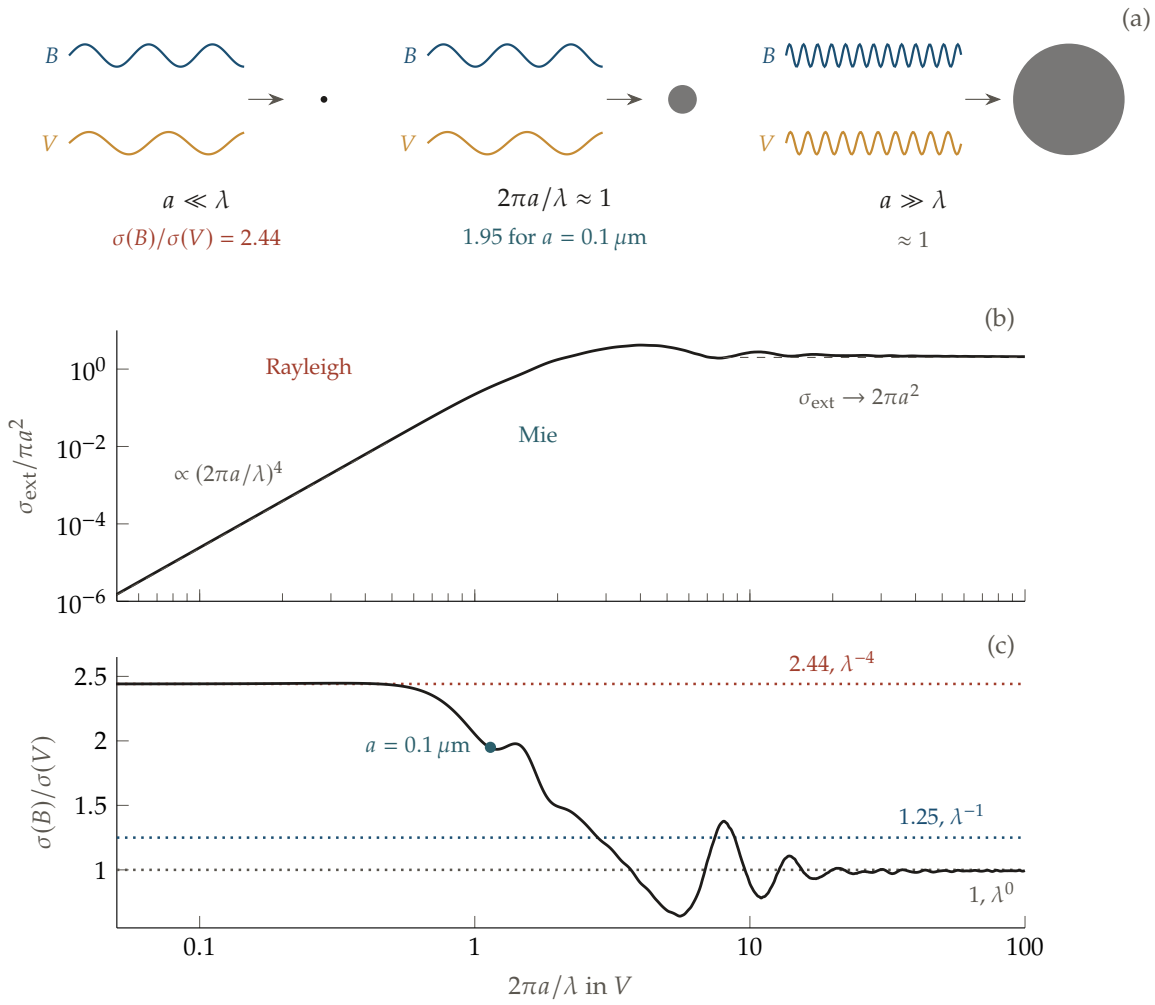


Figure 4.2: (a) A particle much smaller than, near, and much larger than the wavelength, with the B and V waves drawn to the scale of each particle. (b) The extinction cross-section of a sphere of refractive index 1.5, in units of its area πa^2, against $2\pi a/\lambda$, from the Mie series averaged over a ±15% spread in size. (c) The step $\sigma(B)/\sigma(V)$ for the same spheres against $2\pi a/\lambda$ in V, with dotted lines at 2.44, 1.25 and 1, the heights of the curves of Figure 4.1 at the B line. The dot marks a grain of radius 0.1 μm.

(2) Eq. (4.42).

Result 4.7

$$(m_V - M_V)_0 \equiv m_V - M_V - 3.1\,E(B-V) = 5\log_{10} d\,[\mathrm{pc}] - 5. \tag{4.44}$$

The corrected distance modulus $(m_V - M_V)_0$ needs three measurements, m_V, $(B-V)_{\rm obs}$ and the spectral type, which gives both M_V and $(B-V)_0$. Leaving out the dust makes a star seem further away by the factor $10^{0.2A_V}$, which is 2.9 for $A_V = 2.3$ magnitudes (Problem 4.6.1).

4.5 The Radii of Stars

We now have the luminosity of a star, from its flux and its distance, and its effective temperature, from its colour or its spectrum. Almost no star shows a disc in a telescope. Even the nearest

stars look like points, so their sizes cannot be read off the sky with a ruler.

Two electric room heaters glow the same orange. Their surfaces are then at the same temperature, and each square centimetre gives off the same power. If one heater warms the room twice as much as the other, it must have twice the glowing area. Its size follows from its power and its colour, and nobody has to measure it. A star is such a heater. Its temperature fixes the power per square metre, $\sigma_{\mathrm{SB}} T_{\mathrm{eff}}^4$, and its luminosity is the total. The area of the star is the luminosity divided by the power per unit area.

A star at the Sun's temperature with 100 times its luminosity must have 100 times its emitting area and hence 10 times its radius. Aldebaran is cooler, at 3900 K, so each square metre gives only $(3900/5772)^4 = 0.208$ of the Sun's power per square metre. Yet its total power is 439 times the Sun's. Its area must therefore be $439/0.208 = 2106$ times larger, giving a radius 46 times the Sun's. This link between temperature, area and luminosity will give us the radii of stars and their tiny angular sizes on the sky.

4.5.1 Radius from Luminosity and Temperature

Dividing Eq. (4.26) for a star by the same equation for the Sun,

$$\frac{L}{L_\odot} \overset{(1)}{=} \frac{4\pi R^2 \sigma_{\mathrm{SB}} T_{\mathrm{eff}}^4}{4\pi R_\odot^2 \sigma_{\mathrm{SB}} T_{\mathrm{eff},\odot}^4} = \left(\frac{R}{R_\odot}\right)^2 \left(\frac{T_{\mathrm{eff}}}{T_{\mathrm{eff},\odot}}\right)^4 . \tag{4.45}$$

(1) Eq. (4.26) for the star and for the Sun.

Result 4.8

$$\frac{R}{R_\odot} = \sqrt{\frac{L}{L_\odot}} \left(\frac{T_{\mathrm{eff},\odot}}{T_{\mathrm{eff}}}\right)^2 . \tag{4.46}$$

The factors 4π and σ_{SB} cancel, so only ratios to the Sun are needed, with $R_\odot = 6.957 \times 10^8$ m and $T_{\mathrm{eff},\odot} = 5772$ K.

With bolometric magnitudes. Luminosities are often given as absolute bolometric magnitudes. Taking the logarithm,

$$\log_{10} \frac{R}{R_\odot} \overset{(1)}{=} \frac{1}{2} \log_{10} \frac{L}{L_\odot} - 2 \log_{10} \frac{T_{\mathrm{eff}}}{T_{\mathrm{eff},\odot}} \overset{(2)}{=} -0.2\,(M_{\mathrm{bol}} - M_{\mathrm{bol},\odot}) - 2 \log_{10} \frac{T_{\mathrm{eff}}}{T_{\mathrm{eff},\odot}} , \tag{4.47}$$

(1) The logarithm of Eq. (4.46).
(2) Eq. (4.9) gives $\log_{10}(L/L_\odot) = -0.4(M_{\mathrm{bol}} - M_{\mathrm{bol},\odot})$.

with $M_{\mathrm{bol},\odot} = +4.74$ from Section 4.2.

4.5.2 Angular Diameter

A sphere of radius R at distance d is seen as a disc of **angular diameter** θ. The lines of sight that just graze its edge touch the sphere, so each meets the radius drawn to the point of contact

at a right angle. In the right triangle formed by the observer, the centre of the star and that point, the hypotenuse is d and the side opposite the half-angle $\theta/2$ is R. So

$$\sin\frac{\theta}{2} = \frac{R}{d} \quad\Longrightarrow\quad \theta \approx \frac{2R}{d} \quad \text{(radians)}, \tag{4.48}$$

where the small-angle form holds because $R \ll d$ for every star. In the units used for stars,

$$\begin{aligned}\theta\,[\text{mas}] &\overset{(1)}{=} 2.0626\times10^{8}\times\frac{2R}{d}\\ &\overset{(2)}{=} 2.0626\times10^{8}\times\frac{2\times6.957\times10^{8}\,\text{m}}{3.0857\times10^{16}\,\text{m}}\,\frac{R}{R_\odot}\,\frac{1\,\text{pc}}{d}.\end{aligned} \tag{4.49}$$

(1) One radian is $(180/\pi)\times3600\times1000\,\text{mas} = 2.0626\times10^{8}\,\text{mas}$.
(2) R in units of $R_\odot$ and d in parsecs.

Result 4.9

$$\theta = 9.30\,\text{mas}\left(\frac{R}{R_\odot}\right)\left(\frac{1\,\text{pc}}{d}\right). \tag{4.50}$$

The Sun seen from 1 pc would be 9.3 mas across.

Worked example 4.2: The angular diameter of Betelgeuse

Betelgeuse (α Orionis) is a red supergiant with $R \approx 764\,R_\odot$ at a distance $d \approx 168\,\text{pc}$ ($p \approx 5.95\,\text{mas}$). Eq. (4.50) gives

$$\theta = 9.30\,\text{mas}\times\frac{764}{168} = 42.3\,\text{mas}.$$

Infrared interferometers measure 42 to 44 mas. Two telescopes separate angles down to about the wavelength divided by their separation (Eq. (5.15)). At 2.2 μm, 42.3 mas = 2.05×10^{-7} rad needs a separation of $2.2\times10^{-6}/2.05\times10^{-7} = 11\,\text{m}$. Most stars are far smaller on the sky than Betelgeuse, and are measured with telescopes hundreds of metres apart (Chapter 5).

4.5.3 Angular Diameter from Flux and Temperature

A second route to the same angle needs no distance at all. The bolometric flux at the Earth is

$$\begin{aligned}F &\overset{(1)}{=} F_{\text{surf}}\left(\frac{R}{d}\right)^2\\ &\overset{(2)}{=} \sigma_{\text{SB}}T_{\text{eff}}^4\left(\frac{R}{d}\right)^2\\ &\overset{(3)}{=} \sigma_{\text{SB}}T_{\text{eff}}^4\left(\frac{\theta}{2}\right)^2.\end{aligned} \tag{4.51}$$

(1) Eq. (4.3).
(2) $F_{\text{surf}} = L/(4\pi R^2) = \sigma_{\text{SB}}T_{\text{eff}}^4$ by Eq. (4.26).
(3) Eq. (4.48).

$$\boxed{\theta = 2\sqrt{\frac{F}{\sigma_{\text{SB}}T_{\text{eff}}^4}}.} \tag{4.52}$$

The angle depends only on the flux and the temperature, and R and d never enter separately. A star of known flux and colour therefore has a known angular size. Turned round, a measured angular size and a measured flux give $T_{\rm eff}$ without any model of the star. For the Sun, $F = 1361\,\mathrm{W\,m^{-2}}$ and $\theta = 31.97'$ give $T_{\rm eff} = 5772\,\mathrm{K}$ (Problem 4.6.3). For Aldebaran the two routes give the same 20.1 mas (Problem 4.6.2).

4.6 Problem Set: Photometry, Colours and Extinction

4.6.1 Problem 1: Distance and Extinction from a Colour

Problem Statement: An A0 V star has intrinsic colour $(B-V)_0 = 0.00$ and absolute magnitude $M_V = +0.60$. It is seen behind an interstellar cloud with apparent magnitudes $m_B = 14.55$ and $m_V = 13.80$. The dust follows $A_V = 3.1\,E(B-V)$.

(a) What is the colour excess $E(B-V)$?

(b) What is the visual extinction A_V?

(c) What is the true distance d?

(d) What distance would we find if we ignored the dust?

Essential Ideas:

- Dust adds the same colour excess to every star behind it, $(B-V)_{\rm obs} = (B-V)_0 + E(B-V)$, Eq. (4.39). The spectral type gives $(B-V)_0$.
- The dimming in V is $A_V = 3.1\,E(B-V)$, Eq. (4.42).
- The corrected distance modulus is $m_V - M_V - A_V = 5\log_{10} d\,[\mathrm{pc}] - 5$, Eq. (4.44).

Detailed Solutions:

(a) **Colour excess**. The measured colour is $(B-V)_{\rm obs} = m_B - m_V = 14.55 - 13.80 = 0.75$. So

$$\boxed{E(B-V) = (B-V)_{\rm obs} - (B-V)_0 = 0.75 - 0.00 = 0.75.}$$

(b) **Visual extinction**. By Eq. (4.42),

$$\boxed{A_V = 3.1 \times 0.75 = 2.325 \text{ magnitudes.}}$$

The cloud lets through $10^{-0.4\times 2.325} = 12\%$ of the V light.

(c) **True distance**. The corrected distance modulus is

$$m_V - M_V - A_V = 13.80 - 0.60 - 2.325 = 10.875,$$

so

$$\boxed{d = 10^{(10.875+5)/5}\,\mathrm{pc} = 10^{3.175}\,\mathrm{pc} = 1496\,\mathrm{pc} \approx 1.50\,\mathrm{kpc}.}$$

(d) **Distance without the dust correction**. With A_V left out,

$$\boxed{d = 10^{(13.80-0.60+5)/5}\,\mathrm{pc} = 10^{3.64}\,\mathrm{pc} = 4365\,\mathrm{pc}.}$$

This is too large by $10^{0.2A_V} = 10^{0.465} = 2.92$. Dust in the plane of the Milky Way makes uncorrected distances too large in this way, which is why every distance from a magnitude needs a colour as well.

4.6.2 Problem 2: Distance, Magnitudes and Radius of Aldebaran

Problem Statement: Aldebaran (α Tauri) has luminosity $L = 439\,L_\odot$, effective temperature $T_{\rm eff} = 3900\,\mathrm{K}$ and parallax $p = 47\,\mathrm{mas}$.

(a) What is the distance to Aldebaran in parsecs and in metres?

(b) What bolometric flux F arrives at the Earth?

(c) With $M_{\rm bol,\odot} = +4.74$ and $m_{\rm bol,\odot} = -26.83$, find the absolute and apparent bolometric magnitudes of Aldebaran.

(d) What is its radius in solar radii?

(e) What is its angular diameter in milliarcseconds?

Essential Ideas:

- Parallax gives the distance, $d\,[\mathrm{pc}] = 1/p\,['']$, Eq. (3.5), and the flux follows from $F = L/(4\pi d^2)$, Eq. (4.2).
- Absolute magnitudes compare luminosities, $M_{\rm bol} - M_{\rm bol,\odot} = -2.5\log_{10}(L/L_\odot)$, Eq. (4.9). Apparent magnitudes compare fluxes, $m_{\rm bol} - m_{\rm bol,\odot} = -2.5\log_{10}(F/F_\odot)$, Eq. (4.5). The two are linked by $m - M = 5\log_{10} d\,[\mathrm{pc}] - 5$, Eq. (4.11).
- The radius follows from L and $T_{\rm eff}$, $R/R_\odot = \sqrt{L/L_\odot}\,(T_{\rm eff,\odot}/T_{\rm eff})^2$, Eq. (4.46), and the angle from $\theta = 2R/d$, Eq. (4.48), or from $\theta = 2\sqrt{F/\sigma_{\rm SB}T_{\rm eff}^4}$, Eq. (4.52).

Detailed Solutions:

(a) **Distance**. With $p = 47\,\mathrm{mas} = 0.047''$,

$$\boxed{d = \frac{1}{0.047}\,\mathrm{pc} = 21.28\,\mathrm{pc} = 6.565\times10^{17}\,\mathrm{m}.}$$

(b) **Bolometric flux at the Earth**. With $L = 439\times3.828\times10^{26}\,\mathrm{W} = 1.680\times10^{29}\,\mathrm{W}$,

$$\boxed{F = \frac{1.680\times10^{29}\,\mathrm{W}}{4\pi\left(6.565\times10^{17}\,\mathrm{m}\right)^2} = 3.10\times10^{-8}\,\mathrm{W\,m^{-2}}.}$$

(c) **Absolute and apparent bolometric magnitudes**. By Eq. (4.9),

$$\boxed{M_{\rm bol} = 4.74 - 2.5\log_{10}439 = 4.74 - 2.5\times2.6425 = -1.87.}$$

The distance modulus is $5\log_{10}21.28 - 5 = 1.64$, so

$$\boxed{m_{\rm bol} = M_{\rm bol} + 1.64 = -1.87 + 1.64 = -0.23.}$$

The flux gives the same answer by a second route,

$$m_{\rm bol} = -26.83 - 2.5\log_{10}\frac{3.10\times10^{-8}}{1361} = -26.83 + 26.61 = -0.23.$$

The Sun's magnitude in this check must be the bolometric one, -26.83. Its V magnitude, -26.74, would give -0.13.

(d) **Radius**. By Eq. (4.46),

$$\frac{R}{R_\odot} = \sqrt{439}\left(\frac{5772\,\mathrm{K}}{3900\,\mathrm{K}}\right)^2 = 20.95\times2.190, \qquad \boxed{R = 45.9\,R_\odot.}$$

In metres, $R = 45.9\times6.957\times10^8\,\mathrm{m} = 3.19\times10^{10}\,\mathrm{m}$, or $0.213\,\mathrm{AU}$. Placed at the Sun, Aldebaran would reach more than half way to the orbit of Mercury.

(e) **Angular diameter**. By Eq. (4.48),

$$\theta = \frac{2 \times 3.193 \times 10^{10}\,\mathrm{m}}{6.565 \times 10^{17}\,\mathrm{m}} = 9.726 \times 10^{-8}\,\mathrm{rad},$$

and in milliarcseconds

$$\boxed{\theta = 9.726 \times 10^{-8} \times 2.0626 \times 10^{8}\,\mathrm{mas} = 20.1\,\mathrm{mas}.}$$

The flux route of Eq. (4.52) needs neither R nor d,

$$\theta = 2\sqrt{\frac{3.10 \times 10^{-8}}{5.670 \times 10^{-8} \times 3900^4}}\,\mathrm{rad} = 9.72 \times 10^{-8}\,\mathrm{rad} = 20.1\,\mathrm{mas},$$

the same angle. An angle of 20 mas is easily measured with optical interferometers such as CHARA and the VLTI, so for Aldebaran the radius can be checked directly.

4.6.3 Problem 3: The Sun as a Blackbody

Problem Statement: Above the Earth's air the Sun gives the flux $F_\odot = 1361\,\mathrm{W\,m^{-2}}$, and its disc is 31.97′ across. The Earth is $1\,\mathrm{AU} = 1.496 \times 10^{11}$ m from the Sun.

(a) Find the effective temperature of the Sun from these two measurements alone.

(b) At what wavelength does B_λ peak, and at what frequency does B_ν peak? What wavelength belongs to that frequency?

(c) Find the radius and the luminosity of the Sun. Check that they satisfy $L = 4\pi R^2 \sigma_{\mathrm{SB}} T_{\mathrm{eff}}^4$.

Essential Ideas:

- The flux at the Earth is the surface flux diluted by $(\theta/2)^2$, so $\theta = 2\sqrt{F/\sigma_{\mathrm{SB}} T_{\mathrm{eff}}^4}$, Eq. (4.52).
- B_λ peaks at $x = hc/(\lambda k_B T) = 4.965$ (Eq. (4.19)), and B_ν at $x = h\nu/k_B T = 2.821$.
- $R = d\,\theta/2$ (Eq. (4.48)) and $L = 4\pi d^2 F$ (Eq. (4.2)).

Detailed Solutions:

(a) **Effective temperature**. In radians, $\theta = 31.97' = 9.301 \times 10^{-3}$. Solving Eq. (4.52) for the temperature,

$$\begin{aligned} T_{\mathrm{eff}} &\overset{(1)}{=} \left(\frac{F_\odot}{\sigma_{\mathrm{SB}}}\right)^{1/4} \left(\frac{2}{\theta}\right)^{1/2} \\ &= \left(\frac{1361}{5.670 \times 10^{-8}}\right)^{1/4} \left(\frac{2}{9.301 \times 10^{-3}}\right)^{1/2} \mathrm{K} \\ &= 393.6 \times 14.66\,\mathrm{K}. \end{aligned}$$

(1) Square Eq. (4.52), $\theta^2/4 = F_\odot/(\sigma_{\mathrm{SB}} T_{\mathrm{eff}}^4)$, and take the fourth root.

$$\boxed{T_{\mathrm{eff},\odot} = 5772\,\mathrm{K}.}$$

The distance to the Sun is not needed, only the flux and the angle.

(b) **The two peaks**. By Eq. (4.19),

$$\boxed{\lambda_{\max} = \frac{2.898 \times 10^{-3}\,\mathrm{m\,K}}{5772\,\mathrm{K}} = 502\,\mathrm{nm}.}$$

Per unit frequency,

$$\boxed{\nu_{\max} = \frac{2.821\, k_B T}{h} = \frac{2.821 \times 1.381 \times 10^{-23} \times 5772}{6.626 \times 10^{-34}}\,\mathrm{Hz} = 3.39 \times 10^{14}\,\mathrm{Hz},}$$

whose wavelength is $c/\nu_{\max} = 883\,\mathrm{nm}$, in the near infrared. The Sun is "brightest in the green" only when its spectrum is counted per unit wavelength.

(c) **Radius and luminosity**. From the angle and the distance,

$$R_\odot = d\,\frac{\theta}{2} = 1.496 \times 10^{11}\,\mathrm{m} \times 4.650 \times 10^{-3} = 6.957 \times 10^{8}\,\mathrm{m},$$

and from the flux and the distance, as in the solar constant example,

$$L_\odot = 4\pi d^2 F_\odot = 4\pi \left(1.496 \times 10^{11}\,\mathrm{m}\right)^2 \times 1361\,\mathrm{W\,m^{-2}} = 3.828 \times 10^{26}\,\mathrm{W}.$$

Then

$$\boxed{4\pi R_\odot^2\, \sigma_{\mathrm{SB}} T_{\mathrm{eff},\odot}^4 = 4\pi \left(6.957 \times 10^{8}\right)^2 \times 5.670 \times 10^{-8} \times 5772^4\,\mathrm{W} = 3.828 \times 10^{26}\,\mathrm{W} = L_\odot.}$$

The check must work, since T_{eff} in part (a) was found from the same flux and angle. It shows that Eqs. (4.2), (4.26) and (4.52) are one statement seen three ways. Any two of the flux, the angle and the temperature give the third.

Chapter 5

Angular Resolution: Baselines, Diffraction, and Pixel Sampling

Photometry tells us how much light reaches a telescope, while astrometry tells us the direction it came from. Neither measurement guarantees that two nearby sources can be told apart. Even a perfect telescope spreads a point source into a small patch, and two overlapping patches may merge into one image.

Resolution begins with the wavelength compared with the mirror diameter or the separation of two radio dishes. The focal length maps the resulting angle onto the detector, whose pixels must sample it finely enough. For a ground-based telescope, turbulent air can impose a wider blur. Before considering those limits, we must first find which wavelengths reach the ground at all.

5.1 Windows on the Universe

Chapter 4 measures the flux of a star in a band of wavelengths, and takes for granted that the light arrives. Most of it does not. Above the telescope there are about 10^4 kg of air over every square metre, and at most wavelengths that is enough to absorb every photon.

Ordinary window glass shows the trick. It lets visible light through, so the room is bright, but it stops most of the ultraviolet that causes sunburn. Glass is not simply clear or dark. Light of each colour either finds something in the glass that can take its energy, or passes straight through. The air is a window of the same kind, with only a few panes open. The idea is that a photon gets through when its energy matches no transition of the molecules in the air, and a radio wave gets through when its frequency is above the plasma frequency of the ionosphere.

A photon of wavelength λ carries energy $hc/\lambda = 1239.84\,\text{eV nm}/\lambda$. Visible photons, from 400 to 700 nm, carry 3.10 to 1.77 eV. Ozone absorbs shorter wavelengths below 300 nm, above 4.13 eV, by electronic transitions. Water vapour absorbs near 2.7 and 6.3 μm, at 0.46 and 0.20 eV, while carbon dioxide absorbs near 4.3 and 15 μm, at 0.29 and 0.083 eV, through molecular vibrations. Visible light falls between these strong absorption bands. We will trace the windows this leaves in the atmosphere and then explain why the ionosphere closes the low-frequency end of the radio window.

Band by band.

- **Gamma rays and X-rays** ionise the atoms of the air and are absorbed high up. This is what protects life on the ground from them. They can be observed only from orbit (Chandra, NuSTAR).

- **Ultraviolet light** below about 300 nm is absorbed by ozone (O_3) in the stratosphere, and below about 200 nm also by O_2. Both absorb by electronic transitions. So ultraviolet astronomy below the ozone cutoff is done from space.
- **Visible light** (350–800 nm) has too little energy for the electronic transitions of O_2 and O_3, and too much for the vibration bands of H_2O, CO_2 and CH_4. Only a few weak bands remain, such as the oxygen A band at 760 nm. Rayleigh scattering by the air molecules, whose cross-section goes as λ^{-4} (Figure 4.1), removes some of the blue and makes the day sky glow, but it does not close the window. This gap is the **optical window**, and our eyes are tuned to it.
- **Infrared light** passes only through narrow gaps (the J, H, K, L and M bands) between the strong bands of water vapour and carbon dioxide. So infrared observatories sit at high, dry sites (Mauna Kea, the Atacama), or go above most of the water vapour on aircraft (SOFIA, which flew until 2022), on balloons, or in space (JWST).
- **Radio waves** pass from about 10 MHz up to a few hundred GHz. This is the **radio window**. Below about 10 MHz the ionosphere reflects the waves back into space, Eq. (5.2). The same reflection lets short-wave radio signals travel round the globe. At the millimetre and submillimetre end, water vapour absorbs again. This is why ALMA is built at 5000 m in the Atacama desert, above most of the water.

The plasma frequency. The ionosphere is a thin gas of free electrons, n_e of them per cubic metre. We want the refractive index n it has for a wave of frequency ν, and so whether the wave can enter it. The electric field of the wave, $E\cos(2\pi\nu t)$, shakes each electron. Newton's law for the electron, of mass m_e and charge $-e$, is $m_e\ddot{x} = -eE\cos(2\pi\nu t)$. Its steady solution is $x = x_0\cos(2\pi\nu t)$, so $\ddot{x} = -(2\pi\nu)^2 x$, and

$$-m_e(2\pi\nu)^2 x = -eE\cos(2\pi\nu t) \quad\Longrightarrow\quad x \overset{(1)}{=} \frac{eE\cos(2\pi\nu t)}{m_e(2\pi\nu)^2},$$

$$n^2 \overset{(2)}{=} 1 + \frac{n_e(-e)x}{\varepsilon_0 E\cos(2\pi\nu t)} \overset{(3)}{=} 1 - \frac{n_e e^2}{\varepsilon_0 m_e(2\pi\nu)^2} \overset{(4)}{=} 1 - \frac{\nu_p^2}{\nu^2}. \tag{5.1}$$

(1) Divide by $-m_e(2\pi\nu)^2$. The electron moves in step with the field and opposite to the force on a positive charge.
(2) For a non-magnetic medium n^2 is the relative permittivity, $1 + P/(\varepsilon_0 E)$, where the polarisation P is the dipole moment per volume, n_e times $(-e)x$.
(3) The value of x from the first line.
(4) The definition of the **plasma frequency** ν_p below.

$$\boxed{\nu_p = \frac{1}{2\pi}\sqrt{\frac{n_e e^2}{\varepsilon_0 m_e}} = 8.98\,\text{Hz} \times \sqrt{n_e\,[\text{m}^{-3}]}\,.} \tag{5.2}$$

Above ν_p, n^2 is positive and the wave travels on. Below ν_p, n^2 is negative, n is imaginary, and the wave dies away within a short distance and is reflected. At its densest the ionosphere has $n_e \approx 10^{12}\,\text{m}^{-3}$, which gives $\nu_p \approx 9$ MHz. That is the lower edge of the radio window. The result uses only the electron density, so any plasma reflects waves below its own ν_p, whatever it is made of.

Two windows matter most for everyday astronomy, the optical and the radio. They favour two different ways of getting a sharp image. Radio astronomers can combine dishes that stand a continent apart, because the air barely disturbs radio wavefronts over such distances. Optical astronomers are mostly held to a single mirror, whole or made of segments. Its sharpness is set

by diffraction, and from the ground it is spoiled further by turbulence in the air, even though the light itself is not absorbed.

5.2 Interferometric Baselines

The radio window lets waves from about 10 MHz to a few hundred GHz reach the ground, and Chapter 2 gives any place on a sphere a unit vector. A single radio dish is a poor camera. At the 21 cm line of hydrogen a 25 m dish blurs every source over about half a degree, the size of the full Moon.

Close your eyes and listen to a friend calling from one side. You know the direction at once. The sound reaches the nearer ear first, by up to $0.2\,\mathrm{m}/(343\,\mathrm{m\,s^{-1}}) = 0.58\,\mathrm{ms}$ for ears 0.2 m apart, and the brain turns that delay into an angle. A wider head would hear a smaller change of angle as the same change of delay. Two radio dishes far apart are two ears. A wave from a source off to one side reaches one dish first. The idea is that the finest angle two receivers can tell apart is the wavelength divided by their separation, and it does not depend on the size of each receiver.

At 21 cm, two dishes 1 km apart resolve roughly $0.21/1000 = 2.1\times10^{-4}\,\mathrm{rad} = 43''$. A single 25 m dish gives about $1.22\times0.21/25\,\mathrm{rad} = 35'$, around fifty times coarser. The improvement comes from comparing the arrival phase at two places separated by the **baseline**. We first calculate that separation on the Earth, then find how a source's direction changes the delay between the dishes.

5.2.1 The Baseline Between Two Stations

The station vectors. Let two stations sit at latitudes ϕ_1, ϕ_2 and longitudes Λ_1, Λ_2 on a sphere of radius $R_E = 6371.0\,\mathrm{km}$, the mean radius of the Earth. The longitude gets a capital letter because λ is the wavelength here. In the frame with its z axis through the North Pole and its x axis through longitude 0°, each station has the unit vector of Eq. (2.2), with latitude in place of declination and longitude in place of right ascension,

$$\hat{\mathbf{r}}_i = \begin{pmatrix} \cos\phi_i \cos\Lambda_i \\ \cos\phi_i \sin\Lambda_i \\ \sin\phi_i \end{pmatrix}, \qquad i \in \{1,2\}. \tag{5.3}$$

The central angle. The angle γ between the two station vectors, at the centre of the Earth, is the angle between two unit vectors. Chapter 2 found it as their dot product, Eq. (2.35). With $\delta \to \phi$ and $\alpha \to \Lambda$,

$$\cos\gamma = \hat{\mathbf{r}}_1 \cdot \hat{\mathbf{r}}_2 = \sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos(\Lambda_2 - \Lambda_1). \tag{5.4}$$

For stations a few degrees apart, $\cos\gamma$ is close to 1 and loses digits. The haversine form, Eq. (2.37), keeps them,

$$\sin^2\frac{\gamma}{2} = \sin^2\frac{\phi_2-\phi_1}{2} + \cos\phi_1\cos\phi_2\sin^2\frac{\Lambda_2-\Lambda_1}{2}. \tag{5.5}$$

The chord. The baseline B is the straight line through the Earth from one station to the other, the length of $R_E\hat{\mathbf{r}}_1 - R_E\hat{\mathbf{r}}_2$,

$$\begin{aligned} B^2 &= R_E^2\,(\hat{\mathbf{r}}_1 - \hat{\mathbf{r}}_2)\cdot(\hat{\mathbf{r}}_1 - \hat{\mathbf{r}}_2) \overset{(1)}{=} R_E^2\,(1 + 1 - 2\,\hat{\mathbf{r}}_1\cdot\hat{\mathbf{r}}_2) \\ &\overset{(2)}{=} 2R_E^2(1-\cos\gamma) \overset{(3)}{=} 4R_E^2\sin^2\frac{\gamma}{2}. \end{aligned} \tag{5.6}$$

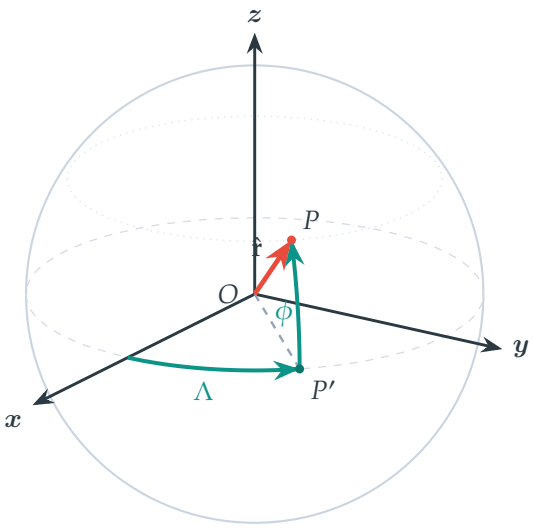


Figure 5.1: A point P on a sphere with centre O, its unit vector, its foot P' in the equatorial plane, its latitude ϕ, and its longitude (marked λ) measured from the x axis.

(1) Multiply out. Each unit vector has $\hat{\mathbf{r}}_i \cdot \hat{\mathbf{r}}_i = 1$.
(2) Eq. (5.4).
(3) $1 - \cos\gamma = 2\sin^2(\gamma/2)$.

$$\boxed{B = 2R_E \sin\left(\frac{\gamma}{2}\right) = R_E\sqrt{2(1-\cos\gamma)}.} \tag{5.7}$$

The same result follows from the isosceles triangle of Figure 5.2. The line from the centre to the midpoint M of the chord cuts it at a right angle, so each half is $R_E \sin(\gamma/2)$. Together with Eq. (5.5), $\sin(\gamma/2)$ comes straight from the coordinates, and no inverse cosine is needed.

5.2.2 Two Sources Seen by Two Antennas

We now want the finest angle a pair of antennas a distance B apart can tell apart, at wavelength λ. First we find what the pair measures for any sky, then put in two point sources.

One source. A source at angle θ from the normal to the baseline sends a flat wavefront. It reaches A_2 first. To reach A_1 it must travel further by the side of the right-angled triangle in Figure 5.3. For the two sources of the figure,

$$\Delta L_1 = B\sin\theta_1, \tag{5.8}$$
$$\Delta L_2 = B\sin\theta_2. \tag{5.9}$$

Each antenna turns the wave into a voltage. For one source of frequency $\nu = c/\lambda$, take $v_2(t) = \cos(2\pi\nu t)$ at A_2. At A_1 the same wave arrives later by $\Delta L/c$, so $v_1(t) = \cos(2\pi\nu t - 2\pi\Delta L/\lambda)$. The **correlator** multiplies the two voltages and averages over many periods,

$$\langle v_1 v_2\rangle \overset{(1)}{=} \frac{1}{2}\left\langle \cos\left(4\pi\nu t - \frac{2\pi\Delta L}{\lambda}\right)\right\rangle + \frac{1}{2}\cos\frac{2\pi\Delta L}{\lambda} \overset{(2)}{=} \frac{1}{2}\cos\frac{2\pi B\sin\theta}{\lambda}. \tag{5.10}$$

(1) The product rule $\cos x\cos y = \frac{1}{2}\cos(x+y) + \frac{1}{2}\cos(x-y)$.
(2) The first term swings at twice the wave frequency and averages to zero. Then $\Delta L = B\sin\theta$.

As the Earth turns, θ changes, $B\sin\theta$ runs through many wavelengths, and the output swings between $+\frac{1}{2}$ and $-\frac{1}{2}$. These swings are the **fringes**.

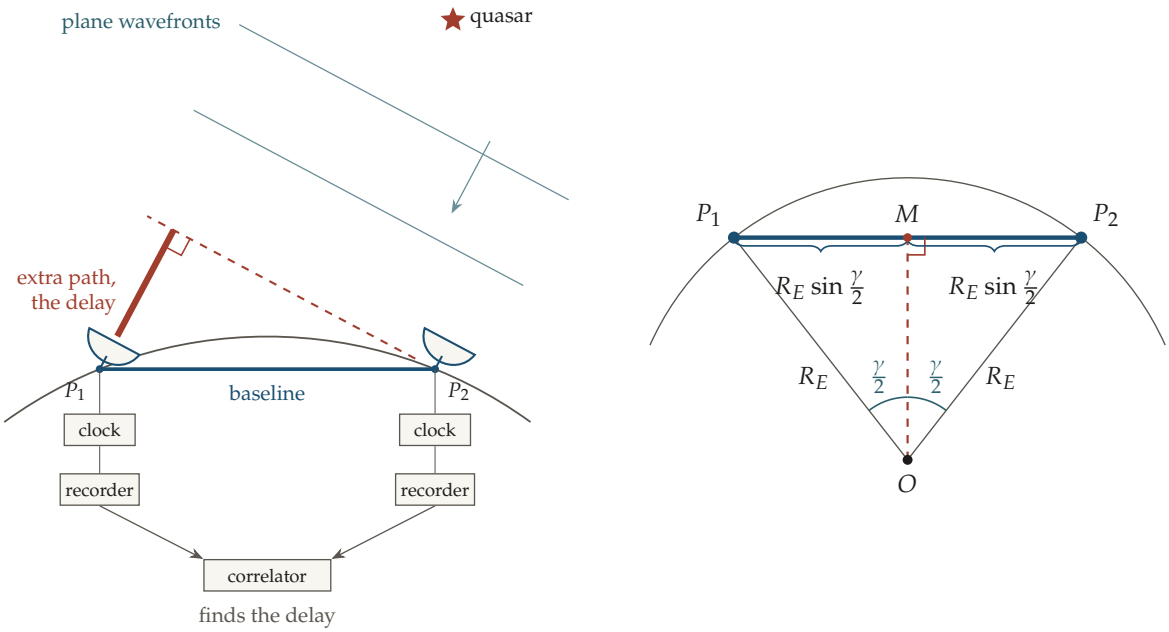


Figure 5.2: Left, a distant quasar sends plane wavefronts to two dishes on the curved Earth. The wavefront reaches the second dish first, and the extra path to the first dish is the delay. Each dish has a clock and a recorder, and a correlator compares the two recordings. Right, two stations P_1 and P_2 on a sphere of radius R_E and centre O, the central angle γ, and the chord between them, cut in half at M.

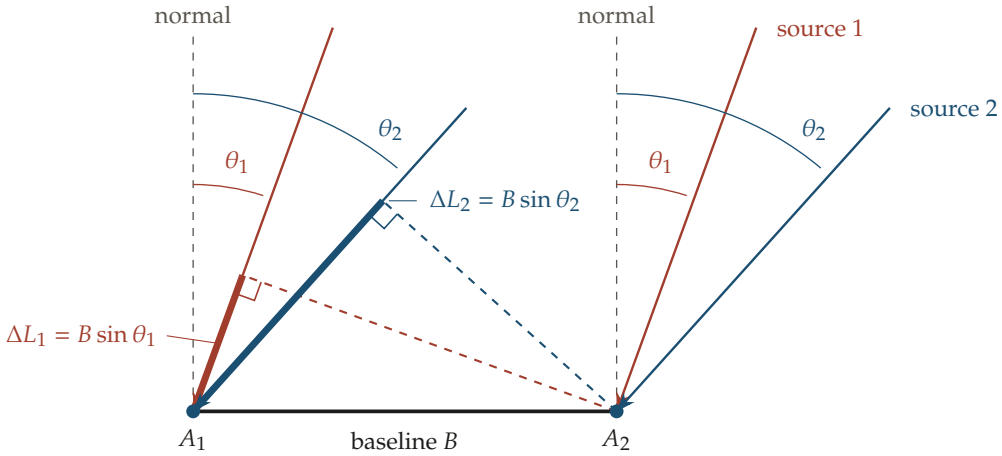


Figure 5.3: Two antennas A_1 and A_2 a baseline B apart, and two distant sources at angles θ_1 and θ_2 from the normal to the baseline. For each source the dashed line is the wavefront through A_2, and the thick segment is the extra path ΔL_1 or ΔL_2 to A_1.

Many sources. Different sources send waves with unrelated phases. In the product of the summed voltages, every cross term between two sources averages to zero, so the outputs of the sources simply add, each weighted by its power. For a sky of brightness $I_{\text{sky}}(\theta)$ along the baseline, near the normal where $\sin\theta \approx \theta$,

$$\langle v_1 v_2 \rangle \propto \int I_{\text{sky}}(\theta) \cos\left(2\pi \frac{B}{\lambda}\,\theta\right) d\theta. \tag{5.11}$$

One pair of antennas measures one Fourier component of the sky, the one with B/λ cycles per radian. Real correlators also form the product with one voltage shifted by a quarter period, which gives the sine component at the same B/λ.

Two equal sources. Put two sources of equal power at θ_1 and θ_2. By Eq. (5.10) the output is

$$\cos\frac{2\pi\Delta L_1}{\lambda} + \cos\frac{2\pi\Delta L_2}{\lambda} \overset{(1)}{=} 2\cos\left(\frac{\pi(\Delta L_1 + \Delta L_2)}{\lambda}\right)\cos\left(\frac{\pi(\Delta L_2 - \Delta L_1)}{\lambda}\right). \tag{5.12}$$

(1) $\cos x + \cos y = 2\cos\frac{x+y}{2}\cos\frac{y-x}{2}$.

The first factor is the fringe of a single source at the mean position. The second factor changes slowly and sets the size of the fringe. It depends on the difference of the two path lags,

$$\Delta L_2 - \Delta L_1 = B\,(\sin\theta_2 - \sin\theta_1) \overset{(1)}{=} 2B\cos\frac{\theta_1+\theta_2}{2}\,\sin\frac{\theta_2-\theta_1}{2}. \tag{5.13}$$

(1) $\sin y - \sin x = 2\cos\frac{x+y}{2}\sin\frac{y-x}{2}$.

Two close sources have a small separation $\Delta\theta = \theta_2 - \theta_1$, so $\sin(\Delta\theta/2) \approx \Delta\theta/2$. Near the normal the cosine is close to 1 as well,

$$\Delta L_2 - \Delta L_1 \approx B\cos\left(\frac{\theta_1+\theta_2}{2}\right)\Delta\theta \approx B\,\Delta\theta. \tag{5.14}$$

Away from the normal the baseline counts only through $B\cos\theta$, its length as seen from the source.

When the pair is resolved. By Eqs. (5.12) and (5.14), the fringe of the pair has the size $|\cos(\pi B\Delta\theta/\lambda)|$ times that of a single source of the same total power. For $\Delta\theta \ll \lambda/B$ it is 1, and the pair cannot be told from a single source. At $B\Delta\theta = \lambda/2$ the two fringe patterns are half a cycle apart and cancel. So the pair shows itself once its extra path $B\Delta\theta$ is a fair part of a wavelength. The angle λ/B is the period of the fringe pattern on the sky, the finest ripple one baseline records. It is quoted as the resolution,

$$B\,\Delta\theta \approx \lambda. \tag{5.15}$$

At exactly $B\Delta\theta = \lambda$ the patterns are a whole cycle apart and agree again. A real array removes this ambiguity with many baselines of different lengths and directions, and the turning Earth changes each projected baseline through the night. With $1\,\text{rad} = 648\,000''/\pi = 206\,265''$,

Result 5.1

$$\Delta\theta \approx \frac{\lambda}{B}\ rad = 206\,265\,\frac{\lambda}{B}\ arcsec. \tag{5.16}$$

Only the ratio of the wavelength to the baseline enters. The size of each dish does not. It sets how much power is collected, and the width λ/D_{dish} of the patch of sky each dish sees, inside which the fine detail λ/B is measured. For the GMRT and the Ooty Radio Telescope, $B = 901$ km and at 1 GHz the resolution is 68.6 mas (Problem 5.7.1).

5.3 Telescope Optics

A radio dish records its own wave, so dishes can stand far apart and be combined later (Section 5.2). Visible light cannot be recorded wave by wave, and above any telescope the air keeps the wavefront flat only over tens of centimetres (Section 5.6). So an optical telescope collects light with one large mirror, whole or made of segments, and brings it to a point.

A magnifying glass held in sunshine makes a small bright image of the Sun on a sheet of paper. Move the glass away from the paper and the image gets sharp at one distance, the focal length. The image is a disc. Its size is the angle the Sun covers, times the focal length. A telescope does the same for every star. The idea is that each direction on the sky goes to one point in the focal plane, at a distance from the axis equal to the focal length times the angle.

The Sun spans $0.533^\circ = 9.30 \times 10^{-3}$ rad on the sky. A lens of focal length 1 m therefore forms an image about 9.3 mm across. The same mapping turns the separation of two stars into a separation on the detector. We will follow that geometry through the telescope and finally ask how many pixels must cover the image to retain its detail.

5.3.1 Reflectors versus Refractors

The main collecting element of a telescope is either a lens (a *refractor*) or a mirror (a *reflector*). Nearly all research telescopes are reflectors. Every reason comes from the light bouncing off the front surface of the mirror instead of passing through glass.

- The light never enters the mirror, so there is no chromatic aberration and almost no absorption. A lens sends different wavelengths to different foci, and thick glass absorbs the ultraviolet and parts of the infrared.
- Only one surface must be polished to the right shape, against two for a lens, and a mirror can be made much lighter than a lens of the same size.
- A mirror can be supported across its whole back. A lens can be held only at its edge, which sags under gravity for large sizes.
- The back has no optical job, so it can carry actuators that bend the mirror to cancel its sag. This is *active optics*, a slow correction of the whole shape, related to the fast adaptive optics of Section 5.6.

The price of a reflector is that the secondary mirror, or the detector itself, must hang *in front of* the primary, in the incoming beam, and block part of it (Section 5.3.3).

5.3.2 Focus Configurations: Prime, Newtonian, Cassegrain, and Coudé

One primary mirror can send its light to a detector in several standard ways. They differ in what stops the beam before the primary focus.

- **Prime focus.** The detector sits at the primary's own focus, with no second mirror. It is the simplest layout and loses no light to a second reflection, but the detector blocks the beam and is hard to reach on a large telescope.
- **Newtonian.** A flat mirror at 45°, just before the primary focus, turns the beam out through the side of the tube. It suits small telescopes.
- **Cassegrain.** A secondary mirror sends the converging beam back through a hole in the primary, so the instrument sits compactly behind the primary.
- **Coudé.** A chain of extra flat mirrors, often along the axes of the mount, carries the beam to a fixed focus in a separate, stable room. Some light is lost at each reflection, but very large and heavy instruments, notably high-resolution echelle spectrographs, can then stand still while the telescope moves.

Whatever the focus, the mount must turn the telescope to follow the daily motion of the sky. It tracks at the sidereal rate in hour angle, Eq. (2.11), either about one axis parallel to the Earth's (an equatorial mount) or about two axes under computer control (an altitude–azimuth mount).

5.3.3 Obstruction Effects and Observatory Domes

The secondary mirror, or a prime-focus detector, is held by a frame of thin vanes across the tube, the *spider*. The vanes are thin opaque strips inside the opening, so they change the aperture mask $t(\xi, \eta)$ of Eq. (5.29). Each strip adds its own diffraction pattern, a streak at right angles to the vane, on top of the star's Airy pattern (Section 5.4 works this out). These are the four- or six-pointed *diffraction spikes* on bright stars in reflector images. The straight edges of a segmented mirror, like the hexagons of JWST, add more spikes in the same way.

Telescopes stand inside a turning *dome*. It keeps off wind, rain and changes of temperature, and it blocks stray light from the site. Moving air of different temperatures inside an open tube would spoil the image in the same way as the seeing of Section 5.6.

5.3.4 Telescope Technologies Across the Electromagnetic Spectrum

The atmosphere decides from where each band can be observed. How each band is focused into an image is a separate question, because ordinary mirrors stop working as the photon energy rises.

Figure 5.4 puts the windows of Section 5.1 on one axis, with the bands and their frequencies above and the cause of each closed range below. A photon is stopped when its energy matches something the air can do. Gamma rays and X-rays carry enough energy to knock electrons out of the atoms of N_2 and O_2, and they are used up high in the air. Ultraviolet photons below 300 nm lift the electrons of O_3 and O_2 to higher levels. Visible light matches no level and passes, which is the optical window. In the infrared the photons match the vibrations of H_2O and CO_2, which leave only narrow gaps, the J, H, K, L, M and N bands between about 1.2 and 13 μm. Further out the photons match the slower rotations of H_2O and O_2, and the air stays opaque until the millimetre band, where only a few rotation lines remain, of O_2 at 60 and 118 GHz and of H_2O at 22 and 183 GHz. The radio window runs from there to about 30 m. Below about 10 MHz the frequency falls under the plasma frequency of the ionosphere, and the waves are reflected back into space, Eq. (5.2).

Ultraviolet and infrared. Ultraviolet and infrared telescopes use the same geometry as optical reflectors. Only the reflecting coating changes, chosen to reflect well at the wavelength wanted. They must still go to space (AstroSat-UVIT in the ultraviolet, JWST in the infrared) or to high, dry sites, because the air is the limit, not the optics.

X-rays and grazing incidence. At X-ray energies a mirror at near-normal incidence reflects almost nothing. The photon frequency is far above the plasma frequency of the electrons in the mirror metal, so by Eq. (5.1) the metal has a refractive index slightly below one,

$$n = \sqrt{1 - \frac{\nu_p^2}{\nu^2}} \approx 1 - \frac{\nu_p^2}{2\nu^2}, \tag{5.17}$$

where every electron is treated as free, which holds roughly when the photon energy is well above the binding energies of most of the electrons. The X-rays travel faster in the metal than in vacuum. Light going from glass into air is totally reflected beyond a critical angle, and X-rays going from vacuum into the metal are totally reflected in the same way, provided they skim the surface.

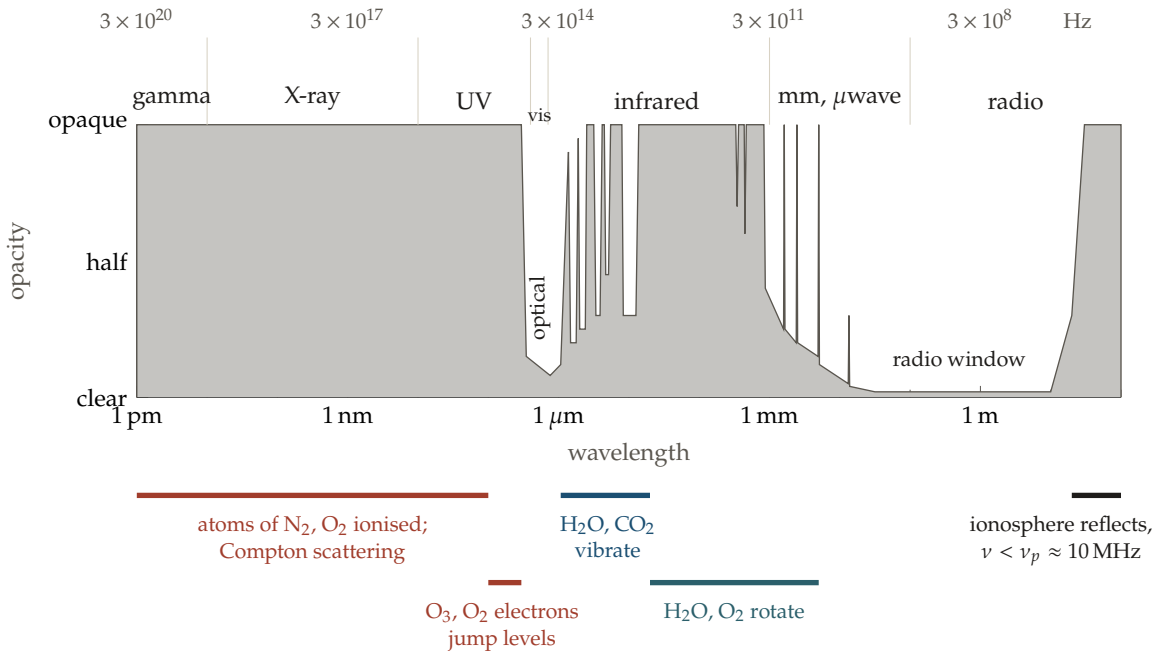


Figure 5.4: Schematic opacity of the air above a high, dry site against wavelength, from clear to opaque, with the names of the bands and their frequencies above. The bars below mark the wavelength range closed by each cause. The edges follow the values of Section 5.1, and the heights between them are drawn by hand. The narrow clear gaps between 1 and 13 μm are the infrared windows J, H, K, L, M and N.

Let the X-ray meet the mirror at the grazing angle α, measured from the surface, so its angle from the normal is $90° - \alpha$. Snell's law gives the angle i from the normal inside the metal,

$$1 \cdot \sin(90° - \alpha) = n \sin i \quad \Longrightarrow \quad \sin i \overset{(1)}{=} \frac{\cos \alpha}{n}. \tag{5.18}$$

(1) $\sin(90° - \alpha) = \cos \alpha$.

No ray enters the metal when this would need $\sin i > 1$, that is when $\cos \alpha > n$. The critical grazing angle α_c has $\cos \alpha_c = n$. For small angles,

$$1 - \frac{\alpha_c^2}{2} \overset{(1)}{=} 1 - \frac{\nu_p^2}{2\nu^2} \quad \Longrightarrow \quad \boxed{\alpha_c = \frac{\nu_p}{\nu}}. \tag{5.19}$$

(1) $\cos \alpha \approx 1 - \alpha^2/2$ on the left, and n from above on the right.

The critical angle is the ratio of the two frequencies, P6 again. The ionosphere reflects radio waves below its ν_p at any angle. A metal reflects X-rays far above its ν_p only at grazing angles below ν_p/ν.

Worked example 5.1: Grazing angle for iridium

The Chandra mirrors are coated with iridium, which has a density of 22.56 g cm^{-3}, an atomic mass of 192.22 and 77 electrons per atom. So

$$n_e = \frac{22\,560 \text{ kg m}^{-3}}{192.22 \times 1.661 \times 10^{-27} \text{ kg}} \times 77 = 5.44 \times 10^{30} \text{ m}^{-3}.$$

By Eq. (5.2), $\nu_p = 2.09 \times 10^{16}$ Hz, a photon energy $h\nu_p = 86.6$ eV. A 10 keV X-ray has

$\nu = 2.42 \times 10^{18}$ Hz, so

$$\alpha_c = \frac{86.6\,\text{eV}}{10\,000\,\text{eV}} = 8.7 \times 10^{-3}\,\text{rad} = 0.50^\circ.$$

Harder X-rays need even shallower angles, since $\alpha_c \propto 1/\nu$.

X-ray telescopes are therefore built of long, nearly cylindrical shells that the X-rays skim. The *Wolter type I* design reflects each X-ray twice, off a paraboloid and then a hyperboloid, which keeps images sharp away from the axis. Many shells are nested inside one another to collect more area, as in Chandra and XMM-Newton.

Gamma rays. At still higher energies, in hard X-rays and gamma rays, reflection is too weak at any angle, and no mirror can focus. Imaging uses a *coded mask* instead. A plate with a known pattern of opaque and open cells stands in front of a detector that records where each photon lands. Each source casts a shifted copy of the pattern's shadow, and the sky image is found by computing which set of shifted shadows adds up to the recorded one. AstroSat-CZTI and Swift-BAT work this way.

Radio. At the other end of the spectrum, a radio telescope is a metal dish with a receiver at its prime focus. The surface may be a mesh, as long as its holes are small compared with the wavelength of centimetres to metres. This simplicity makes the baselines of Section 5.2 possible. A radio wave oscillates slowly enough (periods of nanoseconds to microseconds) to be digitised directly. Each dish records its signal against an atomic clock (Chapter 1), and the recordings are multiplied in a computer afterwards, Eq. (5.10). Visible light oscillates at about 5×10^{14} Hz, far too fast to digitise, so an optical interferometer must bring the beams together with mirrors while the light arrives.

5.3.5 Cassegrain Reflector Telescope Geometry

In a Cassegrain telescope (Figure 5.5), parallel light from a star falls on a concave primary mirror of diameter D and focal length F. Before the light reaches the primary focus, a secondary mirror at distance d_2 in front of the primary catches the converging cone. It sends the light back through a hole in the primary to a detector a distance a behind the primary. We want the smallest diameter b of the secondary, the place a of the detector, and the focal length of the whole system.

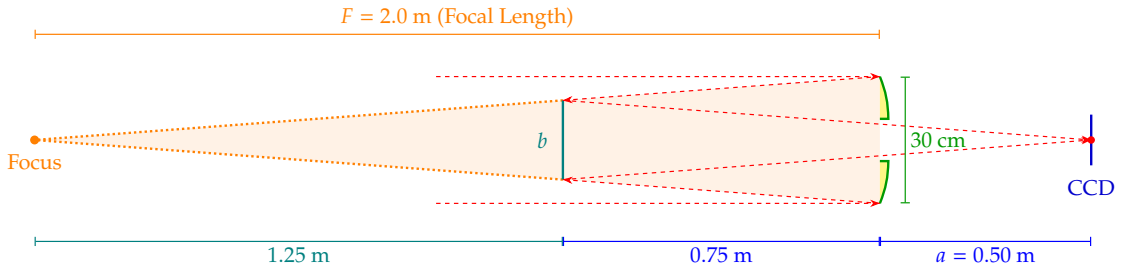


Figure 5.5: A Cassegrain telescope with a 30 cm primary of focal length 2.0 m, a flat secondary of diameter b at 0.75 m from the primary and 1.25 m from the primary focus, and the CCD a distance a behind the primary.

Size of the secondary. The rays from the edge of the primary converge on the primary focus. The cone has diameter D at the primary, a distance F from the focus, and its diameter falls in

proportion to the distance from the focus. The secondary is $F - d_2$ from the focus, so by similar triangles

$$\frac{b}{D} = \frac{F - d_2}{F} \qquad \Longrightarrow \qquad \boxed{b = D\left(1 - \frac{d_2}{F}\right).} \tag{5.20}$$

A smaller secondary catches some of the on-axis light only. Placing the secondary closer to the focus makes it smaller, and so it blocks less of the primary.

Place of the focus. The cone aims at the primary focus, a distance $F - d_2$ behind the secondary. A flat secondary forms the image the same distance in front of itself, $F - d_2$ from it. Back from the secondary towards the primary that is d_2 to the primary, and the rest lies behind it,

$$a = (F - d_2) - d_2 = F - 2d_2 \qquad \text{(flat secondary)}. \tag{5.21}$$

A convex secondary, as in a true Cassegrain, sends the image further back, a distance larger than $F - d_2$ from the secondary.

The effective focal length. After the secondary the cone converges from the diameter b over the distance from the secondary to the final focus. Its ratio of length to width is what a single mirror of some focal length F_{eff} and diameter D would give,

$$\begin{aligned} \frac{F_{\text{eff}}}{D} &= \frac{\text{secondary to final focus}}{b} \overset{(1)}{=} \frac{\text{secondary to final focus}}{D\,(F - d_2)/F}, \\ F_{\text{eff}} &= F \times \frac{\text{secondary to final focus}}{\text{secondary to primary focus}}. \end{aligned} \tag{5.22}$$

(1) b from Eq. (5.20).

A flat secondary leaves the focal length unchanged, $F_{\text{eff}} = F$. A convex one multiplies it by its magnification, which is how a short tube holds a long focal length. The numbers for the telescope of the figure are in Problem 5.7.2.

5.3.6 Optical Ray Diagram and Angular Resolution on the Focal Plane

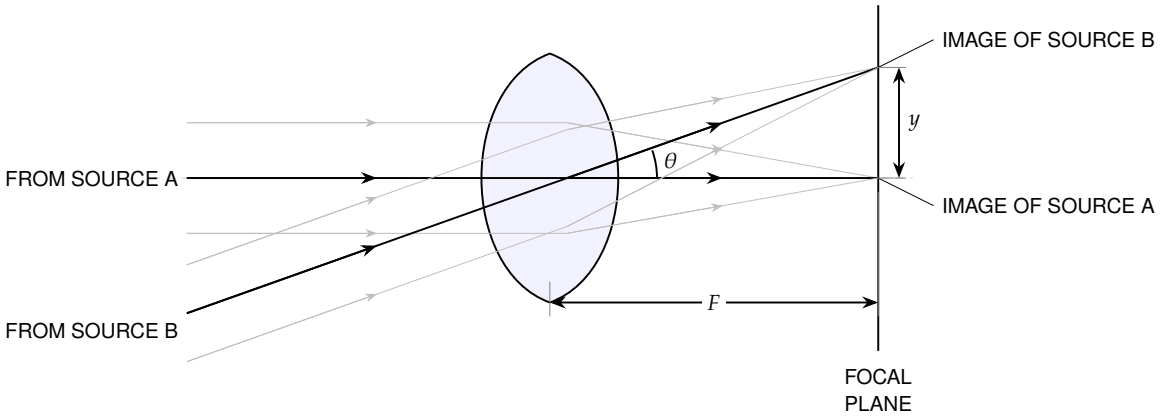


Figure 5.6: Two point sources A and B at the angle θ, a lens of focal length F, and their two images a height y apart in the focal plane. The chief rays through the centre of the lens are dark, and the edge rays are faded.

5.3.7 Plate Scale and Pixel Scale

A detector measures distances in the focal plane, and the astronomer wants angles on the sky. The **plate scale** turns a distance into an angle, and the **pixel scale** is the angle covered by one pixel. Both come from Figure 5.6, where a source at the angle θ from the axis forms its image a height y from the axis.

The exact relation. The chief ray passes through the centre of the lens, or the vertex of the mirror, without bending. It meets the focal plane, a distance F away, at the height

$$y = F\tan\theta \quad\Longrightarrow\quad \theta = \arctan\left(\frac{y}{F}\right). \tag{5.23}$$

With a secondary mirror, F is the effective focal length of Eq. (5.22).

The plate scale. The plate scale is the angle per unit length in the focal plane, the derivative of Eq. (5.23),

$$\frac{\mathrm{d}\theta}{\mathrm{d}y} \overset{(1)}{=} \frac{1}{F}\,\frac{1}{1+(y/F)^2} \overset{(2)}{=} \frac{1}{F}\,\frac{1}{1+\tan^2\theta} \overset{(3)}{=} \frac{\cos^2\theta}{F}. \tag{5.24}$$

(1) $\frac{\mathrm{d}}{\mathrm{d}x}\arctan x = 1/(1+x^2)$, with $x = y/F$.
(2) $y/F = \tan\theta$.
(3) $1+\tan^2\theta = 1/\cos^2\theta$.

So the plate scale is not quite the same everywhere. Away from the axis a millimetre covers slightly fewer seconds of arc. Fields of view rarely exceed a few degrees, and even at $\theta = 1°$, $\cos^2\theta = 0.99970$. In this small-angle limit $\theta \approx y/F$, and the plate scale is a single number,

$$\boxed{\frac{\mathrm{d}\theta}{\mathrm{d}y} = \frac{\theta}{y} = \frac{1}{F},} \tag{5.25}$$

quoted in seconds of arc per millimetre, $206\,265/F$ [mm]. Only wide-field survey cameras need the $\cos^2\theta$ correction.

The pixel scale. A pixel of width y_{pix} covers the angle

$$\theta_{\mathrm{pix}} = \frac{y_{\mathrm{pix}}}{F}, \tag{5.26}$$

which is Eq. (5.25) read at one pixel,

Result 5.2

$$\theta_{\mathrm{pix}} = \left.\frac{y}{F}\right|_{y=y_{\mathrm{pix}}} \textit{rad}, \tag{5.27}$$

or, in seconds of arc per pixel,

$$\theta_{\mathrm{pix}} = \frac{206\,265 \times y_{\mathrm{pix}}\,[\mathrm{mm}]}{F\,[\mathrm{mm}]}\ \textit{arcsec per pixel}. \tag{5.28}$$

For pixels 20 μm wide at $F = 2$ m, $\theta_{\mathrm{pix}} = 10^{-5}$ rad $= 2.06''$. A longer focal length spreads the sky over more pixels.

5.4 Diffraction

A telescope maps each direction to a point of the focal plane, Eq. (5.23), and two antennas a distance B apart resolve about λ/B, Eq. (5.16). Ray optics says a perfect mirror brings a star to a point. Every real telescope, even a perfect one in space such as Hubble or JWST, shows a star as a small disc with faint rings, the **point spread function** (PSF).

Stand beside an open door and you hear people talking in the next room, though you cannot see them. Sound bends round the edge of the door and light does not. The door is about 1 m wide. Sound at 343 Hz has a wavelength of 1 m, the size of the door, so it spreads into every direction. Light has a wavelength two million times smaller, so it spreads by only two millionths of a radian. The edge of a telescope mirror is the door frame. The idea is that a wave passing an opening of size D spreads into an angle of about λ/D.

For sound passing a doorway, $\lambda/D \sim 1$, so it spreads widely. For 500 nm light through a 30 cm mirror, $\lambda/D = 1.7 \times 10^{-6}$ rad; the exact round-aperture result $1.22\lambda/D$ is 0.42″. To see where that factor and the image pattern come from, we will add the waves leaving every open point of an aperture. The same calculation will lead from a slit to a grating and then to the diffraction pattern of a circular mirror.

5.4.1 Diffraction as a Sum of Wavelets

By the **Huygens–Fresnel principle**, every open point of an opening acts as the source of a small spherical wave, a *wavelet*. The waves obey a linear equation, so the field at any point behind the opening is the sum of all the wavelets that reach it, each with the phase of the path it travelled.

The field in the opening. Let the opening lie in the plane $z = 0$, with coordinates (ξ, η). Light from an unresolved star arrives along the z axis as a flat wave of uniform amplitude E_0. The opening is described by its **transmission** $t(\xi, \eta)$, equal to 1 where light passes and 0 where the mirror edge, a baffle or a spider vane blocks it. Just behind the opening the field is

$$E_0\, t(\xi, \eta) = \begin{cases} E_0, & \text{inside the opening,} \\ 0, & \text{where it is blocked.} \end{cases} \tag{5.29}$$

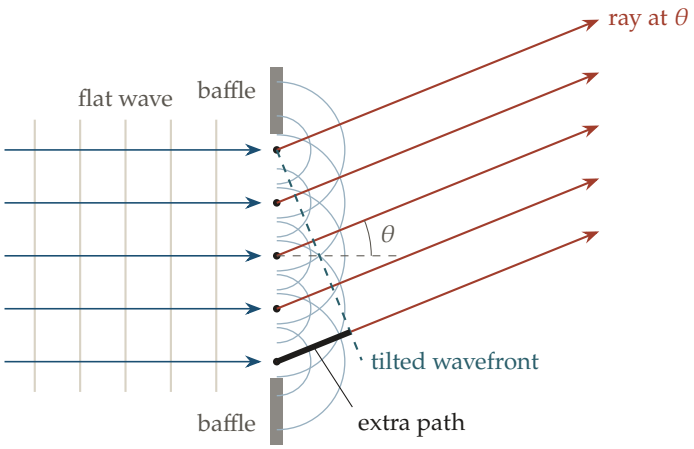


Figure 5.7: A flat wave meeting an opening between two opaque baffles, the wavelets sent out by points of the opening, the rays leaving at the angle θ, and the tilted wavefront. The thick segment marks the extra path of the lowest wavelet.

The path to a far point. Look in the direction of the unit vector

$$\hat{\mathbf{k}} = \begin{pmatrix} \sin\theta\cos\chi \\ \sin\theta\sin\chi \\ \cos\theta \end{pmatrix} \equiv \begin{pmatrix} \theta_x \\ \theta_y \\ \cos\theta \end{pmatrix}, \tag{5.30}$$

where θ is the angle from the axis and χ the angle around it. Its first two components, θ_x and θ_y, are the angles on the sky along the ξ and η directions, for the small angles of astronomy. A point at distance s in this direction is $s\hat{\mathbf{k}}$. The wavelet from the point $\vec{r} = (\xi, \eta, 0)$ of the opening, with $|\vec{r}| = r$, travels to it the distance

$$\begin{aligned} |s\hat{\mathbf{k}} - \vec{r}| &= \sqrt{s^2 - 2s\,\vec{r}\cdot\hat{\mathbf{k}} + r^2} \overset{(1)}{=} s\left(1 - \frac{\vec{r}\cdot\hat{\mathbf{k}}}{s} + \frac{r^2}{2s^2} - \frac{(\vec{r}\cdot\hat{\mathbf{k}})^2}{2s^2} + \dots\right) \\ &\overset{(2)}{=} s - (\xi\theta_x + \eta\theta_y) + \frac{r^2 - (\vec{r}\cdot\hat{\mathbf{k}})^2}{2s} + \dots \end{aligned} \tag{5.31}$$

(1) $\sqrt{1+\epsilon} = 1 + \epsilon/2 - \epsilon^2/8 + \dots$ with $\epsilon = -2\vec{r}\cdot\hat{\mathbf{k}}/s + r^2/s^2$, kept to second order in r/s.
(2) $\vec{r}\cdot\hat{\mathbf{k}} = \xi\theta_x + \eta\theta_y$, since $\vec{r}$ has no z component.

The first term is the same for every wavelet. The second is the path difference that makes the pattern. The last term, at most $(D/2)^2/(2s)$ for an opening of diameter D, changes the phase by at most $\pi D^2/(4\lambda s)$. It can be dropped when

$$s \gg \frac{D^2}{\lambda}, \tag{5.32}$$

a comparison of two lengths (P6). This is the **Fraunhofer** or far-field limit. For a 30 cm mirror at 500 nm, $D^2/\lambda = 180$ km. A telescope brings the far field to its focal plane. A perfect focusing mirror delays each wavelet by just the amount that removes the last term, so the pattern at the focus is the pattern at infinity, with the angle θ mapped to the height $F\theta$.

The integral. A wavelet with extra path $-(\xi\theta_x + \eta\theta_y)$ has the phase factor $\exp\left[-ik(\xi\theta_x + \eta\theta_y)\right]$, with $k = 2\pi/\lambda$. Summing over the opening, and dropping the factor e^{iks}/s and the constant of the Huygens–Fresnel integral, which are the same for every wavelet,

Result 5.3

$$E(\theta_x, \theta_y) = E_0 \iint t(\xi, \eta)\, e^{-ik(\xi\theta_x + \eta\theta_y)}\, d\xi\, d\eta, \qquad I(\theta_x, \theta_y) = |E(\theta_x, \theta_y)|^2. \tag{5.33}$$

The field at the focus is the two-dimensional Fourier transform of the opening, with ξ/λ and η/λ as the frequencies. A wide opening gives a narrow pattern, and a narrow one a wide pattern. Everything below is this one integral for different shapes of t.

5.4.2 Single-Slit Diffraction

The simplest opening is a slit of width a along η and much longer along ξ. The long side makes the pattern narrow along θ_x, so we look only in the plane across the slit, $\theta_x = 0$ and $\theta_y = \sin\theta$, and take the field per unit length of slit.

Pairing the two halves. The dark bands can be found without any integral. Split the slit into a lower half, $-a/2 < \eta < 0$, and an upper half, $0 < \eta < a/2$. Each point η of the lower half has a

partner $\eta + a/2$ in the upper half. At the angle θ the wavelets of the two partners differ in path by $(a/2)\sin\theta$. When that is half a wavelength, every pair cancels at once,

$$\frac{a}{2}\sin\theta = \frac{\lambda}{2} \quad\Longrightarrow\quad a\sin\theta = \lambda \quad\Longrightarrow\quad \theta \approx \frac{\lambda}{a}. \tag{5.34}$$

For a slit the number in front of λ/a is exactly 1.

The full pattern. Eq. (5.33) with $t = 1$ for $|\eta| < a/2$ gives

$$\begin{aligned} E(\theta) &= E_0 \int_{-a/2}^{a/2} e^{-ik\eta\sin\theta}\, d\eta \overset{(1)}{=} E_0\, \frac{e^{-ik(a/2)\sin\theta} - e^{ik(a/2)\sin\theta}}{-ik\sin\theta} \\ &\overset{(2)}{=} E_0\, \frac{2\sin(k(a/2)\sin\theta)}{k\sin\theta} \overset{(3)}{=} E_0\, a\, \frac{\sin\beta_a}{\beta_a}, \qquad \beta_a \equiv \frac{\pi a\sin\theta}{\lambda}. \end{aligned} \tag{5.35}$$

(1) The integral of an exponential, between the two edges.
(2) $e^{-ix} - e^{ix} = -2i\sin x$.
(3) $k(a/2)\sin\theta = \pi a\sin\theta/\lambda = \beta_a$, so $k\sin\theta = 2\beta_a/a$.

The intensity, with $I_0 = E_0^2 a^2$ in the straight-through direction, is

$$\boxed{I(\theta) = I_0 \left(\frac{\sin\beta_a}{\beta_a}\right)^2 = I_0 \left[\frac{\sin(\pi a\sin\theta/\lambda)}{\pi a\sin\theta/\lambda}\right]^2.} \tag{5.36}$$

The first zero is at $\beta_a = \pi$, so $\sin\theta = \lambda/a$, as the pairing found. The integral $\int_{-\infty}^{\infty}(\sin\beta/\beta)^2 d\beta = \pi$ measures the whole pattern. The part between the first zeros, $(1/\pi)\int_{-\pi}^{\pi}(\sin\beta/\beta)^2 d\beta$, is 0.903, so the central band carries 90.3% of the light.

5.4.3 N-Slit Diffraction Grating: Dispersion and Resolving Power

A diffraction grating is N equal parallel slits of width a, with centres a distance d apart, over a total width $W = Nd$. It spreads light into its colours, and it is the heart of most astronomical spectrographs.

Summing the slits. The slit numbered $n = 0, \ldots, N-1$ has its centre at $\eta = nd$. Changing the variable to $\eta' = \eta - nd$ turns its integral into that of a slit at the centre, Eq. (5.35), times the phase factor $e^{-iknd\sin\theta}$. Write $\beta_d \equiv \pi d\sin\theta/\lambda$, so that this factor is $e^{-2in\beta_d}$. Then

$$\begin{aligned} E(\theta) &\overset{(1)}{=} E_0 a\, \frac{\sin\beta_a}{\beta_a} \sum_{n=0}^{N-1} e^{-2in\beta_d} \overset{(2)}{=} E_0 a\, \frac{\sin\beta_a}{\beta_a}\, \frac{1 - e^{-2iN\beta_d}}{1 - e^{-2i\beta_d}} \\ &\overset{(3)}{=} E_0 a\, \frac{\sin\beta_a}{\beta_a}\, \frac{e^{-iN\beta_d}\left(e^{iN\beta_d} - e^{-iN\beta_d}\right)}{e^{-i\beta_d}\left(e^{i\beta_d} - e^{-i\beta_d}\right)} \overset{(4)}{=} E_0 a\, \frac{\sin\beta_a}{\beta_a}\, \frac{\sin N\beta_d}{\sin\beta_d}\, e^{-i(N-1)\beta_d}. \end{aligned} \tag{5.37}$$

(1) Each slit gives Eq. (5.35) times its phase factor $e^{-iknd\sin\theta} = e^{-2in\beta_d}$.
(2) The geometric sum $\sum_{n=0}^{N-1} q^n = (1 - q^N)/(1 - q)$ with $q = e^{-2i\beta_d}$.
(3) Take $e^{-iN\beta_d}$ out of the top and $e^{-i\beta_d}$ out of the bottom.
(4) $e^{ix} - e^{-ix} = 2i\sin x$ in both, and $e^{-iN\beta_d}/e^{-i\beta_d} = e^{-i(N-1)\beta_d}$.

The phase factor drops out of $|E|^2$, and with $I_0 = E_0^2 a^2$ again,

$$\boxed{I(\theta) = I_0 \underbrace{\left(\frac{\sin\beta_a}{\beta_a}\right)^2}_{\text{one slit}} \underbrace{\left(\frac{\sin N\beta_d}{\sin\beta_d}\right)^2}_{N\text{ slits}}.} \tag{5.38}$$

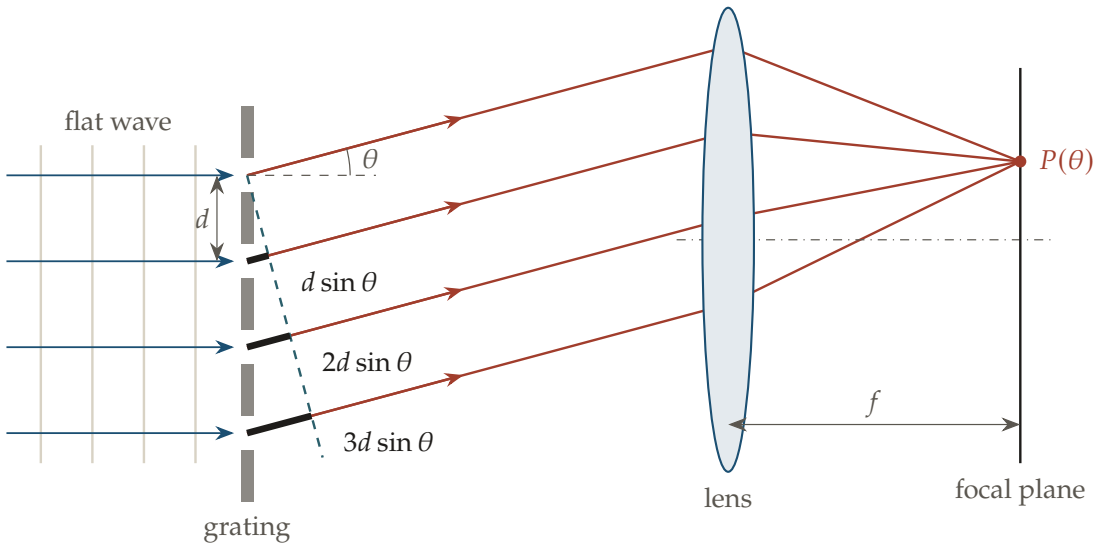


Figure 5.8: A flat wave meeting an N-slit grating, the rays leaving at the angle θ with path lags $d\sin\theta$, $2d\sin\theta$ and $3d\sin\theta$, and a lens of focal length f bringing them to a point $P(\theta)$ in its focal plane.

The width of each slit sets a broad envelope, and their spacing sets sharp peaks inside it.

The principal maxima. When $\beta_d = m\pi$ for a whole number m, all N slits are in phase. The ratio $\sin N\beta_d/\sin\beta_d$ is then $0/0$, and its limit is $\pm N$ (the derivatives of top and bottom give $N\cos Nm\pi/\cos m\pi$). So the peak intensity is N^2 times that of one slit. The condition $\beta_d = m\pi$ reads $d\sin\theta = m\lambda$. If the light arrives at the angle θ_i to the normal, the incoming path to slit n also differs, by $nd\sin\theta_i$. Counting θ_i on the opposite side of the normal from θ_m, as is usual for reflection gratings, this lag adds to the lag on the way out, and the peaks of order m are at

$$d\,(\sin\theta_m + \sin\theta_i) = m\lambda \qquad (m \in \mathbb{Z}). \tag{5.39}$$

This is the **grating equation**. For $m \neq 0$ the angle depends on λ, so each colour goes its own way.

The width of a peak. Next to a peak, $\sin N\beta_d$ first vanishes when $N\beta_d$ has moved from $Nm\pi$ to $Nm\pi + \pi$, a step $\Delta\beta_d = \pi/N$. Since $\beta_d = \pi d\sin\theta/\lambda$, a small step in θ gives $\Delta\beta_d = (\pi d\cos\theta_m/\lambda)\Delta\theta$, and

$$\Delta\theta = \frac{\lambda}{Nd\cos\theta_m} = \frac{\lambda}{W\cos\theta_m}. \tag{5.40}$$

Here $W\cos\theta_m$ is the width of the grating seen from the direction θ_m. The peak is as narrow as the diffraction pattern of a single slit as wide as the whole grating.

Resolving power. Two lines at λ and $\lambda + \Delta\lambda$ are just resolved, by the Rayleigh rule of Section 5.4.5, when the peak of one falls on the first zero of the other. In terms of $d\sin\theta$, the peak of $\lambda + \Delta\lambda$ is at $m(\lambda + \Delta\lambda)$, and the first zero beside the peak of λ is at $m\lambda + \lambda/N$. Setting them equal,

$$m(\lambda + \Delta\lambda) = m\lambda + \frac{\lambda}{N} \quad\Longrightarrow\quad m\Delta\lambda = \frac{\lambda}{N}. \tag{5.41}$$

With Eq. (5.39) for m, the **resolving power** is

Result 5.4

$$\mathcal{R} \equiv \frac{\lambda}{\Delta\lambda} = mN = \frac{W\,(\sin\theta_m + \sin\theta_i)}{\lambda}. \tag{5.42}$$

The last form is the difference in path between light from the two ends of the grating, counted in wavelengths. So a high-resolution spectrograph needs a large grating used at a steep angle in a high order. HARPS and ESPRESSO, which reach $\mathcal{R} > 100\,000$ to measure the wobble of stars with planets, use echelle gratings tens of centimetres long.

Worked example 5.2: Resolving the sodium D lines

The two sodium D lines lie at 589.59 and 588.99 nm, so $\Delta\lambda = 0.60$ nm about the mean 589.29 nm. They need

$$\mathcal{R} = \frac{589.29\,\text{nm}}{0.60\,\text{nm}} = 982.$$

In order $m = 2$, Eq. (5.42) asks for $N \geq 982/2 = 491.1$, so at least 492 lit grooves. On a grating with 600 grooves per millimetre that is a beam only $W = 492/600$ mm $= 0.82$ mm wide.

5.4.4 Circular Apertures and the Airy Disk

Telescope mirrors are round, of diameter D. We want the pattern of Eq. (5.33) for a disc, and the angle of its first dark ring.

Why more than one. In a slit every strip parallel to the edges has the same length, so the two halves cancel each other one to one at $\theta = \lambda/a$. In a disc the strips are short near the rim and long, up to D, across the middle. The middle sends more light than the edges. When the path difference between the two edges is λ, the weak wavelets of the narrow edges cannot cancel the strong ones of the wide middle. The angle must open by about 22% more before the whole disc cancels.

The disc integral. Take polar coordinates in the opening, $\xi = r\cos\psi$ and $\eta = r\sin\psi$. A disc looks the same from every direction around its axis, so the pattern depends only on the angle θ from the axis, and we may look along $\theta_x = \theta$, $\theta_y = 0$. Then

$$\begin{aligned} E(\theta) &\overset{(1)}{=} E_0 \int_0^{D/2}\!\!\int_0^{2\pi} e^{-ikr\theta\cos\psi}\, d\psi\; r\, dr \overset{(2)}{=} 2\pi E_0 \int_0^{D/2} J_0(kr\theta)\, r\, dr \\ &\overset{(3)}{=} E_0 \underbrace{\frac{\pi D^2}{4}}_{\text{area of the mirror}} \frac{2J_1(\beta_D)}{\beta_D}, \qquad \beta_D \equiv \frac{\pi D\theta}{\lambda}. \end{aligned} \tag{5.43}$$

(1) Eq. (5.33) with $d\xi\, d\eta = r\, dr\, d\psi$ and $\xi\theta_x + \eta\theta_y = r\theta\cos\psi$.

(2) The angle integral is $2\pi J_0(z)$ with $z = kr\theta$, where $J_0(z) = (1/2\pi)\int_0^{2\pi} e^{-iz\cos\psi}\, d\psi$ is the Bessel function of order zero. Bessel functions also solve Kepler's equation in Section 6.12.

(3) With $z = kr\theta$ the edge $r = D/2$ becomes $\beta_D = kD\theta/2$, and $\int_0^u z J_0(z)\, dz = u\, J_1(u)$, a property of the Bessel function of order one that follows from their power series.

The variable β_D is the same kind as β_a for the slit, with the radius in place of the half-width. For $\theta \to 0$, $2J_1(\beta_D)/\beta_D \to 1$. Straight ahead every wavelet is in phase, and the field is E_0 times the area.

The Airy pattern. Squaring, with $I_0 = E_0^2(\pi D^2/4)^2$, gives the **Airy pattern**,

$$I(\theta) = I_0 \left[\frac{2J_1(\beta_D)}{\beta_D}\right]^2, \qquad \beta_D = \frac{\pi D \sin\theta}{\lambda} \approx \frac{\pi D\theta}{\lambda}. \tag{5.44}$$

Half the peak intensity is reached at $\beta_D = 1.616$, so the full width at half maximum (FWHM) is $2 \times 1.616/\pi = 1.03\,\lambda/D$.

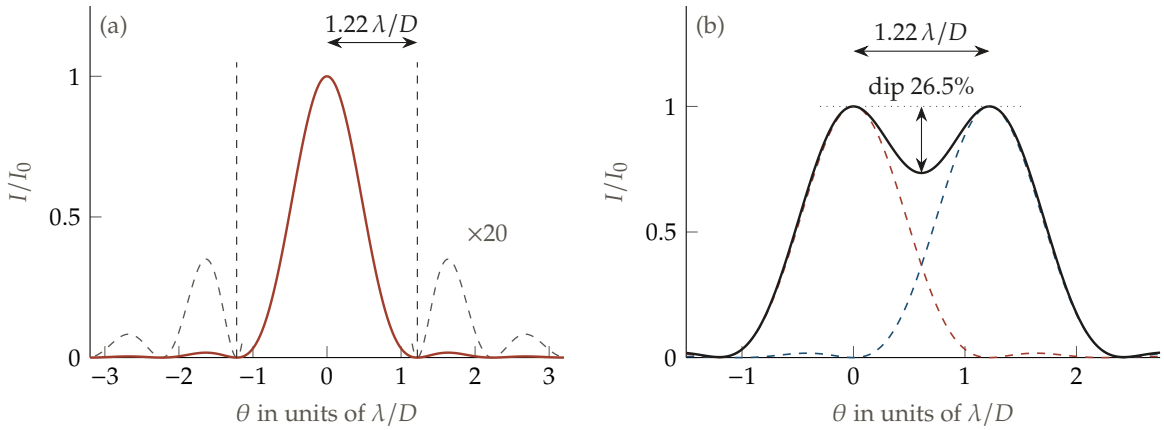


Figure 5.9: (a) The Airy pattern of a round mirror against the angle θ in units of λ/D (red), and the same pattern twenty times higher outside the central disc (grey dashed), with the first dark ring marked. (b) Two equal stars $1.22\lambda/D$ apart, their two patterns (dashed) and the sum (black).

The first dark ring and the factor 1.22. The intensity first vanishes where $J_1(\beta_D) = 0$ for $\beta_D > 0$. The first positive zero of J_1 is $\beta_D = 3.8317$. So the angular radius of the first dark ring is

$$\frac{\pi D\theta}{\lambda} = 3.8317 \quad\Longrightarrow\quad \theta = \frac{3.8317}{\pi}\,\frac{\lambda}{D} = 1.2197\,\frac{\lambda}{D}. \tag{5.45}$$

This is where the factor 1.22 comes from. A second star placed on this ring is the **Rayleigh limit** θ_R of a round telescope,

Result 5.5

$$\theta_R = 1.22\frac{\lambda}{D}\ \text{rad} = 206\,265 \times 1.22\frac{\lambda}{D}\ \text{arcsec}. \tag{5.46}$$

Only the ratio λ/D enters. The focal length, the design of the telescope and the brightness of the star do not.

How the light is shared among the rings. Adding up $I(\theta)\,2\pi\theta\,d\theta$ gives the share of the light inside each ring. The central disc, out to the first dark ring, holds 83.8% of the light, and the disc out to the second dark ring holds 91.0%. So the first bright ring holds only about 7%, and its peak is 1.75% of the central intensity, which is why Figure 5.9(a) draws the rings twenty times higher. The whole pattern covers a solid angle of about $\lambda^2/(\pi D^2/4)$, the wavelength squared over the collecting area, a rule radio engineers use for the beam of any antenna.

A central obstruction. A Cassegrain secondary of diameter b blocks the middle of the mirror. The open ring is the full disc minus a small disc, so its field is Eq. (5.43) for D minus the same for b. The central disc becomes a little smaller and more light goes into the rings. For $b/D = 0.33$

the first dark ring moves in to $1.10\lambda/D$ and the central disc holds 65% of the light, against 84% without it. Brighter rings lower the contrast near a bright star, which matters when imaging faint planets beside it (Section 5.6.4). The vanes that hold the secondary add the spikes of Section 5.3.3.

5.4.5 Resolution Criteria for Point-Source Imaging

Two stars, such as the two of a binary, send light with unrelated phases. Their intensities add, not their fields. For two equal stars a separation $\Delta\theta$ apart, with the profile $I(\theta)$ of one star,

$$I_{\text{tot}}(\theta) = I\left(\theta + \frac{\Delta\theta}{2}\right) + I\left(\theta - \frac{\Delta\theta}{2}\right). \tag{5.47}$$

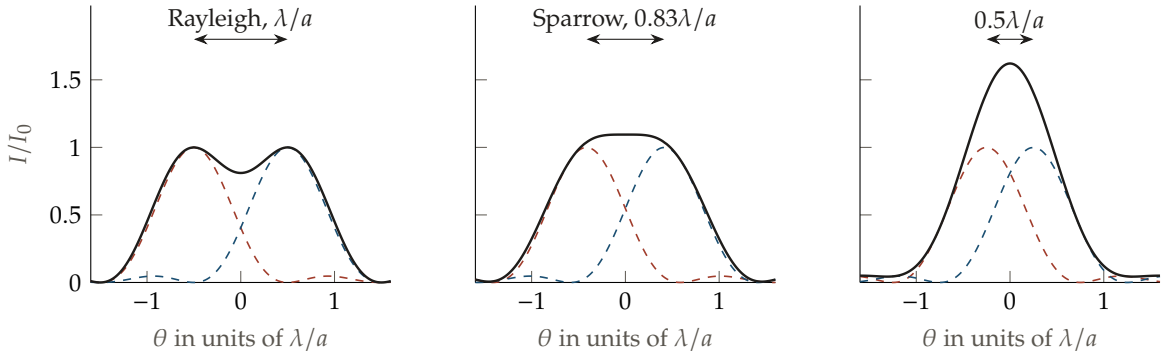


Figure 5.10: The sum of two equal slit patterns at the separations λ/a (Rayleigh, with a 19% dip), $0.83\lambda/a$ (Sparrow, with a flat top) and $0.5\lambda/a$ (a single peak). Dashed curves show the two patterns and solid curves their sum.

The Rayleigh criterion. Rayleigh called two stars resolved when the peak of one falls on the first dark ring of the other, $\Delta\theta = \theta_R$. At the place of each star the sum is then $I_0 + 0 = I_0$. Midway, each pattern is at its half-separation point, and the sum is $2I(\Delta\theta/2)$.

- For a slit, $\Delta\theta = \lambda/a$ and each pattern is at $\beta_a = \pi/2$. The middle is at $2(\sin(\pi/2)/(\pi/2))^2 = 8/\pi^2 = 0.811$ of the peak, a dip of 19%.
- For a round mirror, $\Delta\theta = 1.22\lambda/D$ and each pattern is at $\beta_D = 3.8317/2 = 1.916$. The middle is at $2\,[2J_1(1.916)/1.916]^2 = 0.735$ of the peak, a dip of 26.5% (Figure 5.9(b)).

Rayleigh's choice is convenient because the dark ring is easy to compute. A dip of this size is easy to see by eye at the eyepiece.

A point or a disc. An object smaller than θ_R looks the same as a point. All its light falls inside one Airy core, so its image is the Airy pattern of a star, and nothing in it shows the object's size or shape. The same criterion tells when a planet shows a disc. The two opposite edges of the disc act as the two points, and they are one angular diameter apart, so the disc is resolved when its angular diameter, not its radius, exceeds θ_R. Jupiter at opposition is $5.20 - 1.00 = 4.20\,\text{AU} = 6.28 \times 10^{11}\,\text{m}$ away, and its diameter of $1.43 \times 10^8\,\text{m}$ spans

$$\frac{1.43 \times 10^8\,\text{m}}{6.28 \times 10^{11}\,\text{m}} = 2.28 \times 10^{-4}\,\text{rad} = 46.9''.$$

A pupil about 5 mm wide at night has $\theta_R = 1.22 \times 550 \times 10^{-9}/(5 \times 10^{-3})\,\text{rad} = 27.7''$ at 550 nm. The diameter is larger, so diffraction alone would let the eye see Jupiter as a disc. The radius,

23.5″, is smaller, and comparing it would wrongly call Jupiter a point. In practice the eye does not reach its diffraction limit. The cones at the centre of the retina are about 2.5 μm apart behind a lens of focal length about 17 mm, which is 30″ per cone, and the eye resolves about 1′. The sampling rule of Section 5.5 would ask for $\lambda/2D = 11''$ per cone, so the retina, not the pupil, sets the limit. To the naked eye Jupiter looks like a bright star, and a small telescope shows its disc.

The Sparrow criterion. Two stars stop showing a dip at all when the sum has a flat top in the middle. By Eq. (5.47) the curvature there is $2I''(\Delta\theta/2)$, since the profile is even, so the top turns flat when each profile is at its inflection point. The **Sparrow limit** is twice that angle, $0.83\,\lambda/a$ for a slit and $0.95\,\lambda/D$ for a round mirror. Neither limit is a hard wall. With a high signal-to-noise ratio and a well-measured PSF, two overlapping stars can be separated by fitting, down to the Sparrow limit and below.

Worked example 5.3: Diffraction limit of a 30 cm telescope at 500 nm

For $D = 0.30$ m and $\lambda = 500$ nm, Eq. (5.46) gives

$$\theta_R = 1.2197 \times \frac{500 \times 10^{-9}}{0.30}\ \text{rad} = 2.033 \times 10^{-6}\ \text{rad} = 0.42''.$$

5.5 Comparing Pixel Scale to the Diffraction Limit: Nyquist Sampling

A telescope of diameter D blurs a star into an Airy pattern of radius $\theta_R = 1.22\lambda/D$, Eq. (5.46), and each pixel covers the angle θ_{pix}, Eq. (5.26).

A detector keeps one number per pixel. If the pixels are too coarse, fine detail is lost. Worse, detail that is not there can appear.

In old films the wheels of a moving wagon sometimes seem to turn slowly backwards. The camera takes 24 frames a second. If a spoke moves almost one spoke gap between frames, the eye joins each spoke to the one just behind it, and the wheel seems to creep the wrong way.

The pixels of a detector are the frames, and the star image is the wheel. The idea is that a pattern is recorded faithfully only if there are at least two samples in each period of its finest ripple.

A wheel with 12 spokes turning 1.9 times a second has spokes passing at $12 \times 1.9 = 22.8$ a second. At 24 frames a second this is read as $22.8 - 24 = -1.2$ a second, a slow backward turn.

The picture is a smooth curve with dots at equal steps along it, and a fast wiggle that passes through the very same dots as a slow one.

Compact discs sample sound 44 100 times a second because people hear up to about 20 000 Hz. A digital oscilloscope obeys the same rule. Here the pixel is compared with the finest period in the image, P6 once more. It gives the finest ripple a telescope image can contain, the sampling theorem, and the largest pixel that keeps the image whole.

5.5.1 The Finest Detail in the Image

We want the finest ripple in the image of a telescope. Describe the image by its spatial frequencies $\vec{u} = (u_x, u_y)$, in cycles per radian, through the Fourier transform $\tilde{I}(\vec{u}) = \int I(\vec{\theta})\, e^{-2\pi i \vec{u}\cdot\vec{\theta}}\, d^2\theta$, with $\vec{\theta} = (\theta_x, \theta_y)$. Eq. (5.33) gives $I = EE^*$ as a double integral over two points $\vec{r}$ and $\vec{r}'$ of the

opening. Then

$$\tilde{I}(\vec{u}) \overset{(1)}{=} E_0^2 \iint t(\vec{r})\, t(\vec{r}') \left[\int e^{-2\pi i \vec{\theta}\cdot[(\vec{r}-\vec{r}')/\lambda + \vec{u}]}\, d^2\theta \right] d^2r\, d^2r'$$
$$\overset{(2)}{=} E_0^2 \iint t(\vec{r})\, t(\vec{r}')\, \lambda^2\, \delta^2\big(\vec{r} - \vec{r}' + \lambda\vec{u}\big)\, d^2r\, d^2r' \overset{(3)}{=} E_0^2 \lambda^2 \int t(\vec{r})\, t(\vec{r} + \lambda\vec{u})\, d^2r. \tag{5.48}$$

(1) E and E^* from Eq. (5.33), with $k(\xi\theta_x + \eta\theta_y) = 2\pi\vec{r}\cdot\vec{\theta}/\lambda$, and the order of integration swapped.
(2) $\int e^{-2\pi i\vec{\theta}\cdot\vec{q}}\, d^2\theta = \delta^2(\vec{q})$, and $\delta^2(\vec{x}/\lambda) = \lambda^2\delta^2(\vec{x})$ in two dimensions.
(3) The delta function sets $\vec{r}' = \vec{r} + \lambda\vec{u}$.

Since t is 1 or 0, the last integral is the area where the opening overlaps a copy of itself shifted by $\lambda\vec{u}$ (Figure 5.11a). Two discs of diameter D stop overlapping once their centres are D apart, so the overlap falls to zero at $u = D/\lambda$ (Figure 5.11b). So the image of any scene through a round mirror contains no ripple finer than

$$\boxed{u_{\max} = \frac{D}{\lambda} \text{ cycles per radian.}} \tag{5.49}$$

Each pair of points in the mirror a distance λu apart acts as the pair of antennas of Eq. (5.11), with baseline λu. A mirror measures every baseline inside it at once, up to its diameter D. An interferometer measures only the few its antennas give, but out to the full B.

5.5.2 The Nyquist Sampling Criterion

Take one row of pixels along θ, a distance θ_{pix} apart. The recorded image is the true one times a row of spikes at the pixel centres, $I(\theta)\sum_j \delta(\theta - j\theta_{\text{pix}})$. The row of spikes repeats every θ_{pix}, so it is a Fourier series,

$$\sum_{j=-\infty}^{\infty} \delta(\theta - j\theta_{\text{pix}}) = \frac{1}{\theta_{\text{pix}}} \sum_{j'=-\infty}^{\infty} e^{2\pi i j'\theta/\theta_{\text{pix}}}, \tag{5.50}$$

where each coefficient is $(1/\theta_{\text{pix}})\int_{-\theta_{\text{pix}}/2}^{\theta_{\text{pix}}/2} \delta(\theta)\, e^{-2\pi i j'\theta/\theta_{\text{pix}}} d\theta = 1/\theta_{\text{pix}}$. The transform of the recorded image is

$$\int I(\theta) \sum_j \delta(\theta - j\theta_{\text{pix}})\, e^{-2\pi i u\theta}\, d\theta \overset{(1)}{=} \frac{1}{\theta_{\text{pix}}} \sum_{j'} \int I(\theta)\, e^{-2\pi i(u - j'/\theta_{\text{pix}})\theta}\, d\theta \overset{(2)}{=} \frac{1}{\theta_{\text{pix}}} \sum_{j'} \tilde{I}\left(u - \frac{j'}{\theta_{\text{pix}}}\right). \tag{5.51}$$

(1) Eq. (5.50), and the two exponentials combined.
(2) The definition of $\tilde{I}$, shifted by j'/θ_{pix}.

Sampling makes copies of the true spectrum, spaced $1/\theta_{\text{pix}}$ apart. Each copy runs from $-u_{\max}$ to $+u_{\max}$. They do not overlap if $1/\theta_{\text{pix}} \geq 2u_{\max}$ (Figure 5.11c). Then the copy with $j' = 0$ is the true spectrum, untouched, between $\pm 1/(2\theta_{\text{pix}})$, and the image can be rebuilt exactly. Keeping only that copy and transforming back gives

$$I(\theta) = \sum_j I(j\theta_{\text{pix}})\, \frac{\sin\big(\pi(\theta - j\theta_{\text{pix}})/\theta_{\text{pix}}\big)}{\pi(\theta - j\theta_{\text{pix}})/\theta_{\text{pix}}}, \tag{5.52}$$

since a flat band of width $1/\theta_{\text{pix}}$ transforms into the slit pattern of Eq. (5.35). If the copies overlap, a ripple at a frequency u just above $1/(2\theta_{\text{pix}})$ is recorded as one at $u - 1/\theta_{\text{pix}}$. At every

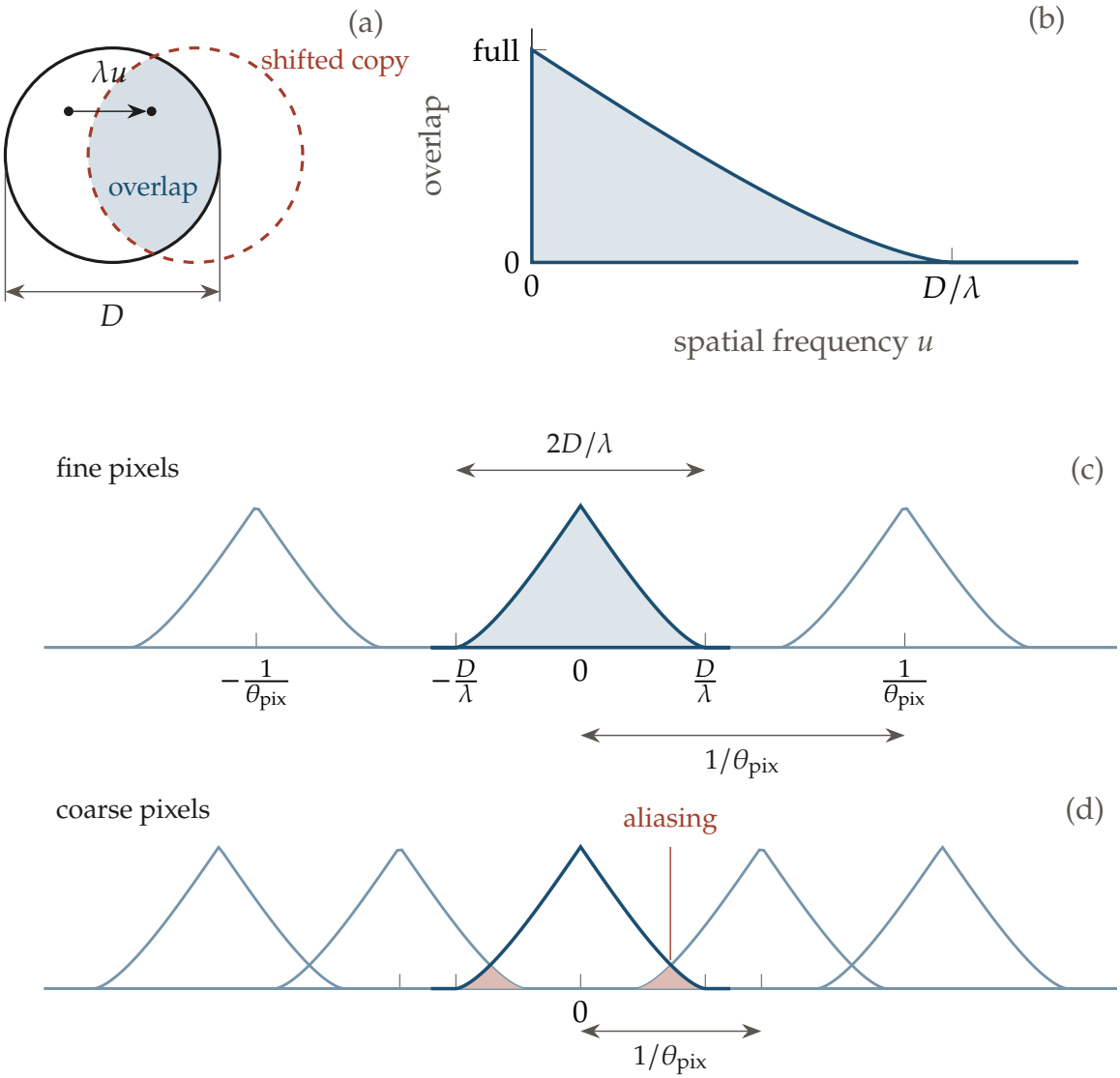


Figure 5.11: (a) A round mirror of diameter D (solid) and a copy of it shifted by λu (dashed), with their overlap shaded. (b) The overlap area against the spatial frequency u, for the image of a star. (c) The spectrum of the sampled image, with copies of (b) every $1/\theta_{\text{pix}}$, for pixels finer than $\lambda/2D$. (d) The same for pixels coarser than $\lambda/2D$, with the overlap of neighbouring copies shaded.

pixel, $\cos\bigl(2\pi u j\theta_{\text{pix}}\bigr) = \cos\bigl(2\pi(u - 1/\theta_{\text{pix}})j\theta_{\text{pix}}\bigr)$, because the two phases differ by $2\pi j$, a whole number of turns. This false ripple is **aliasing**, the backward wagon wheel (Figure 5.11d). Real pixels collect light over their whole width rather than at a point. That multiplies the spectrum by a slit pattern, which lowers the fine ripples a little but does not change the condition.

With $u_{\text{max}} = D/\lambda$ from Eq. (5.49), the **Nyquist–Shannon** condition $\theta_{\text{pix}} \le 1/(2u_{\text{max}})$ becomes

Result 5.6

$$\theta_{\text{pix}} \le \frac{\lambda}{2D} = \frac{\theta_R}{2.44}. \tag{5.53}$$

The FWHM of the Airy pattern is $1.03\lambda/D$, so this asks for about **two pixels across the FWHM**. The common rule of two pixels across θ_R itself is a little too coarse. A row along the other axis

of the detector obeys the same condition.

5.5.3 Configuration Critique: Comparison of Pixel Scale and Diffraction Limit

The 30 cm telescope of Figure 5.5, with a flat secondary, $F = 2\,\text{m}$ and 20 μm pixels, has

$$\theta_{\text{pix}} = 2.06'' \text{ per pixel}, \qquad \theta_R = 0.42'' \text{ at } 500\,\text{nm}.$$

Their ratio is

$$\boxed{\frac{\theta_{\text{pix}}}{\theta_R} = \frac{2.06''}{0.42''} = 4.9.} \tag{5.54}$$

One pixel spans almost five Airy radii. Eq. (5.53) asks for $\theta_{\text{pix}} \le \lambda/(2D) = 0.17''$, twelve times finer than the pixels give. The same comparison in the focal plane reads more simply. The first dark ring has the radius $1.22\lambda F/D = 4.1\,\mu$m, and Nyquist sampling would need pixels of $\lambda F/(2D) = 1.7\,\mu$m. The whole Airy disc falls inside one 20 μm pixel. The image is badly *undersampled*. Its fine detail is aliased, and the position of a star cannot be measured to a fraction of the diffraction blur.

Keeping these pixels, Nyquist sampling at 500 nm would need an effective focal length of $20\,\mu\text{m}/(\lambda/2D) = 24\,\text{m}$, which a convex secondary of magnification 12 could give, Eq. (5.22). From the ground, the seeing of Section 5.6, about 1″, is the real blur. Two pixels across 1″ need $\theta_{\text{pix}} \le 0.5''$, so $F_{\text{eff}} \ge 8.3\,\text{m}$. Even then this telescope at $F = 2\,\text{m}$ undersamples by a factor of four.

Sampling regime	Condition	Consequences
Critically sampled	$\theta_{\text{pix}} \approx \lambda/(2D)$	Keeps every spatial frequency without wasting detector area.
Undersampled	$\theta_{\text{pix}} > \lambda/(2D)$	Aliasing, blocky star images, and poorer star positions.
Oversampled	$\theta_{\text{pix}} \ll \lambda/(2D)$	Spreads the light over many pixels, each adding read noise, $\sqrt{n_{\text{pix}}}\,\sigma_{\text{RON}}$ in total (Eq. (7.18)).

Table 5.1: The three sampling regimes of a detector, for a diffraction-limited image.

5.6 Atmospheric Seeing, PSF Profiles, and Adaptive Optics

The optical window lets visible light through the air (Section 5.1), and a telescope resolves $1.22\lambda/D$ when the wavefront reaching it is flat, Eq. (5.46). From the ground the wavefront is not flat. Transparent air is not still air. On an ordinary night an 8 m telescope shows stars no sharper than a 10 cm one.

Look at a coin at the bottom of a swimming pool while the surface ripples. The coin wobbles and blurs. Each ripple acts as a weak lens and bends the light a little, and the ripples keep moving. Over a hot road in summer the far cars shimmer for the same reason. The air above a telescope is a pool of warm and cool pockets, each with a slightly different refractive index. The idea is that the sharpness is set by the size r_0 of the patch of air over which the wavefront stays flat to about a radian, and not by the size of the mirror.

At 500 nm, air with $r_0 = 10\,\text{cm}$ gives a blur of about $0.98\lambda/r_0 = 1.0''$. An 8 m mirror would reach $1.22\lambda/D = 0.016''$ in steady air, about sixty times sharper. Across a patch of size r_0 the arriving wavefront is nearly flat, but different patches are tilted differently. A mirror larger than r_0 gathers more light without recovering its diffraction-limited sharpness unless those

tilts are corrected. We will see how the blur depends on wavelength and how adaptive optics makes that correction.

5.6.1 Atmospheric Seeing and the Fried Parameter

Turbulence mixes air of slightly different temperatures and so of different refractive index. A wavefront that left the star flat reaches the ground with small random extra paths across it. The **Fried parameter** r_0 is the distance across the wavefront over which the phase difference grows to about one radian. Typical values are 10 to 20 cm in the V band at the best mountain sites. Patches closer than r_0 still add in phase as in an undisturbed Airy pattern. Patches farther apart have unrelated phases. So a mirror larger than r_0 acts like about $(D/r_0)^2$ separate small mirrors of size r_0, and the blur is that of diffraction by r_0, not by D.

How r_0 changes with wavelength. Kolmogorov's theory of turbulence, which we quote, gives the mean square difference in extra path between two points a distance r apart as $C\,r^{5/3}$. The constant C measures the strength of the turbulence, and it hardly depends on wavelength, because the refractive index of air changes little across the optical and infrared. The phase is $k = 2\pi/\lambda$ times the path. Fried wrote the mean square phase difference as $6.88\,(r/r_0)^{5/3}$, with the number chosen so that r_0 is the diameter of a telescope whose diffraction blur matches the seeing. Then

$$6.88\left(\frac{r}{r_0}\right)^{5/3} \overset{(1)}{=} \left(\frac{2\pi}{\lambda}\right)^2 C\,r^{5/3} \quad\Longrightarrow\quad r_0 \overset{(2)}{=} \left(\frac{6.88\,\lambda^2}{4\pi^2 C}\right)^{3/5} \propto \lambda^{6/5}. \tag{5.55}$$

(1) The mean square of k times the path difference.
(2) Cancel $r^{5/3}$ and raise both sides to the power $-3/5$.

For a mirror larger than r_0, the full width at half maximum of the long-exposure star image is, from the same theory,

Result 5.7

$$\theta_{\text{see}} \approx 0.98\frac{\lambda}{r_0} \approx 0.5''\text{–}1.5''. \tag{5.56}$$

It is a width at half maximum, to be compared with the FWHM $1.03\lambda/D$ of the Airy pattern, not with its first dark ring. With $r_0 \propto \lambda^{6/5}$, $\theta_{\text{see}} \propto \lambda^{-1/5}$, so the seeing improves slowly toward the infrared. If $r_0 = 10$ cm at 500 nm, then at 2.2 μm (the K band) $r_0 = 10 \times 4.4^{6/5}$ cm $= 59$ cm and $\theta_{\text{see}} = 0.75''$, against $1.0''$ at 500 nm.

5.6.2 Point Spread Function (PSF) Profiles: The Moffat Function

Seeing turns the Airy pattern into a wider round blob with broad wings. Astronomers fit it with the **Moffat function**,

$$I(\theta) = I_0\left[1 + \left(\frac{\theta}{\theta_{\text{M}}}\right)^2\right]^{-\beta_{\text{M}}}, \tag{5.57}$$

with a core width θ_{M} and a power β_{M}. Fits to real stars give β_{M} between about 2.5 and 4.5. Far from the core the profile falls as the power law $\theta^{-2\beta_{\text{M}}}$, much more slowly than a Gaussian. The

width at half maximum follows from setting $I = I_0/2$, at the half width $\theta_{1/2}$,

$$\left(\frac{\theta_{1/2}}{\theta_{\mathrm{M}}}\right)^2 \overset{(1)}{=} 2^{1/\beta_{\mathrm{M}}} - 1,$$

$$\mathrm{FWHM} = 2\theta_{1/2} \overset{(2)}{=} 2\theta_{\mathrm{M}}\sqrt{2^{1/\beta_{\mathrm{M}}} - 1}. \tag{5.58}$$

(1) $[1 + (\theta_{1/2}/\theta_{\mathrm{M}})^2]^{-\beta_{\mathrm{M}}} = \frac{1}{2}$ from Eq. (5.57), raised to the power $-1/\beta_{\mathrm{M}}$, then 1 subtracted.
(2) The square root, and the full width is twice the half width.

As $\beta_{\mathrm{M}} \to \infty$ with $\theta_{\mathrm{M}}^2/\beta_{\mathrm{M}}$ fixed, $(1 + \theta^2/\theta_{\mathrm{M}}^2)^{-\beta_{\mathrm{M}}} \to \exp\left(-\beta_{\mathrm{M}}\theta^2/\theta_{\mathrm{M}}^2\right)$, a Gaussian. A finite β_{M} keeps the wings.

5.6.3 Adaptive Optics (AO)

Adaptive optics removes the blur of the air as it happens.

1. A **wavefront sensor**, such as a Shack–Hartmann array of small lenses, measures the tilt of the wavefront over each patch, using a bright natural guide star or a laser guide star. The laser excites sodium atoms in a layer at about 90 km height, at 589 nm, one of the D lines resolved above.
2. A fast computer works out the shape that cancels the measured errors.
3. A **deformable mirror**, with hundreds to thousands of actuators, takes that shape and removes the extra paths.

The numbers follow from r_0. An 8 m mirror with $r_0 = 10$ cm has $(8/0.1)^2 = 6400$ patches, which is why so many actuators are needed, and fewer in the infrared, where r_0 is larger. The wind blows the pattern past at about $10\,\mathrm{m\,s^{-1}}$, so it changes in about $0.31\,r_0/(10\,\mathrm{m\,s^{-1}}) = 3$ ms, and the loop must run at about a kilohertz. With the correction on, 8 to 30 m telescopes recover a core of width close to λ/D from the ground.

5.6.4 Imaging Planets Directly

Most planets are found through their star, by its wobble or by the dip in its light as the planet crosses it (Sections 6.19 and 6.20). To take a picture of the planet itself, two things must hold at once. The planet must lie outside the blur of the star, and its light must stand out against the star's light at that place. The first is a comparison of two angles and the second a comparison of two brightnesses, P6 twice.

The separation. Let the planet orbit at the distance a from its star, the semi-major axis of Chapter 6 (not the slit width of Section 5.4). Let the system lie at the distance d from us. Seen face on, the planet is the angle a/d from the star. With a in AU and d in parsecs,

$$\theta = \frac{a}{d}\ \mathrm{rad} \overset{(1)}{=} \frac{a/(1\,\mathrm{AU})}{d/(1\,\mathrm{pc})}\ \mathrm{arcsec}. \tag{5.59}$$

(1) One parsec is the distance at which 1 AU covers one second of arc, Eq. (3.2).

Jupiter, at 5.20 AU, seen from 10 pc is 0.52″ from the Sun. The Earth from 10 pc is 0.10″ from it. The planet β Pictoris b, about 9 to 10 AU from a star 19.4 pc away, is at most about 0.5″ from it.

All three lie inside the seeing blur of about 1″, Eq. (5.56). So without adaptive optics a planet is lost in the star's light before brightness even matters. With adaptive optics an 8 m mirror reaches its diffraction limit. At 1.6 μm that is $\lambda/D = 1.6 \times 10^{-6}/8 = 2.0 \times 10^{-7}$ rad = 0.041″, and at 3.8 μm it is 0.098″. Jupiter at 10 pc then sits $0.52/0.041 = 12.6$ times λ/D from its star at 1.6 μm, and β Pic b sits $0.46/0.098 = 4.7$ times λ/D away at 3.8 μm (for $a = 9$ AU). The Earth at 10 pc is only $2.4\,\lambda/D$ out at 1.6 μm, on the first bright ring of the Airy pattern, Eq. (5.44).

Reflected light. A planet with no heat of its own shines by the starlight it reflects. As for the equilibrium temperature, Eq. (6.245), a star of luminosity L_* sends the flux $L_*/(4\pi a^2)$ onto the planet's cross-section πR_p^2, and the planet sends back the fraction A_B, its Bond albedo. Both reach us from the same distance d, so the ratio of the planet's flux to the star's is (P1, P6)

$$\frac{F_p}{F_*} \overset{(1)}{=} \frac{A_\mathrm{B}\,\dfrac{L_*}{4\pi a^2}\,\pi R_p^2 \Big/ \left(4\pi d^2\right)}{L_* \Big/ \left(4\pi d^2\right)} \overset{(2)}{=} A_\mathrm{B}\,\frac{\pi R_p^2}{4\pi a^2} \overset{(3)}{=} A_\mathrm{B}\left(\frac{R_p}{2a}\right)^2. \tag{5.60}$$

(1) The reflected power is spread evenly over all directions, so at the distance d it is divided by $4\pi d^2$, like the star's light, Eq. (4.3).
(2) L_* and $4\pi d^2$ cancel.
(3) $\pi R_p^2/(4\pi a^2) = R_p^2/(4a^2)$.

$$\boxed{\frac{F_p}{F_*} = A_\mathrm{B}\left(\frac{R_p}{2a}\right)^2} \tag{5.61}$$

The ratio is the fraction of the star's light that the planet's disc catches, times its albedo. It does not depend on the distance to us or on the star. A real planet sends more light back towards the star than to the side, and shows phases, which changes the ratio by a factor of order one.

For Jupiter, with $A_\mathrm{B} = 0.34$, $R_p = 7.15 \times 10^7$ m and $a = 5.20\,\mathrm{AU} = 7.78 \times 10^{11}$ m,

$$\frac{F_p}{F_*} = 0.34 \times \left(\frac{7.15 \times 10^7}{2 \times 7.78 \times 10^{11}}\right)^2 = 0.34 \times \left(4.60 \times 10^{-5}\right)^2 = 0.34 \times 2.11 \times 10^{-9} = 7.2 \times 10^{-10}.$$

For the Earth, with $A_\mathrm{B} = 0.30$, $R_p = 6.37 \times 10^6$ m and $a = 1.496 \times 10^{11}$ m,

$$\frac{F_p}{F_*} = 0.30 \times \left(\frac{6.37 \times 10^6}{2 \times 1.496 \times 10^{11}}\right)^2 = 0.30 \times \left(2.13 \times 10^{-5}\right)^2 = 0.30 \times 4.53 \times 10^{-10} = 1.4 \times 10^{-10}.$$

This is the "billion times fainter" of Section 6.19. A planet further out is easier to separate from its star but reflects less, as $1/a^2$.

The planet's own heat. A giant planet is born hot. Like a young star, it shines by slowly contracting (Section 10.5), and it cools over hundreds of millions of years. A giant a few tens of millions of years old is still at 1000 to 2000 K. A surface at temperature T sends out $\pi B_\lambda(T)$ per unit area and wavelength, Eq. (4.20), and a sphere of radius R seen from the distance d gives the flux $\pi B_\lambda(T)(R/d)^2$, Eq. (4.3). At one wavelength λ, for a planet at T_p and a star of radius R_* at T_*,

$$\frac{F_{p,\lambda}}{F_{*,\lambda}} \overset{(1)}{=} \frac{\pi B_\lambda(T_p)\,(R_p/d)^2}{\pi B_\lambda(T_*)\,(R_*/d)^2} \overset{(2)}{=} \left(\frac{R_p}{R_*}\right)^2 \frac{B_\lambda(T_p)}{B_\lambda(T_*)} \overset{(3)}{=} \left(\frac{R_p}{R_*}\right)^2 \frac{e^{x_*}-1}{e^{x_p}-1}, \qquad x_p = \frac{hc}{\lambda k_B T_p}, \quad x_* = \frac{hc}{\lambda k_B T_*}. \tag{5.62}$$

(1) The flux of each body at the Earth.
(2) π and d^2 cancel.
(3) The Planck function per unit wavelength, Eq. (4.15). The factor $2hc^2/\lambda^5$ is the same for both and cancels. x is the photon energy over $k_B T$, as in Chapter 4.

Only the two radii and the two temperatures enter. At long wavelengths both x are small, $e^x - 1 \approx x$, and the ratio rises to its ceiling $(R_p/R_*)^2\, T_p/T_*$. At short wavelengths x_p is large and the ratio falls as e^{-x_p}. By Wien's law, Eq. (4.19), a planet at 1700 K peaks at $2898/1700 = 1.70\,\mu$m and a star at 8000 K at $0.36\,\mu$m. So the planet is best seen in the near infrared, where it is near its peak and the star is far past its own.

Take a young giant like β Pic b, with $T_p = 1700$ K and 1.5 times Jupiter's radius, $R_p = 1.07\times10^8$ m, beside an A star with $T_* = 8000$ K and $R_* = 1.5\,R_\odot = 1.04\times10^9$ m. Then $(R_p/R_*)^2 = 0.1027^2 = 1.06\times10^{-2}$. With $hc/k_B = 0.014388$ m K, at $\lambda = 3.8\,\mu$m,

$$x_p = \frac{0.014388}{3.8\times10^{-6}\times1700} = 2.227, \qquad x_* = \frac{0.014388}{3.8\times10^{-6}\times8000} = 0.473,$$

$$\frac{F_{p,\lambda}}{F_{*,\lambda}} = 1.06\times10^{-2}\times\frac{e^{0.473}-1}{e^{2.227}-1} = 1.06\times10^{-2}\times\frac{0.605}{8.27} = 7.7\times10^{-4}.$$

This is below the ceiling $1.06\times10^{-2}\times1700/8000 = 2.2\times10^{-3}$, as it must be. In visible light, at $0.55\,\mu$m, the same steps give $x_p = 15.4$, $x_* = 3.27$ and a ratio of only 5.5×10^{-8}. The same planet in reflected light, by Eq. (5.61) with $A_B = 0.3$ at 9 AU, gives 4.8×10^{-10}. Its own heat makes it about a million times brighter.

Jupiter today is at about 125 K. Around the Sun, at 5772 K, Eq. (5.62) gives 7×10^{-16} at $3.8\,\mu$m, so at a few microns an old cold planet shows only its reflected light, near 10^{-9}. Its own heat shows only near $10\,\mu$m, where the ratio is 3.0×10^{-8}. This is why direct imaging finds young, hot, massive planets far from their stars. Young and massive planets are still hot and large, so they glow at 10^{-4} to 10^{-3} of their star in the near infrared, and a wide orbit puts them outside the star's blur.

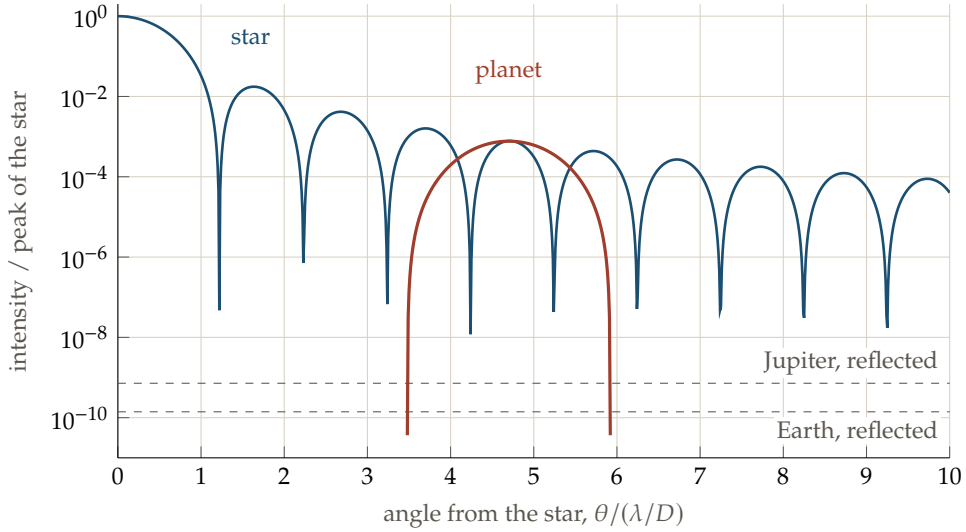


Figure 5.12: The Airy profile of a star (blue) and of a planet $4.7\,\lambda/D$ from it with 7.7×10^{-4} of its peak (red), on a logarithmic intensity axis. The dashed lines mark Jupiter and the Earth around the Sun in reflected light.

Removing the star's light. Figure 5.12 shows the problem left after the angles work. Even a perfect Airy pattern has rings near 10^{-3} of the peak at $5\,\lambda/D$, as bright as the planet. Three steps remove most of the star's light.

1. **Adaptive optics** (Section 5.6.3) gathers the star's light into the Airy core, so that the pattern is the one in the figure and not a seeing blob 1″ wide.

2. A **coronagraph** blocks the core and the rings. A small opaque spot in the focal plane covers the star's core. A lens behind it then forms an image of the mirror. The field at the focus is the Fourier transform of the opening, Eq. (5.33), so removing the core removes the smooth part of the light in this image of the mirror, and what is left gathers in a bright ring along its edge. A stop a little smaller than the mirror's image blocks that ring. The planet's image falls beside the spot, so its light passes, and the stop trims it only a little. A planet closer than the spot's edge, a few λ/D, is hidden. This smallest usable angle is the **inner working angle**.
3. **Angular differential imaging** removes most of what remains. Small errors in the optics leave *speckles*, spots of starlight about λ/D wide that look just like a planet. On an altitude–azimuth telescope with the image rotator switched off, the sky turns on the detector through the night by the parallactic angle (Section 2.10.2), while the speckles from the telescope's own optics stay put. The median of all the frames holds the speckles but not the moving planet. It is subtracted from each frame, each frame is turned back to the same orientation of the sky, and the frames are added. The planet adds up, and what is left of the speckles averages down.

β Pictoris b. β Pictoris is a young A-type star, about 20 million years old and 19.4 pc away, with a large disc of dust seen edge on. Images taken with the NACO camera of the VLT in the L' band at 3.8 μm showed a faint point beside the star, reported in 2008. In 2009 it was seen again, on the other side of the star. Seeing it on both sides traces its orbit. It lies about 9 to 10 AU from the star. By Eq. (5.59) that is at most about 0.5″ on the sky, or about five times λ/D for an 8 m mirror at 3.8 μm. Its brightness and age, through models of how giant planets cool, give a mass of about 10 Jupiter masses. It is the kind of planet the estimates above favour, young, hot, heavy and far from its star, seen in the infrared at a contrast near 10^{-3}. Around the young star HR 8799, four giant planets have been imaged at once, between about 15 and 70 AU from it.

5.7 Problem Set: Interferometric Baselines, Optics, and Resolution

5.7.1 Problem 1: GMRT to Ooty Radio Interferometer Baseline and Resolution

Problem Statement: Find the latitude and longitude of the Giant Metrewave Radio Telescope (GMRT) and of the Ooty Radio Telescope from a map.

(a) What is the straight-line distance between them?

(b) If we do very-long-baseline interferometry between these telescopes at 1000 MHz, what is the best resolution we can get?

Essential Ideas:

- Each station is a unit vector, Eq. (5.3). The central angle γ follows from Eq. (5.4), or better from its haversine form, Eq. (5.5).
- The shortest distance in three dimensions is the chord through the Earth, $B = 2R_E \sin(\gamma/2)$, Eq. (5.7).
- The baseline sets the resolution, Eq. (5.16),

$$\theta \simeq \frac{\lambda}{B}. \tag{5.63}$$

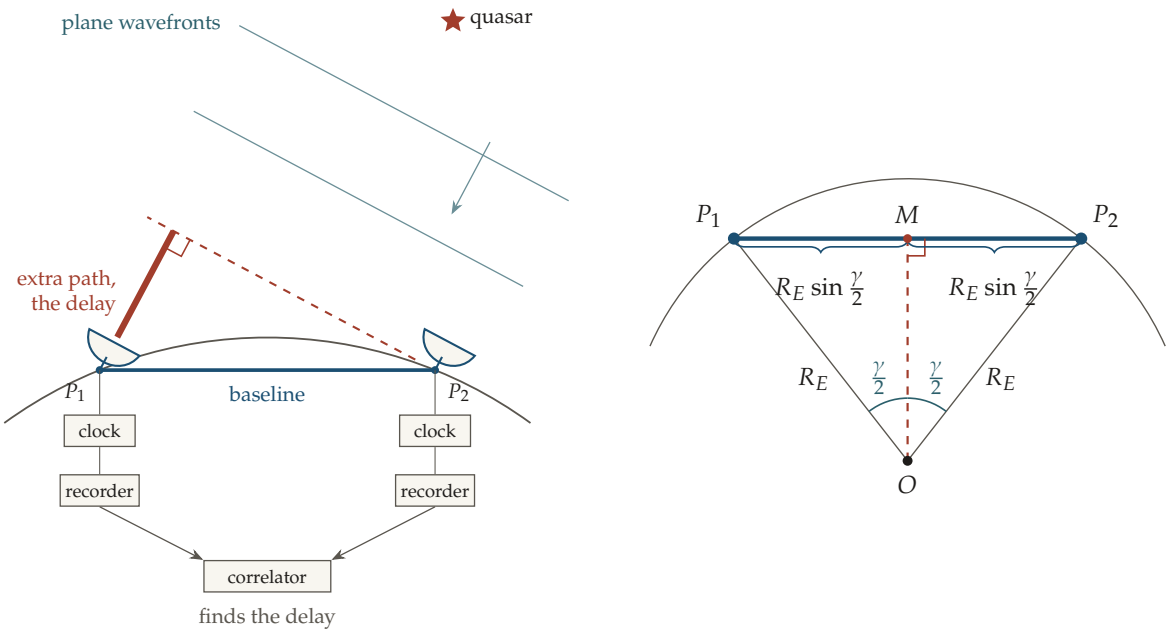


Figure 5.13: Left, a distant quasar sends plane wavefronts to two dishes on the curved Earth, and the extra path to the farther dish is the delay. Right, the chord between two stations P_1 and P_2 on a sphere of radius R_E, with the central angle γ and the midpoint M.

Detailed Solutions:

(a) **Baseline distance**. The coordinates (ϕ, Λ) are about

$$\text{GMRT } (19.09^\circ\text{N},\ 74.05^\circ\text{E}), \qquad \text{Ooty } (11.38^\circ\text{N},\ 76.66^\circ\text{E}).$$

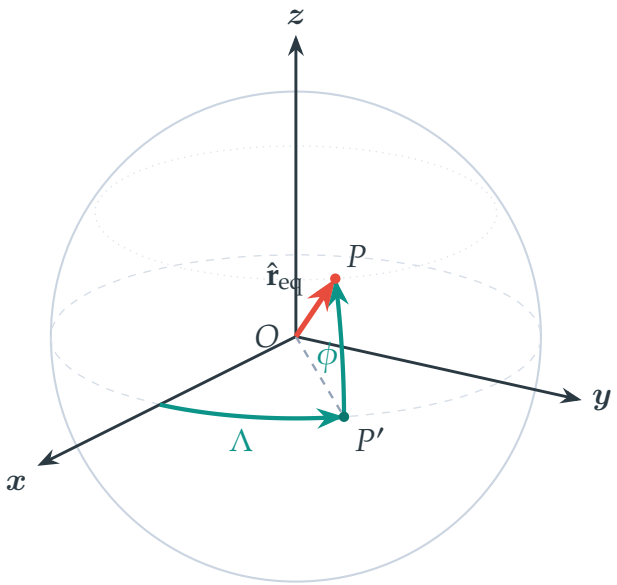


Figure 5.14: A point P on a sphere, its unit vector, its foot P' in the equatorial plane, its latitude ϕ and its longitude (marked λ).

The differences are $\phi_1 - \phi_2 = 7.71^\circ$ and $\Lambda_2 - \Lambda_1 = 2.61^\circ$. With $\cos\phi_1 \cos\phi_2 = 0.94501 \times$

$0.98034 = 0.92643$, Eq. (5.5) gives

$$\sin^2\frac{\gamma}{2} = \sin^2(3.855^\circ) + 0.92643\,\sin^2(1.305^\circ)$$
$$= 0.0045201 + 0.92643 \times 0.00051868 = 0.0050006,$$

so $\sin(\gamma/2) = 0.070715$ and $\gamma = 8.110^\circ$. By Eq. (5.7) with $R_E = 6371\,\text{km}$,

$$\boxed{B = 2 \times 6371\,\text{km} \times 0.070715 = 901.1\,\text{km}.}$$

The cosine form, Eq. (5.4), gives $\cos\gamma = 0.9899988$. Rounded to 0.9899 it would give $B = 905.5\,\text{km}$, 4 km too long. A cosine near 1 loses the digits that matter, which is why the haversine form is used.

(b) **Angular resolution**. At $\nu = 1000\,\text{MHz} = 10^9\,\text{Hz}$ the wavelength is

$$\lambda = \frac{c}{\nu} = \frac{2.998 \times 10^8\,\text{m s}^{-1}}{10^9\,\text{Hz}} = 0.2998\,\text{m}.$$

With $B = 9.011 \times 10^5\,\text{m}$, Eq. (5.63) gives

$$\theta \simeq \frac{0.2998}{9.011 \times 10^5} = 3.327 \times 10^{-7}\,\text{rad}.$$

With $1\,\text{rad} = 2.06265 \times 10^8\,\text{mas}$,

$$\boxed{\theta \simeq 68.6\,\text{mas}.}$$

A single dish would need to be 900 km wide to do as well. The pair records only the sky's ripples at this one baseline, Eq. (5.11), so a real image needs many such pairs.

5.7.2 Problem 2: Cassegrain Reflector Optics, Pixel Scale, and Sampling

Problem Statement: A Cassegrain telescope has a primary mirror of diameter 30 cm and focal length 2 m, and a flat secondary 0.75 m in front of the primary (Figure 5.15).

(a) What are the lengths shown as b and a?

(b) We put a CCD with pixels 20 μm wide at the focal plane. What is the pixel scale?

(c) What is the resolution of this telescope at 500 nm?

(d) Compare the diffraction-limited resolution with the pixel scale. Is this a good configuration?

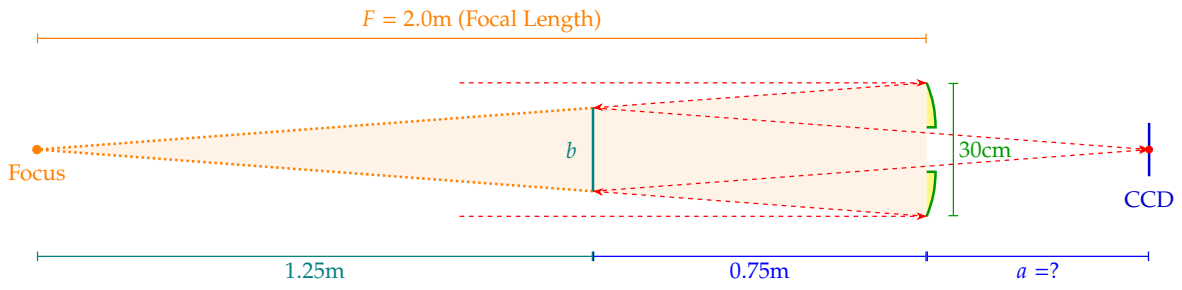


Figure 5.15: The Cassegrain telescope of the problem, with the secondary of diameter b at 0.75 m from the primary and 1.25 m from the primary focus, and the unknown distance a from the primary to the CCD.

Essential Ideas:

- Similar triangles in the converging cone give the size of the secondary, $b = D(1 - d_2/F)$, Eq. (5.20). A flat secondary puts the focus at $a = F - 2d_2$ behind the primary, Eq. (5.21), and leaves the focal length unchanged.
- A pixel of width y_{pix} covers $\theta_{\text{pix}} = y_{\text{pix}}/F$, Eq. (5.27).
- A round mirror resolves $\theta_R = 1.22\lambda/D$, Eq. (5.46), and Nyquist sampling needs $\theta_{\text{pix}} \leq \lambda/(2D)$, Eq. (5.53).

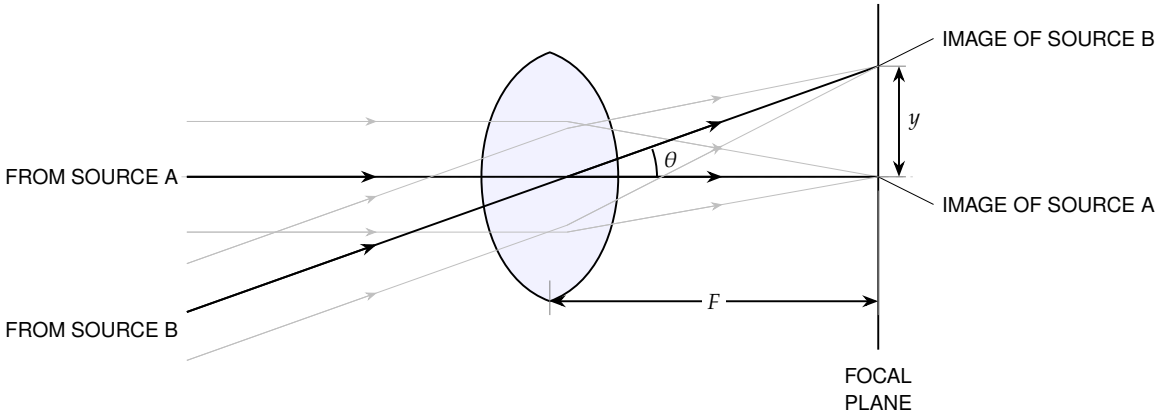


Figure 5.16: Two point sources A and B at the angle θ, a lens of focal length F, and their images a height y apart. The chief rays are dark, and the edge rays faded.

Detailed Solutions:

(a) **The lengths b and a.** With $D = 30\,\text{cm}$, $F = 2.0\,\text{m}$ and $d_2 = 0.75\,\text{m}$, the secondary is $F - d_2 = 1.25\,\text{m}$ from the primary focus. By Eq. (5.20),

$$b = 30\,\text{cm} \times \frac{1.25}{2.0} = 18.75\,\text{cm}.$$

The flat secondary forms the image 1.25 m in front of itself. The primary is 0.75 m away, so by Eq. (5.21)

$$a = 1.25\,\text{m} - 0.75\,\text{m} = 0.50\,\text{m}.$$

$$\boxed{b = 18.75\,\text{cm}, \qquad a = 0.50\,\text{m}.}$$

The secondary blocks $(b/D)^2 = (0.625)^2 = 39\%$ of the mirror's area. A convex secondary closer to the primary focus would block less and lengthen the focal length.

(b) **Pixel scale.** The flat secondary leaves $F = 2\,\text{m}$. With $y_{\text{pix}} = 20 \times 10^{-6}\,\text{m}$, Eq. (5.27) gives

$$\theta_{\text{pix}} = \frac{20 \times 10^{-6}\,\text{m}}{2\,\text{m}} = 1.0 \times 10^{-5}\,\text{rad}.$$

With $1\,\text{rad} = 206\,265''$,

$$\boxed{\theta_{\text{pix}} = 2.06'' \text{ per pixel}.}$$

(c) **Diffraction-limited resolution.** Figure 5.9(b) shows two stars at this separation. With $\lambda = 500 \times 10^{-9}\,\text{m}$ and $D = 0.30\,\text{m}$, Eq. (5.46) gives

$$\theta_R = 1.22 \times \frac{500 \times 10^{-9}}{0.30}\,\text{rad} = 2.03 \times 10^{-6}\,\text{rad}, \qquad \boxed{\theta_R = 2.03 \times 10^{-6} \times 206\,265'' = 0.42''.}$$

(d) **Comparison**. The ratio of the pixel scale to the resolution is

$$\frac{\theta_{\text{pix}}}{\theta_R} = \frac{2.06''}{0.42''} = 4.9.$$

Nyquist sampling needs $\theta_{\text{pix}} \le \lambda/(2D)$, Eq. (5.53),

$$\frac{\lambda}{2D} = \frac{500 \times 10^{-9}}{0.60}\ \text{rad} = 8.3 \times 10^{-7}\ \text{rad} = 0.17''.$$

$$\boxed{\theta_{\text{pix}} = 2.06'' \text{ is 12 times larger than the } 0.17'' \text{ needed.}}$$

It is not a good configuration. One pixel covers almost five Airy radii, so the image is badly undersampled, its fine detail is aliased, and star positions are poor. Even the 1″ seeing on the ground would need $\theta_{\text{pix}} \le 0.5''$, four times finer. A convex secondary that lengthens the focal length to about 8 m would match the pixels to the seeing, and 24 m to the diffraction limit (Section 5.5).

5.7.3 Problem 3: Jupiter and the Earth Seen from Ten Parsecs

Problem Statement: An observer 10 pc away looks at the Sun, with Jupiter at 5.20 AU ($R_p = 7.15 \times 10^7$ m, Bond albedo 0.34) seen face on.

(a) What is the largest angle between Jupiter and the Sun?

(b) What is the ratio of Jupiter's reflected light to the Sun's light?

(c) What mirror diameter puts Jupiter outside $3\lambda/D$ at 1.6 μm?

(d) Repeat (a) and (c) for the Earth at 1 AU.

Essential Ideas:

- The angle is a/d, which in seconds of arc is a in AU over d in parsecs, Eq. (5.59).
- In reflected light the planet catches the fraction $(R_p/2a)^2$ of the star's light and returns the fraction A_{B} of that, Eq. (5.61).
- A coronagraph hides everything within a few λ/D of the star, so the planet needs $\theta \ge 3\lambda/D$, that is $D \ge 3\lambda/\theta$.

Detailed Solutions:

(a) **Separation**. By Eq. (5.59),

$$\theta = \frac{5.20}{10}\ \text{arcsec}, \qquad \boxed{\theta = 0.52'' = \frac{0.52}{206\,265}\ \text{rad} = 2.52 \times 10^{-6}\ \text{rad}.}$$

This is half the typical seeing, so from the ground adaptive optics is needed.

(b) **Contrast**. With $a = 5.20 \times 1.496 \times 10^{11}$ m $= 7.78 \times 10^{11}$ m, Eq. (5.61) gives

$$\frac{F_p}{F_*} = 0.34 \times \left(\frac{7.15 \times 10^7}{2 \times 7.78 \times 10^{11}}\right)^2 = 0.34 \times 2.11 \times 10^{-9}, \qquad \boxed{\frac{F_p}{F_*} = 7.2 \times 10^{-10}.}$$

(c) **Mirror size**. The condition $\theta \ge 3\lambda/D$ gives

$$D \ge \frac{3\lambda}{\theta} = \frac{3 \times 1.6 \times 10^{-6}\ \text{m}}{2.52 \times 10^{-6}} = \frac{4.8 \times 10^{-6}\ \text{m}}{2.52 \times 10^{-6}}, \qquad \boxed{D \ge 1.9\ \text{m}.}$$

The angle is easy. The contrast of part (b) is what makes Jupiter hard.

(d) **The Earth**. The angle is $1/10 = 0.10'' = 4.85 \times 10^{-7}$ rad, and

$$D \geq \frac{4.8 \times 10^{-6}\,\mathrm{m}}{4.85 \times 10^{-7}}, \qquad \boxed{\theta = 0.10'', \qquad D \geq 9.9\,\mathrm{m}.}$$

By Eq. (5.61) the Earth's contrast is 1.4×10^{-10} (Section 5.6.4). Only the largest telescopes reach the angle of an Earth at 10 pc, and then the star must still be dimmed ten billion times, which is why no Earth-like planet around a Sun-like star has yet been imaged.

Chapter 6

The Two-Body Problem

6.1 Central Force Problem and Conservation Laws

We know how to solve Newton's equations for one body in a given force. A stone thrown across a field, or a mass on a spring, moves in a force that does not care where anything else is. Two stars are different. Each one pulls the other, so each moves the source of the other's force. Newton's laws then give six coupled second-order equations, and neither star can be followed alone.

Think of two skaters on ice holding the ends of a rope and pulling on it. Seen from the side, both move in complicated loops. But the rope pulls them with equal and opposite forces, so the balance point between them, nearer the heavier skater, never speeds up. And the rope only pulls along itself, so it cannot change how fast the pair turns about that point. Only the length of the rope and the direction it points are left to work out. A binary star is the pair of skaters, with gravity as the rope. The idea is to split off the motion of the balance point and then use the conserved energy and angular momentum, so that the two-body problem becomes one body moving in a fixed central force.

For skaters of 60 kg and 40 kg standing 5 m apart, the balance point lies $5 \times 40/100 = 2\,\text{m}$ from the heavier one and 3 m from the lighter one. Their separation behaves like the motion of one body with the **reduced mass** $60 \times 40/100 = 24\,\text{kg}$. This separation of centre-of-mass motion from relative motion also describes the Earth and Moon: their common centre lies about 4700 km from the Earth's centre. We will use the same split for any gravitational pair, then use conservation of energy and angular momentum to solve its relative orbit.

6.1.1 Central force problem

A central force acts along the line joining two particles and depends only on their separation $r = |\vec{x}_1 - \vec{x}_2|$. Such a force comes from a scalar potential $V(r)$. We write ∇ for the gradient with respect to $\vec{r} = \vec{x}_1 - \vec{x}_2$, so that $\nabla V = V'(r)\,\hat{r}$ with $V' \equiv dV/dr$. The force on particle 1 is $-\nabla V$, and by Newton's third law the force on particle 2 is $+\nabla V$. Newton's equations for the pair are

$$\frac{d^2}{dt^2}\begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix}\begin{bmatrix} \vec{x}_1 \\ \vec{x}_2 \end{bmatrix} = \begin{bmatrix} -\nabla V(r) \\ \nabla V(r) \end{bmatrix}. \tag{6.1}$$

These are six second-order equations, so the general solution has 12 constants of integration.

6.1.2 Separating centre-of-mass motion

With $M = m_1 + m_2$ and the **reduced mass** $\mu = m_1 m_2/M$, we change to the relative and centre-of-mass coordinates,

$$\begin{bmatrix} \vec{r} \\ \vec{X} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ m_1/M & m_2/M \end{bmatrix} \begin{bmatrix} \vec{x}_1 \\ \vec{x}_2 \end{bmatrix}, \qquad \begin{bmatrix} \vec{x}_1 \\ \vec{x}_2 \end{bmatrix} = \begin{bmatrix} m_2/M & 1 \\ -m_1/M & 1 \end{bmatrix} \begin{bmatrix} \vec{r} \\ \vec{X} \end{bmatrix}.$$

We substitute the second form into (6.1) and multiply both sides from the left by the matrix

$$\begin{bmatrix} m_2/M & -m_1/M \\ 1 & 1 \end{bmatrix},$$

chosen so that it turns the mass matrix diagonal,

$$\begin{bmatrix} m_2/M & -m_1/M \\ 1 & 1 \end{bmatrix} \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \begin{bmatrix} m_2/M & 1 \\ -m_1/M & 1 \end{bmatrix} \frac{d^2}{dt^2} \begin{bmatrix} \vec{r} \\ \vec{X} \end{bmatrix} = \begin{bmatrix} m_2/M & -m_1/M \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -\nabla V \\ \nabla V \end{bmatrix}$$

$$\begin{bmatrix} m_2/M & -m_1/M \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \mu & m_1 \\ -\mu & m_2 \end{bmatrix} \frac{d^2}{dt^2} \begin{bmatrix} \vec{r} \\ \vec{X} \end{bmatrix} \overset{(1)}{=} \begin{bmatrix} -\nabla V \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} \mu & 0 \\ 0 & M \end{bmatrix} \frac{d^2}{dt^2} \begin{bmatrix} \vec{r} \\ \vec{X} \end{bmatrix} \overset{(2)}{=} \begin{bmatrix} -\nabla V \\ 0 \end{bmatrix}. \tag{6.2}$$

(1) On the left, the mass matrix times the inverse transformation has the entries $m_1 m_2/M = \mu$, m_1, $-\mu$ and m_2. On the right, the first row is $-(m_2 + m_1)\nabla V/M = -\nabla V$ and the second is $-\nabla V + \nabla V = 0$.

(2) The first row of the product is $(\mu m_2 + \mu m_1)/M = \mu$ and $(m_1 m_2 - m_2 m_1)/M = 0$. The second row is $\mu - \mu = 0$ and $m_1 + m_2 = M$.

The mass matrix is now diagonal, and the two rows decouple,

$$\mu\, \ddot{\vec{r}} = -\nabla V(r), \tag{6.3}$$

$$M\, \ddot{\vec{X}} = 0. \tag{6.4}$$

Equation (6.4) integrates at once to $\vec{X}(t) = \vec{V}_{\mathrm{CM}}\, t + \vec{X}_0$. The six constants $\vec{V}_{\mathrm{CM}}, \vec{X}_0$ fix an inertial frame in which the centre of mass is at rest at the origin. All the dynamics is in the *relative* coordinate $\vec{r}$, governed by (6.3). It is a one-body central-force problem with the reduced mass μ.

Back to the two bodies. Once $\vec{X}(t)$ and $\vec{r}(t)$ are known, the inverse matrix above gives each body's path. Its rows read $\vec{x}_1 = \vec{X} + (m_2/M)\,\vec{r}$ and $\vec{x}_2 = \vec{X} - (m_1/M)\,\vec{r}$, and $m_1/M = 1 - m_2/M$, so

$$\boxed{\vec{x}_1(t) = \vec{X}(t) + \frac{m_2}{M}\,\vec{r}(t), \qquad \vec{x}_2(t) = \vec{X}(t) - \left(1 - \frac{m_2}{M}\right)\vec{r}(t).} \tag{6.5}$$

Each body follows the centre of mass and adds a copy of the relative orbit, shrunk by a mass fraction. Body 1 takes the share m_2/M of $\vec{r}$, and body 2 takes the rest, $1 - m_2/M$, pointing the other way. The heavier body takes the smaller share. For the Earth and the Moon $M/m_{\mathrm{Moon}} = 82.3$, so the Earth's centre moves on a copy of the Moon's orbit 82.3 times smaller, of radius $384\,400/82.3 \approx 4700$ km. That is the monthly wobble of the Earth.

The relative orbit and the total mass are not enough to find the two paths. Eq. (6.5) needs one more number, the mass fraction m_2/M. The relative orbit cannot supply it. For gravity, $V = -Gm_1 m_2/r$ and $\mu = m_1 m_2/M$, so Eq. (6.3) reads $\ddot{\vec{r}} = -GM\hat{r}/r^2$. The masses enter only through their sum. Move some mass from one body to the other at fixed M and $\vec{r}(t)$ does not

change at all. So no measurement of the relative orbit, however precise, can tell how M is split. The mass fraction has to be measured some other way. It is the place of the centre of mass on the line between the bodies, $m_2/M = |\vec{x}_1 - \vec{X}|/r$, a purely geometric ratio. We can see it by tracking each body against the far background stars (Problem 6.21.1), or by comparing the two Doppler swings of a spectroscopic binary (Problem 6.21.9). When one body is far lighter than the other, the centre of mass sits too close to the heavy body to measure, and the split must come from outside the orbit (Problem 6.21.10).

6.1.3 Conserved quantities

Angular momentum. Define $\vec{L} = \vec{r} \times \vec{p} = \mu\, \vec{r} \times \dot{\vec{r}}$, with $\vec{p} = \mu \dot{\vec{r}}$. Then

$$\begin{aligned} \dot{\vec{L}} &\overset{(1)}{=} \mu \left[\dot{\vec{r}} \times \dot{\vec{r}} + \vec{r} \times \ddot{\vec{r}} \right] \\ &\overset{(2)}{=} -\vec{r} \times \nabla V(r) \\ &\overset{(3)}{=} -V'(r)\, r\, \hat{r} \times \hat{r} = 0. \end{aligned} \tag{6.6}$$

(1) The product rule.
(2) $\dot{\vec{r}} \times \dot{\vec{r}} = 0$, and $\mu \ddot{\vec{r}} = -\nabla V$ from (6.3).
(3) $\vec{r} = r\hat{r}$ and $\nabla V = V'(r)\hat{r}$. The force is central.

$\vec{L}$ supplies three constants of motion. They are its magnitude L and two angles fixing its direction, the **inclination** ι and the **longitude of the ascending node** Ω (Section 6.2). Since $\vec{r} \cdot \vec{L} = \mu\, \vec{r} \cdot (\vec{r} \times \dot{\vec{r}}) = 0$, the body stays in the plane through the origin perpendicular to $\vec{L}$.

Energy. $E = K + V$ with $K = \frac{1}{2}\mu \dot{\vec{r}}^2$. The kinetic energy of the centre of mass is a separate constant and drops out. Writing $\dot{r} = d|\vec{r}|/dt$, we have $\hat{r} \cdot \dot{\vec{r}} = \dot{r}$, because $d(\vec{r} \cdot \vec{r})/dt = 2r\dot{r} = 2\vec{r} \cdot \dot{\vec{r}}$. Then

$$\begin{aligned} \dot{E} &\overset{(1)}{=} \mu \dot{\vec{r}} \cdot \ddot{\vec{r}} + V'(r)\, \dot{r} \\ &\overset{(2)}{=} \mu \dot{\vec{r}} \cdot \ddot{\vec{r}} + \nabla V \cdot \dot{\vec{r}} \\ &= \left(\mu \ddot{\vec{r}} + \nabla V \right) \cdot \dot{\vec{r}} \overset{(3)}{=} 0. \end{aligned} \tag{6.7}$$

(1) The product rule, and the chain rule $dV/dt = V'(r)\, \dot{r}$.
(2) $\nabla V \cdot \dot{\vec{r}} = V'(r)\, \hat{r} \cdot \dot{\vec{r}} = V'(r)\, \dot{r}$.
(3) The equation of motion (6.3).

So E is conserved for *any* central potential $V(r)$, not just gravity. The **total energy and angular momentum** give four constants of integration. There is one more conserved quantity, the Runge–Lenz vector of Section 6.13. It points towards the periapsis and fixes the orientation of the orbit within its plane. The last constant is the time t_0 of a passage through periapsis, which fixes where on the orbit the body is at a given time. The count is

$$\underbrace{\vec{X}_0, \vec{V}_{\mathrm{CM}}}_{6} + \underbrace{a, e, \iota, \Omega, \omega, t_0}_{6} = 12.$$

6.2 Coordinate Systems and Orbital Elements

The angular momentum fixes a plane in space, and the body stays in it. A telescope does not see that plane face on. It sees the orbit squashed onto the sky, and to compare the two we must know how the orbital plane sits relative to the line of sight.

Take a hoop and tilt it in front of a wall. Seen from the front, the round hoop looks like an ellipse. To say how it is tilted, we name the line on the wall about which it was turned, like a hinge, and the angle of the turn. To say where a mark on the hoop is, we give one more angle measured round the hoop from the hinge. An orbit is the hoop and the plane of the sky is the wall. The idea is that the orbit's own frame is the sky frame turned twice, first by Ω about the line of sight and then by ι about the hinge, the line of nodes. One more angle, ω, places the ellipse within its plane.

A circle of radius 1 tilted by $\iota = 60°$ about a line through its centre appears as an ellipse with semi-axes 1 and $\cos 60° = 0.5$. Distances along the hinge stay the same; distances across it shrink by $\cos\iota$. For an orbit, that hinge is the line where the orbital plane meets the sky plane, the line of nodes. We will use its direction and the tilt to turn between the two planes, then locate periapsis within the orbital plane.

The sky frame. Fix an equatorial (reference) frame with the vernal equinox $\hat{\gamma}$ along X and celestial north $\hat{N}$ along Z, and let $\hat{Y}$ complete the right-handed triad. For a source at right ascension α and declination δ, the line of sight from the observer to the source is (Chapter 2, Eq. (2.2))

$$\hat{R} = \cos\delta\cos\alpha\,\hat{\gamma} + \cos\delta\sin\alpha\,\hat{Y} + \sin\delta\,\hat{N}.$$

The local axes in the plane of the sky point along increasing α and δ. They are the normalised derivatives of $\hat{R}$ (Chapter 3, Eq. (3.11)),

$$\hat{\alpha} = \frac{\frac{\partial\hat{R}}{\partial\alpha}}{\left|\frac{\partial\hat{R}}{\partial\alpha}\right|} = -\sin\alpha\,\hat{\gamma} + \cos\alpha\,\hat{Y}, \qquad \hat{\delta} = \frac{\frac{\partial\hat{R}}{\partial\delta}}{\left|\frac{\partial\hat{R}}{\partial\delta}\right|} = -\sin\delta\cos\alpha\,\hat{\gamma} - \sin\delta\sin\alpha\,\hat{Y} + \cos\delta\,\hat{N}.$$

The triad $(\hat{\alpha}, \hat{\delta}, \hat{R})$ is right-handed, since $\hat{\alpha}\times\hat{\delta} = \hat{R}$.

The orbital frame. The orbital plane has its own frame. Its origin is the centre of mass, its X axis lies along the **line of nodes**, where the orbital plane cuts the sky plane, and its Z axis lies along $\vec{L}$. We write $\hat{n}$ for the unit vector towards the **ascending node**, where the body crosses the sky plane moving away from us. The frame is reached by two turns. A turn by Ω about $\hat{R}$ brings the first axis from $\hat{\alpha}$ to $\hat{n}$, and a turn by ι about $\hat{n}$ brings the third axis from $\hat{R}$ to $\hat{L}$. Each turn is a matrix whose rows are the new axes (Chapter 2, Eqs. (2.3), (2.6) and (2.8)), so

$$\begin{pmatrix}\hat{n}\\ \hat{y}\\ \hat{L}\end{pmatrix}_{\text{Orbital Plane}} = \begin{pmatrix}1 & 0 & 0\\ 0 & \cos\iota & \sin\iota\\ 0 & -\sin\iota & \cos\iota\end{pmatrix}\begin{pmatrix}\cos\Omega & \sin\Omega & 0\\ -\sin\Omega & \cos\Omega & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}\hat{\alpha}\\ \hat{\delta}\\ \hat{R}\end{pmatrix}_{\text{Sky plane}}. \tag{6.8}$$

Multiplying out the two matrices,

$$\begin{pmatrix}\hat{n}\\ \hat{y}\\ \hat{L}\end{pmatrix} \overset{(1)}{=} \begin{pmatrix}\cos\Omega & \sin\Omega & 0\\ -\sin\Omega\cos\iota & \cos\Omega\cos\iota & \sin\iota\\ \sin\Omega\sin\iota & -\cos\Omega\sin\iota & \cos\iota\end{pmatrix}\begin{pmatrix}\hat{\alpha}\\ \hat{\delta}\\ \hat{R}\end{pmatrix}. \tag{6.9}$$

(1) Each row of $\mathbf{R}_x(\iota)$ times each column of $\mathbf{R}_z(\Omega)$. The first row of $\mathbf{R}_x$ copies the first row of $\mathbf{R}_z$. The second and third rows mix the second and third rows of $\mathbf{R}_z$ with weights $(\cos\iota, \sin\iota)$ and $(-\sin\iota, \cos\iota)$.

Reading off the rows,

$$\hat{n} = \cos\Omega\,\hat{\alpha} + \sin\Omega\,\hat{\delta}, \tag{6.10}$$

$$\hat{y} = -\sin\Omega\cos\iota\,\hat{\alpha} + \cos\Omega\cos\iota\,\hat{\delta} + \sin\iota\,\hat{R}, \tag{6.11}$$

$$\hat{L} = \sin\Omega\sin\iota\,\hat{\alpha} - \cos\Omega\sin\iota\,\hat{\delta} + \cos\iota\,\hat{R}. \tag{6.12}$$

The third row gives $\iota = \arccos\left(\hat{L} \cdot \hat{R}\right)$, the angle between the orbital plane and the sky plane.

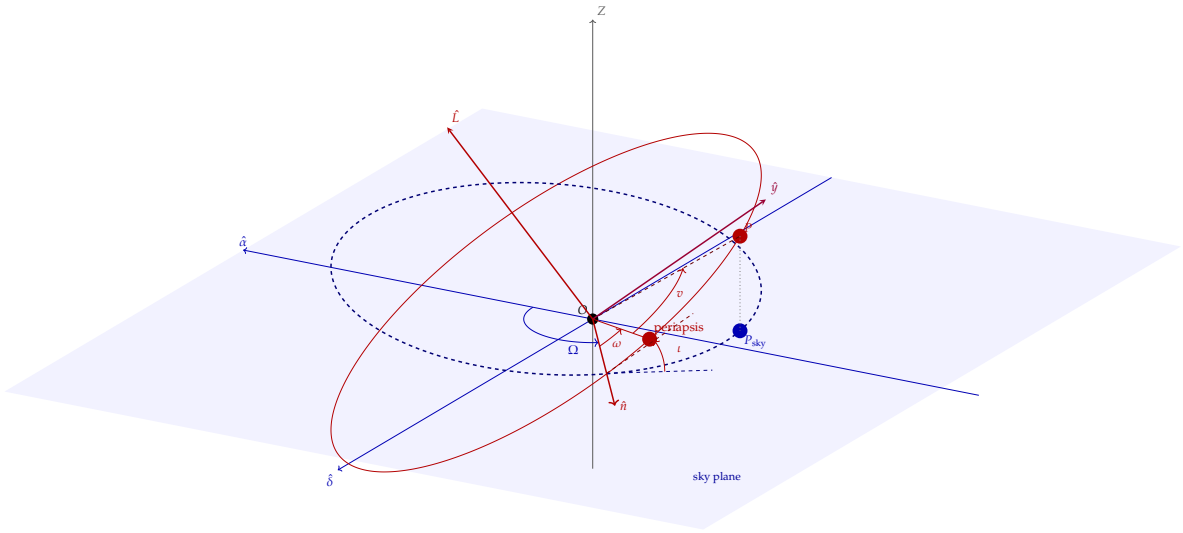


Figure 6.1: An orbit (red) and the plane of the sky (blue), with the sky axes $\hat{\alpha}$, $\hat{\delta}$ and the orbital axes $\hat{n}$, $\hat{y}$, $\hat{L}$. Ω is the angle in the sky plane from $\hat{\alpha}$ to the ascending node. ω is the angle in the orbital plane from the ascending node to the periapsis. ι is the angle between the two planes at the node. v is the angle in the orbital plane from the periapsis to the body P, and P_{sky} is the projection of P on the sky plane.

The rows of a rotation are orthonormal, so $(\hat{n}, \hat{y}, \hat{L})$ is a right-handed orthonormal triad. Its first two vectors span the orbital plane and the third is its normal. In the plane we use polar coordinates (r, ϕ), with ϕ measured from $\hat{n}$. The radial and angular unit vectors are $(\hat{n}, \hat{y})$ turned by ϕ,

$$\begin{pmatrix} \hat{r} \\ \hat{\phi} \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} \hat{n} \\ \hat{y} \end{pmatrix}, \qquad \text{that is} \qquad \hat{r} = \cos\phi\,\hat{n} + \sin\phi\,\hat{y}, \quad \hat{\phi} = -\sin\phi\,\hat{n} + \cos\phi\,\hat{y}. \tag{6.13}$$

The projected orbit in the sky plane. The body sits in its plane at the polar angle $\phi = \omega + v$. Here ω is the angle from the node to the periapsis and v, the **true anomaly**, is the angle from the periapsis to the body. Both are found in Section 6.5. The body's orbital components are $(r\cos(\omega+v),\ r\sin(\omega+v),\ 0)$. To get its sky components (X, Y, Z) along $(\hat{\alpha}, \hat{\delta}, \hat{R})$ we undo Eq. (6.8). A rotation is undone by its transpose (Chapter 2, Eq. (2.4)), and a product is undone in reverse order, so

$$\begin{aligned} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}_{\text{Sky Plane}} &= \begin{pmatrix} \cos\Omega & -\sin\Omega & 0 \\ \sin\Omega & \cos\Omega & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\iota & -\sin\iota \\ 0 & \sin\iota & \cos\iota \end{pmatrix} \begin{pmatrix} r\cos(\omega+v) \\ r\sin(\omega+v) \\ 0 \end{pmatrix}_{\text{Orbital Plane}} \\ &\overset{(1)}{=} \begin{pmatrix} \cos\Omega & -\sin\Omega & 0 \\ \sin\Omega & \cos\Omega & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} r\cos(\omega+v) \\ r\sin(\omega+v)\cos\iota \\ r\sin(\omega+v)\sin\iota \end{pmatrix} \\ &\overset{(2)}{=} r \begin{pmatrix} \cos(\omega+v)\cos\Omega - \sin(\omega+v)\sin\Omega\cos\iota \\ \cos(\omega+v)\sin\Omega + \sin(\omega+v)\cos\Omega\cos\iota \\ \sin(\omega+v)\sin\iota \end{pmatrix}. \end{aligned} \tag{6.14}$$

(1) $\mathbf{R}_x(\iota)^T$ times the column.
(2) $\mathbf{R}_z(\Omega)^T$ times the column.

X and Y give the position on the sky. Z is the distance along the line of sight, beyond the centre of mass, and it returns in the timing of pulsars (Section 6.18).

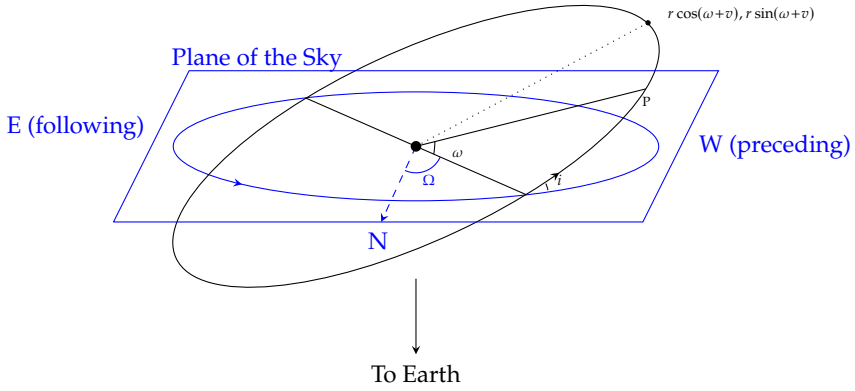


Figure 6.2: An orbit (black) and its projection (blue) on the plane of the sky. N, E and W mark north, east and west on the sky, and the arrow points towards the Earth. Ω is measured in the sky plane from north to the node line, and ω in the orbital plane from the node line to the periapsis. P marks a position $(r\cos(\omega+v), r\sin(\omega+v))$ on the orbit.

Figure 6.2 takes north as the first sky axis and west as the second. This pair is $(\hat{\alpha}, \hat{\delta})$ turned by 90° about $\hat{R}$, so Eq. (6.14) holds unchanged with Ω counted from north. In the same way we may choose the first sky axis along the line of nodes. That sets $\Omega = 0$, and the sky position becomes

$$X = r\cos(\omega + v), \tag{6.15}$$
$$Y = r\sin(\omega + v)\cos\iota. \tag{6.16}$$

Lengths along the line of nodes are unchanged, and every length across it shrinks by the factor $\cos\iota$. A circular orbit therefore appears as an ellipse with axis ratio $\cos\iota$, and it appears as a circle only when $\iota = 0$ (Problem 1).

The six orbital elements. Six constants appear as the orbit is solved. The angles Ω and ι give the orientation of the plane and come from $\vec{L}$. The semi-major axis a and the eccentricity e give the size and shape and come from E and L. The **argument of periapsis** ω gives the orientation of the ellipse *within* the plane. The time t_0 of periapsis passage fixes where on the ellipse the body is at a given time. Figure 6.1 shows them together.

6.3 Planar Motion and Kepler's Second Law

We have a conserved angular momentum, and with it an orbit that stays in one plane. Conservation alone does not say how fast the body moves at each point of its path.

A ball slides on a smooth table, tied to a string that passes down through a small hole. Pull the string from below and the ball circles faster as its circle shrinks. The string pulls only towards the hole, so it never pushes the ball sideways. What stays fixed is the distance from the hole times the sideways speed. A planet does the same. It moves fastest where it is nearest the Sun. The idea is that the area swept by the line from the Sun to the planet grows at a fixed rate, $L/(2\mu)$.

At perihelion in January the Earth is 0.983 AU from the Sun; at aphelion in July it is 1.017 AU away. At both turning points its velocity is perpendicular to the radius, so angular momentum makes distance times speed the same. The Earth therefore moves $1.017/0.983 = 1.035$ times faster in January, about 3.5 per cent. We can express that same conservation as equal areas swept by the Sun–Earth line in equal times: Kepler's second law.

Areal velocity. In a time dt the radius vector turns by $d\phi = \dot\phi\, dt$ and sweeps a thin wedge. In polar coordinates a small patch at distance r' from the focus has sides dr' and $r'\, d\phi$, so its area is $r'\, dr'\, d\phi$. The wedge adds up these patches from the focus out to the orbit at $r(\phi)$,

$$dA \overset{(1)}{=} \left(\int_0^{r(\phi)} r'\, dr' \right) d\phi \overset{(2)}{=} \tfrac{1}{2}\, r^2(\phi)\, d\phi,$$
$$\frac{dA}{dt} \overset{(3)}{=} \frac{r^2\dot\phi}{2} \overset{(4)}{=} \frac{L}{2\mu} = \text{const.} \tag{6.17}$$

(1) The patches $r'\, dr'\, d\phi$ summed along the wedge, from $r' = 0$ to $r' = r(\phi)$.
(2) $\int_0^r r'\, dr' = r^2/2$.
(3) Divide by dt, with $d\phi/dt = \dot\phi$.
(4) $r^2\dot\phi = L/\mu$, from $\vec L = \mu\, \vec r \times \dot{\vec r}$, as Eq. (6.22) below shows.

The same area is half the size of the cross product of the two sides of the thin triangle, $dA = \frac{1}{2}\left|\vec r \times \dot{\vec r}\, dt\right| = (L/2\mu)\, dt$. This is **Kepler's second law**, equal areas in equal times. It follows only from $\vec L$ being conserved, that is, from the force being central. It does not depend on the form of $V(r)$.

Equations of motion in (r, ϕ). With $Z \parallel \vec L$ and $X \parallel \hat n$, we have $\vec r = r\hat r$. The unit vectors (6.13) turn with ϕ. Differentiating their matrix,

$$\frac{d}{dt}\begin{pmatrix}\hat r\\ \hat\phi\end{pmatrix} = \dot\phi \begin{pmatrix}-\sin\phi & \cos\phi\\ -\cos\phi & -\sin\phi\end{pmatrix}\begin{pmatrix}\hat n\\ \hat y\end{pmatrix} = \dot\phi \begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\begin{pmatrix}\hat r\\ \hat\phi\end{pmatrix}, \tag{6.18}$$

so $\dot{\hat r} = \dot\phi\, \hat\phi$ and $\dot{\hat\phi} = -\dot\phi\, \hat r$. The velocity and acceleration follow by the product rule,

$$\dot{\vec r} = \dot r\, \hat r + r\dot{\hat r} \overset{(1)}{=} \dot r\, \hat r + r\dot\phi\, \hat\phi,$$
$$\ddot{\vec r} = \ddot r\, \hat r + \dot r\dot\phi\, \hat\phi + \dot r\dot\phi\, \hat\phi + r\ddot\phi\, \hat\phi + r\dot\phi\, \dot{\hat\phi} \overset{(1)}{=} \big(\ddot r - r\dot\phi^2\big)\hat r + \big(2\dot r\dot\phi + r\ddot\phi\big)\hat\phi. \tag{6.19}$$

(1) $\dot{\hat r} = \dot\phi\, \hat\phi$ and $\dot{\hat\phi} = -\dot\phi\, \hat r$, from Eq. (6.18).

The same rates follow from the Christoffel symbols of polar coordinates, $\nabla_\mu \mathbf{e}_\nu = \Gamma^\alpha_{\mu\nu}\mathbf{e}_\alpha$. With $\Gamma^\phi_{\phi r} = 1/r$ and the coordinate basis vector $\mathbf{e}_\phi = r\hat\phi$, we get $\dot{\hat r} = \dot\phi\, \Gamma^\mu_{\phi r}\mathbf{e}_\mu = \dot\phi\, \mathbf{e}_\phi/r = \dot\phi\, \hat\phi$. Equation (6.3), with $\nabla V = V'(r)\hat r$, splits into radial and tangential parts,

$$\mu\big(\ddot r - r\dot\phi^2\big) = -V'(r), \tag{6.20}$$
$$\mu\big(2\dot r\dot\phi + r\ddot\phi\big) = 0 \implies \frac{\mathrm{d}(r^2\dot\phi)}{\mathrm{d}t} = r\big(2\dot r\dot\phi + r\ddot\phi\big) = 0. \tag{6.21}$$

These look like two coupled equations. But Eq. (6.21) is only the conservation of angular momentum, Kepler's second law again, because

$$\vec L = \mu\vec r \times \dot{\vec r} = \mu\, r\hat r \times \big(\dot r\, \hat r + r\dot\phi\, \hat\phi\big) = \mu r^2\dot\phi\, \hat L.$$

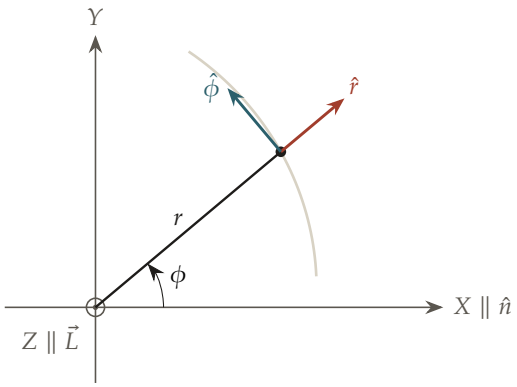


Figure 6.3: The orbit plane, with X along $\hat{n}$ and Z along $\vec{L}$ out of the page, the body at distance r and angle ϕ on a piece of its orbit (grey), and the unit vectors $\hat{r}$ and $\hat{\phi}$ at the body.

So

$$\dot{\phi} = \frac{L}{\mu r^2}, \tag{6.22}$$

and Eq. (6.20) becomes, eliminating $\dot{\phi}$,

$$\mu\ddot{r} - \frac{L^2}{\mu r^3} = -V'(r). \tag{6.23}$$

This is one-dimensional motion in r alone. The term $L^2/(\mu r^3)$ is the centrifugal push, and Problem 5 treats it as an effective potential.

6.4 The Orbit Equation

We have one radial equation, (6.23), with time as the variable. For a closed orbit the first thing we want is its shape, the curve $r(\phi)$. Time only gets in the way.

A hiker's track drawn on a map shows the whole route without saying how fast she walked. Each point of the route is named by where it is, not by when she was there. The angle ϕ of an orbit grows steadily, since $\dot{\phi} = L/(\mu r^2) > 0$ whenever $L \neq 0$. So ϕ can replace t as the clock. The idea is to change the variable from t to ϕ and to write $\rho = 1/r$. For gravity the equation then becomes linear.

A circle is the simplest possible shape: ρ is constant and $d^2\rho/d\phi^2 = 0$. For gravity, the orbit equation below then gives $\rho = GM\mu^2/L^2$, or radius $(L/\mu)^2/GM$. Earth's $L/\mu = r\,|\dot{\vec{r}}| = 1.496\times10^{11}\,\mathrm{m}\times29.78\,\mathrm{km\,s^{-1}} = 4.455\times10^{15}\,\mathrm{m^2\,s^{-1}}$ and $GM_\odot = 1.327\times10^{20}\,\mathrm{m^3\,s^{-2}}$ give 1.000 AU. We now derive the equation for $\rho(\phi)$ so that the same method also describes noncircular paths.

From time to angle. Using Eq. (6.22), $\dfrac{dr}{dt} = \dot{\phi}\dfrac{dr}{d\phi} = \dfrac{L}{\mu r^2}\dfrac{dr}{d\phi}$, and

$$\begin{aligned}\frac{d^2r}{dt^2} &\overset{(1)}{=} \frac{L}{\mu r^2}\frac{d}{d\phi}\left[\frac{L}{\mu r^2}\frac{dr}{d\phi}\right] \\ &\overset{(2)}{=} \frac{L^2}{\mu^2 r^4}\left[\frac{d^2r}{d\phi^2} - \frac{2}{r}\left(\frac{dr}{d\phi}\right)^2\right].\end{aligned} \tag{6.24}$$

(1) $d/dt = \dot{\phi}\,d/d\phi$, applied to dr/dt.

(2) The product rule, with $\dfrac{d}{d\phi}\dfrac{1}{r^2} = -\dfrac{2}{r^3}\dfrac{dr}{d\phi}$.

Substituting into Eq. (6.23) and dividing by L^2/μ,

$$\frac{1}{r^4}\left[\frac{d^2r}{d\phi^2} - \frac{2}{r}\left(\frac{dr}{d\phi}\right)^2\right] - \frac{1}{r^3} = -\frac{\mu}{L^2}V'(r). \tag{6.25}$$

Binet's substitution. We set $r = 1/\rho$. Then

$$\frac{dr}{d\phi} = -\frac{1}{\rho^2}\frac{d\rho}{d\phi}, \qquad \frac{d^2r}{d\phi^2} = -\frac{1}{\rho^2}\frac{d^2\rho}{d\phi^2} + \frac{2}{\rho^3}\left(\frac{d\rho}{d\phi}\right)^2,$$

so the left side of Eq. (6.25) is

$$\rho^4\left[-\frac{1}{\rho^2}\frac{d^2\rho}{d\phi^2} + \frac{2}{\rho^3}\left(\frac{d\rho}{d\phi}\right)^2 - 2\rho\cdot\frac{1}{\rho^4}\left(\frac{d\rho}{d\phi}\right)^2\right] - \rho^3 \overset{(1)}{=} -\rho^2\frac{d^2\rho}{d\phi^2} - \rho^3. \tag{6.26}$$

(1) The two $(d\rho/d\phi)^2$ terms cancel.

Setting this equal to $-(\mu/L^2)V'(1/\rho)$ and dividing by $-\rho^2$ gives the **orbit equation**,

$$\boxed{\frac{d^2\rho}{d\phi^2} + \rho = \frac{\mu}{L^2}\frac{1}{\rho^2}V'\left(\frac{1}{\rho}\right).} \tag{6.27}$$

This is an equation for the orbit's shape alone, valid for any central potential.

6.5 Keplerian Orbits

The orbit equation (6.27) holds for any central force. For gravity we want the actual shape, and we want to know whether the orbit closes on itself.

Hang a weight on a spring. Gravity pulls it down with a constant force, and the weight bobs up and down exactly as it would without gravity. The only change is that it bobs about a lower rest point. A constant force on an oscillator just moves the centre of the oscillation. For gravity the orbit equation is that spring, with ϕ as the time. The idea is that $\rho = 1/r$ oscillates about a constant value, once per turn of ϕ, so the orbit closes and is a conic section.

A 1 kg mass on a spring of stiffness $100\,\mathrm{N\,m^{-1}}$ hangs 9.8 cm below its unstretched position yet oscillates with period $2\pi\sqrt{1/100}\,\mathrm{s} = 0.63\,\mathrm{s}$. Likewise, gravity shifts the equilibrium value of $\rho = 1/r$ in the orbit equation while leaving a cosine variation about it. One full variation takes one turn in ϕ, so the orbit closes. We now turn that cosine into the equation of a conic section.

The conic. For Newtonian gravity $V(r) = -GM\mu/r$, so $V'(r) = GM\mu/r^2 = GM\mu\rho^2$. Equation (6.27) becomes

$$\frac{d^2\rho}{d\phi^2} + \rho = \frac{\mu}{L^2}\frac{1}{\rho^2}GM\mu\rho^2 = \frac{GM\mu^2}{L^2}.$$

Shifting $\tilde\rho = \rho - GM\mu^2/L^2$ turns this into the simple-harmonic-oscillator equation $d^2\tilde\rho/d\phi^2 + \tilde\rho = 0$, with solution $\tilde\rho = C\cos(\phi - \omega)$ for constants $C \geq 0$ and ω. In terms of $r = 1/\rho$,

$$r(\phi) = \left(\frac{GM\mu^2}{L^2} + C\cos(\phi-\omega)\right)^{-1} = \frac{L^2/(GM\mu^2)}{1 + \dfrac{L^2C}{GM\mu^2}\cos(\phi-\omega)}.$$

We define the **semi-latus rectum** $\lambda \equiv L^2/(GM\mu^2)$, which is the value of r at $\phi = \omega + 90°$, and the **eccentricity** $e \equiv \lambda C$. Then

$$\boxed{r(\phi) = \frac{\lambda}{1 + e\cos(\phi - \omega)}.} \tag{6.28}$$

This is the polar equation of a conic section with one focus at the origin, semi-latus rectum λ and eccentricity e. Its major axis is turned by ω, the **argument of periapsis**, and the periapsis is at $\phi = \omega$. The angle from the periapsis, $v \equiv \phi - \omega$, is the **true anomaly**.

The axes. For $e < 1$ the orbit is closed. Its nearest point, the periapsis, is at $r_\mathrm{p} = \lambda/(1+e)$, and its farthest, the apoapsis, is at $r_\mathrm{a} = \lambda/(1-e)$. The major axis is their sum, so the **semi-major axis** is

$$a = \frac{r_\mathrm{p} + r_\mathrm{a}}{2} = \frac{\lambda}{2}\,\frac{(1-e)+(1+e)}{1-e^2} = \frac{\lambda}{1-e^2}.$$

The centre lies midway between the two ends, at $a - r_\mathrm{p} = a - a(1-e) = ae$ from the focus. To

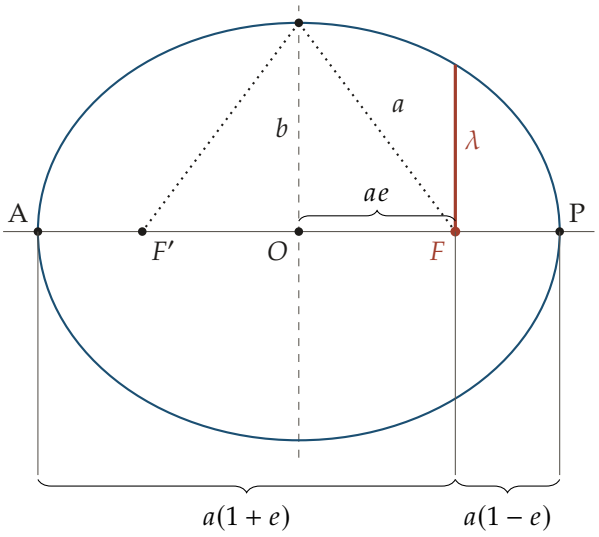


Figure 6.4: An ellipse with semi-axes a and b, centre O and foci F and F', each at distance ae from O. The periapsis P and the apoapsis A lie at distances $a(1-e)$ and $a(1+e)$ from F. The dotted lines join the end of the minor axis to the two foci. The semi-latus rectum λ is drawn at F, perpendicular to the major axis.

see that the conic is the ellipse $x^2/a^2 + y^2/b^2 = 1$ with **semi-minor axis** $b = a\sqrt{1-e^2}$, we put the origin of (x, y) at the centre and the periapsis on the $+x$ axis (taking $\omega = 0$). The focus is then at $(ae, 0)$, and a point at (r, ϕ) about the focus has

$$x = ae + r\cos\phi, \tag{6.29}$$

$$y = r\sin\phi. \tag{6.30}$$

Substituting into the ellipse with $b^2 = a^2(1-e^2)$,

$$\begin{aligned} 1 &= \frac{(ae + r\cos\phi)^2}{a^2} + \frac{r^2\sin^2\phi}{a^2(1-e^2)} \\ 0 &\overset{(1)}{=} r^2(1 - e^2\cos^2\phi) + 2ae(1-e^2)\,r\cos\phi - a^2(1-e^2)^2 \\ r &\overset{(2)}{=} \frac{a(1-e^2)(1-e\cos\phi)}{(1-e\cos\phi)(1+e\cos\phi)} = \frac{a(1-e^2)}{1+e\cos\phi}. \end{aligned} \tag{6.31}$$

(1) Multiply by $a^2(1-e^2)$, expand, and use $\sin^2\phi = 1-\cos^2\phi$ and $(1-e^2)a^2e^2 - a^2(1-e^2) = -a^2(1-e^2)^2$.
(2) The quadratic formula. The discriminant is $4a^2(1-e^2)^2\left[e^2\cos^2\phi + 1 - e^2\cos^2\phi\right] = 4a^2(1-e^2)^2$, and the root with $r > 0$ is kept. Then $1 - e^2\cos^2\phi$ is factored.

This is the conic (6.28) with $\lambda = a(1-e^2)$, so the orbit is this ellipse. The point $(0, b)$ at the end of the minor axis is at distance $\sqrt{b^2 + a^2e^2} = a$ from each focus (Figure 6.4). At $x = ae$ the ellipse has $y = b\sqrt{1-e^2} = b^2/a = \lambda$, which is the semi-latus rectum again.

Eccentricity	Conic section
$e = 0$	circle
$0 < e < 1$	ellipse ($E < 0$, bound)
$e = 1$	parabola ($E = 0$)
$e > 1$	hyperbola ($E > 0$, unbound)

The sign of E in each row is shown in Section 6.6. Bound, closed elliptical orbits ($E < 0$) are **Kepler's first law**. In the centre-of-mass frame, $\vec{X} = 0$, Eq. (6.5) gives the two bodies' positions,

$$\vec{x}_1 = \frac{m_2}{M}\vec{r}, \qquad \vec{x}_2 = -\frac{m_1}{M}\vec{r}.$$

6.6 Energy and Angular Momentum in Terms of a, e

We have the shape of the orbit, with constants λ and e, and two conserved numbers, E and L. The shape constants came out of solving a differential equation. We do not yet know how they depend on E and L.

A child on a swing stops for a moment at the top of each swing. There all of the energy is height, so the height of the swing tells us the energy without any need to know the speed elsewhere. An orbit has two such turning points of r, the periapsis and the apoapsis, where $\dot{r} = 0$. The idea is to evaluate the conserved E at a turning point. The size a then depends on E alone, and e on E and L.

A 20 kg child who rises 0.5 m on a swing gains $E = 20 \times 9.8 \times 0.5 = 98\,\text{J}$ of potential energy and passes the bottom at $\sqrt{2 \times 9.8 \times 0.5} = 3.1\,\text{m s}^{-1}$. The turning point reveals the energy because the motion in that direction stops there. An orbit has radial turning points at periapsis and apoapsis. Evaluating its conserved energy at those points will connect the measurable size and shape of the ellipse to E and L.

6.6.1 Energy

Writing the energy in polar coordinates,

$$\begin{aligned} E &= \frac{1}{2}\mu\dot{\vec{r}}^2 - \frac{GM\mu}{r} \\ &\overset{(1)}{=} \frac{1}{2}\mu\left(\dot{r}^2 + r^2\dot{\phi}^2\right) - \frac{GM\mu}{r} \\ &\overset{(2)}{=} \frac{1}{2}\mu\dot{r}^2 + \frac{L^2}{2\mu r^2} - \frac{GM\mu}{r}. \end{aligned} \tag{6.32}$$

(1) $\dot{\vec{r}} = \dot{r}\,\hat{r} + r\dot{\phi}\,\hat{\phi}$ with $\hat{r} \perp \hat{\phi}$.
(2) $\dot{\phi} = L/(\mu r^2)$, Eq. (6.22).

Since E is conserved, we evaluate it at periapsis ($\phi = \omega$). There $r = \lambda/(1+e)$, and $\dot{r} = 0$ because r has a minimum,

$$\begin{aligned} E &\overset{(1)}{=} \frac{L^2(1+e)^2}{2\mu\lambda^2} - \frac{GM\mu(1+e)}{\lambda} \\ &\overset{(2)}{=} \frac{GM\mu(1+e)^2}{2\lambda} - \frac{GM\mu(1+e)}{\lambda} \\ &\overset{(3)}{=} -\frac{GM\mu}{2\lambda}(1-e^2) \overset{(4)}{=} -\frac{GM\mu}{2a}. \end{aligned} \tag{6.33}$$

(1) Eq. (6.32) with $\dot{r} = 0$ and $r = \lambda/(1+e)$.
(2) $L^2 = GM\mu^2\lambda$, the definition of λ.
(3) $(1+e)^2 - 2(1+e) = (1+e)(e-1) = -(1-e^2)$.
(4) $a = \lambda/(1-e^2)$.

$$\boxed{E = -\frac{GM\mu}{2a}.} \tag{6.34}$$

The energy depends on the size of the orbit only, not on its shape. Since $\lambda > 0$ always, the sign of E follows e. Bound orbits ($E < 0$) have $e < 1$ and $a > 0$. Unbound orbits ($E \geq 0$) have $e \geq 1$, and hyperbolic orbits formally have $a < 0$.

6.6.2 Angular momentum

At the apoapsis, $r = a + ae = \lambda/(1-e)$, and at the periapsis, $r = a - ae = \lambda/(1+e)$, we have $\dot{r} = 0$. Equating the energy at the two points as a check,

$$\begin{aligned} \frac{L^2(1+e)^2}{2\mu\lambda^2} - \frac{GM\mu(1+e)}{\lambda} &= \frac{L^2(1-e)^2}{2\mu\lambda^2} - \frac{GM\mu(1-e)}{\lambda} \\ \frac{L^2\big[(1+e)^2 - (1-e)^2\big]}{2\mu\lambda^2} &\overset{(1)}{=} \frac{GM\mu\big[(1+e) - (1-e)\big]}{\lambda} \\ \frac{2eL^2}{\mu\lambda^2} &\overset{(2)}{=} \frac{2eGM\mu}{\lambda}. \end{aligned} \tag{6.35}$$

(1) Collect the L^2 terms on the left and the $GM\mu$ terms on the right.
(2) $(1+e)^2 - (1-e)^2 = 4e$.

So $L^2 = GM\mu^2\lambda$, the definition of λ again. The two turning points agree, as they must. With $\lambda = a(1-e^2)$ the angular momentum in terms of a and e is

$$L = \sqrt{GM\mu^2\lambda}, \qquad \text{i.e.} \qquad L^2 = GM\mu^2 a(1-e^2), \tag{6.36}$$

and, with the direction $\hat{L}$ from Eq. (6.12),

$$\vec{L} = \sqrt{GM\mu^2\lambda}\,\big(\sin\Omega\sin\iota\,\hat{\alpha} - \cos\Omega\sin\iota\,\hat{\delta} + \cos\iota\,\hat{R}\big).$$

6.6.3 Instantaneous kinetic energy

From $E = K + V$ and Eq. (6.34),

$$K = \tfrac{1}{2}\mu\dot{\vec{r}}^2 = E - V = GM\mu\left(\frac{1}{r} - \frac{1}{2a}\right), \qquad \text{or} \qquad |\dot{\vec{r}}|^2 = GM\left(\frac{2}{r} - \frac{1}{a}\right). \tag{6.37}$$

This is the *vis-viva* equation. The speed at any point depends only on the distance and on a. The kinetic energy is shared between the two bodies through $\vec{x}_1 = (m_2/M)\vec{r}$ and $\vec{x}_2 = -(m_1/M)\vec{r}$,

$$K_1 = \tfrac{1}{2}m_1|\dot{\vec{x}}_1|^2 = \tfrac{1}{2}m_1\frac{m_2^2}{M^2}\dot{\vec{r}}^2 = \frac{m_2}{M}\cdot\tfrac{1}{2}\mu\dot{\vec{r}}^2 = \frac{m_2}{M}K, \qquad K_2 = \frac{m_1}{M}K,$$

using $\mu = m_1m_2/M$. The lighter body carries the larger share.

Inverting for a, e. Equations (6.34) and (6.36) give the orbital elements directly from the conserved energy and angular momentum. The first gives $a = -GM\mu/(2E)$. Then

$$e^2 = 1 - \frac{L^2}{GM\mu^2 a} \overset{(1)}{=} 1 + \frac{2EL^2}{(GM)^2\mu^3}, \tag{6.38}$$

(1) $1/a = -2E/(GM\mu)$.

so

$$a = -\frac{GM\mu}{2E}, \qquad e = \sqrt{1 + \frac{2EL^2}{(GM)^2\mu^3}}. \tag{6.39}$$

6.7 True and Eccentric Anomaly; Kepler's Equation

We have the shape $r(\phi)$ and the size of the orbit from its energy. The shape alone does not say where the body is at a given time, and that is what an observer needs to point a telescope or predict an eclipse.

Draw a clock face on a rubber sheet and stretch the sheet in one direction. The hour marks move, and the angles measured on the stretched sheet are no longer 30° apart. But each mark can still be named by its old angle on the unstretched face, and with that name the arithmetic stays simple. An ellipse is a circle of radius a squashed by the factor b/a across its major axis. A point on the ellipse at $(a\cos u,\ b\sin u)$ from the centre is named by the angle u on that circle, the **eccentric anomaly**. The idea is that with $r = a(1 - e\cos u)$ the time integral of the orbit becomes elementary.

For $e = 0.5$ and eccentric anomaly $u = 90°$, the body is at $r = a(1 - 0) = a$. From the focus, its true anomaly is already $v = 120°$, since Eq. (6.44) gives $\cos v = (0 - 0.5)/(1 - 0) = -0.5$. The two angles name the same point from different centres. The auxiliary circle makes the geometry and time integral simpler, so we will use u to find both the distance and the elapsed time.

The time integral. We now want $r(t)$, and hence $\phi(t)$, not just $r(\phi)$. Solving the energy (6.32) for $\dot{r}$,

$$\dot{r} \overset{(1)}{=} \sqrt{\frac{2}{\mu}\left(E - \frac{L^2}{2\mu r^2} + \frac{GM\mu}{r}\right)}$$
$$\overset{(2)}{=} \sqrt{-\frac{GM}{a} - \frac{GM\lambda}{r^2} + \frac{2GM}{r}} \overset{(3)}{=} \sqrt{GM}\,\frac{1}{r}\left(2r - \frac{r^2}{a} - \lambda\right)^{1/2}. \tag{6.40}$$

(1) Eq. (6.32), on the outbound half of the orbit where $\dot{r} > 0$.
(2) $E = -GM\mu/(2a)$ and $L^2 = GM\mu^2\lambda$, Eqs. (6.34) and (6.36).
(3) Take GM/r^2 out of the square root.

We separate the variables and integrate from the periapsis, which the body passes at the time t_0 with $r = a(1-e)$,

$$t - t_0 = \frac{1}{\sqrt{GM}} \int_{a(1-e)}^{r} \frac{r'\,dr'}{\left(2r' - r'^2/a - \lambda\right)^{1/2}} \overset{(1)}{=} \sqrt{\frac{a}{GM}} \int_{a(1-e)}^{r} \frac{r'dr'}{\sqrt{a^2e^2 - (r'-a)^2}}. \tag{6.41}$$

(1) Multiply the bracket by a. With $\lambda = a(1-e^2)$, $2ar - r^2 - a^2(1-e^2) = a^2e^2 - (r-a)^2$.

6.7.1 Elliptical orbits: eccentric anomaly

For $a > 0$ we substitute

$$r = a\big(1 - e\cos u\big), \tag{6.42}$$

where u is the **eccentric anomaly**. This choice keeps $r > 0$ throughout, since $-1 \le \cos u \le 1$, and it makes the square root a perfect square,

$$t - t_0 \overset{(1)}{=} \sqrt{\frac{a}{GM}} \int_0^u \frac{a(1-e\cos u')\,ae\sin u'\,du'}{ae\sin u'} \overset{(2)}{=} \frac{a^{3/2}}{\sqrt{GM}} \int_0^u (1-e\cos u')\,du' = \frac{a^{3/2}}{\sqrt{GM}}\big[u - e\sin u\big]. \tag{6.43}$$

(1) $dr = ae\sin u\,du$, and $a^2e^2 - (r-a)^2 = a^2e^2 - a^2e^2\cos^2 u = a^2e^2\sin^2 u$, whose root is $ae\sin u$ for $0 < u < \pi$. The periapsis $r = a(1-e)$ is $u = 0$.

(2) Cancel $ae\sin u'$.

On the inbound half, $\pi < u < 2\pi$, both $\dot r$ and $\sin u$ change sign, so the same formula holds for all u. The time t_0 of periapsis passage, where $u = 0$, is the sixth and final constant of integration.

Relating the true anomaly v and the eccentric anomaly u. The true anomaly $v = \phi - \omega$ gives $r = \lambda/(1 + e\cos v)$ from Eq. (6.28). Equating this with Eq. (6.42),

$$1 + e\cos v \overset{(1)}{=} \frac{1-e^2}{1-e\cos u}$$

$$e\cos v = \frac{1 - e^2 - 1 + e\cos u}{1 - e\cos u} = \frac{e(\cos u - e)}{1 - e\cos u}, \qquad \text{so} \qquad \cos v = \frac{\cos u - e}{1 - e\cos u}. \tag{6.44}$$

(1) $a(1 - e\cos u) = a(1-e^2)/(1 + e\cos v)$, with $\lambda = a(1-e^2)$.

The sine follows from $\sin^2 v = 1 - \cos^2 v$,

$$\sin^2 v = \frac{(1 - e\cos u)^2 - (\cos u - e)^2}{(1 - e\cos u)^2} \overset{(1)}{=} \frac{(1-e^2)(1-\cos^2 u)}{(1-e\cos u)^2}, \tag{6.45}$$

(1) The numerator is $1 - 2e\cos u + e^2\cos^2 u - \cos^2 u + 2e\cos u - e^2 = (1-e^2) - (1-e^2)\cos^2 u$.

and v and u lie in the same half of the orbit, so the positive root is taken where $\sin u > 0$. Together,

$$\sin v = \frac{\sqrt{1-e^2}\sin u}{1 - e\cos u}, \qquad \tan v = \frac{\sqrt{1-e^2}\sin u}{\cos u - e}, \qquad \tan\frac{v}{2} = \sqrt{\frac{1+e}{1-e}}\,\tan\frac{u}{2}. \tag{6.46}$$

The half-angle form is proved in Problem 4. At $u = 0$ and $u = \pi$ we get $v = u$. The periapsis and the apoapsis are the same points for the two anomalies, as they must be by symmetry.

Multiplying $\cos v$ and $\sin v$ by $r = a(1 - e\cos u)$ gives the position in the orbital plane, with x towards the periapsis,

$$r\cos v = a(\cos u - e), \qquad r\sin v = a\sqrt{1-e^2}\,\sin u. \tag{6.47}$$

This is the point $(a\cos u,\ b\sin u)$ of the opening, measured from the focus at $(ae, 0)$.

6.7.2 Hyperbolic trajectories

For $a < 0$ and $e > 1$, the matching substitution is $r = a(1 - e\cosh u) = (-a)(e\cosh u - 1)$. Since $\cosh u \geq 1$ and $a < 0$, it keeps $r > 0$. Multiplying the bracket of the time integral by $-a > 0$ instead of a,

$$\begin{aligned} t - t_0 &\overset{(1)}{=} \sqrt{\frac{-a}{GM}} \int_{-a(e-1)}^{r} \frac{r'\,dr'}{\sqrt{(r'-a)^2 - a^2e^2}} \\ &\overset{(2)}{=} \frac{(-a)^{3/2}}{\sqrt{GM}} \int_0^u (e\cosh u' - 1)\,du' = \frac{(-a)^{3/2}}{\sqrt{GM}}\big[e\sinh u - u\big]. \end{aligned} \tag{6.48}$$

(1) $-a\big(2r - r^2/a - \lambda\big) = (r-a)^2 - a^2e^2$, with $\lambda = a(1-e^2)$. The periapsis is at $r = -a(e-1)$.

(2) $dr = (-a)e\sinh u\,du$, and $(r-a)^2 - a^2e^2 = a^2e^2(\cosh^2 u - 1) = a^2e^2\sinh^2 u$, whose root is $(-a)e\sinh u$ for $u > 0$.

This is the **hyperbolic** anomaly relation. The true anomaly follows as for the ellipse, with $\cosh u$ in place of $\cos u$. Equating $\lambda/(1 + e\cos v)$ with $a(1 - e\cosh u)$ gives $\cos v = (\cosh u - e)/(1 - e\cosh u)$. Then $1 - \cos^2 v = (e^2-1)\sinh^2 u/(e\cosh u - 1)^2$, and the half-angle form follows from

$$\frac{1-\cos v}{1+\cos v} = \frac{(e+1)(\cosh u - 1)}{(e-1)(\cosh u + 1)} = \frac{e+1}{e-1}\tanh^2\frac{u}{2}.$$

The true-anomaly relations are

$$\cos v = \frac{e - \cosh u}{e\cosh u - 1}, \qquad \sin v = \frac{\sqrt{e^2-1}\,\sinh u}{e\cosh u - 1}, \qquad \tan\frac{v}{2} = \sqrt{\frac{e+1}{e-1}}\tanh\frac{u}{2}. \tag{6.49}$$

6.8 Kepler's Third Law

We have the time $t(u)$ along an elliptical orbit, Eq. (6.43). What we usually measure first for a binary or a planet is the period, the one time scale of the orbit.

A pendulum clock is set by the length of its pendulum alone. A longer pendulum swings more slowly, in a fixed way, whatever the size of its swing. An orbit is set the same way by its semi-major axis alone. The idea is that one full turn of u is one period, and the eccentricity drops out.

For Earth, $a = 1$ AU and $GM_\odot = 1.327\times10^{20}\ \mathrm{m^3\,s^{-2}}$ give a period of 365.26 days from $2\pi\sqrt{a^3/GM}$. Mars, with $a = 1.524$ AU, takes $1.524^{3/2} = 1.881$ years. The period depends on the orbit's size and the total mass, but not its eccentricity. That independence lets us weigh a pair even when its orbit is far from circular.

The period. In one complete orbit u advances by 2π, while t advances by the orbital period P. We define the **mean motion** $n \equiv 2\pi/P$, the mean angular rate. From Eq. (6.43),

$$\frac{2\pi}{n} = \frac{a^{3/2}}{\sqrt{GM}}\big[2\pi - e\sin(2\pi)\big] = \frac{a^{3/2}}{\sqrt{GM}}\cdot 2\pi \quad\Longrightarrow\quad \boxed{n = \sqrt{\frac{GM}{a^3}}, \qquad n^2a^3 = GM.} \tag{6.50}$$

The period depends on a and the total mass M only. It does not depend on e or on how M is split between the two bodies.

6.9 Kepler's Equation

We have t as a function of u, and the mean motion n. An observer has the time t and needs the position, so the relation must be run the other way.

Two runners start together on a circular track. One keeps a steady pace. The other speeds up on one side of the track and slows down on the other, and both finish the lap together. The steady runner's position is easy to give at any time, and the other's position can be named by how far ahead or behind the steady one she is. For an orbit the steady runner is a fictitious body moving round at the mean rate n, like the mean Sun of Chapter 1. Its angle is the **mean anomaly** $\ell = n(t - t_0)$. The idea is that Kepler's equation ties ℓ to the real body's eccentric anomaly u.

For $e = 0.5$, a quarter of a period after periapsis the mean anomaly is $\ell = 90°$. Kepler's equation, $u - 0.5 \sin u = \pi/2$, gives $u = 115.8°$; Eq. (6.44) then gives the true anomaly $v = 140.2°$. The body is already more than a third of the way around because it moved fastest near periapsis. We now derive the equation that connects the uniform clock angle ℓ to the body's position.

Mean anomaly. Defining the **mean anomaly** $\ell \equiv n(t - t_0)$,

$$\ell \overset{(1)}{=} n \frac{a^{3/2}}{\sqrt{GM}} (u - e \sin u) \overset{(2)}{=} u - e \sin u, \tag{6.51}$$

(1) Eq. (6.43).
(2) $n = \sqrt{GM/a^3}$, Eq. (6.50).

so that

$$\boxed{\ell = u - e \sin u} \qquad \text{(elliptical Kepler's equation).} \tag{6.52}$$

This transcendental equation must be solved for $u(\ell)$, that is $u(t)$, numerically (Problem 3) or by a series (Section 6.12). The hyperbolic analogue, with $n \equiv \sqrt{GM/(-a)^3}$, is

$$\ell = e \sinh u - u. \tag{6.53}$$

6.10 Keplerian Parametrization

With Kepler's equation the position follows at any time. We collect the formulas, then add the rates $\dot{u}$, $\dot{r}$ and $\dot{\phi}$ that the perturbation theory and the timing of binaries need.

6.10.1 Elliptical orbits

Collecting Eqs. (6.42), (6.46) and (6.52),

$$r = a(1 - e \cos u), \qquad \phi = v + \omega, \tag{6.54}$$

$$\tan \frac{v}{2} = \sqrt{\frac{1+e}{1-e}} \tan \frac{u}{2}, \qquad \ell = u - e \sin u. \tag{6.55}$$

Differentiating Kepler's equation (6.52) in t gives $n = \dot{u}(1 - e \cos u)$, since $\dot{\ell} = n$. So

$$\dot{u} = \frac{n}{1 - e \cos u}. \tag{6.56}$$

Differentiating $r = a(1 - e \cos u)$,

$$\dot{r} = ae \sin u \, \dot{u} = \frac{an \, e \sin u}{1 - e \cos u}. \tag{6.57}$$

For $\dot\phi$ we need dv/du. Differentiating $\cos v$ from Eq. (6.44),

$$-\sin v\,\frac{dv}{du} \overset{(1)}{=} \frac{-\sin u\,(1-e\cos u)-(\cos u-e)\,e\sin u}{(1-e\cos u)^2} = -\frac{(1-e^2)\sin u}{(1-e\cos u)^2},$$
$$\frac{dv}{du} \overset{(2)}{=} \frac{(1-e^2)\sin u}{(1-e\cos u)^2}\cdot\frac{1-e\cos u}{\sqrt{1-e^2}\sin u} = \frac{\sqrt{1-e^2}}{1-e\cos u}. \tag{6.58}$$

(1) The quotient rule.
(2) Divide by $-\sin v$, with $\sin v$ from Eq. (6.46).

So

$$\dot\phi = \dot v = \frac{dv}{du}\dot u = \frac{n\sqrt{1-e^2}}{(1-e\cos u)^2}. \tag{6.59}$$

As a check, Eq. (6.22) gives $\dot\phi = L/(\mu r^2) = \sqrt{GMa(1-e^2)}/\left[a^2(1-e\cos u)^2\right]$, which is the same since $\sqrt{GM/a^3} = n$.

6.10.2 Hyperbolic trajectories

In the same way, with $r = -a(e\cosh u - 1)$, $\phi = v + \omega$, $\tan(v/2) = \sqrt{(e+1)/(e-1)}\tanh(u/2)$ and $\ell = e\sinh u - u$,

$$\dot u = \frac{n}{e\cosh u - 1}, \quad \dot r = -an\,\frac{e\sinh u}{e\cosh u - 1}, \quad \dot\phi = \dot v = \frac{n\sqrt{e^2-1}}{(e\cosh u-1)^2}. \tag{6.60}$$

The first is the derivative of $\ell = e\sinh u - u$, the second is $\dot r = (-a)e\sinh u\,\dot u$, and the third uses $dv/du = \sqrt{e^2-1}/(e\cosh u - 1)$, found from $\cos v$ as for the ellipse.

6.11 Hyperbolic-Trajectory Parameters

A body that passes a planet or a star once, on an unbound orbit with $e > 1$, is described by a few numbers read off the hyperbola. We find the speed far away, the angle of the asymptotes, the closest approach, the miss distance and the angle of the velocity. Problem 12 uses them for a flyby of Jupiter.

Hyperbolic excess speed v_∞. Far away $r \to \infty$ and $V \to 0$, so the energy is all kinetic. From $\frac{1}{2}\mu\dot{\vec r}_\infty^{\,2} = E = -GM\mu/(2a)$, with $a < 0$,

$$v_\infty \equiv |\dot{\vec r}_\infty| = \sqrt{-\frac{GM}{a}}. \tag{6.61}$$

Asymptote angle θ_∞. As $u \to \infty$ the body runs off to infinity at the true anomaly v_{asy}, with $\cos v_{\text{asy}} = \lim(e-\cosh u)/(e\cosh u - 1) = -1/e$. The asymptote makes the angle $\theta_\infty = \pi - v_{\text{asy}}$ with the axis on the far side from the periapsis, which is half the angle between the two asymptotes,

$$\cos\theta_\infty = -\lim_{u\to\infty}\frac{e-\cosh u}{e\cosh u - 1} = \frac{1}{e}. \tag{6.62}$$

The body is turned by the **deflection angle** $\delta = \pi - 2\theta_\infty$, so $\sin(\delta/2) = \cos\theta_\infty = 1/e$.

Periapsis distance. The smallest r is at $v = 0$,

$$r_{\text{p}} = \min_v \frac{a(1-e^2)}{1+e\cos v} = \frac{a(1-e^2)}{1+e} = -a(e-1). \tag{6.63}$$

Semi-minor axis and impact parameter. The semi-minor axis $b = -a\sqrt{e^2-1}$ satisfies $b^2 = a^2(e^2-1) = -a\lambda$, since $\lambda = a(1-e^2)$. The impact parameter b' is the distance from the focus to the incoming asymptote, the miss distance had there been no pull. The asymptote passes through the centre of the hyperbola, which is $-a + r_\mathrm{p} = -ae$ from the focus, and it makes the angle θ_∞ with the axis, so

$$b' = (-a + r_\mathrm{p})\sin\theta_\infty = -ae\sin\theta_\infty \overset{(1)}{=} -ae\sqrt{1-\frac{1}{e^2}} = -a\sqrt{e^2-1} = b,$$

(1) $\cos\theta_\infty = 1/e$ from Eq. (6.62).

so the impact parameter and the semi-minor axis are the same length.

Flight-path angle ψ, the angle between $\dot{\vec{r}}$ and the in-plane direction $\hat{\phi}$ perpendicular to $\hat{r}$,

$$\sin\psi = \frac{\dot{\vec{r}}\cdot\hat{r}}{|\dot{\vec{r}}|} = \frac{\dot{r}}{\sqrt{\dot{r}^2 + r^2\dot{\phi}^2}}. \tag{6.64}$$

It is zero at the periapsis, where $\dot{r} = 0$.

6.12 Fourier Solution of Kepler's Equation

Kepler's equation gives ℓ from u in one line, but going from ℓ to u takes an iteration (Problem 3). For work by hand, in perturbation theory or for small e, we want u as an explicit function of ℓ.

A musician who hears a steady beat that is slightly uneven can describe it as the beat plus a few overtones. Anything that repeats can be written as a sum of sines of the repeat frequency and its multiples. The difference $u - \ell = e\sin u$ repeats every orbit, and it changes sign with ℓ. So it is a sum of $\sin(s\ell)$. The idea is to find the coefficients by changing the variable of the integral from ℓ to u, where Kepler's equation makes it simple.

For $e = 0.1$, the first correction is $2J_1(0.1)\sin\ell = 0.0999\sin\ell$, nearly $e\sin\ell$. The second is $J_2(0.2)\sin 2\ell = 0.0050\sin 2\ell$, nearly $\frac{1}{2}e^2\sin 2\ell$. Successive harmonics become smaller as powers of e, so a short series can approximate the orbital position when the eccentricity is modest. We now derive the coefficients of that series from Kepler's equation.

The coefficients. Kepler's equation $u - \ell = e\sin u$ is solved as a Fourier series in ℓ,

$$e\sin u = \sum_{s=1}^{\infty} c_s \sin(s\ell).$$

The function $u(\ell) - \ell$ is odd and has period 2π, and it vanishes at $\ell = 0$ and $\ell = \pi$, where $u = \ell$. So

$$\begin{aligned}
c_s &\overset{(1)}{=} \frac{2}{\pi}\int_0^\pi (u-\ell)\sin(s\ell)\,d\ell \\
&\overset{(2)}{=} \frac{2}{s\pi}\Big[-(u-\ell)\cos(s\ell)\Big]_0^\pi + \frac{2}{s\pi}\int_0^\pi \cos(s\ell)\left(\frac{du}{d\ell}-1\right)d\ell \\
&\overset{(3)}{=} \frac{2}{s\pi}\int_0^\pi \cos(s\ell)\,du \\
&\overset{(4)}{=} \frac{2}{s\pi}\int_0^\pi \cos(su - se\sin u)\,du \overset{(5)}{=} \frac{2}{s}J_s(se).
\end{aligned} \tag{6.65}$$

(1) The Fourier sine coefficient of an odd function of period 2π.
(2) Integration by parts.
(3) The boundary term vanishes since $u = \ell$ at both ends. $\int_0^\pi \cos(s\ell)\,d\ell = 0$ for a whole number $s \geq 1$. Then $(du/d\ell)\,d\ell = du$, and u runs from 0 to π as ℓ does.
(4) $\ell = u - e\sin u$, Kepler's equation.
(5) The integral form of the Bessel function of the first kind, $J_s(x) = \frac{1}{\pi}\int_0^\pi \cos(s\tau - x\sin\tau)\,d\tau$.

So

$$u(\ell) = \ell + \sum_{s=1}^{\infty} \frac{2}{s} J_s(se)\sin(s\ell). \tag{6.66}$$

This is the classical Bessel-function series solution of Kepler's equation. Since $J_s(se) \sim (se/2)^s/s!$ for small e, the s-th term is of order e^s.

6.13 The Laplace–Runge–Lenz Vector

The energy and the angular momentum fix four of the six elements, a, e, ι and Ω. They do not say where the periapsis points within the plane. The angle ω came out of solving the orbit equation, not from anything conserved. Yet the ellipse of a $1/r$ potential does not turn, so something must hold its long axis in place.

A marble rolling round the inside of a salad bowl traces a loop that slowly turns, a rosette. In a bowl where the restoring force is exactly proportional to the distance from the centre, the loop closes on itself and its long axis stays put. Energy and angular momentum are the same in both bowls, so in the second something more is conserved. The Kepler ellipse does not turn either. The idea is that a conserved vector $\vec{A}$ points from the focus to the periapsis, and its direction gives ω.

Earth's eccentricity is $e = 0.0167$. In units of $GM\mu^2$, the vector $\vec{A}$ has that length and points toward perihelion, reached in early January. If the force is exactly inverse-square, the arrow stays fixed while Earth goes around. An additional force slowly turns it, making periapsis precess. We will first show why $\vec{A}$ is conserved in the Kepler problem, then use its direction to describe what changes under a perturbation.

Conservation. Define $\vec{A} \equiv \vec{p}\times\vec{L} - GM\mu^2\hat{r}$, with $\vec{p} = \mu\dot{\vec{r}}$. For the pure Kepler problem,

$$\begin{aligned}
\dot{\vec{A}} &\overset{(1)}{=} \dot{\vec{p}}\times\vec{L} - GM\mu^2\dot{\hat{r}} \\
&\overset{(2)}{=} -\frac{GM\mu}{r^2} L\big(\hat{r}\times\hat{L}\big) - GM\mu^2\dot{\phi}\,\hat{\phi} \\
&\overset{(3)}{=} \frac{GM\mu L}{r^2}\,\hat{\phi} - \frac{GM\mu L}{r^2}\,\hat{\phi} = 0.
\end{aligned} \tag{6.67}$$

(1) The product rule, with $\dot{\vec{L}} = 0$.
(2) $\dot{\vec{p}} = -GM\mu\hat{r}/r^2$ from Eq. (6.3), $\vec{L} = L\hat{L}$, and $\dot{\hat{r}} = \dot{\phi}\,\hat{\phi}$ from Eq. (6.18).
(3) $(\hat{r}, \hat{\phi}, \hat{L})$ is right-handed, so $\hat{r}\times\hat{L} = -\hat{\phi}$. And $\dot{\phi} = L/(\mu r^2)$ from Eq. (6.22).

So $\vec{A}$ is conserved for the unperturbed problem. It is a constant of the Kepler problem beyond E and $\vec{L}$, and it reflects an extra hidden symmetry of the $1/r$ potential. For bound orbits, $\vec{L}$ and a rescaled $\vec{A}$ together generate the rotation group $SO(4)$ of four dimensions.

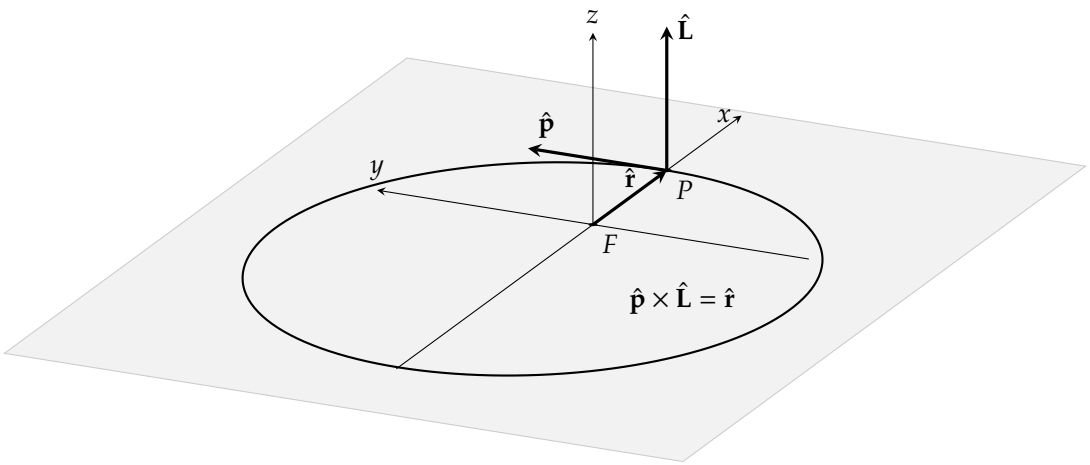


Figure 6.5: An orbit in its plane, with the focus F at the origin and the body at periapsis P. The unit vectors $\hat{\mathbf{r}}$, $\hat{\mathbf{p}}$ and $\hat{\mathbf{L}}$ are drawn at P, where $\hat{\mathbf{p}} \times \hat{\mathbf{L}} = \hat{\mathbf{r}}$.

Direction and magnitude of $\vec{A}$. The vector $\vec{A}$ lies in the orbital plane, since $\vec{A} \cdot \vec{L} = (\vec{p} \times \vec{L}) \cdot \vec{L} - GM\mu^2 \hat{r} \cdot \vec{L} = 0$. So we may write $\vec{A} = A_r \hat{r} + A_\phi \hat{\phi}$.

At periapsis the velocity is purely tangential, $\vec{p} \perp \hat{r}$. Then

$$\hat{r} \times \vec{A}\big|_{\text{peri}} = \hat{r} \times (\vec{p} \times \vec{L})\big|_{\text{peri}} = \big[\vec{p}\,(\hat{r} \cdot \vec{L}) - \vec{L}\,(\hat{r} \cdot \vec{p})\big]_{\text{peri}} = 0,$$

because both dot products vanish there. So $\vec{A}$ is parallel to $\hat{r}$ at periapsis (Figure 6.5). To find which way it points, and how long it is, we dot $\vec{A}$ into $\vec{r}$,

$$\begin{aligned} \vec{A} \cdot \vec{r} &= (\vec{p} \times \vec{L}) \cdot \vec{r} - GM\mu^2 r \\ &\overset{(1)}{=} \epsilon_{kij}\, p_i L_j\, r_k - GM\mu^2 r \overset{(2)}{=} \epsilon_{jki}\, r_k p_i\, L_j - GM\mu^2 r \\ &= \underbrace{(\vec{r} \times \vec{p})}_{\vec{L}} \cdot \vec{L} - GM\mu^2 r = L^2 - GM\mu^2 r. \end{aligned} \tag{6.68}$$

(1) $(\vec{p} \times \vec{L})_k = \epsilon_{kij}\, p_i L_j$, summed over repeated indices.
(2) The cyclic symmetry $\epsilon_{kij} = \epsilon_{jki}$, and $(\vec{r} \times \vec{p})_j = \epsilon_{jki}\, r_k p_i$.

At periapsis $r = \lambda/(1+e)$ and $L^2 = GM\mu^2\lambda$, so $\vec{A} \cdot \vec{r} = GM\mu^2\lambda\, e/(1+e) > 0$ for $e > 0$. Hence $\vec{A}$ points from the focus *towards the periapsis*, and the angle between $\vec{A}$ and $\vec{r}$ is the true anomaly v. Writing $\vec{A} \cdot \vec{r} = Ar\cos v$ and using

$$r = \frac{L^2/(GM\mu^2)}{1 + e\cos v} \implies GM\mu^2 r = \frac{L^2}{1 + e\cos v},$$

we find

$$\begin{aligned} Ar\cos v &= L^2 - GM\mu^2 r \\ A\,\frac{L^2/(GM\mu^2)}{1+e\cos v}\cos v &\overset{(1)}{=} L^2\left(1 - \frac{1}{1+e\cos v}\right) = \frac{L^2 e\cos v}{1+e\cos v}. \end{aligned} \tag{6.69}$$

(1) The conic (6.28) for r on the left, and $GM\mu^2 r = L^2/(1+e\cos v)$ on the right.

Cancelling $L^2 \cos v/(1 + e \cos v)$ on both sides,

$$\boxed{A = GM\mu^2 e.} \tag{6.70}$$

A second route runs the other way. Eq. (6.68) with $\vec{A} \cdot \vec{r} = Ar \cos v$ is $r = L^2/(GM\mu^2 + A \cos v) = \lambda/\big(1 + (A/GM\mu^2)\cos v\big)$. This is the conic (6.28) again, found from two conserved vectors without solving any differential equation.

The periapsis is at the polar angle ω, so by Eq. (6.13),

$$\hat{A} = \cos\omega\, \hat{n} + \sin\omega\, \hat{y}. \tag{6.71}$$

Equivalently, the **eccentricity vector**

$$\vec{e} \equiv \frac{\vec{A}}{GM\mu^2} = e\hat{A} = e_x\, \hat{n} + e_y\, \hat{y}, \qquad e_x \equiv e\cos\omega, \quad e_y \equiv e\sin\omega, \tag{6.72}$$

is a conserved vector that holds (e, ω) together. Since e, Ω and ι are already fixed by $(E, \vec{L})$, the only *new* information in $\vec{A}$ is ω, the direction of the periapsis from the ascending node.

How many new constants $\vec{A}$ gives

The three components of $\vec{A}$ obey two conditions set by E and $\vec{L}$. One is $\vec{A} \cdot \vec{L} = 0$. The other fixes its length, since

$$|\vec{A}|^2 = G^2M^2\mu^4 e^2 \overset{(1)}{=} G^2M^2\mu^4\left(1 + \frac{2EL^2}{G^2M^2\mu^3}\right) = G^2M^2\mu^4 + 2\mu EL^2,$$

where (1) is Eq. (6.39). So $\vec{A}$ supplies one new constant, not three. With E (one), $\vec{L}$ (three), $\vec{A}$ (one) and t_0 (one) the count of six for the relative orbit closes.

6.14 Perturbed Orbits

The Kepler orbit is exact for two point masses alone in space. Real orbits feel other forces as well. The other planets tug on each one, a flattened planet pulls its satellites off the inverse square, gas drags on low satellites, an engine pushes a spacecraft, and general relativity adds a small extra pull. The orbit is then no longer a fixed ellipse.

A navigator on a ship in a slow current plots, each hour, the course the ship would follow in still water from where it is now and how it is moving now. The current shifts that course a little from one hour to the next. The ship's true track is a chain of these still-water courses, each one touching the track for a moment. A planet tugged by the others is the ship. At each instant it moves on some Kepler ellipse that fits its present position and velocity, and the elements of that ellipse drift. The idea is to describe the true path by a Kepler ellipse whose elements change, at rates set by the extra force.

Mercury's perihelion turns about 575″ per century relative to the stars: roughly 532″ from the other planets and 43″ from general relativity. Over about 415 orbits per century, that is only 1.4″ per orbit. The ellipse changes a little each turn, so the full path slowly forms a rosette. We will describe it by an ellipse that matches the current position and velocity at each instant, then derive how the extra force changes its orbital elements.

Equation of motion. We add a force $\vec{F}$ to the relative equation of motion,

$$\boxed{\mu\ddot{\vec{r}} = -\frac{GM\mu}{r^2}\hat{r} + \vec{F}.} \tag{6.73}$$

The central Newtonian term defines the Kepler problem. The **perturbing force** $\vec{F}$ need not be central.

The momentum is $\vec{p} = \mu\dot{\vec{r}}$. In the orbital plane, with $\dot{\vec{r}} = \dot{r}\,\hat{r} + r\dot{\phi}\,\hat{\phi}$ and $L = \mu r^2\dot{\phi}$,

$$\vec{p} = \mu\dot{r}\,\hat{r} + \frac{L}{r}\,\hat{\phi}, \qquad L = \mu r^2\dot{\phi}. \tag{6.74}$$

Orbital perturbation theory represents the actual trajectory by Keplerian orbital elements that vary with time, and derives their rates directly from Eq. (6.73).

6.14.1 Osculating Kepler orbit

We first fix which Kepler ellipse belongs to the true path at each instant. For $\vec{F} = 0$, the Kepler solution can be written in terms of six constants

$$Q = (Q_1, \ldots, Q_6) = (a, e, \iota, \Omega, \omega, t_0), \tag{6.75}$$

so that

$$\vec{r} = \vec{r}(Q, t), \qquad \vec{p} = \vec{p}(Q, t). \tag{6.76}$$

For fixed Q, these functions satisfy the unperturbed equations of motion.

For the perturbed problem, let

$$Q_i = Q_i(t). \tag{6.77}$$

The actual trajectory is represented in terms of a set of time-dependent osculating orbital elements $Q(t)$ as

$$\vec{r}(t) = \vec{r}_{\rm osc}\big(Q(t), t\big), \qquad \vec{p}(t) = \vec{p}_{\rm osc}\big(Q(t), t\big). \tag{6.78}$$

Here, for fixed values of Q, the functions $\vec{r}_{\rm osc}(Q, t)$ and $\vec{p}_{\rm osc}(Q, t)$ describe the corresponding Keplerian orbit. The time dependence of $Q(t)$ accounts for the slow evolution of this osculating Keplerian orbit under the perturbing force.

At each instant, the osculating Kepler orbit is chosen to have the same position and momentum as the actual trajectory. Thus,

$$\vec{r}_{\rm osc}\big(Q(t), t\big) = \vec{r}(t), \qquad \vec{p}_{\rm osc}\big(Q(t), t\big) = \vec{p}(t). \tag{6.79}$$

This does not make the actual trajectory Keplerian. At each instant it is represented by the Keplerian orbit that touches it at that instant, the **osculating** orbit.

The corresponding accelerations are different,

$$\dot{\vec{p}}_{\rm actual} = -\frac{GM\mu}{r^2}\hat{r} + \vec{F}, \tag{6.80}$$

$$\dot{\vec{p}}_{\rm osc} = -\frac{GM\mu}{r^2}\hat{r}. \tag{6.81}$$

Hence

$$\boxed{\dot{\vec{p}}_{\rm actual} - \dot{\vec{p}}_{\rm osc} = \vec{F}.} \tag{6.82}$$

The two trajectories therefore agree in position and momentum at the instant at which the osculating orbit is constructed, but they have different subsequent evolution.

6.14.2 Variation of parameters

The osculating condition becomes six equations for the six rates $\dot{Q}_i$. Differentiating Eq. (6.78) by the chain rule,

$$\dot{\vec{r}} = \left.\frac{\partial \vec{r}}{\partial t}\right|_Q + \sum_{i=1}^{6} \frac{\partial \vec{r}}{\partial Q_i}\dot{Q}_i. \tag{6.83}$$

For fixed Q, the first term is the velocity of the Kepler solution. The osculating condition requires this Kepler velocity to coincide with the actual velocity,

$$\dot{\vec{r}} = \left.\frac{\partial \vec{r}}{\partial t}\right|_Q . \tag{6.84}$$

Therefore

$$\boxed{\sum_{i=1}^{6} \frac{\partial \vec{r}}{\partial Q_i}\dot{Q}_i = 0.} \tag{6.85}$$

This is a vector equation and therefore supplies three scalar conditions.

Now differentiate $\vec{p}(Q(t), t)$,

$$\dot{\vec{p}} = \left.\frac{\partial \vec{p}}{\partial t}\right|_Q + \sum_{i=1}^{6} \frac{\partial \vec{p}}{\partial Q_i}\dot{Q}_i. \tag{6.86}$$

For fixed Q, $\vec{p}(Q, t)$ obeys the Kepler equation,

$$\left.\frac{\partial \vec{p}}{\partial t}\right|_Q = -\frac{GM\mu}{r^2}\hat{r}. \tag{6.87}$$

The actual momentum obeys

$$\dot{\vec{p}} = -\frac{GM\mu}{r^2}\hat{r} + \vec{F}. \tag{6.88}$$

Subtracting the Kepler contribution gives

$$\boxed{\sum_{i=1}^{6} \frac{\partial \vec{p}}{\partial Q_i}\dot{Q}_i = \vec{F}.} \tag{6.89}$$

Equations (6.85) and (6.89) form six scalar equations for the six quantities $\dot{Q}_i$. This is the formal variation-of-parameters construction of the osculating elements.

Six equations for six unknowns, so the count closes. The perturbing force enters only through the change it makes in the instantaneous position and momentum, and the Kepler orbit is redefined at each instant from that state.

6.14.3 Perturbed Kepler invariants

The Keplerian energy, angular momentum and eccentricity vector are defined from the instantaneous state by

$$E \equiv \frac{p^2}{2\mu} - \frac{GM\mu}{r}, \tag{6.90}$$

$$\vec{L} \equiv \vec{r} \times \vec{p}, \tag{6.91}$$

$$\vec{A} \equiv \vec{p} \times \vec{L} - GM\mu^2\,\hat{r}. \tag{6.92}$$

For the unperturbed Kepler problem these are constants. In the perturbed problem they become time-dependent quantities.

Their magnitudes determine the instantaneous Keplerian parameters,

$$E = -\frac{GM\mu}{2a(t)}, \tag{6.93}$$

$$L^2 = GM\mu^2 a(t)(1 - e^2(t)), \tag{6.94}$$

$$A \equiv |\vec{A}| = GM\mu^2 e(t). \tag{6.95}$$

Thus a and e may be regarded as functions of the actual state even when the full trajectory is not Keplerian.

6.14.4 Exact Rate of the Energy

We differentiate Eq. (6.90) along the actual motion,

$$\begin{aligned}\dot{E} &\overset{(1)}{=} \frac{\vec{p}}{\mu}\cdot\dot{\vec{p}} + \frac{GM\mu}{r^2}\dot{r} \\ &\overset{(2)}{=} \dot{\vec{r}}\cdot\left(-\frac{GM\mu}{r^2}\hat{r} + \vec{F}\right) + \frac{GM\mu}{r^2}\dot{r} \\ &\overset{(3)}{=} -\frac{GM\mu}{r^2}\dot{r} + \dot{\vec{r}}\cdot\vec{F} + \frac{GM\mu}{r^2}\dot{r} = \dot{\vec{r}}\cdot\vec{F}.\end{aligned} \tag{6.96}$$

(1) $d(p^2)/dt = 2\vec{p}\cdot\dot{\vec{p}}$ and $d(1/r)/dt = -\dot{r}/r^2$.
(2) $\vec{p}/\mu = \dot{\vec{r}}$, and $\dot{\vec{p}}$ from the equation of motion (6.73).
(3) $\dot{\vec{r}}\cdot\hat{r} = \dot{r}$.

$$\boxed{\dot{E} = \dot{\vec{r}}\cdot\vec{F}.} \tag{6.97}$$

The Kepler energy changes only by the work done by the extra force.

6.14.5 Exact Rate of angular momentum

From $\vec{L} = \vec{r}\times\vec{p}$,

$$\dot{\vec{L}} \overset{(1)}{=} \dot{\vec{r}}\times\vec{p} + \vec{r}\times\dot{\vec{p}} \overset{(2)}{=} \vec{r}\times\left(-\frac{GM\mu}{r^2}\hat{r} + \vec{F}\right) \overset{(3)}{=} \vec{r}\times\vec{F}. \tag{6.98}$$

(1) The product rule.
(2) $\dot{\vec{r}}\times\vec{p} = \mu\,\dot{\vec{r}}\times\dot{\vec{r}} = 0$, and $\dot{\vec{p}}$ from Eq. (6.73).
(3) The central term vanishes because $\vec{r} \parallel \hat{r}$.

$$\boxed{\dot{\vec{L}} = \vec{r}\times\vec{F}.} \tag{6.99}$$

Only the part of $\vec{F}$ perpendicular to $\vec{r}$ exerts a torque.

6.14.6 Exact Rate of the Runge–Lenz vector

We need one kinematic fact first. Since $\hat{r} = \vec{r}/r$,

$$\dot{\hat{r}} = \frac{\dot{\vec{r}}}{r} - \frac{\dot{r}\,\vec{r}}{r^2} \overset{(1)}{=} \frac{r^2\dot{\vec{r}} - (\vec{r}\cdot\dot{\vec{r}})\,\vec{r}}{r^3} \overset{(2)}{=} \frac{(\vec{r}\times\dot{\vec{r}})\times\vec{r}}{r^3} = \frac{\vec{L}\times\vec{r}}{\mu r^3}. \tag{6.100}$$

(1) $r\dot{r} = \vec{r}\cdot\dot{\vec{r}}$.
(2) $(\vec{a}\times\vec{b})\times\vec{c} = \vec{b}\,(\vec{a}\cdot\vec{c}) - \vec{a}\,(\vec{b}\cdot\vec{c})$.

It uses only the definitions, so it holds with the perturbation on. Now we differentiate Eq. (6.92),

$$\begin{aligned}\dot{\vec{A}} &\overset{(1)}{=} \dot{\vec{p}}\times\vec{L} + \vec{p}\times\dot{\vec{L}} - GM\mu^2\,\dot{\hat{r}}\\ &\overset{(2)}{=} -\frac{GM\mu}{r^2}\hat{r}\times\vec{L} + \vec{F}\times\vec{L} + \vec{p}\times(\vec{r}\times\vec{F}) - GM\mu^2\,\dot{\hat{r}}\\ &\overset{(3)}{=} -\frac{GM\mu}{r^2}\hat{r}\times\vec{L} + \vec{F}\times\vec{L} + \vec{p}\times(\vec{r}\times\vec{F}) + \frac{GM\mu}{r^2}\,\hat{r}\times\vec{L}\\ &= \vec{F}\times\vec{L} + \vec{p}\times(\vec{r}\times\vec{F}).\end{aligned} \tag{6.101}$$

(1) The product rule.
(2) $\dot{\vec{p}}$ from Eq. (6.73) and $\dot{\vec{L}} = \vec{r}\times\vec{F}$ from Eq. (6.99).
(3) $-GM\mu^2\dot{\hat{r}} = -(GM\mu/r^2)\,\vec{L}\times\hat{r} = (GM\mu/r^2)\,\hat{r}\times\vec{L}$, from Eq. (6.100) with $\vec{r} = r\hat{r}$.

The two central terms cancel. Therefore

$$\boxed{\dot{\vec{A}} = \vec{F}\times\vec{L} + \vec{p}\times(\vec{r}\times\vec{F}).} \tag{6.102}$$

This equation is the starting point for the evolution of the eccentricity and of the direction of the periapsis.

6.15 Planar perturbations

We have exact rates for E, $\vec{L}$ and $\vec{A}$ under any extra force. To follow an orbit we need the rates of the elements a, e and ω themselves, written with the parts of the force along and across the radius.

Pushing a child on a swing adds energy only if the push is along the motion. A push at the top of the swing, where the child is at rest, does no work at all. When and in which direction we push decides what changes. An orbit is the same. A push along the motion changes the energy and so the size a. A push along the radius at the periapsis, where $\dot{r} = 0$, does no work, but it still changes the shape and turns the ellipse. The idea is to split $\vec{F}$ into its radial and angular parts and read off $\dot{a}$, $\dot{e}$ and $\dot{\omega}$ from the exact rates. These are the planar **Gauss equations**.

For example, a spacecraft on a circular Earth orbit of radius $a = 2.03\times10^4$ km fires a thruster along its motion at $0.2\,\mathrm{m\,s^{-2}}$ (Problem 8). The Gauss equation below gives $\dot{a} = 2\,(F_\phi/\mu)\sqrt{a^3/GM_\oplus} = 1.8\,\mathrm{km\,s^{-1}}$ at that instant, so the orbital size initially grows by almost 2 km each second. A drag force along the opposite direction would lower it instead. We now split any in-plane force into radial and transverse parts to see how each changes the size, shape and orientation of the orbit.

Force in the plane. We take a perturbing force confined to the orbital plane,

$$\boxed{\vec{F} = F_r\hat{r} + F_\phi\hat{\phi}.} \tag{6.103}$$

Its torque is $\vec{r}\times\vec{F} = rF_\phi\,\hat{r}\times\hat{\phi} = rF_\phi\,\hat{L}$, along $\hat{L}$. So $\dot{\vec{L}}$ is parallel to $\vec{L} = L\hat{L}$, and only the length of $\vec{L}$ changes. The plane itself stays fixed,

$$\boxed{i = 0, \qquad \dot{\Omega} = 0.} \tag{6.104}$$

In the orbital plane, by Eq. (6.74),

$$\vec{p} = \mu\dot{r}\,\hat{r} + \mu r\dot{\phi}\,\hat{\phi} = \mu\dot{r}\,\hat{r} + \frac{L}{r}\hat{\phi}. \tag{6.105}$$

The equations of motion are the $\hat{r}$ and $\hat{\phi}$ parts of Eq. (6.73), with the acceleration of Section 6.3,

$$\mu(\ddot{r} - r\dot{\phi}^2) = -\frac{GM\mu}{r^2} + F_r, \tag{6.106}$$

$$\mu(r\ddot{\phi} + 2\dot{r}\dot{\phi}) = F_\phi. \tag{6.107}$$

The second equation gives the rate of L,

$$\frac{dL}{dt} = \frac{d}{dt}(\mu r^2\dot{\phi}) = \mu\left(2r\dot{r}\dot{\phi} + r^2\ddot{\phi}\right) = r\,\mu\left(r\ddot{\phi} + 2\dot{r}\dot{\phi}\right) = rF_\phi, \tag{6.108}$$

so

$$\boxed{\dot{L} = rF_\phi.} \tag{6.109}$$

The energy equation (6.97) becomes

$$\dot{E} \overset{(1)}{=} \dot{r}F_r + r\dot{\phi}F_\phi \overset{(2)}{=} \dot{r}F_r + \frac{L}{\mu r}F_\phi. \tag{6.110}$$

(1) $\dot{\vec{r}} \cdot \vec{F}$ with $\dot{\vec{r}} = \dot{r}\,\hat{r} + r\dot{\phi}\,\hat{\phi}$.
(2) $r\dot{\phi} = L/(\mu r)$.

6.15.1 Keplerian geometry of the instantaneous ellipse

We need the osculating ellipse's r, $\dot{r}$ and $\dot{\phi}$ in terms of its elements. The true anomaly is

$$v \equiv \phi - \omega. \tag{6.111}$$

For the instantaneous Kepler orbit,

$$\boxed{r = \frac{\lambda}{1 + e\cos v},} \tag{6.112}$$

where

$$\boxed{\lambda \equiv a(1 - e^2) = \frac{L^2}{GM\mu^2}.} \tag{6.113}$$

The instantaneous angular velocity is

$$\dot{\phi} = \frac{L}{\mu r^2}. \tag{6.114}$$

For the radial velocity we start from the general case, where the elements depend on time,

$$a(t) = a_0 + \delta a(t), \qquad e(t) = e_0 + \delta e(t), \qquad \omega(t) = \omega_0 + \delta\omega(t), \tag{6.115}$$

and the changes δa, δe and $\delta\omega$ are produced by the perturbing force. Along the actual path, by the chain rule,

$$\dot{r} = \left.\frac{\partial r}{\partial t}\right|_Q + \sum_{i=1}^{6} \frac{\partial r}{\partial Q_i}\dot{Q}_i. \tag{6.116}$$

Since $r = \hat{r}\cdot\vec{r}$ and $\hat{r}\cdot\partial\hat{r}/\partial Q_i = 0$, each $\partial r/\partial Q_i$ is $\hat{r}\cdot\partial\vec{r}/\partial Q_i$. The sum is therefore the radial part of $\sum_i(\partial\vec{r}/\partial Q_i)\dot{Q}_i$, which vanishes by the osculating condition (6.85). So $\dot{r}$ is the radial velocity of the osculating ellipse with its elements held fixed. On that ellipse only the true anomaly moves, at the rate $\dot{v} = \dot{\phi}$ of Eq. (6.114), giving

$$\boxed{\dot{r} = \frac{dr}{dv}\dot{v} = \frac{\lambda e\sin v}{(1+e\cos v)^2}\frac{L}{\mu r^2} = \frac{Le\sin v}{\mu\lambda}.} \tag{6.117}$$

The last step uses $1/r^2 = (1+e\cos v)^2/\lambda^2$. The result is exact for the osculating elements.

Since

$$\frac{\lambda}{r} = 1 + e\cos v, \tag{6.118}$$

we also have

$$\boxed{\frac{\mu r\dot{r}}{L} = \frac{e\sin v}{1+e\cos v}.} \tag{6.119}$$

We name this ratio,

$$\boxed{\xi \equiv \frac{e\sin v}{1+e\cos v}.} \tag{6.120}$$

6.15.2 Rate of the semi-major axis

At a given instant the osculating orbit has the same energy as the actual orbit, so its semi-major axis $a(t)$ follows the energy. Its Keplerian energy and its rate are

$$E\Big|_{\text{osculating}} = -\frac{GM\mu}{2a}, \qquad \therefore \qquad \dot{E}\Big|_{\text{osculating}} = \frac{GM\mu}{2a^2}\dot{a}. \tag{6.121}$$

The right side of Eq. (6.110) is evaluated with the osculating r, $\dot{r}$ and L,

$$\dot{E}\Big|_{\text{osculating}} = \dot{r}F_r + \frac{L}{\mu r}F_\phi \overset{(1)}{=} \frac{L}{\mu\lambda}e\sin v\,F_r + \frac{L}{\mu r}F_\phi. \tag{6.122}$$

(1) $\dot{r} = Le\sin v/(\mu\lambda)$, Eq. (6.117).

Every quantity here is the osculating value, so this is exact. Since the perturbing force is small, $F_r, F_\phi = O(F)$, the rates of the elements are $O(F)$ too. To first order the quantities multiplying F_r and F_ϕ may therefore be evaluated on the unperturbed Keplerian orbit, and this is what the orbit averages of Section 6.16 use. Equating with Eq. (6.121),

$$\begin{aligned}\dot{a} &= \frac{2a^2L}{GM\mu^2}\left[\frac{e\sin v}{\lambda}F_r + \frac{1}{r}F_\phi\right] \\ &\overset{(1)}{=} \frac{2a^2\sqrt{GM\lambda}}{GM\mu\lambda}\left[e\sin v\,F_r + (1+e\cos v)F_\phi\right] \overset{(2)}{=} \sqrt{\frac{4a^3}{GM\mu^2(1-e^2)}}\left[e\sin v\,F_r + (1+e\cos v)F_\phi\right].\end{aligned} \tag{6.123}$$

(1) $L = \mu\sqrt{GM\lambda}$ and $1/r = (1+e\cos v)/\lambda$.
(2) $2a^2/(\mu\sqrt{GM\lambda}) = \sqrt{4a^4/(GM\mu^2 a(1-e^2))}$, with $\lambda = a(1-e^2)$.

Here we used

$$L = \mu\sqrt{GM\lambda}, \qquad \lambda = a(1-e^2). \tag{6.124}$$

So

$$\boxed{\dot{a} = \sqrt{\frac{4a^3}{GM\mu^2(1-e^2)}}\left[e\sin v\,F_r + (1+e\cos v)F_\phi\right].} \tag{6.125}$$

6.15.3 Rate of the eccentricity

Since $A = GM\mu^2 e$, it is enough to find $\dot{A}$ from Eq. (6.102). For the planar force, with $\vec{L} = L\hat{L}$, the two terms are

$$\vec{F} \times \vec{L} = L\left(F_r \hat{r} + F_\phi \hat{\phi}\right) \times \hat{L} \overset{(1)}{=} L\left(-F_r \hat{\phi} + F_\phi \hat{r}\right),$$
$$\vec{p} \times (\vec{r} \times \vec{F}) \overset{(2)}{=} \left(\mu \dot{r} \hat{r} + \frac{L}{r}\hat{\phi}\right) \times \left(r F_\phi \hat{L}\right) \overset{(1)}{=} L F_\phi \hat{r} - \mu r \dot{r} F_\phi \hat{\phi}. \tag{6.126}$$

(1) $\hat{r} \times \hat{L} = -\hat{\phi}$ and $\hat{\phi} \times \hat{L} = \hat{r}$, since $(\hat{r}, \hat{\phi}, \hat{L})$ is right-handed.

(2) $\vec{p}$ from Eq. (6.105), and $\vec{r} \times \vec{F} = r F_\phi \hat{L}$.

Adding them,

$$\dot{\vec{A}} = \vec{F} \times \vec{L} + \vec{p} \times (\vec{r} \times \vec{F}) = 2L F_\phi \hat{r} - \left(L F_r + \mu r \dot{r} F_\phi\right) \hat{\phi}. \tag{6.127}$$

With $\mu r \dot{r} = L\xi$ from Eq. (6.120),

$$\boxed{\dot{\vec{A}} = L\left[2F_\phi \hat{r} - (F_r + \xi F_\phi)\hat{\phi}\right].} \tag{6.128}$$

For a purely radial force this is $\dot{\vec{A}} = -F_r L\, \hat{\phi}$.

The unit vector $\hat{A}$ points to the periapsis, and the true anomaly is measured from it. Let $\hat{q}$ be $\hat{A}$ turned by 90° towards increasing v. Then $(\hat{r}, \hat{\phi})$ is $(\hat{A}, \hat{q})$ turned by v, and the inverse turn is the transpose,

$$\begin{pmatrix} \hat{r} \\ \hat{\phi} \end{pmatrix} = \begin{pmatrix} \cos v & \sin v \\ -\sin v & \cos v \end{pmatrix} \begin{pmatrix} \hat{A} \\ \hat{q} \end{pmatrix} \quad \Longrightarrow \quad \begin{pmatrix} \hat{A} \\ \hat{q} \end{pmatrix} = \begin{pmatrix} \cos v & -\sin v \\ \sin v & \cos v \end{pmatrix} \begin{pmatrix} \hat{r} \\ \hat{\phi} \end{pmatrix}. \tag{6.129}$$

The first row is

$$\boxed{\hat{A} = \cos v\, \hat{r} - \sin v\, \hat{\phi}.} \tag{6.130}$$

Writing $\vec{A} = A\hat{A}$,

$$\dot{\vec{A}} = \dot{A}\hat{A} + A\dot{\hat{A}}, \tag{6.131}$$
$$\dot{\vec{A}} \cdot \hat{A} \overset{(1)}{=} \dot{A}, \tag{6.132}$$

(1) $\hat{A} \cdot \hat{A} = 1$, so $2\hat{A} \cdot \dot{\hat{A}} = 0$.

and therefore

$$\dot{A} = \dot{\vec{A}} \cdot \hat{A} \overset{(1)}{=} L\left[2F_\phi \cos v + (F_r + \xi F_\phi)\sin v\right], \tag{6.133}$$

(1) Eq. (6.128) dotted into Eq. (6.130), with $\hat{r} \cdot \hat{r} = \hat{\phi} \cdot \hat{\phi} = 1$ and $\hat{r} \cdot \hat{\phi} = 0$.

that is,

$$\boxed{\dot{A} = L\left[F_r \sin v + F_\phi\left(2\cos v + \xi \sin v\right)\right].} \tag{6.134}$$

Up to this point all relations have been exact. We now write both sides with the osculating elements, $A = GM\mu^2 e$ and $L = \sqrt{GM\mu^2 a(1-e^2)}$,

$$\dot{A}\Big|_{\text{osculating}} = L\left[F_r \sin v + F_\phi\left(2\cos v + \xi \sin v\right)\right]\Big|_{\text{osculating}}$$
$$GM\mu^2 \dot{e} = \sqrt{GM\mu^2 a(1-e^2)}\left[F_r \sin v + F_\phi\left(2\cos v + \xi \sin v\right)\right], \tag{6.135}$$

and obtain

$$\boxed{\dot{e} = \sqrt{\frac{a(1-e^2)}{GM\mu^2}}\left[F_r \sin v + F_\phi\,(2\cos v + \xi \sin v)\right].} \tag{6.136}$$

6.15.4 Rate of periapsis

The direction of $\vec{A}$ gives ω. The unit vector perpendicular to $\hat{A}$, towards increasing true anomaly, is the second row of Eq. (6.129),

$$\boxed{\hat{q} = \sin v\,\hat{r} + \cos v\,\hat{\phi}.} \tag{6.137}$$

Differentiating Eq. (6.130),

$$\begin{aligned}\dot{\hat{A}} &= -\dot{v}\sin v\,\hat{r} + \cos v\,\dot{\hat{r}} - \dot{v}\cos v\,\hat{\phi} - \sin v\,\dot{\hat{\phi}} \\ &\overset{(1)}{=} (\dot{\phi} - \dot{v})\left(\sin v\,\hat{r} + \cos v\,\hat{\phi}\right) \overset{(2)}{=} \dot{\omega}\,\hat{q}.\end{aligned} \tag{6.138}$$

(1) $\dot{\hat{r}} = \dot{\phi}\,\hat{\phi}$ and $\dot{\hat{\phi}} = -\dot{\phi}\,\hat{r}$, Eq. (6.18).
(2) $v = \phi - \omega$ gives $\dot{v} = \dot{\phi} - \dot{\omega}$, and Eq. (6.137).

Since

$$\dot{\vec{A}} = \dot{A}\,\hat{A} + A\dot{\hat{A}} = \dot{A}\,\hat{A} + A\dot{\omega}\,\hat{q}, \tag{6.139}$$

projection onto $\hat{q}$ gives

$$A\dot{\omega} = \dot{\vec{A}} \cdot \hat{q}. \tag{6.140}$$

Using Eqs. (6.128) and (6.137),

$$A\dot{\omega}\Big|_{\text{osculating}} = L\left[2F_\phi \sin v - (F_r + \xi F_\phi)\cos v\right]\Big|_{\text{osculating}}. \tag{6.141}$$

With $A = GM\mu^2 e$,

$$\boxed{\dot{\omega} = -\frac{L}{GM\mu^2 e}\left[F_r \cos v + F_\phi\,(\xi\cos v - 2\sin v)\right].} \tag{6.142}$$

Using

$$\frac{L}{GM\mu^2 e} = \frac{1}{e}\sqrt{\frac{a(1-e^2)}{GM\mu^2}}, \tag{6.143}$$

we obtain

$$\boxed{\dot{\omega} = -\frac{1}{e}\sqrt{\frac{a(1-e^2)}{GM\mu^2}}\left[F_r \cos v + F_\phi\,(\xi\cos v - 2\sin v)\right].} \tag{6.144}$$

6.15.5 Rate of the true anomaly

Since $v = \phi - \omega$, we have $\dot{v} = \dot{\phi} - \dot{\omega}$. The instantaneous Keplerian value of $\dot{\phi}$ is $L/(\mu r^2)$, Eq. (6.114). With the osculating orbit,

$$\begin{aligned}\dot{v} &= \frac{L}{\mu r^2} - \dot{\omega}\Big|_{\text{osculating}} \\ &\overset{(1)}{=} \frac{\mu\sqrt{GMa(1-e^2)}\,(1+e\cos v)^2}{\mu\, a^2(1-e^2)^2} - \dot{\omega} \overset{(2)}{=} \sqrt{\frac{GM}{a^3}}\frac{(1+e\cos v)^2}{(1-e^2)^{3/2}} - \dot{\omega}.\end{aligned} \tag{6.145}$$

(1) $r = a(1-e^2)/(1+e\cos v)$ and $L = \mu\sqrt{GMa(1-e^2)}$.
(2) $\sqrt{a(1-e^2)}/\left(a^2(1-e^2)^2\right) = a^{-3/2}(1-e^2)^{-3/2}$.

6.15.6 Planar Gauss equations

Collecting the results,

$$\dot{a} = \sqrt{\frac{4a^3}{GM\mu^2(1-e^2)}}\left[e\sin v\, F_r + (1+e\cos v)F_\phi\right], \tag{6.146}$$

$$\dot{e} = \sqrt{\frac{a(1-e^2)}{GM\mu^2}}\left[F_r\sin v + F_\phi\,(2\cos v + \xi\sin v)\right], \tag{6.147}$$

$$\dot{\omega} = -\frac{1}{e}\sqrt{\frac{a(1-e^2)}{GM\mu^2}}\left[F_r\cos v + F_\phi\,(\xi\cos v - 2\sin v)\right], \tag{6.148}$$

$$\dot{v} = \sqrt{\frac{GM}{a^3}}\frac{(1+e\cos v)^2}{(1-e^2)^{3/2}} - \dot{\omega}, \tag{6.149}$$

where

$$\xi = \frac{e\sin v}{1+e\cos v}. \tag{6.150}$$

For a perturbation confined to the plane,

$$\boxed{i = \dot{\Omega} = 0.} \tag{6.151}$$

These are equations for the *osculating* elements. The quantities a, e and ω are therefore functions of time, even though the actual position and momentum at each instant are those of a Keplerian orbit. Only the force in the plane enters. A push along the motion (F_ϕ at $e = 0$) changes a, while a radial push changes e and ω but, at the apsides, not a.

6.16 Orbit-averaged evolution

The Gauss equations give the rates of the elements at each instant. Over one orbit most of that change goes forward and then back again. Over many orbits only the net drift matters, and the instantaneous rates hide it.

The sea rises and falls twice a day with the tides. To see the sea level creep up over decades, we average over each day and watch how the daily average changes. The elements of an orbit wobble within each orbit in the same way. The idea is to average the rates over one unperturbed Kepler orbit. To first order in the force, what is left is the slow, **secular** drift.

For the example below, an extra radial force χ/r^3 with $\chi\mu/L^2 = 0.01$ turns periapsis by $-\pi \times 0.01 = -0.031$ rad $= -1.8°$ per orbit. The size a and eccentricity e return to their starting values after each turn. An instantaneous rate can therefore be nonzero even when its average is zero. We will average the Gauss equations over one unperturbed orbit to isolate the changes that accumulate over many turns.

The average. For any quantity X,

$$\boxed{\langle\dot{X}\rangle = \frac{1}{P}\int_0^P \dot{X}\,dt.} \tag{6.152}$$

For an unperturbed Kepler ellipse the elements are fixed, so $\dot{v} = \dot{\phi}$ and

$$L = \mu r^2 \dot{v}, \tag{6.153}$$

and therefore

$$\boxed{dt = \frac{\mu r^2}{L}\, dv.} \tag{6.154}$$

On the true path $\dot{v} = \dot{\phi} - \dot{\omega}$ differs from $\dot{\phi}$ by $\dot{\omega} = O(F)$. Inside an average that is already multiplied by F, the difference between $d\phi$ and dv changes the result only at $O(F^2)$.

The average rate can therefore be written as

$$\left\langle \dot{X} \right\rangle = \frac{1}{P} \int_0^{2\pi} \dot{X} \frac{\mu r^2}{L}\, dv. \tag{6.155}$$

A nonzero value of $\left\langle \dot{X} \right\rangle$ is a secular change. Terms whose integral over $0 \le v \le 2\pi$ vanishes describe only oscillations within each orbit.

6.16.1 Example: Perturbation by an Inverse-Cube Force $F_r = \chi/r^3$

We take the perturbing force

$$\boxed{\vec{F} = \frac{\chi}{r^3}\hat{r},} \tag{6.156}$$

with a constant χ that is positive for a push outward. Thus

$$F_r = \frac{\chi}{r^3}, \qquad F_\phi = 0. \tag{6.157}$$

We find the three rates directly from E, L and $\vec{A}$, and then compare with the Gauss equations.

Rate of the semimajor axis

From Eq. (6.121),

$$\dot{E} = \frac{GM\mu}{2a^2}\dot{a}. \tag{6.158}$$

The exact rate of change of the energy of the actual orbit is, by Eq. (6.97),

$$\dot{E} = \vec{F} \cdot \dot{\vec{r}}. \tag{6.159}$$

The perturbation is radial and $\dot{\vec{r}} = \dot{r}\,\hat{r} + r\dot{\phi}\,\hat{\phi}$, so only the radial velocity does work,

$$\dot{E} = \frac{\chi}{r^3}\dot{r}. \tag{6.160}$$

For the osculating Keplerian orbit, Eq. (6.117) with $\lambda = L^2/(GM\mu^2)$ gives

$$\dot{r} = \frac{Le\sin v}{\mu}\,\frac{GM\mu^2}{L^2} = \frac{GM\mu}{L} e \sin v. \tag{6.161}$$

Therefore

$$\dot{E} = \frac{\chi GM\mu}{Lr^3} e \sin v. \tag{6.162}$$

Comparing this result with Eq. (6.158),

$$\frac{GM\mu}{2a^2}\dot{a} = \frac{\chi GM\mu}{Lr^3} e \sin v, \tag{6.163}$$

so that

$$\boxed{\dot{a} = \frac{2\chi a^2 e \sin v}{Lr^3}}. \tag{6.164}$$

Rate of the eccentricity

The Runge–Lenz vector is

$$\vec{A} = \vec{p} \times \vec{L} - GM\mu^2 \hat{r}, \tag{6.165}$$

and for the osculating Keplerian orbit

$$A = GM\mu^2 e. \tag{6.166}$$

Since the perturbing force is radial, it exerts no torque,

$$\dot{\vec{L}} = \vec{r} \times \vec{F} = \vec{r} \times \frac{\chi}{r^3}\hat{r} = 0. \tag{6.167}$$

Differentiating $L^2 = GM\mu^2 a(1-e^2)$ gives

$$2L\dot{L} = GM\mu^2 \left[(1-e^2)\dot{a} - 2ae\dot{e}\right], \tag{6.168}$$

and the left side is zero. Therefore

$$\dot{e} = \frac{1-e^2}{2ae}\dot{a}. \tag{6.169}$$

Substituting Eq. (6.164),

$$\dot{e} = \frac{1-e^2}{2ae}\,\frac{2\chi a^2 e \sin v}{Lr^3} = \frac{\chi a(1-e^2)\sin v}{Lr^3}. \tag{6.170}$$

As a check, the Gauss equation (6.147) with $F_\phi = 0$ gives $\dot{e} = \sqrt{a(1-e^2)/(GM\mu^2)}\,\chi \sin v / r^3$. With $L = \mu\sqrt{GMa(1-e^2)}$ the two are the same.

Rate of the argument of periapsis

The direction of $\vec{A}$ gives the periapsis. Differentiating Eq. (6.165),

$$\dot{\vec{A}} = \dot{\vec{p}} \times \vec{L} + \vec{p} \times \dot{\vec{L}} - GM\mu^2 \dot{\hat{r}}. \tag{6.171}$$

Since $\dot{\vec{L}} = 0$,

$$\dot{\vec{A}} = \dot{\vec{p}} \times \vec{L} - GM\mu^2 \dot{\hat{r}}. \tag{6.172}$$

The momentum equation is

$$\dot{\vec{p}} = -\frac{GM\mu}{r^2}\hat{r} + \frac{\chi}{r^3}\hat{r}. \tag{6.173}$$

Therefore

$$\dot{\vec{A}} = \left(-\frac{GM\mu}{r^2} + \frac{\chi}{r^3}\right) \underbrace{\hat{r} \times \vec{L}}_{=-L\hat{\phi}} \underbrace{- GM\mu^2 \underbrace{\dot{\hat{r}}}_{=\dot{\phi}\,\hat{\phi}}}_{=\frac{GM\mu L}{r^2}\hat{\phi}}. \tag{6.174}$$

The terms with the central gravitational force cancel, leaving

$$\dot{\vec{A}} = -\frac{\chi L}{r^3}\hat{\phi}. \tag{6.175}$$

We use the periapsis frame $(\hat{A}, \hat{q})$ of Eq. (6.129). Here $\hat{A}$ points toward the periapsis, $\hat{A} = \cos\omega\,\hat{n} + \sin\omega\,\hat{y}$ by Eq. (6.71), and $\hat{q}$ is $\hat{A}$ turned by 90° towards increasing true anomaly. Then

$$\vec{A} = A\hat{A}, \tag{6.176}$$

and, since $\dot{\hat{A}} = \dot{\omega}\,\hat{q}$ (Section 6.15),

$$\dot{\vec{A}} = \dot{A}\,\hat{A} + A\dot{\omega}\,\hat{q}. \tag{6.177}$$

The radial and angular unit vectors in the periapsis frame are, from Eq. (6.129),

$$\hat{r} = \cos v\,\hat{A} + \sin v\,\hat{q}, \tag{6.178}$$

and

$$\hat{\phi} = -\sin v\,\hat{A} + \cos v\,\hat{q}. \tag{6.179}$$

Substituting the latter into Eq. (6.175),

$$\dot{\vec{A}} = -\frac{\chi L}{r^3}\left(-\sin v\,\hat{A} + \cos v\,\hat{q}\right) = \frac{\chi L}{r^3}\sin v\,\hat{A} - \frac{\chi L}{r^3}\cos v\,\hat{q}. \tag{6.180}$$

Comparing this with Eq. (6.177),

$$\dot{A} = \frac{\chi L}{r^3}\sin v, \tag{6.181}$$

and

$$A\dot{\omega} = -\frac{\chi L}{r^3}\cos v. \tag{6.182}$$

With $A = GM\mu^2 e$,

$$\boxed{\dot{\omega} = -\frac{\chi L}{GM\mu^2 e\, r^3}\cos v}. \tag{6.183}$$

6.16.2 Semi-major axis

We now average the three rates over one orbit, starting with a. As a check on Eq. (6.164), the Gauss equation (6.146) gives

$$\dot{a} = \sqrt{\frac{4a^3}{GM\mu^2(1-e^2)}}\,e\sin v\,\frac{\chi}{r^3} \overset{(1)}{=} \frac{2a^2}{\mu\sqrt{GM\lambda}}\,\frac{\mu\lambda\dot{r}}{L}\,\frac{\chi}{r^3}, \tag{6.184}$$

(1) $\sqrt{4a^3/(GM\mu^2(1-e^2))} = 2a^2/(\mu\sqrt{GM\lambda})$, and $e\sin v = \mu\lambda\dot{r}/L$ from Eq. (6.117).

and with $L = \mu\sqrt{GM\lambda}$,

$$\boxed{\dot{a} = \frac{2\chi a^2}{GM\mu r^3}\dot{r}.} \tag{6.185}$$

For the Keplerian ellipse we use the eccentric anomaly u,

$$r = a(1 - e\cos u), \tag{6.186}$$

and, from Eqs. (6.57) and (6.50),

$$\dot{r} = \frac{ane\sin u}{1 - e\cos u}, \qquad n = \sqrt{\frac{GM}{a^3}}. \tag{6.187}$$

Therefore

$$\boxed{\dot{a} = \frac{2\chi e n \sin u}{GM\mu(1 - e\cos u)^4}.} \tag{6.188}$$

Over one unperturbed orbit, from Eq. (6.56),

$$dt = \frac{1 - e\cos u}{n}\, du, \tag{6.189}$$

so

$$\begin{aligned}\Delta a &= \int_0^{2\pi} \dot{a}\, dt = \frac{2\chi e}{GM\mu}\int_0^{2\pi} \frac{\sin u}{(1 - e\cos u)^3}\, du \\ &\overset{(1)}{=} \frac{2\chi e}{GM\mu}\left[-\frac{1}{2e\,(1 - e\cos u)^2}\right]_0^{2\pi} = 0.\end{aligned} \tag{6.190}$$

(1) $\dfrac{d}{du}(1 - e\cos u)^{-2} = -2e\sin u\,(1 - e\cos u)^{-3}$.

The same follows without any integral. Eq. (6.185) is $\dot{a} = -(\chi a^2/GM\mu)\, d(r^{-2})/dt$, and r returns to its starting value after one orbit. Thus

$$\boxed{\langle \dot{a} \rangle = 0} \tag{6.191}$$

to first order in the perturbation.

The osculating semi-major axis oscillates during an orbit but has no secular drift.

6.16.3 Eccentricity

Angular momentum is conserved, since the force is radial, so

$$L^2 = GM\mu^2 a(1 - e^2) \tag{6.192}$$

is constant, and therefore

$$0 = (1 - e^2)\dot{a} - 2ae\dot{e}. \tag{6.193}$$

Hence

$$\boxed{\dot{e} = \frac{1 - e^2}{2ae}\dot{a}.} \tag{6.194}$$

To first order the factor $(1 - e^2)/(2ae)$ is constant over the orbit, so since $\langle \dot{a} \rangle = 0$,

$$\boxed{\langle \dot{e} \rangle = 0.} \tag{6.195}$$

6.16.4 Apsidal precession

For the radial force, Eq. (6.148) with $F_\phi = 0$ and $L/(GM\mu^2 e) = \sqrt{a(1 - e^2)/(GM\mu^2)}/e$ reduces to Eq. (6.183),

$$\boxed{\dot{\omega} = -\frac{\chi L \cos v}{GM\mu^2 e\, r^3}.} \tag{6.196}$$

To first order in χ we evaluate the secular change on the unperturbed Keplerian orbit,

$$r = \frac{\lambda}{1 + e\cos v}, \qquad \lambda = a(1 - e^2), \tag{6.197}$$

together with Eq. (6.154),

$$dt = \frac{\mu r^2}{L}\, dv. \tag{6.198}$$

Then

$$\begin{aligned}\Delta\omega &= \int_0^{2\pi} \dot\omega\, \frac{\mu r^2}{L}\, dv \overset{(1)}{=} -\frac{\chi}{GM\mu e}\int_0^{2\pi} \frac{\cos v}{r}\, dv \\ &\overset{(2)}{=} -\frac{\chi}{GM\mu e}\int_0^{2\pi} \frac{1+e\cos v}{\lambda}\cos v\, dv \\ &= -\frac{\chi}{GM\mu e\lambda}\left[\int_0^{2\pi}\cos v\, dv + e\int_0^{2\pi}\cos^2 v\, dv\right].\end{aligned} \tag{6.199}$$

(1) $\dfrac{\chi L}{GM\mu^2 e\, r^3}\cdot\dfrac{\mu r^2}{L} = \dfrac{\chi}{GM\mu e\, r}$.
(2) $1/r = (1+e\cos v)/\lambda$.

Since

$$\int_0^{2\pi}\cos v\, dv = 0, \qquad \int_0^{2\pi}\cos^2 v\, dv = \pi, \tag{6.200}$$

we obtain

$$\boxed{\Delta\omega = -\frac{\pi\chi}{GM\mu\, a(1-e^2)}.} \tag{6.201}$$

Dividing by the period

$$P = 2\pi\sqrt{\frac{a^3}{GM}}, \tag{6.202}$$

the secular rate is

$$\boxed{\langle\dot\omega\rangle = -\frac{\chi}{2\mu\sqrt{GM}\, a^{5/2}(1-e^2)}.} \tag{6.203}$$

An outward push ($\chi > 0$) turns the periapsis backwards, against the motion, and an extra inward pull ($\chi < 0$) turns it forwards. The size a and the shape e do not drift.

A second route, exact. The same result follows from the exact orbit equation (6.27) for the full radial force

$$F_r = -\frac{GM\mu}{r^2} + \frac{\chi}{r^3}, \tag{6.204}$$

which is $-V'(r)$. With $\rho = 1/r$ the right side of Eq. (6.27) is $-(\mu/L^2)F_r/\rho^2 = GM\mu^2/L^2 - \chi\mu\rho/L^2$, so

$$\frac{d^2\rho}{d\phi^2} + \left(1 + \frac{\chi\mu}{L^2}\right)\rho = \frac{GM\mu^2}{L^2}. \tag{6.205}$$

This is the shifted oscillator of Section 6.5 with a stiffer spring. In ϕ the oscillation of ρ has the angular frequency $\sqrt{1+\chi\mu/L^2}$, so one periapsis follows the next after the angle

$$\Delta\phi_{\text{apsis}} = \frac{2\pi}{\sqrt{1+\chi\mu/L^2}}. \tag{6.206}$$

The exact shift of the periapsis per orbit, relative to 2π, is therefore

$$\Delta\omega_{\text{exact}} = 2\pi\left[\frac{1}{\sqrt{1+\chi\mu/L^2}} - 1\right]. \tag{6.207}$$

For a weak perturbation,

$$\frac{\chi\mu}{L^2} \ll 1, \tag{6.208}$$

so

$$\begin{aligned}\Delta\omega_{\text{exact}} &\overset{(1)}{=} -\frac{\pi\chi\mu}{L^2} + O(\chi^2) \\ &\overset{(2)}{=} -\frac{\pi\chi}{GM\mu\, a(1-e^2)} + O(\chi^2),\end{aligned} \tag{6.209}$$

(1) $(1+x)^{-1/2} = 1 - x/2 + O(x^2)$.
(2) $L^2 = GM\mu^2 a(1-e^2)$, Eq. (6.36).

which agrees with the orbit-averaged result (6.201).

6.17 Nearly circular orbits

The Gauss equation (6.148) for $\dot\omega$ carries a factor $1/e$. As $e \to 0$ the periapsis loses its meaning, because a circle has no nearest point. Then $\dot\omega$ blows up, although nothing physical does.

Ask which point of a round table is nearest the wall. For a table almost centred in the room the answer swings right round with the smallest nudge, while the nudge itself is tiny. It is better to give how far the table's centre has moved east and north. Those two numbers are always well defined and change smoothly. For an orbit the two numbers are the components of the eccentricity vector (6.72). The idea is to use $e_x = e\cos\omega$ and $e_y = e\sin\omega$ in place of e and ω. They go smoothly to zero with e, and the factor $1/e$ disappears.

Take $e = 0.001$ and $\omega = 0$, so $(e_x, e_y) = (0.001, 0)$. A kick adding only 0.001 to e_y turns ω by 45°, while e rises only to 0.0014. The ellipse has changed slightly, but the direction of its poorly defined periapsis has changed greatly. The components (e_x, e_y) describe that small physical change smoothly, so we use them to write the nearly circular orbit and its evolution.

The orbit to first order in e. The components of the eccentricity vector along $\hat n$ and $\hat y$ are

$$\boxed{e_x = e\cos\omega, \qquad e_y = e\sin\omega.} \tag{6.210}$$

For small e we define the **mean longitude**

$$\bar\phi \equiv \omega + \ell, \tag{6.211}$$

the polar angle of a body moving round at the mean rate. For $e = 0$ it is the polar angle itself. To first order in e, $v = \ell + 2e\sin\ell + O(e^2)$, derived from Kepler's second law in Eq. (1.22), where it gives the equation of time. Then

$$\phi = \omega + v \overset{(1)}{=} \bar\phi + 2e\sin\big(\bar\phi - \omega\big) + O(e^2) \overset{(2)}{=} \bar\phi + 2e_x\sin\bar\phi - 2e_y\cos\bar\phi + O(e^2). \tag{6.212}$$

(1) $v = \ell + 2e\sin\ell$ and $\ell = \bar\phi - \omega$.
(2) $e\sin\big(\bar\phi - \omega\big) = e\cos\omega\sin\bar\phi - e\sin\omega\cos\bar\phi$.

In the same way $r = a(1 - e\cos u) = a(1 - e\cos\ell) + O(e^2)$ and $e\cos\big(\bar\phi - \omega\big) = e_x\cos\bar\phi + e_y\sin\bar\phi$, so

$$\boxed{r = a\left(1 - e_x\cos\bar\phi - e_y\sin\bar\phi\right) + O(e^2).} \tag{6.213}$$

For fixed elements $\dot{\bar{\phi}} = \dot{\ell} = n$, with

$$n = \sqrt{\frac{GM}{a^3}}, \tag{6.214}$$

so differentiating,

$$\dot{r} = an\left(e_x \sin\bar{\phi} - e_y \cos\bar{\phi}\right) + O(e^2). \tag{6.215}$$

Also,

$$\boxed{\dot{\phi} = n\left(1 + 2e_x \cos\bar{\phi} + 2e_y \sin\bar{\phi}\right) + O(e^2).} \tag{6.216}$$

We name the first-order offset of the polar angle from the mean longitude,

$$\delta\phi \equiv \phi - \bar{\phi} = 2e_x \sin\bar{\phi} - 2e_y \cos\bar{\phi}. \tag{6.217}$$

Then $\dot{r} = an\,\delta\phi/2$, and to first order $\xi = e\sin v/(1 + e\cos v) = e\sin\big(\bar{\phi} - \omega\big) + O(e^2) = \delta\phi/2$. The variables e_x and e_y are better than ω near a circle because they stay finite and smooth as $e \to 0$.

6.17.1 First-order equations for nearly circular motion

We now write the Gauss equations in a, e_x, e_y to first order in e. For a, Eqs. (6.121) and (6.110) give $\dot{a} = (2a^2/GM\mu)\big[\dot{r}F_r + (L/\mu r)F_\phi\big]$. To first order, $\dot{r} = an\,\delta\phi/2$ and $L/(\mu r) = \sqrt{GMa}\,(1 + e_x\cos\bar{\phi} + e_y\sin\bar{\phi})/a = an(1 + e_x\cos\bar{\phi} + e_y\sin\bar{\phi})$. With $2a^3n/(GM\mu) = \sqrt{4a^3/(GM\mu^2)}$,

$$\boxed{\dot{a} = \sqrt{\frac{4a^3}{GM\mu^2}}\left[\tfrac{1}{2}\delta\phi\,F_r + \left(1 + e_x\cos\bar{\phi} + e_y\sin\bar{\phi}\right)F_\phi\right] + O(e^2F).} \tag{6.218}$$

The eccentricity vector is, exactly,

$$\vec{A} = GM\mu^2\left(e_x\hat{n} + e_y\hat{y}\right). \tag{6.219}$$

The frame $(\hat{n}, \hat{y})$ is fixed for a planar force, so $\dot{e}_x = \dot{\vec{A}}\cdot\hat{n}/(GM\mu^2)$ and $\dot{e}_y = \dot{\vec{A}}\cdot\hat{y}/(GM\mu^2)$. With $\hat{r}\cdot\hat{n} = \cos\phi$, $\hat{\phi}\cdot\hat{n} = -\sin\phi$, $\hat{r}\cdot\hat{y} = \sin\phi$ and $\hat{\phi}\cdot\hat{y} = \cos\phi$ from Eq. (6.13), Eq. (6.128) gives

$$\begin{aligned}\dot{e}_x &\overset{(1)}{=} \frac{L}{GM\mu^2}\Big[F_r\sin\phi + F_\phi\big(2\cos\phi + \xi\sin\phi\big)\Big] \\ &\overset{(2)}{=} \sqrt{\frac{a}{GM\mu^2}}\Big[F_r\big(\sin\bar{\phi} + \delta\phi\cos\bar{\phi}\big) + F_\phi\big(2\cos\bar{\phi} - 2\delta\phi\sin\bar{\phi} + \tfrac{1}{2}\delta\phi\sin\bar{\phi}\big)\Big] + O(e^2F),\end{aligned} \tag{6.220}$$

(1) $\dot{\vec{A}}\cdot\hat{n} = L\big[2F_\phi\cos\phi + (F_r + \xi F_\phi)\sin\phi\big]$.

(2) $L/(GM\mu^2) = \sqrt{a(1-e^2)/(GM\mu^2)} = \sqrt{a/(GM\mu^2)} + O(e^2)$, $\sin\phi = \sin\bar{\phi} + \delta\phi\cos\bar{\phi}$, $\cos\phi = \cos\bar{\phi} - \delta\phi\sin\bar{\phi}$ and $\xi = \delta\phi/2$, all to first order.

and in the same way for $\dot{e}_y$, with $\dot{\vec{A}}\cdot\hat{y} = L\big[-F_r\cos\phi + F_\phi(2\sin\phi - \xi\cos\phi)\big]$. Collecting terms,

$$\boxed{\dot{e}_x = \sqrt{\frac{a}{GM\mu^2}}\left[\sin\bar{\phi}\,F_r + 2\cos\bar{\phi}\,F_\phi + \tfrac{1}{2}\delta\phi\left(2\cos\bar{\phi}\,F_r - 3\sin\bar{\phi}\,F_\phi\right)\right] + O(e^2F),} \tag{6.221}$$

$$\boxed{\dot{e}_y = \sqrt{\frac{a}{GM\mu^2}}\left[-\cos\bar{\phi}\,F_r + 2\sin\bar{\phi}\,F_\phi + \tfrac{1}{2}\delta\phi\left(2\sin\bar{\phi}\,F_r + 3\cos\bar{\phi}\,F_\phi\right)\right] + O(e^2F).} \tag{6.222}$$

Neither equation contains $1/e$.

For the unperturbed orbit $\dot{\bar{\phi}} = n$ exactly. The force changes this rate only by $O(F)$,

$$\boxed{\dot{\bar{\phi}} = n + O(F).} \tag{6.223}$$

The rates above are already $O(F)$, so dividing them by $\dot{\bar{\phi}}$ instead of n changes them only at $O(F^2)$.

Using $\bar{\phi}$ as the independent variable, $d/d\bar{\phi} = n^{-1} d/dt$, with $\sqrt{a/(GM\mu^2)}/n = a^2/(GM\mu)$ and $\sqrt{4a^3/(GM\mu^2)}/n = 2a^3/(GM\mu)$,

$$\boxed{\frac{da}{d\bar{\phi}} = \frac{2a^3}{GM\mu}\left[\tfrac{1}{2}\delta\phi\, F_r + \left(1 + e_x\cos\bar{\phi} + e_y\sin\bar{\phi}\right) F_\phi\right],} \tag{6.224}$$

$$\boxed{\frac{de_x}{d\bar{\phi}} = \frac{a^2}{GM\mu}\left[\sin\bar{\phi}\, F_r + 2\cos\bar{\phi}\, F_\phi + \tfrac{1}{2}\delta\phi\left(2\cos\bar{\phi}\, F_r - 3\sin\bar{\phi}\, F_\phi\right)\right],} \tag{6.225}$$

$$\boxed{\frac{de_y}{d\bar{\phi}} = \frac{a^2}{GM\mu}\left[-\cos\bar{\phi}\, F_r + 2\sin\bar{\phi}\, F_\phi + \tfrac{1}{2}\delta\phi\left(2\sin\bar{\phi}\, F_r + 3\cos\bar{\phi}\, F_\phi\right)\right].} \tag{6.226}$$

6.18 Rømer delay for a Keplerian binary

We have the position of each star in its orbit, and its distance along the line of sight, the Z of Eq. (6.14). For a pulsar in a binary we cannot see the position at all. What we measure is the arrival time of each pulse.

A drummer marching to and fro along the line between us and a distant field beats a perfectly steady rhythm. We hear the beats bunch up as he comes towards us and spread out as he goes away, because the sound from the far end has farther to travel. A pulsar in a binary is the drummer, and its pulses are the beats. The idea is that each pulse arrives late by the extra light-travel distance $\vec{x}_1 \cdot \hat{R}/c$, so the timing measures the line-of-sight coordinate of the orbit.

Light takes 499 s to cross 1 AU. As the Earth moves around the Sun, pulses from a source near the orbital plane can arrive up to 499 s earlier or later than they would at the Solar System's centre. Rømer used a related delay in eclipses of Jupiter's moon Io to estimate the speed of light in 1676. A pulsar moving in its own binary orbit produces the same kind of delay. Its changing arrival time therefore measures its position along our line of sight, which we will express in terms of the orbital elements.

The delay. For a binary of component masses m_1 and m_2, $M = m_1 + m_2$. Let $\vec{r}$ be the relative separation. The primary's position relative to the centre of mass is, by Section 6.1.2,

$$\vec{x}_1 = \frac{m_2}{M}\vec{r}. \tag{6.227}$$

$\hat{R}$ points from the observer towards the source (Section 6.2). A primary that is farther away by $\vec{x}_1 \cdot \hat{R}$ than the centre of mass sends pulses that arrive later by

$$\Delta_R = \frac{\vec{x}_1 \cdot \hat{R}}{c}. \tag{6.228}$$

In the orbital-plane basis,

$$\vec{r} = r\left[\cos\phi\,\hat{n} + \sin\phi\,\hat{y}\right], \tag{6.229}$$

with

$$\phi = \omega + v, \tag{6.230}$$

and, from the third components of Eqs. (6.10) and (6.11),

$$\hat{y} \cdot \hat{R} = \sin \iota, \qquad \hat{n} \cdot \hat{R} = 0, \tag{6.231}$$

so that

$$\Delta_R = \frac{m_2}{Mc}\, r \sin \iota \, \sin(\omega + v). \tag{6.232}$$

This is $(m_2/M)\, Z/c$ with Z from Eq. (6.14).

We define the primary's semi-major axis about the centre of mass,

$$a_1 = \frac{m_2}{M} a, \tag{6.233}$$

and use the eccentric-anomaly relations (6.47),

$$r \cos v = a(\cos u - e), \tag{6.234}$$

$$r \sin v = a\sqrt{1 - e^2} \sin u. \tag{6.235}$$

Then

$$\begin{aligned} \Delta_R &= \frac{m_2}{Mc} \sin \iota \left(r \cos v \, \sin \omega + r \sin v \, \cos \omega \right) \\ &\overset{(1)}{=} \frac{a_1 \sin \iota}{c} \left[(\cos u - e) \sin \omega + \sqrt{1 - e^2} \sin u \cos \omega \right], \end{aligned} \tag{6.236}$$

(1) $\sin(\omega + v) = \sin \omega \cos v + \cos \omega \sin v$ was used in the first line, then Eq. (6.47) and $a_1 = (m_2/M)a$.

so

$$\boxed{\Delta_R = \frac{a_1 \sin \iota}{c} \left[(\cos u - e) \sin \omega + \sqrt{1 - e^2} \sin u \cos \omega \right].} \tag{6.237}$$

Only the projected size $a_1 \sin \iota$ enters, the same combination that a radial-velocity curve gives (Problem 9). Timing alone does not give ι.

6.19 Radial Velocities and Planets

Problems 6.21.9 and 6.21.11 found the velocity along the line of sight of a star in a binary. A star with a planet is a binary whose second member is too faint to see. We see only the star, and only the part of its motion along our line of sight. The planet's own light is lost in the glare, a billion times fainter in visible light for a Jupiter beside a Sun, Eq. (5.61).

Think of a hammer thrower spinning with the hammer at arm's length. The thrower does not turn about their own middle. They lean back and circle a point between themselves and the hammer. Watching only the thrower from far away, you can still tell how heavy the hammer is, from how far and how fast the thrower swings. The star circles the centre of mass on a small copy of the relative orbit, Eq. (6.5). As it swings toward and away from us its spectral lines shift back and forth (Chapter 3). The shifts give its velocity along the line of sight against time. The shape of that curve fixes the period, the eccentricity, the orientation of the ellipse in its plane and the time of periapsis. Its height, with the star's own mass, fixes the planet's mass times $\sin \iota$.

Jupiter makes the Sun's radial velocity swing by $12.5\,\mathrm{m\,s^{-1}}$; Earth produces only $9\,\mathrm{cm\,s^{-1}}$. The Jupiter-sized speed changes each wavelength by the tiny fraction 4.2×10^{-8}. Repeated

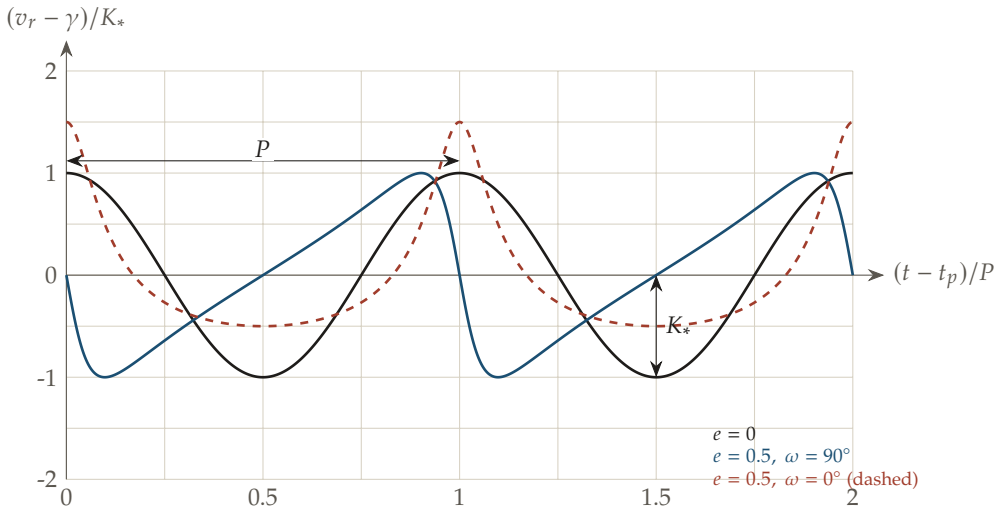


Figure 6.6: The star's velocity along the line of sight, $(v_r - \gamma)/K_*$, against time in units of the period P, for a circular orbit and for $e = 0.5$ with $\omega = 0°$ and $\omega = 90°$. Time is counted from periapsis. The arrows mark the semi-amplitude K_* and the period P.

measurements reveal the periodic shift. Figure 6.6 shows how a circular orbit gives a smooth wave while eccentricity changes its shape.

The same method weighed the first planet found around a Sun-like star, 51 Pegasi b, in 1995, and the stars of every spectroscopic binary. It gives what the spectrum measures, what the curve fixes, and what stays hidden.

From the spectrum to a velocity. A star's spectrum holds thousands of absorption lines. Each is shifted by the same fraction,

$$\frac{\lambda_{\text{obs}}}{\lambda_{\text{lab}}} = 1 + \frac{v_r}{c}$$

(Eq. (3.25)), because all of them come from the same moving star. So the spectrograph does not measure one line. It slides a template of the whole spectrum along in wavelength and finds the one fraction that lines up all the lines at once. A shift of $1\,\mathrm{m\,s^{-1}}$ moves a line by 3×10^{-9} of its wavelength, about a thousandth of a pixel of the detector. Only the average over thousands of lines, against a wavelength scale held fixed by a lamp or by a cell of iodine gas in the beam, can find it. The Earth's own motion, up to $30\,\mathrm{km\,s^{-1}}$, is removed as in Section 3.4. Repeated over many nights, this gives the star's $v_r(t)$ relative to the centre of mass of the Solar System.

The model. Let the star be body 1 and the planet body 2, with masses M_* and m_p and $M = M_* + m_p$, and $\vec{r}$ the vector from the planet to the star. By Eq. (6.5) the star sits at $\vec{X} + (m_p/M)\vec{r}$. Its velocity along the line of sight $\hat{R}$ is

$$\begin{aligned} v_r &\overset{(1)}{=} \dot{\vec{X}} \cdot \hat{R} + \frac{m_p}{M}\, \dot{\vec{r}} \cdot \hat{R} \\ &\overset{(2)}{=} \gamma + \frac{m_p}{M}\, \sin\iota \sqrt{\frac{GM}{a(1-e^2)}} \Big[\cos(\omega + v) + e\cos\omega \Big]. \end{aligned} \tag{6.238}$$

(1) Differentiate the star's position and project onto $\hat{R}$.

(2) $\gamma = \dot{\vec{X}} \cdot \hat{R}$ is constant (Eq. (6.4)), and $\dot{\vec{r}} \cdot \hat{R}$ is Problem 6.21.11(c), with a, e, ι, ω the elements of the relative orbit.

$$\boxed{v_r(t) = \gamma + K_*\Big[\cos\big(\omega + v(t)\big) + e\cos\omega\Big], \qquad K_* = \frac{m_p}{M}\sin\iota\sqrt{\frac{GM}{a(1-e^2)}}.} \tag{6.239}$$

The star moves on the same ellipse as the relative orbit, shrunk by m_p/M and pointing the same way, so it shares ω, e and P. The true anomaly $v(t)$ follows from the time through the mean anomaly $\ell = 2\pi(t - t_p)/P$ and Kepler's equation (Section 6.9).

What each part of the curve fixes. Eq. (6.239) has six numbers, and each leaves its own mark on the curve.

- The **period** P is the time after which the curve repeats.
- The **eccentricity** e is how far the curve is from a sine. On an ellipse the star rushes through periapsis and crawls through apoapsis (Kepler's second law), so the curve swings sharply in a short time and then drifts back slowly. For $e = 0$ the bracket is $\cos(\omega + v)$ with v growing evenly, a pure sine.
- The **argument of periapsis** ω is which way the lopsided curve leans. It says where the periapsis lies relative to the line of sight, and Figure 6.6 shows the same ellipse seen with $\omega = 0°$ and $\omega = 90°$.
- The **time of periapsis** t_p is when the sharp swing happens.
- The **semi-amplitude** K_* is half the distance from the highest to the lowest point. The bracket runs from $-1 + e\cos\omega$ to $1 + e\cos\omega$, a range of exactly 2.
- The **systemic velocity** γ is the average of v_r over a whole period, since the star returns to where it started and its average $\dot{\vec{r}}\cdot\hat{R}$ is zero. It is not the midpoint of the range, which is $\gamma + K_* e\cos\omega$.

Fitting Eq. (6.239) to the measured points gives all six. Of the six elements of the relative orbit, $(a, e, \iota, \Omega, \omega, t_p)$, the curve fixes e, ω and t_p directly and P in place of a. The rest is tied up with the masses in K_*.

From K_* to the planet's mass. Kepler's third law, Eq. (6.50) with $n = 2\pi/P$, gives $a = (GMP^2/4\pi^2)^{1/3}$. Putting it into K_*,

$$\begin{aligned} K_* &\overset{(1)}{=} \frac{m_p\sin\iota}{M}\,\frac{(GM)^{1/2}}{\sqrt{1-e^2}}\left(\frac{GMP^2}{4\pi^2}\right)^{-1/6} \\ &\overset{(2)}{=} \left(\frac{2\pi G}{P}\right)^{1/3}\frac{m_p\sin\iota}{M^{2/3}\sqrt{1-e^2}}. \end{aligned} \tag{6.240}$$

(1) $a^{-1/2}$ from Kepler's third law.
(2) Collect the powers, $(GM)^{1/2-1/6} = (GM)^{1/3}$ and $(P^2/4\pi^2)^{-1/6} = (2\pi/P)^{1/3}$.

Cubing and rearranging puts everything measured on one side,

$$\boxed{\frac{(m_p\sin\iota)^3}{(M_* + m_p)^2} = \frac{P\,K_*^3\,(1-e^2)^{3/2}}{2\pi G}.} \tag{6.241}$$

The right side is the **mass function**, measured from the curve alone. The left side still holds two unknown masses. This is the lesson of Section 6.1.2 again. The curve of one body gives its share m_p/M of the relative motion times $\sin\iota$, but not how the total mass is split. The extra fact comes from the star itself. Its spectrum gives its temperature and surface gravity, its parallax

and brightness give its luminosity, and with models of stars (Chapter 12) these give M_*. A planet is far lighter than its star, $m_p \ll M_*$, so $M \approx M_*$ and

$$\boxed{m_p \sin\iota \approx \left(\frac{P}{2\pi G}\right)^{1/3} M_*^{2/3}\, K_* \sqrt{1-e^2}, \qquad a \approx \left(\frac{GM_*P^2}{4\pi^2}\right)^{1/3}.} \tag{6.242}$$

A heavier planet swings its star harder, and a closer one, with a shorter period, swings it faster, $K_* \propto m_p P^{-1/3}$. So the method finds heavy planets close to their stars most easily.

What stays hidden. The curve cannot see Ω at all, because $\hat{R}$ does not depend on it (Problem 6.21.11(a)). Turning the orbit about our line of sight changes nothing along it. Nor can it separate m_p from $\sin\iota$. A light planet on an orbit seen edge-on and a heavy one on an orbit seen nearly face-on give the same curve. So the method gives a **minimum mass**, $m_p \sin\iota \le m_p$. The minimum is usually close to the truth. For orbits oriented at random, $\cos\iota$ is spread evenly between -1 and 1, so the true mass exceeds twice the minimum, $\sin\iota < \frac{1}{2}$, only when $|\cos\iota| > \frac{\sqrt{3}}{2}$, in $1 - \frac{\sqrt{3}}{2} = 13\%$ of systems. The hidden ι is found when the planet also passes in front of its star. Then $\iota \approx 90°$, the mass is exact, and the fraction of starlight blocked, $(R_p/R_*)^2$, gives the planet's radius and so its density. Astrometry of the star's tiny wobble on the sky (Problem 6.21.1) gives ι and Ω directly.

51 Pegasi b. Its star, of $M_* = 1.11\,M_\odot$, swings with $P = 4.23\,\mathrm{d}$, $K_* = 55.9\,\mathrm{m\,s^{-1}}$ and $e \approx 0$. Eq. (6.241) gives $m_p \sin\iota = 0.48$ Jupiter masses at $a = 0.053\,\mathrm{AU}$, a planet like Jupiter seven times closer to its star than Mercury is to the Sun. No such planet was expected. Finding the Earth's twin needs $9\,\mathrm{cm\,s^{-1}}$, below the jitter of the star's own surface, about $1\,\mathrm{m\,s^{-1}}$, which is why Earth-like planets are found mostly by their transits instead.

6.20 Other Ways to Find Planets

The radial velocity of Section 6.19 weighs a planet but leaves its size, its orbit's tilt and its nature unknown. A planet sends almost no light of its own, so every method must catch its effect on something bright, its star or a star behind it. Each effect is tiny, and each method is blind to some planets the others see.

Think of guessing what is in a dark room from outside. You can listen for footsteps, watch for a shadow passing the window, or wait for a lamp behind it to flicker. Each tells you something different, and together they tell you a lot. The same orbit leaves several marks. The star wobbles, which shifts its lines (radial velocity), moves it on the sky (astrometry) and changes the arrival time of any regular signal it sends (timing). If the orbit is seen nearly edge-on, the planet blocks part of the star once per orbit (transit). And its gravity can bend the light of a more distant star (microlensing).

The signals are small but tell us different things. Jupiter crossing the Sun would dim it by 1%, while Earth would dim it by only 0.008%. An Earth-mass planet around a $1.4\,M_\odot$ pulsar would shift its pulse arrival times by about a millisecond, detectable when pulses can be timed to a microsecond. A transit gives a planet's size; a timing shift traces its star's orbit. We will see how these methods complement the radial-velocity measurement.

Words. A **planet** orbits a star, is massive enough for its gravity to pull it nearly round and to clear its orbit of smaller bodies, and is not massive enough to burn hydrogen or deuterium by fusion (Chapter 12). An **exoplanet** orbits a star other than the Sun. A **gas giant**, like Jupiter and Saturn, is mostly hydrogen and helium. A **rocky** or terrestrial planet, like the Earth and

Mars, is mostly silicate rock and metal, and in the Solar System these lie close to the Sun. A **super-Earth** is a rocky planet heavier than the Earth but lighter than the gas giants. The Solar System has no example, and they turn out to be common.

The first planets. In 1992 Aleksander Wolszczan and Dale Frail found planets around the pulsar PSR B1257+12, 2300 light years away, from the timing of its pulses. In 1995 Michel Mayor and Didier Queloz found 51 Pegasi b, the first planet around a star like the Sun, 50 light years away, from its radial velocity. The Kepler space telescope, launched in 2009 on an orbit trailing the Earth around the Sun, watched the brightness of 145 000 stars in Cygnus, Lyra and Draco without a break for four years with a 0.95 m telescope and 42 CCDs (Chapter 7), and found thousands of planets by their transits.

Transits. If the orbit is seen nearly edge-on, the planet passes in front of its star once per orbit. While it is in front it blocks a disk of radius R_p out of the star's disk of radius R_*. For a star of uniform brightness the fraction of light lost is the ratio of the two areas,

$$\boxed{\frac{\Delta F}{F} = \left(\frac{R_p}{R_*}\right)^2 .} \tag{6.243}$$

Jupiter in front of the Sun gives $(0.10)^2 = 1\%$, and the Earth gives $(0.0092)^2 = 8.4 \times 10^{-5}$, which needs the steady photometry of space (Chapter 7). The transit lasts while the planet crosses the star. For a central crossing of a circular orbit, with $R_p \ll R_* \ll a$, the planet moves a chord $2R_*$ at the orbital speed $2\pi a/P$,

$$t_{\text{tr}} = \frac{2R_*}{2\pi a/P} = \frac{P}{\pi}\,\frac{R_*}{a},$$

13 hours for the Earth. A transit happens only if the orbit is tilted from edge-on by less than about R_*/a radians. For orbits oriented at random, $\cos\iota$ is spread evenly, so the chance is about R_*/a, 0.5% for an Earth at 1 AU and 9% for a hot Jupiter at 0.05 AU. A transiting planet gives what the radial velocity cannot. Its orbit is seen edge-on, $\sin\iota \approx 1$, so Eq. (6.242) gives its true mass, the depth gives its radius, and together they give its mean density, $3m_p/(4\pi R_p^3)$, which says whether it is rock, ice or gas.

Timing. A pulsar sends a pulse at fixed intervals, like the ticks of a clock. A pulsar with nothing around it moves at a steady speed or not at all, so its pulses reach us at one fixed interval, one after another, and the signal stays periodic. Now give it a planet. The pulsar starts to orbit the centre of mass of the pair, and its distance from us changes with the orbit. The pulses then arrive a little late for half an orbit and a little early for the other half, and the pattern repeats once per orbit. A pulse train that runs late and early in turn, with a fixed period, is the sign of an unseen companion. A star that sends a regular signal is a clock carried around the centre of mass. On the far side of its orbit its signals have farther to travel and arrive late, on the near side early. This is the Rømer delay of Section 6.18, with amplitude $a_* \sin\iota/c$, where $a_* = (m_p/M)\,a$ is the size of the star's own orbit (Eq. (6.5)). For an Earth-mass planet at 1 AU around a pulsar of $1.4\,M_\odot$,

$$\frac{a_*}{c} = \frac{m_p}{M}\,\frac{a}{c} = \frac{5.97 \times 10^{24}}{2.78 \times 10^{30}} \times \frac{1.496 \times 10^{11}\ \text{m}}{3.00 \times 10^{8}\ \text{m s}^{-1}} = 1.1\ \text{ms}.$$

Pulsars keep time to about a microsecond, so even a planet of the Moon's mass shows. The planets of PSR B1257+12 were found this way. The radial velocity measures the star's velocity along the line of sight, and timing measures its position along it, the same orbit seen through

its derivative or itself. Any steady clock works. The planet V391 Pegasi b was found from the timing of the regular pulsations of its star, a hot subdwarf.

Figure 6.7 shows what such data look like. What is plotted is the arrival time of a pulse, or of a pulsation maximum, minus the time a lone, steady clock would predict. A star with no companion gives a flat line at zero, with only noise around it. A star with a planet gives a wave. The period of the wave is the orbital period P. Its height is the light-travel time across the star's own orbit, $a_* \sin \iota / c$.

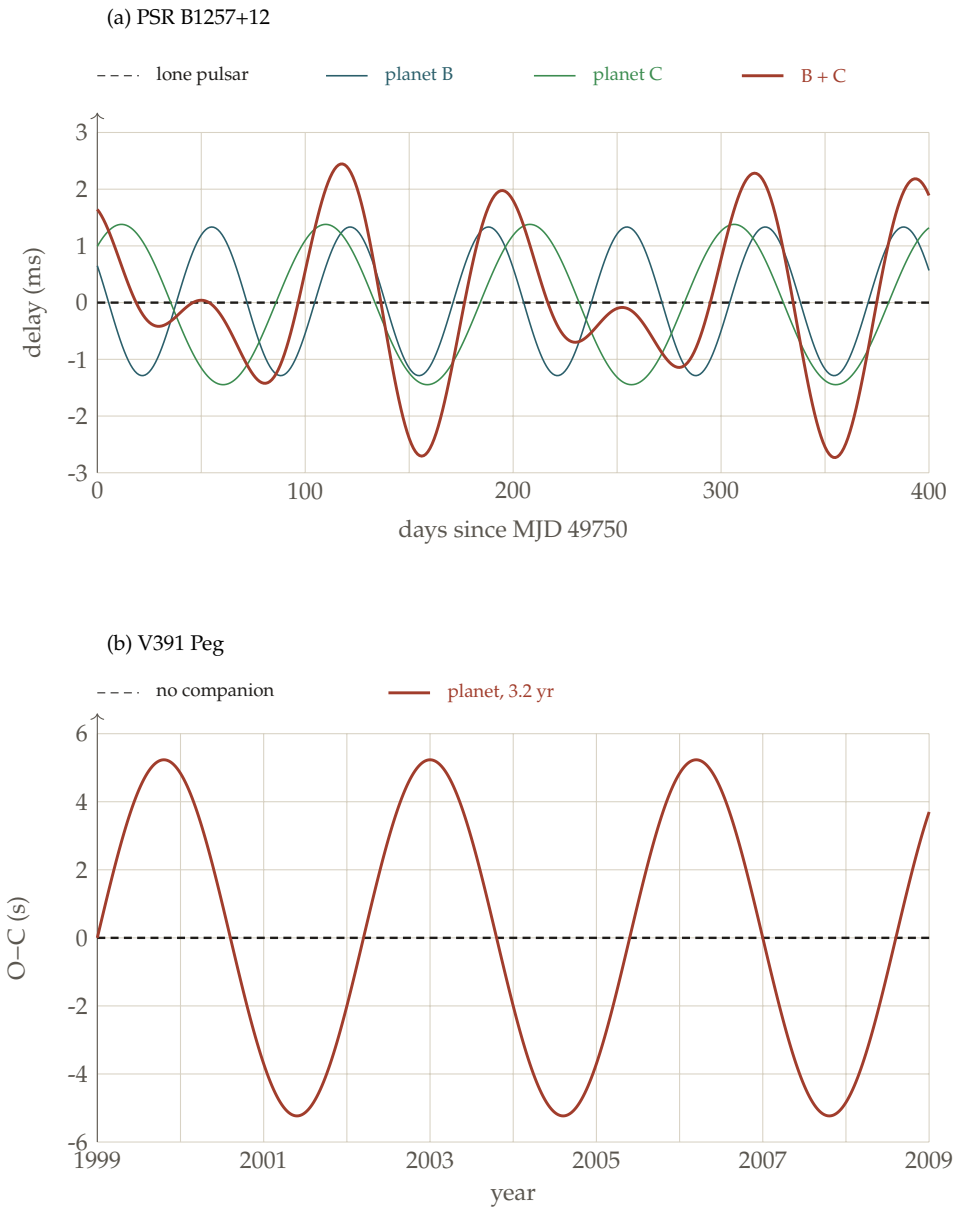


Figure 6.7: Timing signals of planets, computed from published orbits and not from raw timing data. (a) The delay of the pulses of PSR B1257+12 from its planets B and C (orbital elements of Konacki and Wolszczan 2003, pulsar mass 1.4 $M_\odot$), and the two planets alone. Planet A is too small to show. (b) The O–C of the pulsation maxima of V391 Peg for a planet of period 3.2 yr, with the slow change of the pulsation period taken out and an arbitrary phase. The dashed line in each panel is a lone pulsar or a star without a companion.

The two readings give the planet in two steps. The period gives the planet's distance by Kepler's third law, $a^3 = GMP^2/4\pi^2$. The height gives $a_* \sin \iota$ as c times the height, and Eq. (6.5) turns

it into $m_p \sin\iota = M\, a_* \sin\iota / a$ when $m_p \ll M$. Take planet B of PSR B1257+12 with a height of 1.311 ms and a period of 66.54 d. Then $a_* \sin\iota = 393$ km, the law gives $a = 0.36$ AU for a pulsar of $1.4\, M_\odot$, and $m_p \sin\iota = 3.4\, M_\oplus$. Planet C, with 1.413 ms and 98.21 d, gives $a = 0.47$ AU and $m_p \sin\iota = 2.8\, M_\oplus$. Planet A, at 0.003 ms, is about 400 times smaller than the other two and cannot be seen in the figure. The periods of B and C are close to the ratio 3:2 (1.48), so their two waves drift in and out of step. The sum is lopsided and nearly repeats every 196 d, with swings up to 2.7 ms.

In V391 Peg the pulsation period is about 349.5 s, and the wave in the pulsation maxima has a period of 3.2 yr and a height of about 5.2 s. Take a star of $0.5\, M_\odot$, which is assumed and not measured. The period gives $a = 1.7$ AU and the height gives $a_* = 2.3\, R_\odot$, so $m_p \sin\iota \approx 3.2$ Jupiter masses. The height is only 1.5% of the pulsation period, so the signal is small and the data must cover many pulsation cycles. Later and longer data (Silvotti et al. 2018) fit the planet less well, and the planet needs another method to confirm it.

Microlensing, direct imaging and atmospheres. A planet around a star that passes in front of a more distant star bends that star's light and adds a short spike to the brightening caused by its own star (Chapter 14, Eq. (14.41)). The planet OGLE-2005-BLG-390Lb, about five Earth masses, was found this way. A young, hot giant far from its star, such as β Pictoris b, can be photographed directly, if the glare of the star can be blocked (Section 5.6.4). And when a planet transits, a little of the starlight passes through its atmosphere on the way to us, so the depth of the transit changes with wavelength where the atmosphere's molecules absorb, Eq. (6.243) with R_p larger at those wavelengths (Section 11.7.1). No sign of life has been found in any of them so far.

Where water can be liquid. A planet at distance a from a star of luminosity L_* catches the flux $L_*/(4\pi a^2)$ on its cross-section πR_p^2 and reflects a fraction A_B, its albedo. It glows from its whole surface $4\pi R_p^2$ like a blackbody at temperature T_{eq} (Eq. (4.24)). In a steady state the two balance (P1, P4),

$$\begin{aligned} (1 - A_B)\, \frac{L_*}{4\pi a^2}\, \pi R_p^2 &\overset{(1)}{=} 4\pi R_p^2\, \sigma_{SB} T_{eq}^4 \\ \Longrightarrow \qquad T_{eq} &\overset{(2)}{=} \left(\frac{(1 - A_B)\, L_*}{16\pi \sigma_{SB}\, a^2} \right)^{1/4} . \end{aligned} \tag{6.244}$$

(1) Sunlight absorbed on the cross-section equals heat radiated from the whole surface. A planet that turns fast spreads the heat over its surface.
(2) Solve for T_{eq}. The planet's radius has dropped out.

$$\boxed{T_{eq} \propto \frac{L_*^{1/4}}{a^{1/2}}, \qquad a_{HZ} \propto \sqrt{L_*}\,.} \tag{6.245}$$

For the Earth, $A_B = 0.30$ gives $T_{eq} = 255$ K. The real surface is 288 K because the atmosphere traps part of the outgoing heat. The **habitable zone** is the range of distances where a planet with an atmosphere can keep liquid water on its surface. A fixed T_{eq} needs $a \propto \sqrt{L_*}$. Around a red dwarf of $0.01\, L_\odot$ the zone lies ten times closer than around the Sun, near 0.1 AU, where transits are likely and the star's swing is large, so such planets are the easiest Earth-sized ones to find.

6.21 Applied Orbital Mechanics: Binary Stars, Transfer Orbits, and Perturbations

6.21.1 Problem 1: Astrometric Binary White Dwarf System

Problem Statement: Two white dwarfs in a binary system appear on the sky plane as concentric circles of radii 10 μas and 5 μas, respectively, moving anticlockwise. The observed orbital period is 1 day. The system has a parallax of 1 mas.

(a) What is the semi-major axis of the relative orbit?

(b) The stars appear to move with constant angular speed along circular trajectories. What are the orbital eccentricity e and inclination ι?

(c) Find the individual masses m_1 and m_2 and the total mass M.

(d) If the orbit were inclined, what would be the orientation of the orbit in space? Specify the angles Ω and ω.

(e) What is the magnitude of the Runge–Lenz vector for this system?

Essential Ideas:

- **Center of mass**: For a binary system, both stars orbit their mutual center of mass with identical period P. Their individual barycentric semi-major axes satisfy

$$a = a_1 + a_2, \qquad m_1 a_1 = m_2 a_2,$$

  where a is the semi-major axis of the relative orbit. Both paths are copies of the relative orbit, and the heavier body's copy is smaller by the mass ratio, as Eq. (6.5) shows.

- **Projected orbit**: A circular orbit tilted with respect to the sky plane projects to an ellipse with axis ratio $\cos\iota$ (Section 6.2). Therefore an observed circle requires $\iota = 0°$. Constant orbital speed then implies a circular Kepler orbit, so $e = 0$.

- **Runge–Lenz vector**: The Runge–Lenz vector points towards the periapsis with magnitude

$$|\vec{A}| = GM\mu^2 e.$$

  For a circular orbit ($e = 0$), $\vec{A} = \vec{0}$.

Detailed Solutions:

(a) **Semi-major axis of the relative orbit.**

The distance d to the binary follows from the parallax of 1 mas $= 10^{-3}{}''$ (Chapter 3, Eq. (3.5)),

$$d = \frac{1\,\mathrm{pc}}{10^{-3}} = 10^3\,\mathrm{pc}.$$

An angle of 1″ at 1 pc is 1 AU, so the physical barycentric orbital radii are

$$a_1 = \theta_1 d = (10\,\mu\mathrm{as})(10^3\,\mathrm{pc}) = (10^{-5}{}'')(10^3\,\mathrm{pc}) = 10^{-2}\,\mathrm{AU} \simeq 1.496 \times 10^9\,\mathrm{m},$$
$$a_2 = \theta_2 d = (5\,\mu\mathrm{as})(10^3\,\mathrm{pc}) = (5 \times 10^{-6}{}'')(10^3\,\mathrm{pc}) = 0.5 \times 10^{-2}\,\mathrm{AU} \simeq 7.48 \times 10^8\,\mathrm{m}.$$

Therefore the relative separation has the semi-major axis

$$\boxed{a = a_1 + a_2 = 1.5 \times 10^{-2}\,\mathrm{AU} = 2.244 \times 10^9\,\mathrm{m}}.$$

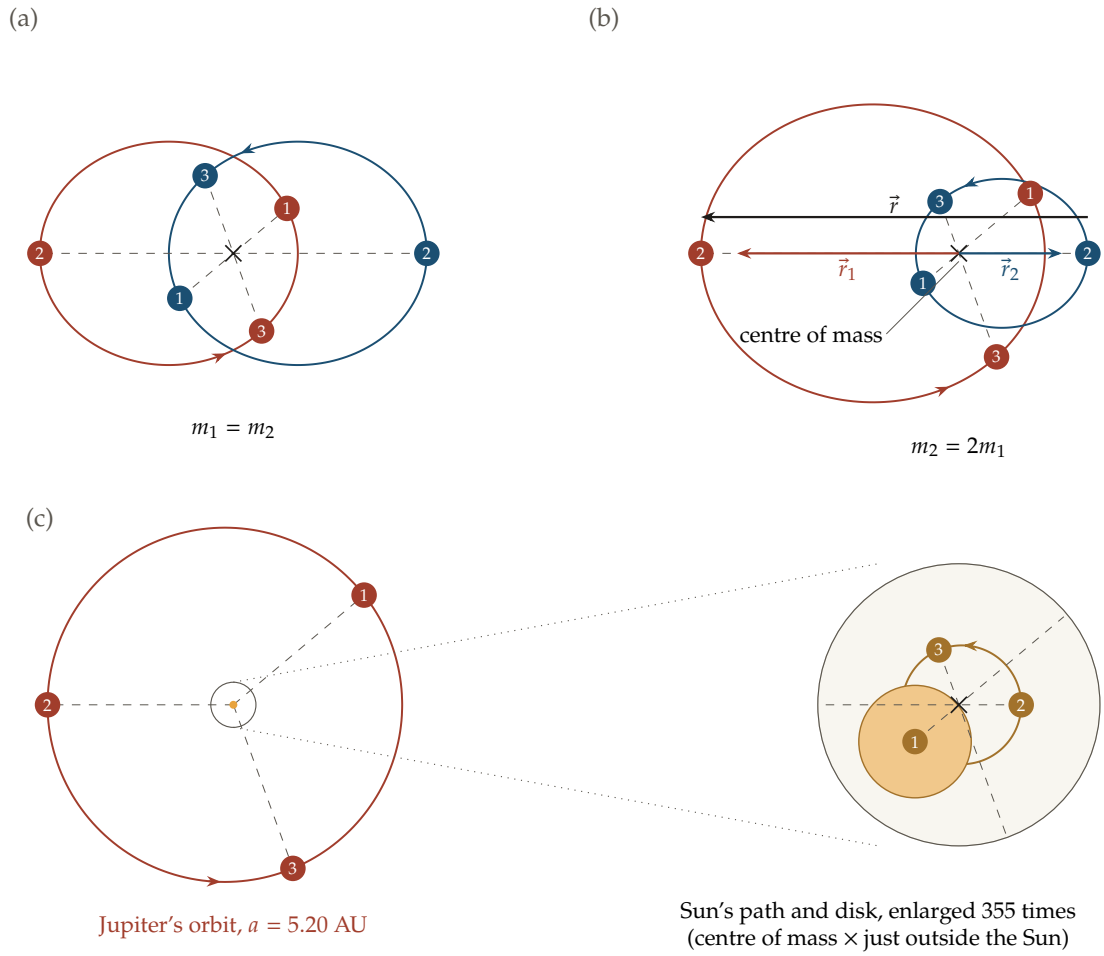


Figure 6.8: The paths of two bodies about their common centre of mass (cross), at the focus of every path, for relative orbit eccentricity $e = 0.5$. Body 1 (red) follows $(m_2/M)\vec{r}$ and body 2 (blue) follows $-(m_1/M)\vec{r}$. Equal numbers mark the same moment, and the dashed line joining them passes through the centre of mass. (a) Equal masses. (b) $m_2 = 2m_1$, with $\vec{r}$, $\vec{r}_1$ and $\vec{r}_2$ at moment 2. (c) The Sun and Jupiter, with $m_{\text{Jup}}/M_\odot \approx 1/1048$, $a = 5.20$ AU and $e = 0.048$. The inset enlarges the Sun's path 355 times, with the Sun's disk to scale.

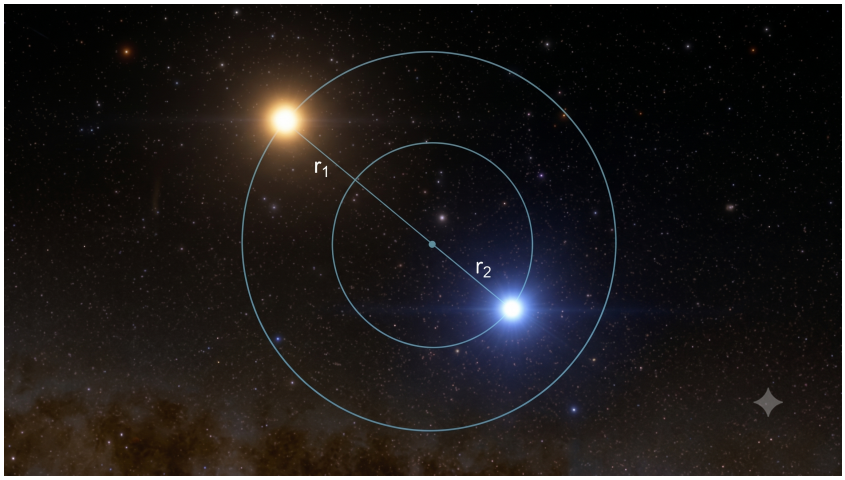


Figure 6.9: Two stars on circular orbits about their common centre of mass, with the barycentric radii r_1 and r_2 marked.

(b) **Eccentricity and inclination.**

Since the observed trajectories are circles, the projection factor satisfies $\cos\iota = 1$, so $\boxed{\iota = 0^\circ}$. Since the orbital speed is constant, the orbit is circular, $\boxed{e = 0}$.

(c) **Masses.**

The mass ratio is the inverse of the ratio of the barycentric semi-major axes,

$$\frac{m_1}{m_2} = \frac{a_2}{a_1} = \frac{5}{10} = \frac{1}{2} \implies m_2 = 2m_1.$$

From Kepler's third law, Eq. (6.50),

$$P^2 = \frac{4\pi^2 a^3}{G(m_1 + m_2)} = \frac{4\pi^2 a^3}{3Gm_1}.$$

With $P = 1\,\text{day} = 86400\,\text{s}$ and $a = 2.244 \times 10^9\,\text{m}$,

$$M = m_1 + m_2 = \frac{4\pi^2 a^3}{GP^2} \simeq 8.95 \times 10^{29}\,\text{kg} \simeq 0.450\,M_\odot.$$

Thus

$$\boxed{m_1 \simeq 0.150\,M_\odot}, \qquad \boxed{m_2 \simeq 0.300\,M_\odot}.$$

The same follows in solar units, $M/M_\odot = (a/\text{AU})^3/(P/\text{yr})^2 = (1.5{\times}10^{-2})^3{\times}365.25^2 = 0.450$. Both masses are well below the Chandrasekhar mass and are typical of the low-mass white dwarfs found in close double-degenerate binaries. Their separation, $3.2\,R_\odot$, is small enough for a one-day orbit and far larger than either white dwarf. The orbit spans $10\,\mu$as on the sky, near the limit of what Gaia can measure.

(d) **Orientation angles Ω and ω.**

For a circular face-on orbit ($\iota = 0^\circ$, $e = 0$), the line of nodes is undefined, since the orbital and sky planes do not intersect, and the periapsis is undefined, since the distance is constant. Hence neither Ω nor ω is defined. If the same circular orbit were inclined, the node line would exist and Ω would be defined, but ω would still be undefined because a circle has no periapsis.

(e) **Runge–Lenz vector.**

Because the orbit is circular ($e = 0$), $\vec{A} = \vec{0}$. A circle has no distinguished periapsis direction, which is why the Runge–Lenz vector vanishes.

6.21.2 Problem 2: Bounds on the Symmetric Mass Ratio

Problem Statement: Prove that the symmetric mass ratio $\eta = \frac{m_1 m_2}{M^2}$ (where $M = m_1 + m_2$) satisfies $\eta \in (0, 0.25]$.

Solution: Since $m_1, m_2 > 0$, $\eta > 0$. For the upper bound,

$$\begin{aligned}(m_1 - m_2)^2 \geq 0 &\implies m_1^2 - 2m_1 m_2 + m_2^2 \geq 0 \\ &\implies (m_1 + m_2)^2 \geq 4m_1 m_2 \\ &\implies \eta = \frac{m_1 m_2}{(m_1 + m_2)^2} \leq \frac{1}{4}.\end{aligned}$$

Equality holds if and only if $m_1 = m_2$. Hence $\boxed{\eta \in (0, 0.25]}$. Since $\eta = \mu/M$, the reduced mass is at most a quarter of the total mass.

6.21.3 Problem 3: Numerical Solution of Kepler's Equation

Problem Statement: Solve Kepler's equation $\ell = u - e\sin u$ using Newton's method and plot $u(\ell)$ for $e = 0, 0.1, 0.3, 0.5, 0.7, 0.9$.

Newton's Iteration Scheme: For fixed mean anomaly ℓ and eccentricity e, let $g(u) = u - e\sin u - \ell$. Then $g'(u) = 1 - e\cos u$, and Newton's method gives the recurrence

$$u_{k+1} = u_k - \frac{u_k - e\sin u_k - \ell}{1 - e\cos u_k}. \tag{6.246}$$

The derivative $1 - e\cos u \geq 1 - e > 0$ never vanishes for $e < 1$, so each step is defined. Starting from $u_0 = \ell$, the error is already $e\sin\ell$, and Newton's method then roughly squares it at each step. The listing below implements it (the code calls the mean anomaly `l`).

```
import matplotlib.pyplot as plt
import numpy as np

def solver(l_vals, e, tol=1e-12, max_iter=100):
    u_vals = np.zeros_like(l_vals)
    for i, l in enumerate(l_vals):
        u = l
        for _ in range(max_iter):
            f = u - e * np.sin(u) - l
            fp = 1 - e * np.cos(u)
            u = u - f / fp
            if abs(f / fp) < tol:
                break
        u_vals[i] = u
    return u_vals

l_grid = np.linspace(0, 2 * np.pi, 500)
eccentricities = [0, 0.1, 0.3, 0.5, 0.7, 0.9]

plt.figure(figsize=(8, 6))
for e in eccentricities:
    u_grid = solver(l_grid, e)
    plt.plot(l_grid, u_grid, label=f"e = {e}")
plt.xlabel(r"Mean anomaly $l$ (rad)")
plt.ylabel(r"Eccentric anomaly $u$ (rad)")
plt.title(r"Kepler equation: $u(l)$")
plt.grid(True)
plt.legend()
plt.show()
```

6.21.4 Problem 4: Relation Between True and Eccentric Anomalies

Problem Statement: Prove that the eccentric anomaly u and true anomaly v satisfy

$$\tan\frac{v}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{u}{2}. \tag{6.247}$$

Proof: Equating the two expressions for the orbital radius, Eqs. (6.42) and (6.28),

$$r = a(1 - e\cos u) = \frac{a(1-e^2)}{1 + e\cos v} \implies 1 + e\cos v = \frac{1-e^2}{1 - e\cos u}.$$

Solving for $\cos v$ as in Eq. (6.44),

$$\cos v = \frac{\cos u - e}{1 - e\cos u}.$$

Then

$$1 - \cos v = \frac{1 - e\cos u - \cos u + e}{1 - e\cos u} = \frac{(1+e)(1-\cos u)}{1 - e\cos u},$$
$$1 + \cos v = \frac{1 - e\cos u + \cos u - e}{1 - e\cos u} = \frac{(1-e)(1+\cos u)}{1 - e\cos u}.$$

Dividing the two equations, and using $\tan^2(x/2) = (1-\cos x)/(1+\cos x)$,

$$\frac{1-\cos v}{1+\cos v} = \frac{1+e}{1-e}\frac{1-\cos u}{1+\cos u} \implies \tan^2\frac{v}{2} = \frac{1+e}{1-e}\tan^2\frac{u}{2}.$$

The half-angles $u/2$ and $v/2$ lie in the same quadrant, since u and v are on the same side of the periapsis, so taking the positive root gives

$$\boxed{\tan\frac{v}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{u}{2}}. \tag{6.248}$$

6.21.5 Problem 5: Effective Potential and Hohmann Transfer Orbits

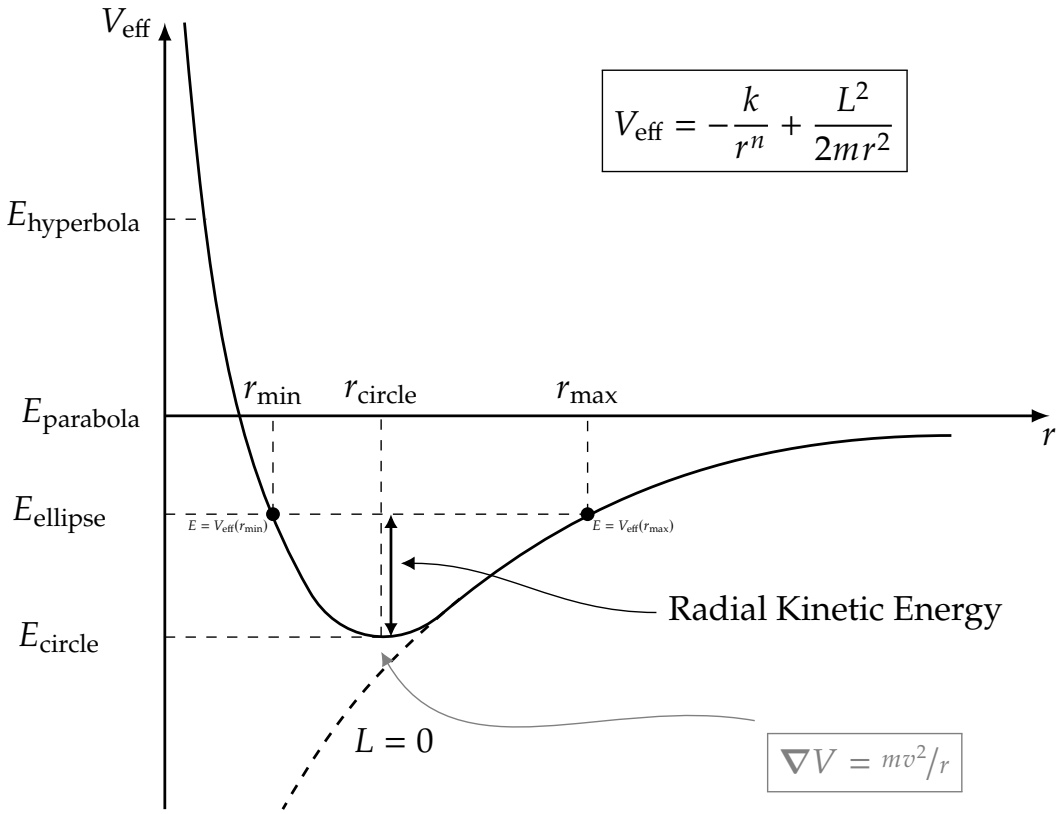


Figure 6.10: The effective potential $V_{\text{eff}}(r) = -k/r^n + L^2/(2mr^2)$ against r, with the energy levels of parabolic, circular, elliptical and hyperbolic orbits, the turning points r_{min} and r_{max} at one bound energy, and the curve for $L = 0$.

Problem Statement: The relative-orbit energy (6.32), divided by μ as in Problem 10, reads

$$\mathcal{E} \equiv \frac{E}{\mu} = \frac{1}{2}\dot{r}^2 + V_{\text{eff}}(r), \qquad V_{\text{eff}}(r) \equiv \frac{h^2}{2r^2} - \frac{GM}{r}, \qquad h \equiv \frac{L}{\mu}.$$

This is one-dimensional motion in the *effective potential* $V_{\text{eff}}(r)$ per unit mass. It is Newtonian gravity's case of the general power-law form $V_{\text{eff}} = -k/r^n + L^2/(2mr^2)$ of Figure 6.10, with $n = 1$, $k = GM$ and $m = 1$. Since $\frac{1}{2}\dot{r}^2 \geq 0$ always, the radii reachable at fixed $\mathcal{E}$ are exactly those with $V_{\text{eff}}(r) \leq \mathcal{E}$.

(a) Find the circular-orbit radius r_{circle}, where V_{eff} is stationary, and the corresponding energy $\mathcal{E}_{\text{circle}} \equiv V_{\text{eff}}(r_{\text{circle}})$, in terms of h and GM.

(b) For $\mathcal{E}_{\text{circle}} < \mathcal{E} < 0$, solve $V_{\text{eff}}(r) = \mathcal{E}$ for its two roots $r_{\min}, r_{\max}$, and show they are exactly the periapsis and apoapsis $a(1 \mp e)$ already implied by Eqs. (6.34) and (6.36).

(c) Show that $\mathcal{E} = \mathcal{E}_{\text{circle}}$, $\mathcal{E} = 0$, and $\mathcal{E} > 0$ correspond to the circular, parabolic, and hyperbolic rows of the table of conic sections in Section 6.5.

(d) An object orbits the Sun ($GM_\odot = 1.327 \times 10^{20}\,\text{m}^3\,\text{s}^{-2}$) with $h = 4.089 \times 10^{15}\,\text{m}^2\,\text{s}^{-1}$ and $\mathcal{E} = -4.423 \times 10^{8}\,\text{J}\,\text{kg}^{-1}$. Find $r_{\min}$, $r_{\max}$, and e.

Detailed Solutions:

(a) Setting $V'_{\text{eff}}(r) = -h^2/r^3 + GM/r^2 = 0$,

$$\boxed{r_{\text{circle}} = \frac{h^2}{GM}}, \qquad \mathcal{E}_{\text{circle}} = \frac{h^2}{2r_{\text{circle}}^2} - \frac{GM}{r_{\text{circle}}} = \frac{GM}{2r_{\text{circle}}} - \frac{GM}{r_{\text{circle}}} \implies \boxed{\mathcal{E}_{\text{circle}} = -\frac{G^2M^2}{2h^2}.}$$

(b) Multiplying $V_{\text{eff}}(r) = \mathcal{E}$ through by $2r^2$ gives the quadratic $2\mathcal{E}r^2 + 2GMr - h^2 = 0$, with roots

$$r = \frac{-GM \pm \sqrt{G^2M^2 + 2\mathcal{E}h^2}}{2\mathcal{E}}.$$

Substituting $\mathcal{E} = -GM/(2a)$ and $h^2 = GMa(1 - e^2)$, which are Eqs. (6.34) and (6.36) divided by μ and μ^2, gives $2\mathcal{E}h^2 = -G^2M^2(1 - e^2)$. The discriminant becomes $G^2M^2e^2$, and with $1/(2\mathcal{E}) = -a/GM$,

$$\boxed{r_{\min} = a(1 - e), \qquad r_{\max} = a(1 + e),}$$

the periapsis and apoapsis distances. The turning points of the effective potential and the geometry of the conic are two descriptions of the same orbit.

(c) From $\mathcal{E} = -GM/(2a)$. $\mathcal{E} = \mathcal{E}_{\text{circle}}$ forces $r_{\min} = r_{\max}$, so $e = 0$ (circular). $\mathcal{E} = 0$ forces $a \to \infty$, so $e = 1$ (parabolic). $\mathcal{E} > 0$ forces $a < 0$, so $e > 1$ (hyperbolic). This is the classification in the table of Section 6.5.

(d) With $GM = 1.327 \times 10^{20}\,\text{m}^3\,\text{s}^{-2}$, $r_{\text{circle}} = h^2/GM = 1.260 \times 10^{11}\,\text{m}$. From $\mathcal{E} = -GM/(2a)$, $a = -GM/(2\mathcal{E}) = 1.500 \times 10^{11}\,\text{m}$. Then $h^2 = GMa(1 - e^2)$ gives $1 - e^2 = h^2/(GMa) = 0.8400$, so

$$\boxed{e \simeq 0.400, \qquad r_{\min} = a(1 - e) \simeq 9.00 \times 10^{10}\,\text{m}, \qquad r_{\max} = a(1 + e) \simeq 2.10 \times 10^{11}\,\text{m}.}$$

6.21.6 Problem 6: Hohmann Transfer Impulses and Energy Budget

Problem Statement: A satellite of mass $m = 1000\,\text{kg}$ transfers from an initial circular orbit of period $P_1 = 6\,\text{h}$ to a geostationary orbit ($P_{\text{GEO}} \approx 86164\,\text{s}$), using $GM_\oplus = 3.986 \times 10^{14}\,\text{m}^3\,\text{s}^{-2}$.

(a) Find the intermediate transfer eccentricity e.

(b) Find the impulsive velocity boosts $\Delta v_1, \Delta v_2$ and the required impulses $m\Delta v_1$, $m\Delta v_2$.

(c) Find the total energy and angular momentum supplied.

Solution:

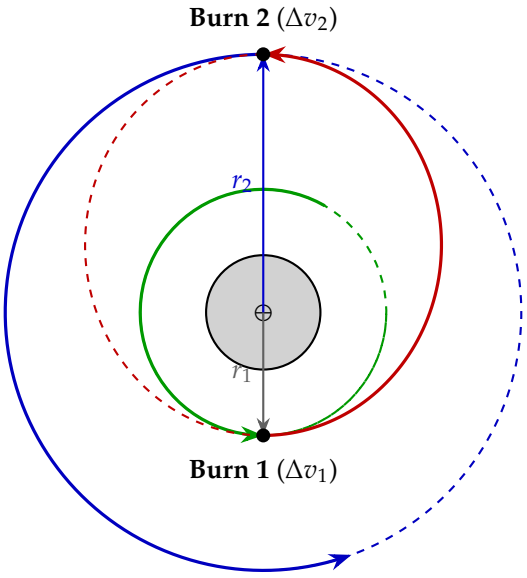


Figure 6.11: A Hohmann transfer ellipse between an initial circular orbit of radius r_1 and a geostationary orbit of radius r_2.

(a) By Kepler's third law (6.50), the initial and final radii are

$$r_1 = \left(\frac{GM_\oplus P_1^2}{4\pi^2}\right)^{1/3} \simeq 1.676 \times 10^7\,\text{m}, \qquad r_2 = \left(\frac{GM_\oplus P_{\text{GEO}}^2}{4\pi^2}\right)^{1/3} \simeq 4.216 \times 10^7\,\text{m}.$$

The transfer ellipse touches both circles, with periapsis $r_1 = a_t(1-e)$ and apoapsis $r_2 = a_t(1+e)$. So $a_t = (r_1+r_2)/2$, and the eccentricity is

$$\boxed{e = \frac{r_2 - r_1}{r_2 + r_1} \simeq 0.431}.$$

(b) The circular speeds are $v_{c1} = \sqrt{GM_\oplus/r_1} \simeq 4.876\,\text{km/s}$ and $v_{c2} = \sqrt{GM_\oplus/r_2} \simeq 3.075\,\text{km/s}$. At perigee of the transfer orbit, the vis-viva equation (6.37) gives

$$v_p = \sqrt{GM_\oplus(2/r_1 - 1/a_t)} \simeq 5.833\,\text{km/s}.$$

Thus

$$\Delta v_1 = v_p - v_{c1} \simeq 0.957\,\text{km/s} \implies \boxed{m\Delta v_1 \simeq 9.57 \times 10^5\,\text{N s}}.$$

At apogee, $v_a = \sqrt{GM_\oplus(2/r_2 - 1/a_t)} \simeq 2.319\,\text{km/s}$. Thus

$$\Delta v_2 = v_{c2} - v_a \simeq 0.755\,\text{km/s} \implies \boxed{m\Delta v_2 \simeq 7.55 \times 10^5\,\text{N s}}.$$

(c) The energy of a circular orbit is $-GM_\oplus m/(2r)$ by Eq. (6.34), and its angular momentum is $m\sqrt{GM_\oplus r}$ by Eq. (6.36) with $e = 0$. The total energy and angular momentum supplied are

$$\boxed{\Delta E = -\frac{GM_\oplus m}{2r_2} + \frac{GM_\oplus m}{2r_1} \simeq 7.16 \times 10^9\,\text{J}},$$

$$\boxed{\Delta L = m\sqrt{GM_\oplus r_2} - m\sqrt{GM_\oplus r_1} \simeq 4.79 \times 10^{13}\,\text{kg m}^2\,\text{s}^{-1}}.$$

6.21.7 Problem 7: Determining Orbital Elements from Initial State Vectors

Problem Statement: Given the initial relative position $\vec{r}_0$ and velocity $\dot{\vec{r}}_0$ at $t = 0$, derive the orbital elements $(a, e, \Omega, \iota, \omega, t_0)$ of a bound orbit.

Solution:

(a) **Conserved quantities.** The specific energy $\mathcal{E} = E/\mu = \dot{\vec{r}}_0^2/2 - GM/r_0$, the specific angular momentum $\vec{h} = \vec{L}/\mu = \vec{r}_0 \times \dot{\vec{r}}_0$, and the eccentricity vector $\vec{e} = \vec{A}/(GM\mu^2) = \dot{\vec{r}}_0 \times \vec{h}/GM - \hat{r}_0$, Eq. (6.72).

(b) **Orbital elements.** From the vis-viva equation (6.37) and Eqs. (6.70), (6.50) and (6.12),

$$a = \left(\frac{2}{r_0} - \frac{\dot{\vec{r}}_0^2}{GM} \right)^{-1}, \qquad e = |\vec{e}|, \qquad n = \sqrt{\frac{GM}{a^3}}, \qquad \iota = \arccos\left(\frac{\vec{h} \cdot \hat{R}}{h} \right).$$

The node vector $\vec{N} = \hat{R} \times \vec{h}$ points to the ascending node, since $\hat{n}$ is perpendicular to both $\hat{R}$ and $\hat{L}$. Then

$$\Omega = \operatorname{atan2}(\vec{N} \cdot \hat{\delta},\ \vec{N} \cdot \hat{\alpha}), \qquad \omega = \operatorname{atan2}\left(\frac{\hat{h} \cdot (\vec{N} \times \vec{e})}{Ne}, \frac{\vec{N} \cdot \vec{e}}{Ne} \right),$$

from Eqs. (6.10) and (6.71).

(c) **Time of periapsis.** The initial true anomaly has $\cos v_0 = \vec{e} \cdot \hat{r}_0/e$, with $\sin v_0$ of the same sign as $\dot{r}_0 = \hat{r}_0 \cdot \dot{\vec{r}}_0$. The eccentric anomaly u_0 follows from Eq. (6.46), $\tan(u_0/2) = \sqrt{(1-e)/(1+e)}\tan(v_0/2)$. Kepler's equation at $t = 0$ gives $\ell_0 = -nt_0 = u_0 - e\sin u_0$, so the time of periapsis passage is

$$t_0 = -\frac{u_0 - e\sin u_0}{n}.$$

6.21.8 Problem 8: Continuous Tangential Thrust Orbital Evolution

Problem Statement: A 500 kg satellite initially moves in a circular Earth orbit of period $P = 8\,\mathrm{h}$. At $t = 0$, a thruster of force $F_\mathrm{T} = 100\,\mathrm{N}$ is switched on, always directed along the instantaneous velocity $\dot{\vec{r}}$.

(a) Find the initial orbital radius $a(0)$.

(b) Derive the rate of change of the semi-major axis $\dot{a}$.

(c) Derive the rate of change of the eccentricity $\dot{e}$ from the exact rate of the Laplace–Runge–Lenz vector $\vec{A}$.

(d) Can the coupled evolution equations for (a, e) be solved analytically?

Essential Ideas:

- **Thrust acceleration and initial radius**: The satellite's mass is negligible next to the Earth's, so $\mu = m = 500\,\mathrm{kg}$. The thrust acceleration is $F_\mathrm{T}/\mu = 100\,\mathrm{N}/500\,\mathrm{kg} = 0.2\,\mathrm{m\,s^{-2}}$. For an initially circular orbit ($e = 0$) of period $P = 8\,\mathrm{h} = 28{,}800\,\mathrm{s}$ around the Earth ($GM_\oplus \approx 3.986 \times 10^{14}\,\mathrm{m^3\,s^{-2}}$), Kepler's third law gives

$$a(0) = \left(\frac{GM_\oplus P^2}{4\pi^2} \right)^{1/3} \simeq 2.03 \times 10^7\,\mathrm{m} \approx 20{,}300\,\mathrm{km}.$$

- **Perturbation force vector**: The thrust is tangential to the path, $\vec{F} = F_{\mathrm{T}}\,\dot{\vec{r}}/|\dot{\vec{r}}|$. In polar coordinates the velocity is $\dot{\vec{r}} = \dot{r}\,\hat{r} + r\dot{\phi}\,\hat{\phi} = \dot{r}\,\hat{r} + \frac{L}{\mu r}\,\hat{\phi}$, so $\vec{F} = F_r\hat{r} + F_\phi\hat{\phi}$ has the components

$$F_r = F_{\mathrm{T}}\frac{\dot{r}}{|\dot{\vec{r}}|}, \qquad F_\phi = \frac{F_{\mathrm{T}}L}{\mu r|\dot{\vec{r}}|}.$$

  With $\xi \equiv \mu r\dot{r}/L = e\sin v/(1+e\cos v)$ (Eq. (6.120)), the radial component is $F_r = \xi F_\phi$.

Detailed Solutions:

(a) **Initial semi-major axis**.

$$\boxed{a(0) = \left(\frac{GM_\oplus P^2}{4\pi^2}\right)^{1/3} = \left(\frac{3.986\times10^{14}\times(28800)^2}{4\pi^2}\right)^{1/3} \simeq 2.03\times10^7\ \mathrm{m}}.$$

(b) **Evolution of the semi-major axis $\dot{a}$**. The work done per unit time by the tangential thrust is, by Eq. (6.97),

$$\dot{E} = \vec{F}\cdot\dot{\vec{r}} = F_{\mathrm{T}}|\dot{\vec{r}}|.$$

The energy $E = -GM_\oplus\mu/(2a)$ changes at the rate $\dot{E} = (GM_\oplus\mu/2a^2)\dot{a}$, so

$$\frac{GM_\oplus\mu}{2a^2}\dot{a} = F_{\mathrm{T}}|\dot{\vec{r}}| \implies \dot{a} = \frac{2a^2}{GM_\oplus}\frac{F_{\mathrm{T}}}{\mu}\,|\dot{\vec{r}}|.$$

The vis-viva equation (6.37) with $r = a(1-e^2)/(1+e\cos v)$ gives

$$|\dot{\vec{r}}|^2 = GM_\oplus\left(\frac{2}{r}-\frac{1}{a}\right) = GM_\oplus\,\frac{2(1+e\cos v)-(1-e^2)}{a(1-e^2)} = \frac{GM_\oplus}{a}\,\frac{1+2e\cos v+e^2}{1-e^2},$$

so the exact rate of change of the semi-major axis is

$$\boxed{\frac{da}{dt} = \frac{2F_{\mathrm{T}}}{\mu}\sqrt{\frac{a^3}{GM_\oplus}}\sqrt{\frac{1+2e\cos v+e^2}{1-e^2}}}. \tag{6.249}$$

(c) **Evolution of the eccentricity $\dot{e}$ via the Laplace–Runge–Lenz vector**. The Laplace–Runge–Lenz vector $\vec{A} = \vec{p}\times\vec{L} - GM\mu^2\hat{r}$ has magnitude $A = GM_\oplus\mu^2 e$ and points to the periapsis, $\hat{A} = \cos v\,\hat{r} - \sin v\,\hat{\phi}$ (Eq. (6.130)).

Under a general planar perturbing force $\vec{F} = F_r\hat{r} + F_\phi\hat{\phi}$, the exact rate of change of $\vec{A}$ is Eq. (6.128),

$$\dot{\vec{A}} = \vec{F}\times\vec{L} + \vec{p}\times(\vec{r}\times\vec{F}) = L\left[2F_\phi\hat{r} - (F_r+\xi F_\phi)\hat{\phi}\right].$$

Since $\hat{A}\cdot\hat{A} = 1$ gives $\hat{A}\cdot\dot{\hat{A}} = 0$, the rate $\dot{A}$ is the projection of $\dot{\vec{A}}$ on $\hat{A}$,

$$\begin{aligned}\dot{A} &= \dot{\vec{A}}\cdot\hat{A} = L\left[2F_\phi\hat{r} - (F_r+\xi F_\phi)\hat{\phi}\right]\cdot(\cos v\,\hat{r} - \sin v\,\hat{\phi})\\ &= L\left[2F_\phi\cos v + (F_r+\xi F_\phi)\sin v\right]\\ &= L\left[F_r\sin v + F_\phi(2\cos v + \xi\sin v)\right].\end{aligned}$$

Evaluating this on the instantaneous osculating orbit with $A = GM_\oplus\mu^2 e$ and $L = \mu\sqrt{GM_\oplus a(1-e^2)}$ gives Gauss's equation for the eccentricity, Eq. (6.136),

$$\dot{e} = \sqrt{\frac{a(1-e^2)}{GM_\oplus\mu^2}}\left[F_r\sin v + F_\phi(2\cos v + \xi\sin v)\right]. \tag{6.250}$$

Now we substitute the components for tangential thrust, $F_r = \xi F_\phi$ and $F_\phi = \dfrac{L}{r}\dfrac{F_\mathrm{T}}{\mu|\dot{\vec{r}}|}$,

$$\begin{aligned} F_r \sin v + F_\phi(2\cos v + \xi \sin v) &= \frac{F_\mathrm{T} L}{\mu r|\dot{\vec{r}}|}\left[\xi \sin v + 2\cos v + \xi \sin v\right] \\ &= \frac{2F_\mathrm{T} L}{\mu r|\dot{\vec{r}}|}\left(\cos v + \xi \sin v\right). \end{aligned}$$

Using $\xi = e\sin v/(1 + e\cos v)$,

$$\cos v + \xi \sin v = \cos v + \frac{e\sin^2 v}{1 + e\cos v} = \frac{\cos v(1 + e\cos v) + e\sin^2 v}{1 + e\cos v} = \frac{\cos v + e}{1 + e\cos v}.$$

With $\dfrac{L}{r} = \dfrac{\mu\sqrt{GM_\oplus a(1-e^2)}}{a(1-e^2)/(1+e\cos v)} = \dfrac{\mu\sqrt{GM_\oplus}\,(1 + e\cos v)}{\sqrt{a(1-e^2)}}$,

$$\begin{aligned} \dot{e} &= \sqrt{\frac{a(1-e^2)}{GM_\oplus \mu^2}} \cdot \frac{2F_\mathrm{T}}{\mu|\dot{\vec{r}}|}\left[\frac{\mu\sqrt{GM_\oplus}(1 + e\cos v)}{\sqrt{a(1-e^2)}}\right]\left[\frac{\cos v + e}{1 + e\cos v}\right] \\ &= \frac{2F_\mathrm{T}}{\mu}\frac{\cos v + e}{|\dot{\vec{r}}|}. \end{aligned}$$

Writing $|\dot{\vec{r}}|$ in terms of the elements with the vis-viva equation,

$$\boxed{\frac{de}{dt} = \frac{2F_\mathrm{T}}{\mu}\frac{\cos v + e}{|\dot{\vec{r}}|} = \frac{2F_\mathrm{T}}{\mu}\sqrt{\frac{a(1-e^2)}{GM_\oplus}}\,\frac{\cos v + e}{\sqrt{1 + 2e\cos v + e^2}}}. \tag{6.251}$$

Differentiating $e^2 = 1 + 2\mathcal{E}h^2/(GM_\oplus)^2$ (Eq. (6.39) with $\mathcal{E} = E/\mu$ and $h = L/\mu$) would give $\dot{e}$ divided by e, an undefined $0/0$ at $e = 0$. The rate of the Runge–Lenz vector avoids this and gives the finite initial rate $\dot{e}\big|_{e=0} = (2F_\mathrm{T}/\mu)\sqrt{a/GM_\oplus}\cos v$.

(d) **Analytical solvability**. Equations (6.249) and (6.251) do not form a closed system for a and e, because both rates depend on the orbital phase $v(t)$. The phase itself evolves by Eq. (6.149),

$$\dot{v} = \frac{L}{\mu r^2} - \dot{\omega},$$

where the periapsis rate $\dot{\omega}$ is also driven by the thrust. The thrust changes the orbit throughout each revolution, so the coupled system $(a(t), e(t), v(t))$ has no closed-form solution in elementary functions. In mission design it is solved by numerical integration, or by *orbit averaging* over each orbital period (Section 6.16),

$$\langle \dot{a} \rangle = \frac{1}{2\pi}\int_0^{2\pi} \dot{a}\, d\ell = \frac{1}{P}\int_0^{2\pi} \dot{a}\,\frac{\mu r^2}{L}\, dv.$$

6.21.9 Problem 9: Double-Lined Spectroscopic Binary — Radial Velocities and Masses

Problem Statement: In a *double-lined spectroscopic binary* (SB2), we see absorption lines from *both* stars in the combined spectrum. Each line is Doppler-shifted by that star's own orbital motion around the center of mass. A Doppler shift only measures the *radial velocity* $v_r \equiv \dot{\vec{x}} \cdot \hat{R}$, the part of a star's velocity $\dot{\vec{x}}$ along the line of sight $\hat{R}$. Motion across the sky, perpendicular

to $\hat{R}$, does not stretch or compress the wavelength, so it leaves no Doppler signal at all. That sideways motion only shows up as proper motion (Chapter 3), not in the spectrum. Tracking each star's wavelength shift separately gives two radial-velocity curves, $v_{r,1}(t)$ and $v_{r,2}(t)$. Both share the same period P and eccentricity e, but usually have different semi-amplitudes K_1 and K_2 (Figures 6.12 and 6.13). A particular system is measured to have $P = 4.20\,\mathrm{d}$, $e = 0.15$, $K_1 = 88\,\mathrm{km\,s^{-1}}$, and $K_2 = 132\,\mathrm{km\,s^{-1}}$.

(a) Without knowing the orbital inclination ι, what can already be said about the mass ratio m_1/m_2?

(b) Derive $(m_1 + m_2)\sin^3\iota$ in terms of P, K_1, K_2, and e, and evaluate it numerically.

(c) Hence find $m_1 \sin^3\iota$ and $m_2 \sin^3\iota$ individually.

(d) This same system is also observed to be eclipsing, so $\iota \simeq 90°$. What are the actual masses m_1 and m_2?

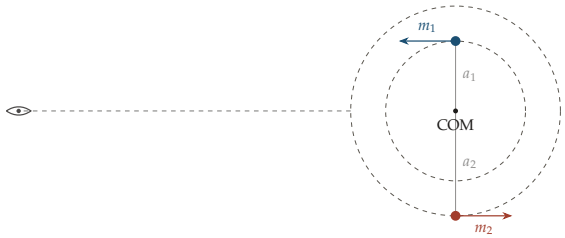


Figure 6.12: A binary at quadrature, with m_1, m_2, and the COM in a line. m_1 moves away from and m_2 towards the observer, along the line of sight $\hat{R}$.

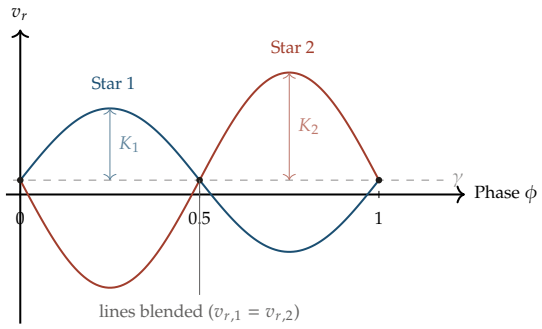


Figure 6.13: Radial-velocity curves of both stars for a circular orbit. Two sine curves in antiphase about γ, with semi-amplitudes $K_1 \neq K_2$.

Essential Ideas:

- Start from $\vec{x}_1 = (m_2/M)\vec{r}$, $\vec{x}_2 = -(m_1/M)\vec{r}$ (Section 6.1.2), shifted by the centre-of-mass motion $\vec{V}_{\mathrm{CM}}t$. Differentiate and project onto $\hat{R}$,

$$v_{r,1} - \gamma = \frac{m_2}{M}\, v_{r,\mathrm{rel}}, \qquad v_{r,2} - \gamma = -\frac{m_1}{M}\, v_{r,\mathrm{rel}}, \qquad \gamma \equiv \vec{V}_{\mathrm{CM}} \cdot \hat{R}.$$

  Here $v_{r,\mathrm{rel}} \equiv v_{r,1} - v_{r,2}$, and γ is constant (Eq. (6.4)). This holds at every phase. So

$$\frac{K_1}{K_2} = \frac{m_2}{m_1},$$

  whatever ι and e are.

- A Keplerian orbit's line-of-sight speed has semi-amplitude $K = 2\pi a \sin\iota/(P\sqrt{1-e^2})$ (Problem 11). Apply this to the relative orbit ($a = a_1 + a_2$). Then $v_{r,\mathrm{rel}} = v_{r,1} - v_{r,2}$ has amplitude $K_1 + K_2$.

Detailed Solutions:

(a) **Mass ratio.**

$$\boxed{\frac{m_1}{m_2} = \frac{K_2}{K_1} = \frac{132}{88} = 1.50.}$$

(b) **Total mass**. Apply $K = 2\pi a \sin\iota/(P\sqrt{1-e^2})$ to the relative orbit, whose amplitude is K_1+K_2, so $a\sin\iota = (K_1+K_2)P\sqrt{1-e^2}/(2\pi)$. Kepler's third law (6.50) is $P^2 = 4\pi^2 a^3/[G(m_1+m_2)]$, so

$$(m_1+m_2)\sin^3\iota = \frac{4\pi^2(a\sin\iota)^3}{GP^2} = \frac{4\pi^2}{GP^2}\,\frac{(K_1+K_2)^3P^3(1-e^2)^{3/2}}{8\pi^3}$$
$$= \frac{P(K_1+K_2)^3(1-e^2)^{3/2}}{2\pi G}.$$

With $P = 3.629\times10^5$ s, $K_1+K_2 = 2.20\times10^5\,\mathrm{m\,s^{-1}}$ and $(1-e^2)^{3/2} = 0.966$,

$$\boxed{(m_1+m_2)\sin^3\iota \simeq 8.91\times10^{30}\,\mathrm{kg} \simeq 4.48\,M_\odot.}$$

(c) **Individual masses**. Use $m_1 = \dfrac{K_2}{K_1+K_2}(m_1+m_2)$ and $m_2 = \dfrac{K_1}{K_1+K_2}(m_1+m_2)$, the same centre-of-mass weighting as in (a).

$$\boxed{m_1\sin^3\iota \simeq 2.69\,M_\odot, \qquad m_2\sin^3\iota \simeq 1.79\,M_\odot.}$$

(d) **Eclipsing case**. $\iota \simeq 90^\circ$, so $\sin^3\iota \simeq 1$ and the factors in (c) drop out.

$$\boxed{m_1 \simeq 2.69\,M_\odot, \qquad m_2 \simeq 1.79\,M_\odot.}$$

Double-lined eclipsing binaries are valuable for exactly this reason. Both masses, and from the light curve both radii, come out with almost no model assumptions.

6.21.10 Problem 10: Dynamical Mass of a Host Halo from a Satellite's Orbit

Problem Statement: A satellite of mass m_s, for example a dwarf galaxy or globular cluster, orbits a much bigger dark-matter halo of mass M_h. We measure its present distance r from the Galactic center and its full relative velocity $\dot{\vec{r}}$ (from proper motion plus radial velocity). We also estimate its orbital eccentricity e independently, for example from similar orbits in cosmological simulations. As in Section 6.1.2, let $M \equiv m_s + M_h$ be the total mass and $\mu \equiv m_s M_h/M$ the reduced mass. If the satellite sits well outside the halo's inner mass, we can treat the halo as a point mass (P2). Then the satellite follows an ordinary Keplerian orbit, with energy $E = -GM\mu/(2a)$ (Eq. (6.34)) and angular momentum $\vec{L} = \mu\,\vec{r}\times\dot{\vec{r}}$, $L^2 = GM\mu^2 a(1-e^2)$ (Eq. (6.36)).

(a) Show that μ cancels out of both the vis-viva equation and the angular-momentum relation, leaving two equations in the directly observable quantities r, $|\dot{\vec{r}}|$, and $h \equiv |\vec{r}\times\dot{\vec{r}}| = L/\mu$. Combine them to show that the total mass M satisfies

$$\frac{G^2(1-e^2)}{h^2}M^2 - \frac{2G}{r}M + \dot{\vec{r}}^{\,2} = 0,$$

and solve for M in terms of $r, |\dot{\vec{r}}|, h, e$.

(b) Suppose instead the satellite happens to be caught exactly at apocenter, where the velocity is purely tangential ($\dot{\vec{r}} \perp \vec{r}$) and $r = r_{\mathrm{apo}} = a(1+e)$. Show that the mass estimate then becomes the much simpler

$$M = \frac{v_{\mathrm{apo}}^2\, r_{\mathrm{apo}}}{G(1-e)},$$

where $v_{\mathrm{apo}} = |\dot{\vec{r}}|$ at apocenter.

(c) A dwarf satellite is observed at apocenter with $r_{\rm apo} = 250\,\mathrm{kpc}$ and $v_{\rm apo} = 95\,\mathrm{km\,s^{-1}}$, and its orbit is modeled to have $e = 0.6$. Evaluate M, in kg and in $M_\odot$, and argue that it is an excellent estimate of M_h alone.

(d) Eccentricity is only one possible choice of the "extra" datum used here. What property of the governing equations makes some third quantity beyond r and $|\dot{\vec{r}}|$ necessary at all, and what else could play that role?

(e) This method returns only the total mass $M = m_s + M_h$, never m_s and M_h separately. Explain why no amount of additional *orbital* data — more precise r, $\dot{\vec{r}}$, or e, or even more satellites on more orbits — could ever split M into m_s and M_h individually, and say what kind of information could.

Essential Ideas:

- μ cancels. In the vis-viva equation $\dot{\vec{r}}^{\,2} = GM(2/r - 1/a)$, Eq. (6.37), the orbit's shape and motion depend on m_s and M_h only through M. μ drops out once E and L are turned into relations among r, $|\dot{\vec{r}}|$ and $h \equiv |\vec{r} \times \dot{\vec{r}}|$.
- There are three unknowns (a, e, M) but only two equations relating r, $|\dot{\vec{r}}|$ and h. So one more independent piece of data is needed, and e is the usual choice.
- At apocenter $\dot{r} = 0$, so the velocity is purely tangential and $r_{\rm apo} = a(1+e)$ directly, without needing h.
- Test-particle limit. Since $m_s \ll M_h$, $M = m_s + M_h$ is almost exactly M_h.
- **Degeneracy, not just ignorance**. μ is the one combination of m_s and M_h that is *not* their sum, and it cancels completely from both relations. So every orbital observable depends on the masses only through M, and any split of M gives the identical orbit. This is different from Problem 9's mass *ratio*, where two tracers gave one comparison. Here there is only one tracer, so only one orbit.

Detailed Solutions:

(a) **General mass estimator**. Divide $E = \frac{1}{2}\mu\dot{\vec{r}}^{\,2} - \dfrac{GM\mu}{r} = -\dfrac{GM\mu}{2a}$ by μ to recover vis-viva. Divide $L^2 = GM\mu^2 a(1-e^2)$ by μ^2 and write $h \equiv L/\mu = |\vec{r} \times \dot{\vec{r}}|$, which needs no mass.

$$\dot{\vec{r}}^{\,2} = GM\left(\frac{2}{r} - \frac{1}{a}\right), \qquad h^2 = GMa(1-e^2) \implies a = \frac{h^2}{GM(1-e^2)}.$$

Substituting $1/a$ into the first,

$$\dot{\vec{r}}^{\,2} = \frac{2GM}{r} - \frac{G^2M^2(1-e^2)}{h^2} \implies \boxed{\frac{G^2(1-e^2)}{h^2}M^2 - \frac{2G}{r}M + \dot{\vec{r}}^{\,2} = 0,}$$

a quadratic in M, giving

$$\boxed{M = \frac{h^2}{G(1-e^2)}\left[\frac{1}{r} \pm \sqrt{\frac{1}{r^2} - \frac{(1-e^2)\dot{\vec{r}}^{\,2}}{h^2}}\right].}$$

The product of the two roots is $h^2\dot{\vec{r}}^{\,2}/[G^2(1-e^2)] > 0$ and their sum is positive, so both roots are positive and both are possible orbits. With $\lambda = h^2/(GM)$ the conic gives $e\cos v = \lambda/r - 1$. The + root, the larger M, has $r > \lambda$ and puts the satellite on the far half of its orbit. The − root puts it on the near half. The sign of $\dot{r}$ does not decide between them, so one more fact about the phase is needed. At apocenter, part (b), it is the + root.

(b) **Apocenter simplification**. Here $\dot{r} = 0$, and $r = r_{\text{apo}} = a(1+e)$, so $1/a = (1+e)/r_{\text{apo}}$.

$$v_{\text{apo}}^2 = GM\left(\frac{2}{r_{\text{apo}}} - \frac{1+e}{r_{\text{apo}}}\right) = \frac{GM(1-e)}{r_{\text{apo}}} \implies \boxed{M = \frac{v_{\text{apo}}^2\, r_{\text{apo}}}{G(1-e)}.}$$

This matches the + root of part (a) at apocenter, where $h = r_{\text{apo}} v_{\text{apo}}$ and the square root becomes e/r_{apo}.

(c) **Numerical evaluation**. Use $r_{\text{apo}} = 7.714 \times 10^{21}$ m, $v_{\text{apo}} = 9.5 \times 10^4\,\text{m s}^{-1}$, $e = 0.6$.

$$M = \frac{(9.5\times10^4)^2(7.714\times10^{21})}{(6.674\times10^{-11})(1-0.6)} \implies \boxed{M \simeq 2.61\times10^{42}\,\text{kg} \simeq 1.31\times10^{12}\,M_\odot,}$$

which is typical for a Milky-Way-sized halo. Dwarf-satellite masses (10^4–$10^8\,M_\odot$) are tiny next to this, at most about 10^{-4} of it, so $M \simeq M_h$ to four figures or better.

(d) **Why a third datum is unavoidable**. r and $|\dot{\vec{r}}|$ give two independent numbers, but the orbit has three unknowns (a, e, M), related by only two equations, vis-viva and $h^2 = GMa(1-e^2)$. So one more independent piece of data is needed to close the system. It could be e, as used here, or h itself, or the time elapsed since a known pericenter passage, found from Kepler's equation (Section 6.9). This last option is the basis of the Local Group timing argument for the Milky Way–M31 mass.

(e) **The barycenter location is the missing parameter**. To specify $\vec{x}_s(t)$ and $\vec{x}_h(t)$ separately, we need the six orbital elements $(a, e, \iota, \Omega, \omega, t_0)$ of the *relative* orbit $\vec{r} \equiv \vec{x}_s - \vec{x}_h$, the total mass M, and one more number. The six elements describe only the shape, orientation, and timing of $\vec{r}(t)$. They say nothing about how the total mass M is split between the two bodies, and it is that split we are still missing. The extra number is the mass fraction M_h/M, which is the same as the barycenter's location on the line joining the two bodies. Parts (a)–(d) give only the relative orbit and M. The relative orbit cannot see M_h/M, and this is an exact symmetry, not just a practical limit. Newton's third law gives $\ddot{\vec{x}}_s = -GM_h\hat{r}/r^2$ and $\ddot{\vec{x}}_h = Gm_s\hat{r}/r^2$, so

$$\ddot{\vec{r}} = \ddot{\vec{x}}_s - \ddot{\vec{x}}_h = -\frac{GM}{r^2}\hat{r},$$

which depends on m_s and M_h only through M. So $\vec{r}(t)$ stays exactly the same under $m_s \to m_s + \delta$, $M_h \to M_h - \delta$ at fixed M. No relative-orbit measurement, however precise, can pin down M_h/M.

The individual position $\vec{x}_s = \vec{X} + (M_h/M)\vec{r}$ does depend on M_h/M. And $M_h/M = |\vec{x}_s - \vec{X}|/r$ is a purely geometric ratio, so it needs no dynamics to measure. This is the missing parameter. It is the same idea as Problem 1, where $m_1/m_2 = a_2/a_1$ came from each star's distance to the visible barycenter.

For a satellite and its halo, the problem is one of scale, not principle. Since $m_s/M_h \sim 10^{-4}$–10^{-8}, the barycenter sits only that tiny fraction of r from the halo's center, far too close to measure astrometrically (unlike Problem 1's pair of comparable masses). Without it, we estimate m_s some other way, for example from its luminosity via a mass-to-light ratio, or from its own internal velocity dispersion via the virial theorem ($M \sim \sigma_v^2 R/G$, Section 10.2). Then $M_h = M - m_s$.

6.21.11 Problem 11: The Line-of-Sight Direction, in Matrices and Without Them

Problem Statement: Problem 9 projected an orbital velocity onto the line of sight $\hat{R}$, through the semi-amplitude K and the systemic velocity $\gamma = \vec{V}_{\text{CM}} \cdot \hat{R}$, without ever pinning down what $\hat{R}$ *is* in terms of the orbital elements (ι, Ω, ω).

(a) Starting from the rotation relating the orbital frame $(\hat{n}, \hat{y}, \hat{L})$ to the observer frame $(\hat{\alpha}, \hat{\delta}, \hat{R})$, Eq. (6.8), invert it to express $\hat{R}$ in terms of $\hat{n}, \hat{y}, \hat{L}$. Show that the result does not depend on Ω at all,

$$\hat{R} = \sin\iota\,\hat{y} + \cos\iota\,\hat{L}.$$

(b) Rederive the same identity without any matrix algebra. Use only two facts. First, the node line $\hat{n}$ is, by construction, the intersection of the orbital plane and the sky plane. Second, the inclination is the angle between the two planes' normals, $\hat{L}$ and $\hat{R}$.

(c) A body follows a Keplerian orbit of elements (a, e, ι, ω) in the plane spanned by $\hat{n}, \hat{y}$, at true anomaly v (so its polar angle from $\hat{n}$ is $\phi = \omega + v$). Using $\hat{R} = \sin\iota\,\hat{y} + \cos\iota\,\hat{L}$ together with the fact that the velocity lies entirely in the orbital plane, show that the line-of-sight (radial) velocity $v_r = \dot{\vec{r}} \cdot \hat{R}$ takes the form

$$v_r = \gamma + K\big[\cos(\omega + v) + e\cos\omega\big],$$

and identify K in terms of G, M, a, e, ι.

(d) Using Kepler's third law, show that this K is the semi-amplitude

$$K = \frac{2\pi a \sin\iota}{P\sqrt{1-e^2}}$$

quoted without proof in Problem 9.

Essential Ideas:

- A rotation matrix is orthogonal, so $\mathbf{R}^{-1} = \mathbf{R}^T$ (Chapter 2, Eq. (2.4)). Undoing it costs no more than transposing and reversing the product order.
- Simple geometric facts about an orthonormal triad, such as which vectors are perpendicular, or what a named angle is the cosine of, can rebuild a rotation result with less algebra than the matrix itself.
- Angular momentum conservation confines $\vec{r}$ and $\dot{\vec{r}}$ to the orbital plane, so $\vec{r} \cdot \hat{L} = \dot{\vec{r}} \cdot \hat{L} = 0$. This drops $\hat{L}$ from $v_r = \dot{\vec{r}} \cdot \hat{R}$ in part (c).

Detailed Solutions:

(a) **Matrix version**. Both 3×3 factors in Eq. (6.8) are rotation matrices, so they are orthogonal, $\mathbf{R}^{-1} = \mathbf{R}^T$. Inverting the product reverses the order and transposes each factor.

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\delta}\\ \hat{R}\end{pmatrix} = \begin{pmatrix}\cos\Omega & \sin\Omega & 0\\ -\sin\Omega & \cos\Omega & 0\\ 0 & 0 & 1\end{pmatrix}^T \begin{pmatrix}1 & 0 & 0\\ 0 & \cos\iota & \sin\iota\\ 0 & -\sin\iota & \cos\iota\end{pmatrix}^T \begin{pmatrix}\hat{n}\\ \hat{y}\\ \hat{L}\end{pmatrix}$$

$$= \begin{pmatrix}\cos\Omega & -\sin\Omega & 0\\ \sin\Omega & \cos\Omega & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\ 0 & \cos\iota & -\sin\iota\\ 0 & \sin\iota & \cos\iota\end{pmatrix}\begin{pmatrix}\hat{n}\\ \hat{y}\\ \hat{L}\end{pmatrix}.$$

The third row of the product is the third row of the first matrix, $(0, 0, 1)$, times the second matrix. That picks out the second matrix's third row, $(0, \sin\iota, \cos\iota)$, so

$$\hat{R} = \underbrace{(0)}_{\hat{n}\text{-coefficient}}\,\hat{n} + \sin\iota\,\hat{y} + \cos\iota\,\hat{L} \implies \boxed{\hat{R} = \sin\iota\,\hat{y} + \cos\iota\,\hat{L},}$$

since Ω enters only the first two rows of the first matrix, which never touch the row producing $\hat{R}$.

(b) **Geometric version**. $\hat{n}$ is the intersection of the orbital and sky planes. So it is perpendicular to both planes' normals, $\hat{n}\cdot\hat{L}=0$ and $\hat{n}\cdot\hat{R}=0$. Since $\{\hat{n},\hat{y},\hat{L}\}$ is orthonormal, any unit vector with zero $\hat{n}$-component is a combination of $\hat{y}$ and $\hat{L}$ alone.

$$\hat{R}=(\hat{R}\cdot\hat{y})\,\hat{y}+(\hat{R}\cdot\hat{L})\,\hat{L}.$$

By definition, ι is the angle between the planes' normals, so $\hat{R}\cdot\hat{L}=\cos\iota$. Since $\hat{R}\cdot\hat{n}=0$ and $\hat{R}$ is a unit vector, $(\hat{R}\cdot\hat{y})^2=\sin^2\iota$. The orientation convention, that $\hat{y}$ leads $\hat{n}$ by 90° toward $\hat{R}$, fixes the sign.

$$\boxed{\hat{R}=\sin\iota\,\hat{y}+\cos\iota\,\hat{L},}$$

the same as part (a), with no matrix multiplied.

(c) **Radial velocity**. Angular momentum conservation confines the orbit to the plane spanned by $\hat{n},\hat{y}$, so $\dot{\vec{r}}\cdot\hat{L}=0$, and

$$v_r=\dot{\vec{r}}\cdot\hat{R}=\sin\iota\,(\dot{\vec{r}}\cdot\hat{y}).$$

With $\vec{r}=r\cos\phi\,\hat{n}+r\sin\phi\,\hat{y}$ and the planar-polar velocity $\dot{\vec{r}}=\dot{r}\,\hat{r}+r\dot{\phi}\,\hat{\phi}$,

$$\dot{\vec{r}}\cdot\hat{y}=\dot{r}\sin\phi+r\dot{\phi}\cos\phi.$$

Substitute $\dot{r}=\dfrac{Le\sin v}{\mu\lambda}$ (Eq. (6.117)) and $r\dot{\phi}=\dfrac{L}{\mu r}=\dfrac{L(1+e\cos v)}{\mu\lambda}$ (Eq. (6.114)), with $\phi=\omega+v$.

$$\dot{\vec{r}}\cdot\hat{y}=\frac{L}{\mu\lambda}\Big[e\sin v\sin(\omega+v)+(1+e\cos v)\cos(\omega+v)\Big]=\frac{L}{\mu\lambda}\Big[\cos(\omega+v)+e\cos\omega\Big],$$

using $\sin v\sin(\omega+v)+\cos v\cos(\omega+v)=\cos\omega$. Since $L=\mu\sqrt{GM\lambda}$ (Eq. (6.36)) and $\lambda=a(1-e^2)$, $L/(\mu\lambda)=\sqrt{GM/[a(1-e^2)]}$. Add back γ, Problem 9's line-of-sight velocity of the centre of mass, as the baseline.

$$v_r=\sin\iota\sqrt{\frac{GM}{a(1-e^2)}}\Big[\cos(\omega+v)+e\cos\omega\Big]+\gamma.$$

Matching to the stated form,

$$\boxed{K=\sin\iota\sqrt{\frac{GM}{a(1-e^2)}}.}$$

(d) **Consistency with Problem 9**. Kepler's third law gives $GM=4\pi^2a^3/P^2$, so

$$K=\sin\iota\sqrt{\frac{4\pi^2a^3/P^2}{a(1-e^2)}}=\sin\iota\sqrt{\frac{4\pi^2a^2}{P^2(1-e^2)}}\implies\boxed{K=\frac{2\pi a\sin\iota}{P\sqrt{1-e^2}},}$$

the radial-velocity semi-amplitude of a spectroscopic binary, Problem 9's K.

6.21.12 Problem 12: The Gravitational Slingshot

Problem Statement: A spacecraft flies by Jupiter. Jupiter's own speed around the Sun is $V_p\simeq 13.1\,\mathrm{km\,s^{-1}}$. Relative to Jupiter, the spacecraft moves on a hyperbolic orbit with $e=1.2$ and $v_\infty=10\,\mathrm{km\,s^{-1}}$ (Eq. (6.61)). Find the change in the spacecraft's speed around the Sun, given that its speed relative to Jupiter cannot change.

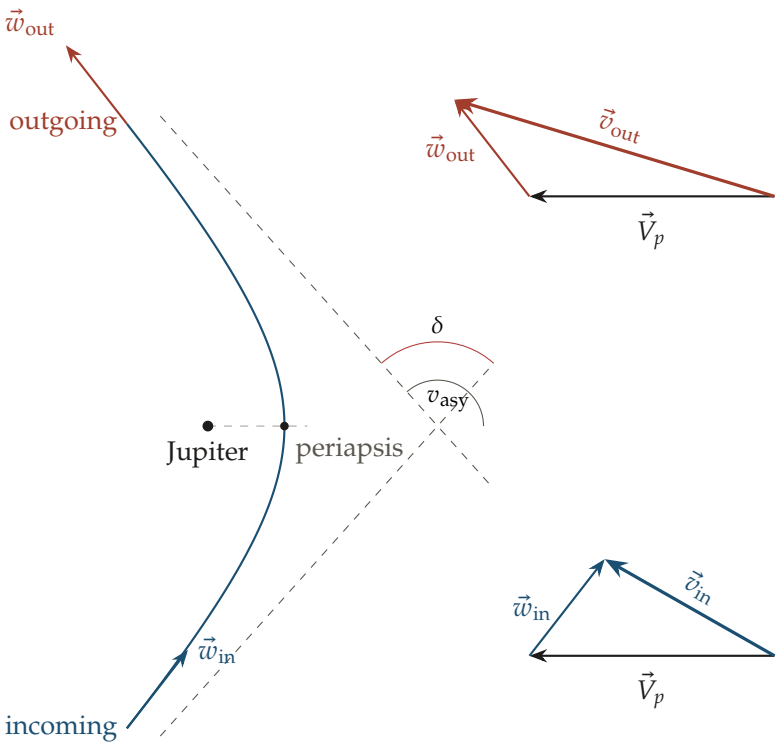


Figure 6.14: The gravitational slingshot. Left, Jupiter's frame, with the incoming and outgoing branches of the hyperbola. The dashed lines are the asymptotes, crossing at the hyperbola's center, with the angle δ between them and the angle v_{asy} from the periapsis direction. $\vec{w}_{\text{in}}$ and $\vec{w}_{\text{out}}$ are the velocities along the asymptotes. Right, $\vec{w}_{\text{in}}$ and $\vec{w}_{\text{out}}$ added to Jupiter's heliocentric velocity $\vec{V}_p$.

Solution. In the solar frame, we drop the Jupiter–Sun terms that do not involve $\vec{x}_{\text{sc}}$, and write the Lagrangian of the spacecraft,

$$\mathcal{L} = \frac{1}{2} m_{\text{sc}} \dot{\vec{x}}_{\text{sc}}^2 + \frac{G M_\odot m_{\text{sc}}}{|\vec{x}_{\text{sc}}|} + \frac{G m_{\text{sc}} M_J}{|\vec{x}_{\text{sc}} - \vec{X}_J(t)|}. \tag{6.252}$$

Let $\vec{r} \equiv \vec{x}_{\text{sc}} - \vec{X}_J(t)$. Over the brief encounter, $\vec{X}_J(t) \approx \vec{X}_{J,0} + \vec{V}_p t$ and $|\vec{x}_{\text{sc}}| \approx X_J(t)$, dropping the small term $O(r/X_J)$. Also $\dot{\vec{x}}_{\text{sc}} = \vec{V}_p + \dot{\vec{r}}$, so

$$\mathcal{L} \approx \underbrace{\tfrac{1}{2} m_{\text{sc}} V_p^2 + \frac{G M_\odot m_{\text{sc}}}{X_J(t)}}_{g(t)} + \underbrace{m_{\text{sc}} \vec{V}_p \cdot \dot{\vec{r}}}_{\frac{d}{dt}(m_{\text{sc}} \vec{V}_p \cdot \vec{r})} + \frac{1}{2} m_{\text{sc}} \dot{\vec{r}}^2 + \frac{G m_{\text{sc}} M_J}{r} \tag{6.253}$$

$$\Longrightarrow \quad \mathcal{L}(\vec{r}, \dot{\vec{r}}) = \frac{1}{2} m_{\text{sc}} \dot{\vec{r}}^2 + \frac{G m_{\text{sc}} M_J}{r}, \tag{6.254}$$

since a function of time alone and a total time derivative do not change the equations of motion. This $\mathcal{L}$ has no explicit t, so by Noether's theorem the energy is conserved,

$$E = \dot{\vec{r}} \cdot \frac{\partial \mathcal{L}}{\partial \dot{\vec{r}}} - \mathcal{L} = \frac{1}{2} m_{\text{sc}} \dot{\vec{r}}^2 - \frac{G m_{\text{sc}} M_J}{r} = \text{const} \tag{6.255}$$

$$\Longrightarrow \quad v_{\infty,\text{in}} = v_{\infty,\text{out}} \equiv v_\infty = \sqrt{-G M_J / a}, \tag{6.256}$$

the Jupiter-relative speed at $r \to \infty$, Eq. (6.61). The solar-frame velocity is $\vec{V}_p + \vec{w}$, where $\vec{w} \equiv \dot{\vec{r}}$ far from Jupiter. Its size $|\vec{w}| = v_\infty$ is fixed, but its direction is free (Figure 6.14).

Put the x-axis along the periapsis direction. The orbit is $r(v) = \lambda/(1 + e\cos v)$, and $r \to \infty$ at $v = \pm v_{\text{asy}}$, where

$$\cos v_{\text{asy}} = -\frac{1}{e}, \tag{6.257}$$

as in Section 6.11. So $\hat{w}_{\text{out}}$ sits at the angle v_{asy} from the axis. At periapsis $\dot{r} = 0$, so the velocity is purely tangential, perpendicular to the axis. The trajectory is a mirror image about the periapsis, so the total turn δ splits evenly on each side. This puts $\hat{w}_{\text{out}}$ at $\delta/2$ past that perpendicular.

$$v_{\text{asy}} = \frac{\pi}{2} + \frac{\delta}{2} \quad \Longrightarrow \quad \sin\frac{\delta}{2} = \sin\Big(v_{\text{asy}} - \frac{\pi}{2}\Big) = -\cos v_{\text{asy}} = \frac{1}{e}, \tag{6.258}$$

the deflection angle $\delta = \pi - 2\theta_\infty$ of Eq. (6.62). For $e = 1.2$ this gives $\delta \simeq 112.9^\circ$. As $e \to 1^+$, $\delta \to \pi$, and as $e \to \infty$, $\delta \to 0$. By the same symmetry, $\hat{w}_{\text{in}}$ sits at $v_{\text{asy}} - \delta = \pi/2 - \delta/2$ from the axis.

Take $\vec{V}_p = -V_p\hat{x}$, so Jupiter moves parallel to the periapsis direction, to the left. $\hat{w}_{\text{in}}$ sits at $\pi - v_{\text{asy}}$ and $\hat{w}_{\text{out}}$ at v_{asy} from the axis, so their angles with $\vec{V}_p$ are v_{asy} and $\pi - v_{\text{asy}}$. By the law of cosines,

$$v_{\text{in}}^2 = V_p^2 + v_\infty^2 + 2V_p v_\infty \cos v_{\text{asy}} = V_p^2 + v_\infty^2 - \frac{2V_p v_\infty}{e} \tag{6.259}$$

$$= 13.1^2 + 10^2 - \frac{2(13.1)(10)}{1.2} \implies v_{\text{in}} \simeq 7.3\,\text{km}\,\text{s}^{-1}, \tag{6.260}$$

$$v_{\text{out}}^2 = V_p^2 + v_\infty^2 + 2V_p v_\infty \cos\big(\pi - v_{\text{asy}}\big) = V_p^2 + v_\infty^2 + \frac{2V_p v_\infty}{e} \tag{6.261}$$

$$= 13.1^2 + 10^2 + \frac{2(13.1)(10)}{1.2} \implies v_{\text{out}} \simeq 22.1\,\text{km}\,\text{s}^{-1}. \tag{6.262}$$

The spacecraft's speed around the Sun grows by $\boxed{v_{\text{out}} - v_{\text{in}} \simeq 14.8\,\text{km}\,\text{s}^{-1}}$.

By momentum conservation, $m_{\text{sc}}\Delta\vec{v}_{\text{sc}} = -M_J\Delta\vec{V}_p$, so for a spacecraft of about a tonne

$$\Delta V_p = \frac{m_{\text{sc}}}{M_J}\,\Delta v_{\text{sc}} < 10^{-20}\,\text{km}\,\text{s}^{-1}, \tag{6.263}$$

undetectable on Jupiter.

Chapter 7

Signal, Noise, and Detection in CCD Photometry

Photometry describes a star's light as a flux or magnitude; a telescope forms that light into an image. A CCD turns the image into counts: photons free electrons in a grid of silicon pixels, each pixel stores its electrons during the exposure, and the detector reads them out at the end. We must work with these counts to decide how accurately the star's brightness was measured.

A count of electrons is a whole number, and it changes from one exposure to the next even when the star does not. How much it changes decides whether a faint star is seen at all, how well its magnitude is known, and whether one long exposure is better than several short ones. All three answers come from one equation for the signal-to-noise ratio, built from the statistics of counting.

7.1 Counting Photoelectrons

Chapter 4 gave the light of a star as a flux F in $\mathrm{W\,m^{-2}}$, and Chapter 5 gave the size of its image on the detector. A CCD does not report watts. It reports a number of electrons in each pixel, and part of that number did not come from the star at all.

Picture a grid of buckets on a lawn, with a garden sprinkler running over one spot during a light drizzle. After ten minutes the buckets near the sprinkler hold most of its water. Every bucket also holds the same depth of drizzle, and a little water that has condensed on its inside. To read a bucket we pour it into a measuring jug and read the scale, which adds a small error of its own. To find what the sprinkler gave, we add up the buckets around it and take away what the drizzle and the condensation would have put there. The pixels are the buckets. The star is the sprinkler, the faint glow of the night sky is the drizzle, electrons freed by heat in the silicon are the condensation, and the electronics that measure each pixel's charge are the jug. The signal is the number of electrons that the star alone put into the pixels around its image.

A star freeing 50 electrons per second contributes $50 \times 60 = 3000$ electrons in a one-minute exposure. If its image covers 20 pixels and the sky contributes 2 electrons per second to each, the sky adds $2 \times 20 \times 60 = 2400$ electrons in the same pixels. Subtracting that background leaves the star's mean signal. We now name each contribution so that we can calculate the scatter that remains after subtraction.

Photons become electrons. A photon that enters a pixel frees one electron with a probability $\eta(\lambda)$, the **quantum efficiency** of the detector. For modern astronomical CCDs η is about 0.5 to 0.85. So

$$\text{number of electrons} = \eta \times \text{number of photons that reach the pixel},$$

and 1000 photons give about 800 electrons when $\eta = 0.8$. The factor η multiplies every rate that comes from light, so we fold it into those rates and count only electrons from here on. Every rate below is a rate of detected electrons.

The rates. The light of the star is spread over several pixels by the image blur. We therefore take R_{src} to be the rate at which the star frees electrons in all the pixels of the measuring aperture together. The sky is the same in every pixel, so R_{sky} is a rate per pixel. Heat frees **dark electrons** at the rate R_{dark} in every pixel, light or no light, and they look exactly like electrons freed by photons. The aperture contains n_{pix} pixels. Detector manuals often quote the dark rate per hour, and dividing by 3600 turns it into a rate per second. All the symbols are collected in Table 7.1.

Symbol	Meaning	Units
R_{src}	Rate of electrons from the star, summed over the aperture	$\mathrm{e^-\,s^{-1}}$
R_{sky}	Rate of electrons from the sky, per pixel	$\mathrm{e^-\,s^{-1}\,pix^{-1}}$
R_{dark}	Rate of dark electrons freed by heat, per pixel	$\mathrm{e^-\,s^{-1}\,pix^{-1}}$
n_{pix}	Number of pixels in the aperture, proportional to $\theta^2_{\mathrm{seeing}}$	pixels
t_{exp}	Length of one exposure	s
N_{exp}	Number of exposures added together	—
t	Total time, $t = N_{\mathrm{exp}} t_{\mathrm{exp}}$	s
S	Signal, the electrons from the star in the total time, $S = R_{\mathrm{src}} t$	$\mathrm{e^-}$
$\bar{n}$	Mean of a count of electrons	$\mathrm{e^-}$
σ_{RON}	Readout noise of one pixel in one reading (standard deviation)	$\mathrm{e^-}$
σ_x	Standard deviation added by the process x	$\mathrm{e^-}$
SNR	Signal-to-noise ratio, S/σ_{tot}	—
m, δm	Apparent magnitude and its error	mag

Table 7.1: Symbols, meanings and units for the CCD noise budget.

The signal. Over a total time t the star gives

$$S = R_{\mathrm{src}}\, t \tag{7.1}$$

electrons. In practice t is often split into N_{exp} exposures of length t_{exp} each, which are added afterwards,

$$t = N_{\mathrm{exp}}\, t_{\mathrm{exp}}, \qquad S = R_{\mathrm{src}}\, N_{\mathrm{exp}}\, t_{\mathrm{exp}}. \tag{7.2}$$

The difference between t and N_{exp} matters. Electrons pile up in proportion to the total time only. The electronics add their error once at every reading, so that part grows with the number of exposures.

7.2 Shot Noise and the Poisson Distribution

We can now write the signal as a mean number of electrons, $S = R_{\mathrm{src}} t$.

A real exposure does not give the mean. It gives one count, and a second exposure of the same steady star gives a different count. To say how far a count can stray from its mean we need the probability of every possible count.

Stand by a quiet road at night and count the cars that pass in each minute. The traffic is steady, about ten cars a minute on average. Yet one minute brings seven and the next thirteen. Nothing

about the road changed. The cars simply come one at a time, each at its own moment, without arranging themselves to arrive evenly.

Photons from a steady star arrive in the same way, one at a time and independently. So the number of electrons in a fixed exposure scatters about its mean. The scatter is called **shot noise**. It comes from counting, not from any fault of the detector, and its size is the square root of the mean count.

In numbers, a mean of 10 cars a minute gives a scatter of about $\sqrt{10} \approx 3.2$, so most minutes bring between 7 and 13 cars. A mean of 100 gives a scatter of 10, which is 10 per cent of the mean. A mean of 10 000 gives 100, only 1 per cent. The scatter grows with the mean, but more slowly, so it becomes a smaller fraction of it.

Repeated counts form a distribution around the mean. As the mean grows, the absolute spread grows as its square root, but the spread becomes a smaller fraction of the signal. We will derive this Poisson distribution from independent arrivals and then show why it approaches a Gaussian shape for large counts.

Three assumptions. The derivation rests on two physical facts about the arrivals and one mild mathematical condition.

- **Independence.** One arrival says nothing about when the next one comes. There is no memory and no bunching.
- **Constant rate.** The source does not change its brightness during the exposure, so the chance of an arrival is the same in every interval of the same length.
- **Smoothness.** The probability $f(\Delta t)$ of an arrival in an interval of length Δt is a smooth function near $\Delta t = 0$, so it has a Taylor series there.

One short interval. An interval of zero length holds nothing, so $f(0) = 0$. For a short interval,

$$\begin{aligned} f(\Delta t) &\overset{(1)}{=} f(0) + f'(0)\,\Delta t + O(\Delta t^2) \\ &\overset{(2)}{=} f'(0)\,\Delta t + O(\Delta t^2). \end{aligned} \tag{7.3}$$

(1) Taylor series about $\Delta t = 0$, allowed by the smoothness condition.
(2) $f(0) = 0$.

So for any short interval the chance of an arrival is proportional to its length. This is true of any probability that starts from zero, not just of light. Two arrivals in the same interval need two independent rare events at once, so their probability is of order Δt^2.

Many short intervals. Divide the exposure t into N equal intervals of length $\Delta t = t/N$. In each one, an arrival happens with probability

$$p = f'(0)\,\frac{t}{N} \equiv \frac{\bar{n}}{N}, \qquad \bar{n} \equiv f'(0)\,t,$$

and two or more arrivals happen with a probability of order $1/N^2$, which we drop as $N \to \infty$. Each interval then holds either no arrival or one. The intervals are independent, so the probability of a particular pattern with arrivals in n chosen intervals is $p^n(1-p)^{N-n}$, and there are $\binom{N}{n}$ ways to choose the n intervals. The count n therefore follows the binomial distribution.

Letting $N \to \infty$ with $\bar{n}$ fixed,

$$\begin{aligned} P(n) &\overset{(1)}{=} \lim_{N\to\infty} \frac{N!}{n!\,(N-n)!} \left(\frac{\bar{n}}{N}\right)^n \left(1-\frac{\bar{n}}{N}\right)^{N-n} \\ &\overset{(2)}{=} \frac{\bar{n}^{\,n}}{n!} \lim_{N\to\infty} \left[\frac{N!}{(N-n)!\,N^n}\right] \left(1-\frac{\bar{n}}{N}\right)^{N} \left(1-\frac{\bar{n}}{N}\right)^{-n} \\ &\overset{(3)}{=} \frac{\bar{n}^{\,n}}{n!} \cdot 1 \cdot e^{-\bar{n}} \cdot 1. \end{aligned} \tag{7.4}$$

(1) The binomial distribution with $p = \bar{n}/N$.
(2) Regroup, taking out the factors that do not depend on N.
(3) $N!/[(N-n)!\,N^n] = \prod_{j=0}^{n-1}(1-j/N)$ is a product of a fixed number n of factors that each tend to 1. The logarithm of the middle factor is $N\ln(1-\bar{n}/N) = N(-\bar{n}/N - \bar{n}^2/2N^2 - \dots) \to -\bar{n}$. The last factor has a fixed power n and tends to 1.

This is the **Poisson distribution**,

$$\boxed{P(n) = \frac{\bar{n}^{\,n}\, e^{-\bar{n}}}{n!}.} \tag{7.5}$$

It depends on the rate and the time only through $\bar{n}$. Its probabilities add up to one (Problem 7.7.1).

Mean. The mean count is

$$\begin{aligned} \langle n \rangle &\overset{(1)}{=} \sum_{n=0}^{\infty} n\, \frac{\bar{n}^{\,n} e^{-\bar{n}}}{n!} \\ &\overset{(2)}{=} e^{-\bar{n}} \sum_{n=1}^{\infty} \frac{\bar{n}^{\,n}}{(n-1)!} \\ &\overset{(3)}{=} \bar{n}\, e^{-\bar{n}} \sum_{j=0}^{\infty} \frac{\bar{n}^{\,j}}{j!} \\ &\overset{(4)}{=} \bar{n}. \end{aligned} \tag{7.6}$$

(1) The definition of the mean, with Eq. (7.5).
(2) The $n = 0$ term is zero, and $n/n! = 1/(n-1)!$.
(3) Put $j = n - 1$ and take one factor $\bar{n}$ out of the sum.
(4) The sum is the series of $e^{\bar{n}}$.

So $\bar{n}$ is the mean count, and $f'(0) = \bar{n}/t$ is the mean number of arrivals per second. For the star it is $R_{\rm src}$, and for the sky and the dark current it is $n_{\rm pix}R_{\rm sky}$ and $n_{\rm pix}R_{\rm dark}$.

Variance. The variance is $\sigma^2 \equiv \big\langle (n - \langle n \rangle)^2 \big\rangle$. The factor $n(n-1)$ cancels cleanly against $n!$, so we first find the average of $n(n-1)$,

$$\begin{aligned} \langle n(n-1) \rangle &\overset{(1)}{=} e^{-\bar{n}} \sum_{n=2}^{\infty} \frac{\bar{n}^{\,n}}{(n-2)!} \\ &\overset{(2)}{=} \bar{n}^2\, e^{-\bar{n}} \sum_{j=0}^{\infty} \frac{\bar{n}^{\,j}}{j!} = \bar{n}^2. \end{aligned} \tag{7.7}$$

(1) The terms $n = 0$ and $n = 1$ are zero, and $n(n-1)/n! = 1/(n-2)!$.
(2) Put $j = n - 2$, take $\bar{n}^2$ out, and sum the series of $e^{\bar{n}}$.

Then

$$\begin{aligned}\sigma^2 &\overset{(1)}{=} \langle n^2 \rangle - \langle n \rangle^2 \\ &\overset{(2)}{=} \langle n(n-1) \rangle + \langle n \rangle - \langle n \rangle^2 \\ &\overset{(3)}{=} \bar{n}^2 + \bar{n} - \bar{n}^2 = \bar{n}.\end{aligned} \tag{7.8}$$

(1) Expand the square. $\langle n \rangle$ is a constant, so $\langle 2n \langle n \rangle \rangle = 2 \langle n \rangle^2$.
(2) $n^2 = n(n-1) + n$.
(3) The two averages just found, Eq. (7.6) and the one above.

The shot noise, the standard deviation of the count, is therefore

$$\boxed{\sigma = \sqrt{\bar{n}}.} \tag{7.9}$$

The noise grows as the square root of the mean, so the fractional noise $\sigma/\bar{n} = 1/\sqrt{\bar{n}}$ falls as the count grows. A computer algebra check of the three sums gives 1, $\bar{n}$ and $\bar{n}$ exactly. Figure 7.1 shows the distribution for three means.

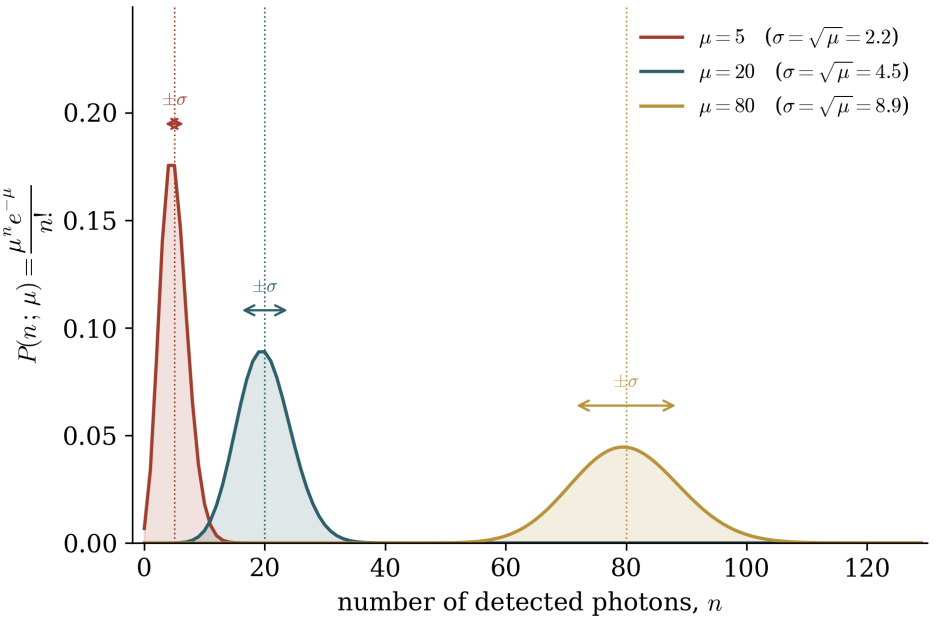


Figure 7.1: The Poisson distribution for mean counts of 5, 20 and 80 (labelled μ in the plot). A dotted line marks each mean, and a double arrow spans one standard deviation on each side of it.

Why the variance grows as the count. The same variance follows from a second route, which shows where the square root comes from. We need one general fact first. Let a quantity $X = \sum_i X_i$ be a sum of independent parts, and write $\delta X_i = X_i - \langle X_i \rangle$ for the deviation of each

part from its mean. Then

$$\begin{aligned}
\sigma_X^2 &\overset{(1)}{=} \left\langle \Big(\sum_i \delta X_i \Big)^2 \right\rangle \\
&\overset{(2)}{=} \sum_i \sum_j \left\langle \delta X_i \, \delta X_j \right\rangle \\
&\overset{(3)}{=} \sum_i \left\langle \delta X_i^2 \right\rangle + \sum_{i \neq j} \left\langle \delta X_i \right\rangle \left\langle \delta X_j \right\rangle \\
&\overset{(4)}{=} \sum_i \sigma_i^2 .
\end{aligned} \tag{7.10}$$

(1) $X - \langle X \rangle = \sum_i \delta X_i$, because the mean of a sum is the sum of the means.
(2) Expand the square of the sum into a double sum.
(3) Split off the terms with $i = j$. For $i \neq j$ the parts are independent, so the average of the product is the product of the averages.
(4) $\langle \delta X_i \rangle = 0$ by its definition, and $\left\langle \delta X_i^2 \right\rangle = \sigma_i^2$.

The variances of independent parts add, whatever the distribution of each part. Standard deviations therefore add **in quadrature**, $\sigma_X = \sqrt{\sum_i \sigma_i^2}$. The cross terms die for the same reason as in the random walk, where they gave $R_{\mathrm{rms}} = \sqrt{N}\, s$ (Eq. (9.41)).

Now let X_i be the number of arrivals in the i-th short interval. It is 1 with probability p and 0 otherwise, so $\langle X_i \rangle = p$ and $\left\langle X_i^2 \right\rangle = p$, which gives $\sigma_i^2 = p - p^2$. The count is $n = \sum_i X_i$, and

$$\begin{aligned}
\sigma^2 &\overset{(1)}{=} \sum_{i=1}^{N} p\,(1-p) \\
&\overset{(2)}{=} \bar{n} \left(1 - \frac{\bar{n}}{N} \right) \\
&\to \bar{n} \quad (N \to \infty).
\end{aligned} \tag{7.11}$$

(1) Eq. (7.10), with the same variance $p(1-p)$ for every interval.
(2) $Np = \bar{n}$.

This agrees with Eq. (7.8). Each interval adds the same small variance, so the variance grows in proportion to the number of intervals, and so in proportion to the mean count. The noise grows only as its square root.

Large counts look Gaussian. For a large mean the Poisson hump becomes the Gaussian bell curve. To see this we write $n = \bar{n} + x$ and use Stirling's formula, $\ln n! \approx n \ln n - n + \frac{1}{2} \ln(2\pi n)$, for large n.

$$\begin{aligned}
\ln P(n) &\overset{(1)}{=} n \ln \bar{n} - \bar{n} - n \ln n + n - \tfrac{1}{2} \ln(2\pi n) \\
&\overset{(2)}{=} -(\bar{n} + x) \ln\!\left(1 + \frac{x}{\bar{n}} \right) + x - \tfrac{1}{2} \ln(2\pi n) \\
&\overset{(3)}{=} -(\bar{n} + x) \left(\frac{x}{\bar{n}} - \frac{x^2}{2\bar{n}^2} + \ldots \right) + x - \tfrac{1}{2} \ln(2\pi \bar{n}) + \ldots \\
&\overset{(4)}{=} -\frac{x^2}{2\bar{n}} - \tfrac{1}{2} \ln(2\pi \bar{n}) + O\!\left(\frac{x^3}{\bar{n}^2}, \frac{x}{\bar{n}} \right).
\end{aligned} \tag{7.12}$$

(1) The logarithm of Eq. (7.5), with Stirling's formula for $\ln n!$.

(2) $n \ln \bar{n} - n \ln n = -n \ln(n/\bar{n})$ with $n = \bar{n} + x$, and $-\bar{n} + n = x$.
(3) Taylor series $\ln(1 + y) = y - y^2/2 + \ldots$, and $\ln(2\pi n) = \ln(2\pi\bar{n}) + \ln(1 + x/\bar{n})$.
(4) Multiply out. The terms in x cancel, and $x^2/2\bar{n} - x^2/\bar{n} = -x^2/2\bar{n}$.

Dropping the small terms,

$$\boxed{P(n) \approx \frac{1}{\sqrt{2\pi\bar{n}}} \exp\left[-\frac{(n-\bar{n})^2}{2\bar{n}}\right].} \tag{7.13}$$

This is a Gaussian with mean $\bar{n}$ and variance $\bar{n}$, as it must be. The dropped term $x^3/\bar{n}^2$ is $1/\sqrt{\bar{n}}$ at one standard deviation, $x = \sqrt{\bar{n}}$. So the Gaussian is good when $\bar{n} \gg 1$ and within a few standard deviations of the mean. The Poisson distribution is defined only at whole numbers and the Gaussian is smooth, but when $\sqrt{\bar{n}} \gg 1$ the steps of one electron are tiny compared with the width, and the difference cannot be seen (Figure 7.8 in Problem 7.7.1). A computer algebra expansion of $\ln P(n)$ to second order in x gives the same leading terms.

7.3 The Noise Budget and the Signal-to-Noise Equation

We now know that a count with mean $\bar{n}$ has a standard deviation $\sqrt{\bar{n}}$. The number we read from a CCD is not one clean count. It is the star's electrons on top of sky and dark electrons, measured by electronics that add an error of their own. We need the noise of the star's share alone, after the rest has been taken away.

Weigh yourself on a bathroom scale while holding a cat, then weigh yourself alone, and subtract. Each reading of the scale wanders by about 0.2 kg. Your own weight cancels in the difference, but the wandering of the second reading does not. The cat's weight carries both. In an image, the star's electrons sit on top of the sky and dark electrons. We subtract the mean sky and dark level, but their scatter stays in the result, and the readout adds more. The idea is that the variances of independent sources of noise add.

For the cat, two independent 0.2 kg errors give $\sqrt{0.2^2 + 0.2^2} = 0.28$ kg of uncertainty in the difference. They do not add directly because either reading can wander up or down. The same variance rule, Eq. (7.10), combines the star's shot noise with the sky, dark current and readout noise. Once we add those variances, we can calculate the signal-to-noise ratio of an exposure.

Signal-to-noise ratio. Let σ_{tot} be the standard deviation of the measured signal S, after the mean sky and dark levels have been subtracted. The measurement is reported as $S \pm \sigma_{\text{tot}}$, and its **signal-to-noise ratio** is

$$\text{SNR} \equiv \frac{S}{\sigma_{\text{tot}}}. \tag{7.14}$$

Its inverse $1/\text{SNR} = \sigma_{\text{tot}}/S$ is the fractional error of the flux. Our goal is σ_{tot} in terms of the rates and times of Table 7.1.

Shot noise of the star. The electrons from the star in time t have the mean $S = R_{\text{src}}t$ (Eq. (7.1)). The rate R_{src} is already summed over the aperture, so no factor n_{pix} appears. Eq. (7.9) gives

$$\sigma_{\text{src}} = \sqrt{S} = \sqrt{R_{\text{src}}\, t}. \tag{7.15}$$

Shot noise of the sky. The sky gives R_{sky} per pixel, so the aperture collects a mean of $n_{\text{pix}}R_{\text{sky}}t$ sky electrons. We assume that the mean sky level is measured over a large empty area nearby,

so that its own error is small. Subtracting it removes the mean, but the Poisson scatter of the sky electrons inside the aperture remains,

$$\sigma_{\text{sky}} = \sqrt{n_{\text{pix}}\, R_{\text{sky}}\, t}. \tag{7.16}$$

The aperture must hold the blurred image, so its area grows as the square of the seeing, $n_{\text{pix}} \propto \theta^2_{\text{seeing}}$ (Eq. (5.56)). Good seeing means fewer pixels and less sky noise.

Shot noise of the dark current. Heat in the silicon frees electrons at random moments, independently and at a steady rate, just as photons arrive. So the dark electrons are also Poisson counts, with mean $n_{\text{pix}} R_{\text{dark}} t$ in the aperture,

$$\sigma_{\text{dark}} = \sqrt{n_{\text{pix}}\, R_{\text{dark}}\, t}. \tag{7.17}$$

Cooling the detector lowers R_{dark}, which is why astronomical CCDs are kept cold.

Readout noise. The readout noise is not a count of anything. Dark current adds extra charge to a pixel before it is measured. Readout noise is an error in the measurement itself, after the charge is fixed. The charge of each pixel is turned into a voltage by an output transistor and then into a number by an analogue-to-digital converter. Both have their own electrical noise, the thermal (Johnson–Nyquist) noise of any resistor and a slowly wandering "$1/f$" noise. Each reading of a pixel therefore gets a random error with a Gaussian distribution and a standard deviation σ_{RON}, whatever charge the pixel holds and however long the exposure was. Every reading is independent, and the aperture is read n_{pix} times in each of N_{exp} exposures, so

$$\begin{aligned} \sigma^2_{\text{read}} &\overset{(1)}{=} \sum_{\text{readings}} \sigma^2_{\text{RON}} \\ &= n_{\text{pix}}\, N_{\text{exp}}\, \sigma^2_{\text{RON}}. \end{aligned}$$

(1) Eq. (7.10), with the same variance for every reading.

$$\boxed{\sigma_{\text{read}} = \sqrt{n_{\text{pix}}\, N_{\text{exp}}}\; \sigma_{\text{RON}}.} \tag{7.18}$$

The readout variance grows with N_{exp}, not with t. Splitting a total time into N_{exp} exposures pays the readout variance N_{exp} times.

Adding the four. The star, the sky, the heat in the silicon and the readout electronics are independent of each other, so Eq. (7.10) adds their variances,

$$\begin{aligned} \sigma^2_{\text{tot}} &\overset{(1)}{=} \sigma^2_{\text{src}} + \sigma^2_{\text{sky}} + \sigma^2_{\text{dark}} + \sigma^2_{\text{read}} \\ &\overset{(2)}{=} R_{\text{src}}\, t + n_{\text{pix}} R_{\text{sky}}\, t + n_{\text{pix}} R_{\text{dark}}\, t + n_{\text{pix}} N_{\text{exp}}\, \sigma^2_{\text{RON}}. \end{aligned} \tag{7.19}$$

(1) Eq. (7.10) for four independent parts.
(2) Eqs. (7.15)–(7.18).

Putting this and $S = R_{\text{src}} t$ into Eq. (7.14) gives the **CCD signal-to-noise equation**,

Result 7.1

$$\mathrm{SNR} = \frac{R_{\mathrm{src}}\, t}{\sqrt{R_{\mathrm{src}}\, t \;+\; n_{\mathrm{pix}} R_{\mathrm{sky}}\, t \;+\; n_{\mathrm{pix}} R_{\mathrm{dark}}\, t \;+\; n_{\mathrm{pix}} N_{\mathrm{exp}}\, \sigma^2_{\mathrm{RON}}}}, \qquad t = N_{\mathrm{exp}} t_{\mathrm{exp}}. \tag{7.20}$$

The signal and the first three variances grow in proportion to t. The last variance grows with N_{exp}. So the signal-to-noise ratio depends on how the time is split as well as on the total time. Figure 7.2 shows the share of each variance in three typical observations.

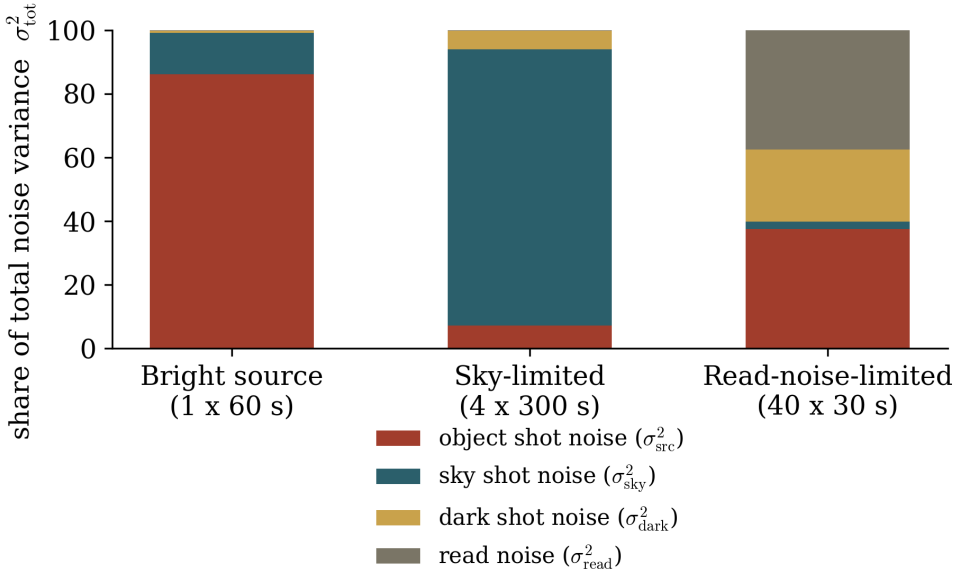


Figure 7.2: Share of each term in the total variance σ^2_{tot}, in per cent, for three observations. A bright source in one 60 s exposure, a sky-limited source in four 300 s exposures, and a faint source in forty 30 s exposures.

Dark current and read noise from two dark frames. The detector's own two terms can be measured with the shutter closed. Such a **dark frame** has no star and no sky. Its variance over an area of n_{pix} pixels, for an exposure of length t, is the last two terms of Eq. (7.19) with $N_{\mathrm{exp}} = 1$,

$$\sigma^2(t) = n_{\mathrm{pix}} R_{\mathrm{dark}}\, t + n_{\mathrm{pix}} \sigma^2_{\mathrm{RON}}.$$

This is a straight line in t. Its slope is the dark variance per unit time and its intercept is the read variance. Two unknowns need two measurements, and two dark frames of lengths t_1 and t_2 with variances σ_1^2 and σ_2^2 give

$$\begin{pmatrix} t_1 & 1 \\ t_2 & 1 \end{pmatrix} \begin{pmatrix} n_{\mathrm{pix}} R_{\mathrm{dark}} \\ n_{\mathrm{pix}} \sigma^2_{\mathrm{RON}} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \end{pmatrix}.$$

The determinant of the matrix is $t_1 - t_2$. Two frames of the same length give the same equation twice and cannot separate the two terms. For $t_1 \neq t_2$,

$$\begin{aligned} \begin{pmatrix} n_{\mathrm{pix}} R_{\mathrm{dark}} \\ n_{\mathrm{pix}} \sigma^2_{\mathrm{RON}} \end{pmatrix} &\overset{(1)}{=} \frac{1}{t_1 - t_2} \begin{pmatrix} 1 & -1 \\ -t_2 & t_1 \end{pmatrix} \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \end{pmatrix} \\ &= \frac{1}{t_2 - t_1} \begin{pmatrix} \sigma_2^2 - \sigma_1^2 \\ t_2 \sigma_1^2 - t_1 \sigma_2^2 \end{pmatrix}. \end{aligned} \tag{7.21}$$

(1) The inverse of a 2×2 matrix, $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \dfrac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

The dark variance per unit time is the rise in variance divided by the rise in time, the slope of the line in Figure 7.3. The read variance is where the line meets $t = 0$. A third frame of another length tests whether the points really lie on a straight line. Problem 7.7.4 puts in numbers.

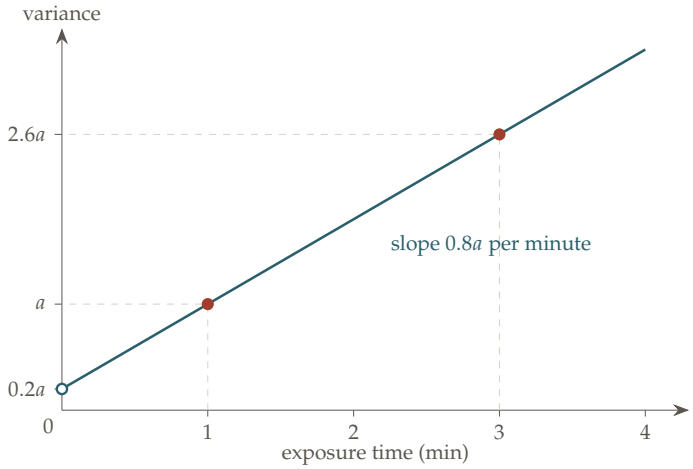


Figure 7.3: Variance of a dark frame against exposure time. The filled points are frames of 1 and 3 minutes with variances a and $2.6a$. The straight line through them meets the vertical axis at $0.2a$ (open circle).

Defects that are not noise. Some flaws of a CCD frame are not random scatter and must be found and removed before Eq. (7.19) means anything.

- **Hot and cold pixels.** Single pixels with a far higher dark rate (hot) or a far lower quantum efficiency (cold). They are found in calibration frames and flagged or filled in from their neighbours.
- **Saturation and bleeding.** A pixel holds only a limited charge, its **full well** (Section 7.4.1). Near it the response stops being linear, and beyond it extra charge spills into neighbouring pixels along a column, leaving a bright *bleed trail*.
- **Cosmic rays.** Fast charged particles leave a sharp spike of charge in one frame only. Comparing several exposures (Section 7.4.2) or the shape of the spike, which is sharper than any star, removes them.
- **Airglow and city lights.** Light emitted by the upper atmosphere and light scattered from towns raise R_{sky} itself, and with it σ_{sky} in Eq. (7.16), above the level of a dark site on a moonless night.

Dithering, moving the telescope by a few arcseconds between exposures, deals with several of these at once. A given point of the sky then falls on a different pixel in each exposure. A bad pixel or a bleed trail stays fixed on the detector and a cosmic ray appears in one frame only, so when the frames are lined up on the sky and combined they can be rejected.

7.4 Three Regimes of the Signal-to-Noise Equation

Eq. (7.20) gives the signal-to-noise ratio of any exposure. As it stands it is hard to plan with. Should we take one long exposure or many short ones? Does good seeing matter? Does a

larger telescope help a faint star more than a bright one? The answer depends on which of the four terms under the square root is largest.

Think of a taxi fare. There is a fixed charge for each ride and a charge for each minute. On one long ride the minutes dominate and the fixed charge hardly matters. Split the same trip into many short rides and the fixed charges pile up. Whether splitting matters depends on which part of the fare is larger. In the signal-to-noise equation the star, the sky and the dark current charge their variance per second, and the readout charges its variance per exposure. When one term is much larger than the others, the equation simplifies and tells us directly what to do. Which term wins is decided by comparing two numbers, principle P6 (Section 9.6).

For a CCD with $\sigma_{\mathrm{RON}} = 5\,\mathrm{e}^-$, a 20-pixel aperture and sky rate $2\,\mathrm{e}^-\,\mathrm{s}^{-1}$ per pixel, each read adds variance $20 \times 25 = 500\,\mathrm{e}^{-2}$. The sky adds $40\,\mathrm{e}^{-2}$ each second. In a 300 s exposure its variance reaches $12\,000\,\mathrm{e}^{-2}$, making one read relatively cheap. In a 10 s exposure it reaches only $400\,\mathrm{e}^{-2}$, so readout noise dominates. This comparison tells us when splitting an observation into many exposures costs sensitivity. We will make it a general exposure-time condition, then examine the other noise limits.

How long one exposure should be. For a faint star, R_{src} is small next to the sky and dark rates. Then the variance in Eq. (7.19) is

$$\begin{aligned}\sigma_{\mathrm{tot}}^2 &\overset{(1)}{\approx} n_{\mathrm{pix}}(R_{\mathrm{sky}} + R_{\mathrm{dark}})\,t + n_{\mathrm{pix}} N_{\mathrm{exp}}\,\sigma_{\mathrm{RON}}^2 \\ &\overset{(2)}{=} n_{\mathrm{pix}} N_{\mathrm{exp}} \left[(R_{\mathrm{sky}} + R_{\mathrm{dark}})\,t_{\mathrm{exp}} + \sigma_{\mathrm{RON}}^2\right] \\ &= n_{\mathrm{pix}}(R_{\mathrm{sky}} + R_{\mathrm{dark}})\,t \left[1 + \frac{\sigma_{\mathrm{RON}}^2}{(R_{\mathrm{sky}} + R_{\mathrm{dark}})\,t_{\mathrm{exp}}}\right]. \end{aligned} \tag{7.22}$$

(1) Drop the star's own shot noise, $R_{\mathrm{src}}t$.
(2) $t = N_{\mathrm{exp}} t_{\mathrm{exp}}$.

The last factor compares the read variance of one pixel with the sky and dark variance that the pixel collects in one exposure. Splitting the time costs little when that ratio is small,

$$\boxed{t_{\mathrm{exp}} \gg \frac{\sigma_{\mathrm{RON}}^2}{R_{\mathrm{sky}} + R_{\mathrm{dark}}}.} \tag{7.23}$$

For the numbers above the right side is $25/2 = 12.5$ s. Exposures of 300 s lower the signal-to-noise ratio by a factor $(1 + 12.5/300)^{-1/2} = 0.98$, only 2 per cent. The bound depends only on the detector and the sky, not on the star.

7.4.1 Regime 1: Bright Source, Limited by Its Own Shot Noise

For a bright star the first term under the root is much larger than the others, $R_{\mathrm{src}}t \gg n_{\mathrm{pix}}(R_{\mathrm{sky}} + R_{\mathrm{dark}})t + n_{\mathrm{pix}} N_{\mathrm{exp}} \sigma_{\mathrm{RON}}^2$. Then

$$\begin{aligned}\mathrm{SNR} &\overset{(1)}{\approx} \frac{R_{\mathrm{src}}\,t}{\sqrt{R_{\mathrm{src}}\,t}} \\ &= \sqrt{R_{\mathrm{src}}\,t},\end{aligned}$$

(1) Keep only the star's own term in Eq. (7.20).

$$\boxed{\mathrm{SNR} \approx \sqrt{R_{\mathrm{src}}\,t} \propto \sqrt{t}.} \tag{7.24}$$

The signal-to-noise ratio is the square root of the number of electrons from the star. It does not depend on the sky, the detector or the way the time is split. Reaching a given ratio takes $t = \mathrm{SNR}^2 / R_{\mathrm{src}}$, so ten times the precision costs a hundred times the time.

Here the difficulty is not to collect enough light but to avoid too much. A pixel saturates at its full well, typically about 10^5 electrons, and its response stops being linear before that. A bright star is therefore observed with shorter exposures, which cost little here, or with the telescope slightly out of focus, which spreads the light over more pixels.

7.4.2 Regime 2: Faint Source on a Bright Sky

For a faint star seen through a broad optical filter, the sky term is the largest, $n_{\mathrm{pix}} R_{\mathrm{sky}} t \gg R_{\mathrm{src}} t,\ n_{\mathrm{pix}} R_{\mathrm{dark}} t,\ n_{\mathrm{pix}} N_{\mathrm{exp}} \sigma_{\mathrm{RON}}^2$. Then

$$\begin{aligned} \mathrm{SNR} &\overset{(1)}{\approx} \frac{R_{\mathrm{src}}\, t}{\sqrt{n_{\mathrm{pix}} R_{\mathrm{sky}}\, t}} \\ &= \frac{R_{\mathrm{src}}}{\sqrt{n_{\mathrm{pix}} R_{\mathrm{sky}}}}\, \sqrt{t}, \end{aligned}$$

(1) Keep only the sky term under the root in Eq. (7.20).

$$\boxed{\mathrm{SNR} \approx \frac{R_{\mathrm{src}} \sqrt{t}}{\sqrt{n_{\mathrm{pix}} R_{\mathrm{sky}}}} \propto \frac{\sqrt{t}}{\sqrt{n_{\mathrm{pix}}}}.} \tag{7.25}$$

The ratio still grows as $\sqrt{t}$, but now in proportion to R_{src}, not its square root. A star four times fainter has a quarter of the signal-to-noise ratio and needs sixteen times the time.

Seeing. Solving Eq. (7.25) for the time needed to reach a given ratio,

$$\begin{aligned} t &\overset{(1)}{=} \mathrm{SNR}^2\, \frac{n_{\mathrm{pix}} R_{\mathrm{sky}}}{R_{\mathrm{src}}^2} \\ &\overset{(2)}{\propto} \theta_{\mathrm{seeing}}^2. \end{aligned} \tag{7.26}$$

(1) Square Eq. (7.25) and solve for t.
(2) The aperture must hold the seeing disc, so $n_{\mathrm{pix}} \propto \theta_{\mathrm{seeing}}^2$, with the star and the sky fixed.

Seeing of 0.6″ instead of 1.2″ needs a quarter of the time for the same result, or doubles the signal-to-noise ratio in the same time. On a bright sky, good seeing is worth as much as a larger telescope.

Splitting is free. Here the signal-to-noise ratio depends on t and not on N_{exp}, as long as each exposure satisfies Eq. (7.23). Splitting 3000 s into ten exposures of 300 s costs almost nothing. It allows cosmic rays to be rejected and the pointing to be dithered, and it keeps bright stars in the field below saturation.

7.4.3 Regime 3: Faint Source Limited by Readout Noise

When the background is very low, as through a narrow filter, in a spectrograph that spreads the light thinly, or in very short exposures, the readout term is the largest, $n_{\mathrm{pix}} N_{\mathrm{exp}} \sigma_{\mathrm{RON}}^2 \gg (R_{\mathrm{src}} + n_{\mathrm{pix}} R_{\mathrm{sky}} + n_{\mathrm{pix}} R_{\mathrm{dark}})\, t$. Then

$$\mathrm{SNR} \overset{(1)}{\approx} \frac{R_{\mathrm{src}}\, t}{\sqrt{n_{\mathrm{pix}} N_{\mathrm{exp}}}\, \sigma_{\mathrm{RON}}},$$

(1) Keep only the readout term under the root in Eq. (7.20).

$$\boxed{\mathrm{SNR} \approx \frac{R_{\mathrm{src}}\, t}{\sqrt{n_{\mathrm{pix}} N_{\mathrm{exp}}}\, \sigma_{\mathrm{RON}}} \propto \frac{t}{\sqrt{n_{\mathrm{pix}} N_{\mathrm{exp}}}}.} \tag{7.27}$$

The noise no longer depends on the light at all, so the ratio grows in proportion to the signal. At a fixed number of exposures it grows as t, faster than in the other two regimes. At a fixed total time it falls as $1/\sqrt{N_{\mathrm{exp}}}$. A star twice as far away gives a quarter of the electrons and a quarter of the signal-to-noise ratio (Problem 7.7.3).

Binning. The charges of a block of 2×2 pixels can be added on the chip before they are read. The block is then read once, so the number of readings in the aperture falls by 4. The read variance $n_{\mathrm{pix}} N_{\mathrm{exp}} \sigma_{\mathrm{RON}}^2$ falls by 4, while the shot-noise terms stay the same, because the electrons are the same. In this regime the signal-to-noise ratio rises by $\sqrt{4} = 2$. Binning coarsens the pixel scale, so it stops being useful once about two binned pixels span the seeing disc, the sampling limit of Section 5.5.

Long exposures. At a fixed total time, fewer and longer exposures give fewer readings. The useful length is limited mainly by cosmic rays, which pile up in a long exposure. For CCDs on the ground the usual limit is about 30 to 45 minutes.

Figure 7.4 shows both cases at a fixed total time.

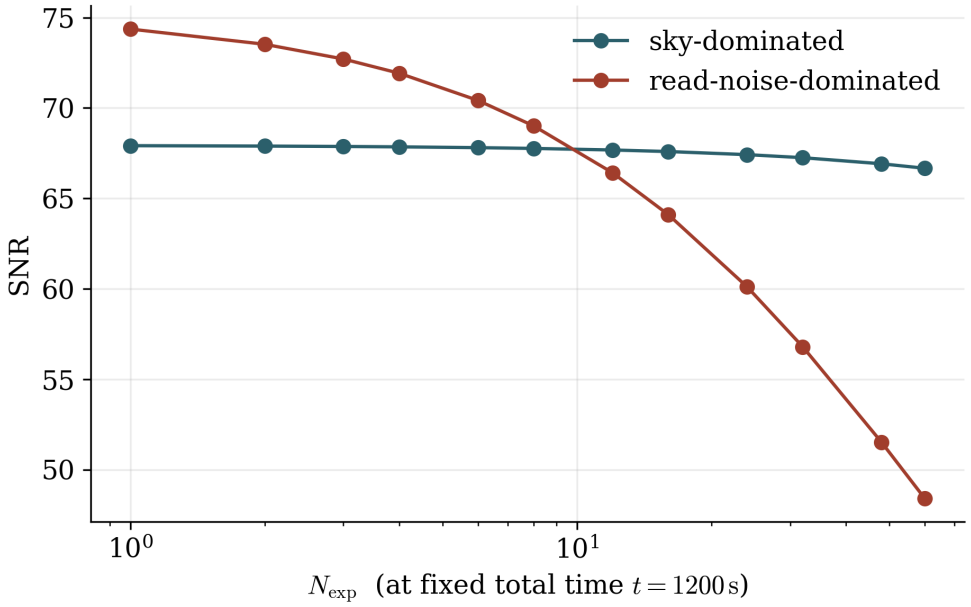


Figure 7.4: Signal-to-noise ratio against the number of exposures N_{exp}, at a fixed total time $t = 1200\,\mathrm{s}$, for a sky-dominated case and a read-noise-dominated case.

The regimes side by side. The term that dominates the variance decides how the ratio scales and how to observe. Table 7.2 lists the three cases.

7.5 From Signal-to-Noise to Magnitudes and Detections

The signal-to-noise equation tells us how well the electrons of a star are counted. Astronomers want two other things. They want the error of a magnitude, the unit in which brightness is

Regime	Largest variance	Scaling	How to observe
Bright source	$\sigma^2_{\mathrm{src}} = R_{\mathrm{src}} t$	$\mathrm{SNR} \propto \sqrt{t}$	Keep below the full well. Use short exposures or defocus.
Bright sky	$\sigma^2_{\mathrm{sky}} = n_{\mathrm{pix}} R_{\mathrm{sky}} t$	$\mathrm{SNR} \propto \dfrac{\sqrt{t}}{\sqrt{n_{\mathrm{pix}}}}$	Use the smallest aperture good seeing allows. Split freely into exposures that satisfy Eq. (7.23).
Readout	$\sigma^2_{\mathrm{read}} = n_{\mathrm{pix}} N_{\mathrm{exp}} \sigma^2_{\mathrm{RON}}$	$\mathrm{SNR} \propto \dfrac{t}{\sqrt{n_{\mathrm{pix}} N_{\mathrm{exp}}}}$	Bin on the chip up to the sampling limit. Use the longest exposures cosmic rays allow.

Table 7.2: The three regimes of the signal-to-noise equation, with the largest variance, the scaling of SNR, and the way of observing in each.

reported. And for a faint smudge in an image they want to know whether it is a star at all, or just the noise happening to be high in one place.

Toss a coin ten times and get eight heads. It is odd, but a fair coin does it once in about 18 tries. Toss it a hundred times and get eighty heads, and the coin is surely bent. A fair coin does that about once in two billion tries. The fraction of heads is the same. What changed is how far the result lies from the mean, measured in standard deviations. A smudge in an image is a count of electrons above the mean sky. If the excess is a few times the noise, chance can produce it. If it is many times the noise, chance almost never does. The idea is that significance is the excess in units of the standard deviation, and the Gaussian shape of the noise turns it into a probability.

Ten coin tosses have a mean of 5 heads and standard deviation $\sqrt{10 \times \frac{1}{4}} = 1.58$. Eight heads is only 1.9 standard deviations high, with probability 5.5 per cent. In a hundred tosses, eighty heads is 6 standard deviations above the mean and has probability 5.6×10^{-10}. A detection threshold works the same way: it asks how likely noise alone is to produce an excess at least as large as the one measured. Before setting that threshold, we turn the signal-to-noise ratio into an error on a magnitude.

7.5.1 The Error of a Magnitude

A magnitude is $m = -2.5 \log_{10} F + \text{constant}$ (Eq. (4.7)). With the same telescope, filter and exposure, the signal S is proportional to the flux F, so $m = -2.5 \log_{10} S + \text{constant}$. A signal that is high by one standard deviation makes the star look brighter by

$$\begin{aligned} \delta m &\equiv m(S) - m(S + \sigma_{\mathrm{tot}}) \\ &\overset{(1)}{=} 2.5 \log_{10} \frac{S + \sigma_{\mathrm{tot}}}{S} \\ &\overset{(2)}{=} 2.5 \log_{10} \left(1 + \frac{1}{\mathrm{SNR}} \right). \end{aligned} \tag{7.28}$$

(1) The constant cancels, and $-\log_{10} a + \log_{10} b = \log_{10}(b/a)$.
(2) Eq. (7.14).

For a large signal-to-noise ratio the logarithm is nearly linear,

$$\delta m \overset{(1)}{=} \frac{2.5}{\ln 10}\,\ln\!\left(1+\frac{1}{\mathrm{SNR}}\right) \overset{(2)}{\approx} \frac{2.5}{\ln 10}\cdot\frac{1}{\mathrm{SNR}}, \tag{7.29}$$

(1) $\log_{10} y = \ln y/\ln 10$.
(2) $\ln(1+y)\approx y$ for $y \ll 1$.

$$\boxed{\delta m \approx \frac{1.0857}{\mathrm{SNR}}.} \tag{7.30}$$

The factor $2.5/\ln 10 = 1.0857$ is the one that turned an optical depth into magnitudes of extinction in Chapter 4 (Eq. (4.36)). Both times a natural logarithm is turned into magnitudes. A signal-to-noise ratio of 10 gives an error of about 0.1 magnitude, and 100 gives 0.01.

The error is not symmetric. A signal low by one standard deviation makes the star look fainter by $-2.5\log_{10}(1-1/\mathrm{SNR})$, which is larger. At SNR = 3 the two shifts are 0.31 and 0.44 magnitude, and Eq. (7.30) gives 0.36, between them. At SNR = 10 they are 0.103 and 0.114, and the linear value 0.109 differs from each by about 5 per cent. Figure 7.5 compares the brighter-side shift with the linear form. Below SNR ≈ 5 the linear form is a poor guide, and both sides should be given. Table 7.3 lists what typical signal-to-noise ratios mean in practice.

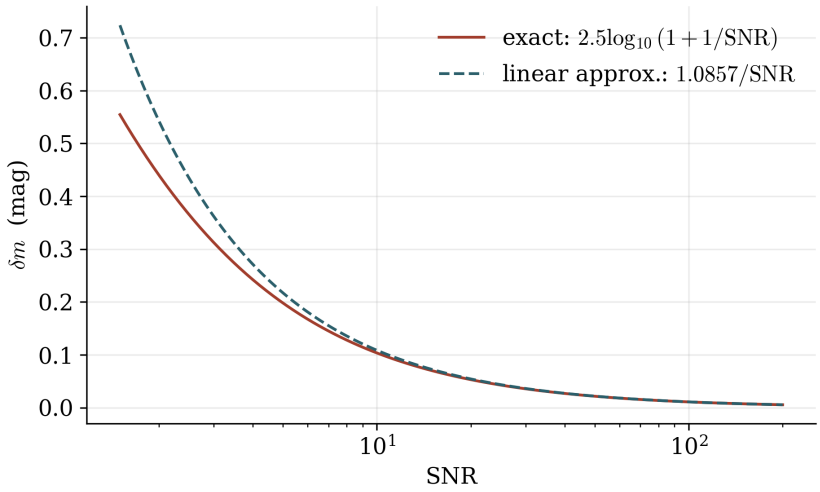


Figure 7.5: Magnitude error δm against SNR. The solid curve is $2.5\log_{10}(1+1/\mathrm{SNR})$ and the dashed curve is $1.0857/\mathrm{SNR}$.

7.5.2 Detection and the False-Alarm Probability

At a signal-to-noise ratio of 2 to 5 the question is whether there is a source at all.

The noise is Gaussian. A large Poisson count has the Gaussian shape of Eq. (7.13), and the readout noise is Gaussian from the start. The total noise is a sum of these independent parts. A sum of independent Gaussian quantities is again Gaussian, and by the central limit theorem a sum of many independent parts of any kind is close to Gaussian. So the measured excess, when no source is present, scatters about zero with a Gaussian distribution of standard deviation σ_{tot}. Measuring the excess in units of σ_{tot}, as $k\sigma_{\mathrm{tot}}$, turns every case into the same standard Gaussian.

SNR	1/SNR	**Meaning**
2–3	33–50%	Barely detected. In a search over a whole image, often a false alarm.
5	20%	The usual threshold for claiming a detection (5σ).
10	10%	A routine measurement, about 0.1 mag.
100	1%	Precise photometry, about 0.01 mag, as needed for planetary transits and variable stars.

Table 7.3: Typical signal-to-noise ratios, the fractional error of each, and what each means for an observation.

Probability of a false alarm. With no source present, the probability that the noise alone rises by at least $k\sigma_{\text{tot}}$ is the area of the standard Gaussian beyond k,

$$\boxed{P_{\text{tail}}(k) = \int_k^\infty \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx.} \tag{7.31}$$

The integral has no closed form, but for large k it is easy to estimate.

$$\begin{aligned} \int_k^\infty e^{-x^2/2}\, dx &\overset{(1)}{=} \int_k^\infty \frac{1}{x}\, d\left(-e^{-x^2/2}\right) \\ &\overset{(2)}{=} \frac{e^{-k^2/2}}{k} - \int_k^\infty \frac{e^{-x^2/2}}{x^2}\, dx \\ &\overset{(3)}{\approx} \frac{e^{-k^2/2}}{k}. \end{aligned} \tag{7.32}$$

(1) $x\, e^{-x^2/2}\, dx = d\big(-e^{-x^2/2}\big)$.
(2) Integration by parts. The boundary term at infinity is zero.
(3) The remaining integral is smaller than the first term by a factor of order $1/k^2$.

$$P_{\text{tail}}(k) \approx \frac{e^{-k^2/2}}{k\sqrt{2\pi}} \qquad (k \gg 1). \tag{7.33}$$

At $k = 5$ this gives 2.97×10^{-7}, against the exact 2.87×10^{-7}. The factor $e^{-k^2/2}$ is why the probability falls so steeply. Going from 5σ to 6σ multiplies it by about $e^{-5.5} \times 5/6$, or $1/294$, against the exact $1/290$. Table 7.4 gives exact values, and Figure 7.6 plots them.

k	$P_{\text{tail}}(k)$	about one in
2	2.28×10^{-2}	44
3	1.35×10^{-3}	741
5	2.87×10^{-7}	3.49×10^{6}
10	7.62×10^{-24}	1.3×10^{23}

Table 7.4: Probability that Gaussian noise alone exceeds k standard deviations in a single trial.

7.5.3 Searching a Whole Image

The probability $P_{\text{tail}}(k)$ applies to one place chosen in advance, such as the position of a known star. A search for new objects, such as supernovae or asteroids, tests every place in the image at

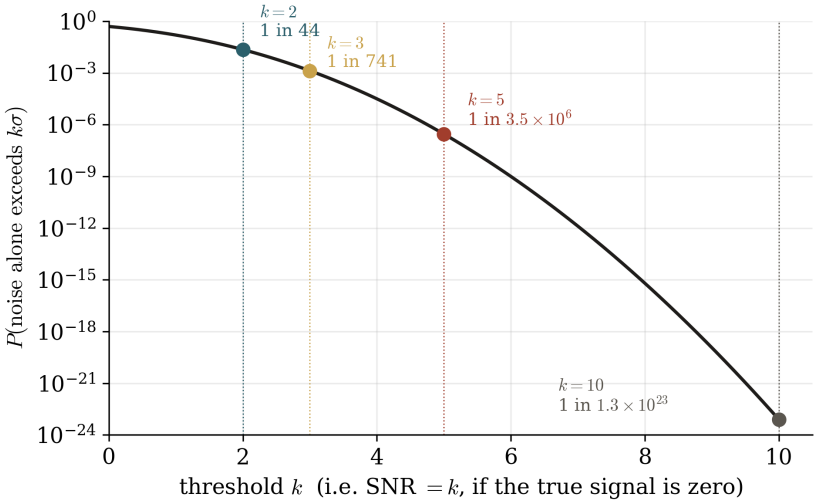


Figure 7.6: Single-trial probability $P_{\text{tail}}(k)$ that Gaussian noise with zero mean exceeds $k\sigma$, on a logarithmic scale, with the values at $k = 2$, 3, 5 and 10 marked.

once. Let the image hold n_{trial} independent places, each of which could hold a source. The probability that at least one of them shows a false excess of $k\sigma$ is

$$P_{\text{any}}(k) \overset{(1)}{=} 1 - \left[1 - P_{\text{tail}}(k)\right]^{n_{\text{trial}}} \overset{(2)}{\approx} n_{\text{trial}}\, P_{\text{tail}}(k). \tag{7.34}$$

(1) The places are independent, so the probability that none of them exceeds $k\sigma$ is the product of n_{trial} equal factors.
(2) $(1 - y)^n \approx 1 - ny$ when $ny \ll 1$.

The product $n_{\text{trial}} P_{\text{tail}}(k)$ is also exactly the mean number of false alarms in the image, whether or not it is small. When it is small it is the chance of even one.

Take a 4000×4000 CCD and count every pixel as a place, $n_{\text{trial}} = 1.6 \times 10^7$. Neighbouring pixels share the light of one star, so the true number of independent places is smaller, and this count gives an upper estimate.

- At $k = 3$ the mean number of false alarms is $1.6 \times 10^7 \times 1.35 \times 10^{-3} \approx 2.2 \times 10^4$. The real objects would be lost among them.
- At $k = 5$ it is $1.6 \times 10^7 \times 2.87 \times 10^{-7} \approx 4.6$ in the whole image.
- At $k = 5.5$ it is about 0.3.

One false alarm per image is expected at $k = 5.3$. This is why large surveys require 5σ or more before they claim a new object.

7.5.4 Finding Planets in Transit

A planet whose orbit we see edge-on passes in front of its star once per orbit and blocks a little of its light (Section 6.20). Finding such a planet is a counting problem. The dip must stand out from the scatter of the counts, and the search over many stars must not be fooled by chance.

The signal. A planet of radius R_p in front of a star of radius R_* removes the fraction $\Delta F/F = (R_p/R_*)^2$ of the light (Eq. (6.243)). For the Earth and the Sun,

$$\frac{\Delta F}{F} = \left(\frac{6.371\times10^6\,\mathrm{m}}{6.957\times10^8\,\mathrm{m}}\right)^2 = (0.00916)^2 = 8.4\times10^{-5},$$

or 84 parts per million (ppm). A central crossing lasts 13 hours (Section 6.20), and it comes once a year. The planet enters and leaves the star's disk in a fraction R_p/R_* of that time, about 7 minutes, so the dip is nearly a box. Jupiter gives 1%, the depth for which Table 7.3 asks for SNR $\approx$ 100. An Earth needs a precision a hundred times finer.

The Kepler mission. Kepler was launched in March 2009. It was a Schmidt telescope with a 0.95 m aperture, a design that gives sharp images over a wide field. Its camera held 42 CCDs of 2200×1024 pixels each, 9.5×10^7 pixels in all. It stared at one field in Cygnus, Lyra and Draco and measured the brightness of about 145 000 stars without a break. An Earth's transit comes once a year and lasts half a day, so a gap of a day can miss it, and only steady staring at the same stars catches several. Kepler did not orbit the Earth. It went round the Sun once in 372.5 days, trailing the Earth and slowly drifting away from it. A telescope in low Earth orbit would have the Earth in front of its field for part of each orbit, and it would pass from sunlight to shadow every hour and a half, which changes its temperature and so its focus. Far from the Earth nothing blocks the field and the temperature stays steady. The planets it found are counted in Section 6.20.

The noise of one transit. To find the signal-to-noise ratio of the depth, we compare two counts from the same star (P5 + P6). Let N_{in} be the number of electrons collected during the transit, in its duration t_{tr}, and N_{out} the number collected outside it, in a time t_{out}. The depth is one minus the ratio of the two count rates,

$$\frac{\Delta F}{F} = 1 - q, \qquad q \equiv \frac{N_{\mathrm{in}}/t_{\mathrm{tr}}}{N_{\mathrm{out}}/t_{\mathrm{out}}}.$$

The times are known exactly, and only the counts scatter. A small change of the counts changes q by

$$\begin{aligned}\frac{\delta q}{q} &\overset{(1)}{=} \delta \ln q \\ &\overset{(2)}{=} \delta \ln N_{\mathrm{in}} - \delta \ln N_{\mathrm{out}} \\ &= \frac{\delta N_{\mathrm{in}}}{N_{\mathrm{in}}} - \frac{\delta N_{\mathrm{out}}}{N_{\mathrm{out}}}.\end{aligned} \tag{7.35}$$

(1) $d\ln q = dq/q$.
(2) $\ln q = \ln N_{\mathrm{in}} - \ln N_{\mathrm{out}} + \ln(t_{\mathrm{out}}/t_{\mathrm{tr}})$, and the last term is fixed.

The two counts are independent, so their variances add,

$$\begin{aligned}\frac{\sigma_q^2}{q^2} &\overset{(1)}{=} \frac{\sigma_{\mathrm{in}}^2}{N_{\mathrm{in}}^2} + \frac{\sigma_{\mathrm{out}}^2}{N_{\mathrm{out}}^2} \\ &\overset{(2)}{=} \frac{1}{N_{\mathrm{in}}} + \frac{1}{N_{\mathrm{out}}}.\end{aligned} \tag{7.36}$$

(1) Eq. (7.10) for the two terms of the last line above.
(2) A Poisson count has a variance equal to its mean, Eq. (7.8).

The error of $\Delta F/F = 1 - q$ is the error of q. So the signal-to-noise ratio of one transit is

$$\begin{aligned}\mathrm{SNR}(1) &\overset{(1)}{=} \frac{\Delta F/F}{\sigma_q} \\ &\overset{(2)}{=} \frac{\Delta F/F}{q\sqrt{1/N_{\mathrm{in}} + 1/N_{\mathrm{out}}}} \\ &\overset{(3)}{\approx} \frac{\Delta F/F}{\sqrt{1/N_{\mathrm{in}} + 1/N_{\mathrm{out}}}} \\ &\overset{(4)}{\approx} \frac{\Delta F}{F}\sqrt{N_{\mathrm{in}}}. \end{aligned} \tag{7.37}$$

(1) Eq. (7.14), with the depth as the signal.
(2) The line above.
(3) $q = 1 - \Delta F/F$ differs from 1 by less than 10^{-4} for an Earth.
(4) A long baseline, $t_{\mathrm{out}} \gg t_{\mathrm{tr}}$, gives $N_{\mathrm{out}} \gg N_{\mathrm{in}}$.

$$\boxed{\mathrm{SNR}(1) \approx \frac{\Delta F}{F}\sqrt{N_{\mathrm{in}}}.} \tag{7.38}$$

The depth is measured in units of $1/\sqrt{N_{\mathrm{in}}}$, the fractional shot noise of the light collected during the transit (Eq. (7.24)). The baseline costs nothing once it is long. Only the light that arrives while the planet is in front sets the noise.

Many transits. Each transit gives a new count, independent of the others. Lining up n_{tr} transits at the period and adding them, as two images were added in Problem 7.7.2, collects $n_{\mathrm{tr}}N_{\mathrm{in}}$ electrons in transit,

$$\begin{aligned}\mathrm{SNR}(n_{\mathrm{tr}}) &\overset{(1)}{=} \frac{\Delta F}{F}\sqrt{n_{\mathrm{tr}}N_{\mathrm{in}}} \\ &= \sqrt{n_{\mathrm{tr}}}\,\mathrm{SNR}(1). \end{aligned} \tag{7.39}$$

(1) Eq. (7.38) with the in-transit count $n_{\mathrm{tr}}N_{\mathrm{in}}$, since the variances of the n_{tr} counts add (Eq. (7.10)).

$$\boxed{\mathrm{SNR}(n_{\mathrm{tr}}) = \sqrt{n_{\mathrm{tr}}}\,\mathrm{SNR}(1).} \tag{7.40}$$

Here n_{tr} is the number of transits added. Four transits double the signal-to-noise ratio of one.

Photons from a twelfth-magnitude star. To put in numbers we need N_{in} for a star like the Sun of magnitude 12, seen through Kepler's aperture of diameter $D = 0.95\,\mathrm{m}$. The zero point of Eq. (4.7) is known for bolometric magnitudes, $F_{\mathrm{ref}} = 2.518 \times 10^{-8}\,\mathrm{W\,m^{-2}}$. We take 12 as the bolometric magnitude, since for a star like the Sun the two differ by less than 0.1 magnitude. We count every photon as one of wavelength $\lambda = 600\,\mathrm{nm}$, near the middle of the visible band, and we let a single throughput of 0.5 stand for the fraction of photons that fall in the band, pass the optics and free an electron. Both are rough assumed values. Then

$$\begin{aligned}R_{\mathrm{src}} &\overset{(1)}{=} 0.5 \times \frac{F\,\pi(D/2)^2}{hc/\lambda} \\ &\overset{(2)}{=} 0.5 \times \frac{F_{\mathrm{ref}}\,10^{-0.4\times 12}\,\pi(D/2)^2}{hc/\lambda} \\ &= 0.5 \times \frac{2.518\times10^{-8}\times 1.585\times10^{-5}\,\mathrm{W\,m^{-2}}\times 0.709\,\mathrm{m^2}}{3.31\times10^{-19}\,\mathrm{J}} \\ &= 4.3\times10^{5}\,\mathrm{e^-\,s^{-1}}. \end{aligned} \tag{7.41}$$

(1) Power collected by the mirror, divided by the energy of one photon, times the throughput.
(2) Eq. (4.7) solved for F.

In 6.5 hours, 23 400 s, the star gives $N = 1.0 \times 10^{10}$ electrons, and the fractional shot noise is $1/\sqrt{N} = 1.0 \times 10^{-5}$, or 10 ppm. The bolometric zero point counts light at all wavelengths, while a real band takes only part of it, so this is the smallest the shot noise can be. During a 13-hour transit, $N_{\text{in}} = 2.0 \times 10^{10}$ and $\sqrt{N_{\text{in}}} = 1.41 \times 10^5$, so Eq. (7.38) gives $\text{SNR}(1) = 8.4 \times 10^{-5} \times 1.41 \times 10^5 = 12$. With photons as the only noise, one transit of an Earth would show.

The star is the floor. The star itself flickers. Hot gas rises and sinks in cells across its surface (granulation), and dark spots turn into and out of view. Together they change a Sun-like star's brightness by tens of ppm over hours. This is one more independent part in Eq. (7.10), and unlike shot noise it does not shrink when more photons are collected. For twelfth-magnitude Sun-like stars, Kepler measured a total noise of about 30 ppm over 6.5 hours, three times our photon estimate. Most of the noise comes from the star, not the photons, and a larger telescope would not remove it. If the noise averages down like independent parts, over a 13-hour transit it is $30\,\text{ppm} \times \sqrt{6.5/13} = 21$ ppm, and

$$\text{SNR}(1) = \frac{84\,\text{ppm}}{21\,\text{ppm}} = 4.0.$$

Flicker that lasts for hours averages down more slowly than this, so 4.0 is an upper value.

How high the threshold must be. A search tries every star, every period and every start time, and each try is a chance for a false alarm (Eq. (7.34)). We count the tries for one star watched for a time T. For a trial period P, the first transit can start at any of P/t_{tr} times that can be told apart. Two trial periods that differ by ΔP put the transits in different places once the T/P transits have drifted apart by t_{tr}, so $\Delta P = t_{\text{tr}} P/T$. The number of tries is

$$\begin{aligned} n_{\text{trial}} &\overset{(1)}{=} \int_{P_{\text{min}}}^{P_{\text{max}}} \frac{P}{t_{\text{tr}}} \frac{dP}{t_{\text{tr}} P/T} \\ &= \frac{T\,(P_{\text{max}} - P_{\text{min}})}{t_{\text{tr}}^2}. \end{aligned} \tag{7.42}$$

(1) Start times per period, times the number of distinct periods in each dP.

Kepler watched for $T = 4\,\text{yr} = 1461$ d. A period must fit three transits into T, so $P_{\text{max}} = 487$ d, and we take $P_{\text{min}} = t_{\text{tr}} = 0.5$ d. Then $n_{\text{trial}} = 1461 \times 486.5/0.25 = 2.8 \times 10^6$ for one star, and 4.3×10^{11} for 1.5×10^5 stars. One false alarm in the whole search is expected when $n_{\text{trial}} P_{\text{tail}}(k) = 1$, or $P_{\text{tail}}(k) = 2.3 \times 10^{-12}$. Eq. (7.33) gives $P_{\text{tail}} = 2.7 \times 10^{-12}$ at $k = 6.9$ and 1.3×10^{-12} at $k = 7.0$. So the threshold is about 7σ, set by the number of tries and not by any one light curve. The Kepler search used 7.1σ, which leaves about 0.3 false alarms by this count.

How many transits. With $\text{SNR}(1) = 4.0$ for an Earth around a twelfth-magnitude Sun-like star, Eq. (7.40) reaches 7.1 when

$$n_{\text{tr}} = \left(\frac{7.1}{4.0}\right)^2 = 3.2,$$

so four transits are needed, and an Earth at 1 AU makes one a year. Three transits are needed in any case to show that the dips repeat with one period, since two dips fix a period and a third tests it. This is why Kepler watched the same stars for four years. Figure 7.7 compares the dip of an Earth with the photon noise and the measured noise.

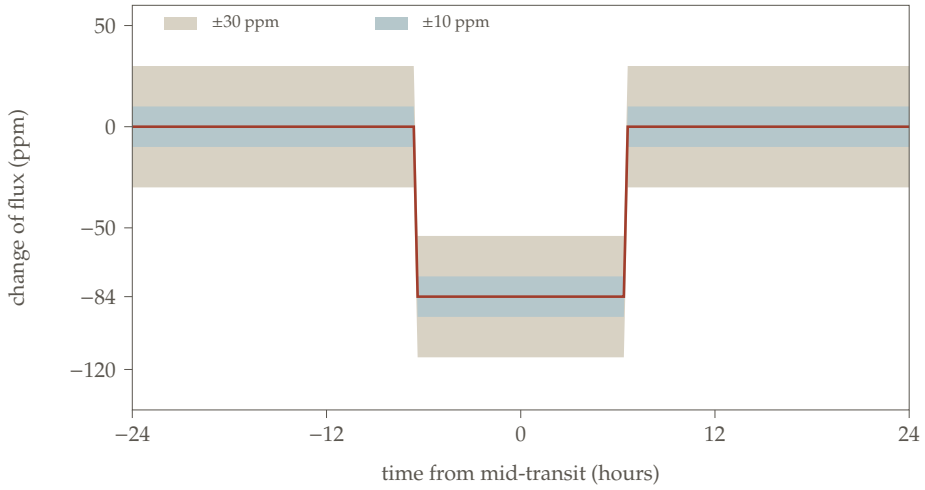


Figure 7.7: Change of a Sun-like star's flux, in ppm, during a central transit of an Earth (red line), 84 ppm deep and 13 hours long. The bands show one standard deviation of the flux averaged over 6.5 hours, ±30 ppm as measured by Kepler (light) and ±10 ppm from shot noise alone (dark).

7.6 Noise Budgets Beyond the CCD

The variance sum of Eq. (7.19) used nothing particular to a CCD. Any measurement made of independent parts has a budget of the same kind, and the part that dominates decides what to improve. Timing the radio pulses of a pulsar is a second example. The measured quantity is the arrival time of a pulse instead of a number of electrons, and its total error is again a sum of independent variances.

- **Radiometer noise.** The noise of the receiver and its electronics, like the read and dark noise of a CCD.
- **Sky noise.** Radio emission from everything else in the beam, the radio form of σ_{sky}.
- **Source noise.** Changes in the pulsar itself, with no CCD counterpart. Single pulses differ in shape (*pulse jitter*), and the spin of the neutron star wanders slightly (*spin noise*).
- **Dispersion noise.** Free electrons in the ionised gas between the stars (Chapter 8) delay a pulse by an amount that depends on the radio frequency. The column of electrons along the line of sight changes as the pulsar and the gas move, so the delay changes too. A CCD image has nothing like it, because the atmosphere and the dust only dim and scatter its light (Section 5.1).
- **Radio-frequency interference.** Man-made radio signals from television, mobile phones, radar and power lines, which must be found and cut out, much as cosmic-ray hits are removed from a CCD frame.

As for a CCD, the term that dominates a given observation decides how it should be planned and reduced.

7.7 Problem Set: Poisson Statistics and CCD Signal-to-Noise

7.7.1 Problem 1: Normalisation, Moments, and the Gaussian Limit of the Poisson Distribution

Problem Statement: A count n has the Poisson distribution $P(n) = \bar{n}^{\,n} e^{-\bar{n}}/n!$ for $n = 0, 1, 2, \ldots$

(a) Prove that it is normalised, $\sum_{n=0}^{\infty} P(n) = 1$.

(b) Prove that its mean and its variance are both equal to $\bar{n}$.

(c) Plot the distribution for $\bar{n} = 0.1$, 1, 10, 100 and 1000, and show that it approaches a Gaussian when $\bar{n}$ is large.

Essential Ideas:

- The exponential series, $\sum_{n=0}^{\infty} \bar{n}^{\,n}/n! = e^{\bar{n}}$, does every sum.
- The mean and variance are derived in Eqs. (7.6) and (7.8). A second route differentiates the exponential series with respect to $\bar{n}$.
- For $\bar{n} \gg 1$, Stirling's formula turns $P(n)$ into a Gaussian of mean $\bar{n}$ and variance $\bar{n}$, Eq. (7.13).

Detailed Solutions:

(a) **Normalisation**.

$$\sum_{n=0}^{\infty} P(n) \overset{(1)}{=} e^{-\bar{n}} \sum_{n=0}^{\infty} \frac{\bar{n}^{\,n}}{n!} \overset{(2)}{=} e^{-\bar{n}}\, e^{\bar{n}}$$

(1) $e^{-\bar{n}}$ does not depend on n.
(2) The exponential series.

$$\boxed{\sum_{n=0}^{\infty} P(n) = 1.}$$

(b) **Mean and variance**. The direct sums are Eqs. (7.6) and (7.8). As a second route, write $\bar{n}\,\frac{\mathrm{d}}{\mathrm{d}\bar{n}}$ for the operator that multiplies the n-th term of the series by n, since $\bar{n}\,\frac{\mathrm{d}}{\mathrm{d}\bar{n}}\bar{n}^{\,n} = n\,\bar{n}^{\,n}$. Applying it once and twice to the exponential series,

$$\sum_{n=0}^{\infty} n\,\frac{\bar{n}^{\,n}}{n!} \overset{(1)}{=} \bar{n}\,\frac{\mathrm{d}}{\mathrm{d}\bar{n}}\, e^{\bar{n}} = \bar{n}\, e^{\bar{n}},$$

$$\sum_{n=0}^{\infty} n^2\,\frac{\bar{n}^{\,n}}{n!} \overset{(2)}{=} \bar{n}\,\frac{\mathrm{d}}{\mathrm{d}\bar{n}}\left(\bar{n}\, e^{\bar{n}}\right) = \bar{n}\,(1+\bar{n})\, e^{\bar{n}}.$$

(1) Differentiate the series term by term and multiply by $\bar{n}$.
(2) Do it again to the first line.

Multiplying by $e^{-\bar{n}}$ gives $\langle n\rangle = \bar{n}$ and $\langle n^2\rangle = \bar{n} + \bar{n}^2$. Then

$$\boxed{\langle n\rangle = \bar{n}, \qquad \sigma^2 = \langle n^2\rangle - \langle n\rangle^2 = \bar{n}.}$$

(c) **Plot and the Gaussian limit**. The distribution can be drawn directly from its definition.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import poisson, norm

lambdas = [0.1, 1, 10, 100, 1000]
fig, axes = plt.subplots(1, len(lambdas), figsize=(18, 3.2))

for ax, lam in zip(axes, lambdas):
```

```
    k_max = int(lam + 6*np.sqrt(lam) + 5)
    k = np.arange(0, k_max + 1)
    ax.bar(k, poisson.pmf(k, lam), color="C0", width=1.0, label="
        Poisson")
    if lam >= 10:
        x = np.linspace(0, k_max, 400)
        ax.plot(x, norm.pdf(x, loc=lam, scale=np.sqrt(lam)), "C3--",
            label="Gaussian")
    ax.set_title(rf"$\lambda={lam}$")
    ax.set_xlabel("$k$")
axes[0].set_ylabel("$P[X=k]$")
axes[-1].legend()
plt.tight_layout()
plt.show()
```

(In the code, `lam` is $\bar{n}$ and `k` is n.) Figure 7.8 shows the five distributions with the Gaussian of the same mean and variance, Eq. (7.13).

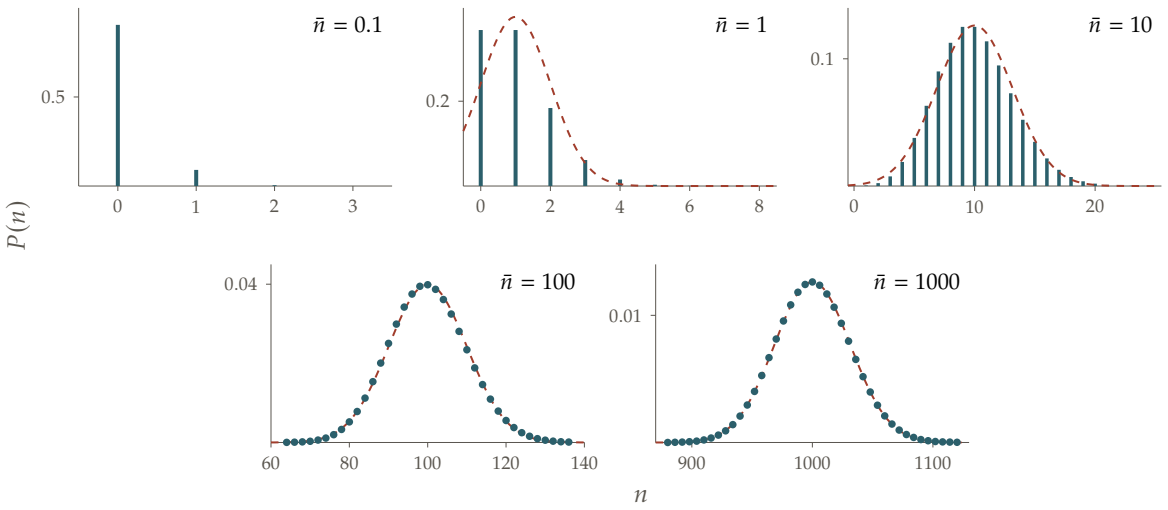


Figure 7.8: Poisson probabilities $P(n)$ for $\bar{n} = 0.1$, 1, 10, 100 and 1000 (bars, or dots at every second and every sixth n in the last two panels), with the Gaussian of mean $\bar{n}$ and variance $\bar{n}$ (dashed) for $\bar{n} \geq 1$.

For $\bar{n} = 0.1$ the count is 0 with probability $e^{-0.1} = 0.905$ and 1 with probability $0.1\,e^{-0.1} = 0.090$, and larger counts almost never happen. No bell curve fits this. For $\bar{n} = 1$ the counts 0 and 1 are equally likely, and the Gaussian puts probability at negative n, which cannot occur. For $\bar{n} = 10$ the shape is close to a bell but lopsided, with a longer tail to the right. For $\bar{n} = 100$ and 1000 the points lie on the Gaussian. The peak height checks this, since Eq. (7.13) gives $1/\sqrt{2\pi\bar{n}} = 0.012616$ at $\bar{n} = 1000$, against the Poisson value 0.012615. The lopsidedness comes from the dropped term of order $(n-\bar{n})^3/\bar{n}^2$, which is $1/\sqrt{\bar{n}}$ at one standard deviation, so

$$\boxed{P(n) \to \frac{1}{\sqrt{2\pi\bar{n}}} \exp\left[-\frac{(n-\bar{n})^2}{2\bar{n}}\right] \quad \text{for } \bar{n} \gg 1.}$$

This is why the noise of large counts may be treated as Gaussian, as in Section 7.5.2.

7.7.2 Problem 2: Signal-to-Noise with Exposure Time and Adding Images

Problem Statement: Two images I_1 and I_2 of the same bright object are taken with the same instrument. Both are dominated by the shot noise of the source. I_1 has an exposure time of 1 min and SNR = 100.

(a) What is the SNR of I_2 if its exposure time is 5 min?

(b) If the two images are added pixel by pixel, what is the SNR of the sum?

Essential Ideas:

- With the source's own shot noise dominant, $\mathrm{SNR} = \sqrt{R_{\mathrm{src}} t} \propto \sqrt{t}$, Eq. (7.24).
- Variances of independent images add, Eq. (7.10), and the variance of a Poisson count equals its mean, Eq. (7.8).

Detailed Solutions:

(a) **Longer exposure**. Only the time differs between the two images.

$$\begin{aligned} \mathrm{SNR}_2 &\overset{(1)}{=} \sqrt{R_{\mathrm{src}} t_2} \\ &\overset{(2)}{=} \mathrm{SNR}_1 \sqrt{\frac{t_2}{t_1}} = 100\sqrt{5} \end{aligned}$$

(1) Eq. (7.24).
(2) Divide by $\mathrm{SNR}_1 = \sqrt{R_{\mathrm{src}} t_1}$.

$$\boxed{\mathrm{SNR}_2 = 223.6.}$$

(b) **Sum of the two images**. The signals add, $S_{1+2} = S_1 + S_2 = R_{\mathrm{src}}(t_1 + t_2)$. The two images are independent, so their variances add, and each variance equals the signal.

$$\begin{aligned} \mathrm{SNR}_{1+2} &\overset{(1)}{=} \frac{S_1 + S_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \\ &\overset{(2)}{=} \frac{S_1 + S_2}{\sqrt{S_1 + S_2}} = \sqrt{R_{\mathrm{src}}(t_1 + t_2)} \\ &\overset{(3)}{=} \mathrm{SNR}_1 \sqrt{1 + \frac{t_2}{t_1}} = 100\sqrt{6} \end{aligned}$$

(1) Eq. (7.14), with the variance of the sum from Eq. (7.10).
(2) $\sigma_i^2 = S_i$ for a Poisson count, Eq. (7.8).
(3) Divide by $\mathrm{SNR}_1 = \sqrt{R_{\mathrm{src}} t_1}$.

$$\boxed{\mathrm{SNR}_{1+2} = 244.9.}$$

As a check by counting electrons, $\mathrm{SNR}_1 = 100$ means $S_1 = 100^2 = 10^4$ electrons, so $S_2 = 5 \times 10^4$ and the sum has 6×10^4 electrons, with $\sqrt{6 \times 10^4} = 244.9$. Adding the images gives the same result as one exposure of 6 min. That stops being exactly true when readout noise matters, because the sum has been read twice and the single exposure only once.

7.7.3 Problem 3: A Faint Star Twice as Far Away

Problem Statement: The image of a faint star has SNR = 5, and its noise is dominated by instrumental noise.

(a) What would the SNR be if the star were twice as far away, with everything else unchanged?

(b) What would it be if the noise had instead been dominated by the star's own shot noise?

Essential Ideas:

- Instrumental noise, the dark and readout terms, does not depend on the star, so SNR $\propto R_{\mathrm{src}}$, Eq. (7.27).
- With the star's own shot noise dominant, SNR $\propto \sqrt{R_{\mathrm{src}}}$, Eq. (7.24).
- The flux, and so R_{src}, falls as $1/d^2$, Eq. (4.2).

Detailed Solutions:

(a) **Instrumental noise**. The variance $n_{\mathrm{pix}} R_{\mathrm{dark}} t + n_{\mathrm{pix}} N_{\mathrm{exp}} \sigma_{\mathrm{RON}}^2$ contains no light from the star, so it is the same at either distance. Only the signal changes.

$$\frac{\mathrm{SNR}(2d)}{\mathrm{SNR}(d)} \overset{(1)}{=} \frac{R_{\mathrm{src}}(2d)}{R_{\mathrm{src}}(d)} \overset{(2)}{=} \left(\frac{d}{2d}\right)^2 = \frac{1}{4}$$

(1) $\mathrm{SNR} = R_{\mathrm{src}} t/\sigma_{\mathrm{tot}}$ with t and σ_{tot} unchanged.
(2) $R_{\mathrm{src}} \propto F \propto 1/d^2$.

$$\boxed{\mathrm{SNR}(2d) = \frac{5}{4} = 1.25.}$$

The star is no longer detected. An excess of 1.25σ is produced by noise alone about once in ten trials.

(b) **Shot-noise limit**. Now $\mathrm{SNR} = \sqrt{R_{\mathrm{src}} t} \propto \sqrt{R_{\mathrm{src}}} \propto 1/d$, so

$$\boxed{\mathrm{SNR}(2d) = \frac{5}{2} = 2.5.}$$

A star limited by its own light loses less when it is moved away, because its noise falls with its signal. The same light loss costs a faint star on a bright sky or on a noisy detector twice as much.

7.7.4 Problem 4: Dark Current and Readout Noise from Two Dark Frames

Problem Statement: Two dark frames are taken with the same telescope and CCD. The first has an exposure time of 1 min and a noise variance a. The second has an exposure time of 3 min and a noise variance $2.6a$.

(a) Find the noise variance of the dark current per unit time.

(b) Find the noise variance of the readout noise.

Essential Ideas:

- A dark frame has no star and no sky. Its variance is $\sigma^2(t) = n_{\mathrm{pix}} R_{\mathrm{dark}} t + n_{\mathrm{pix}} \sigma_{\mathrm{RON}}^2$, a straight line in t. For a single pixel, $n_{\mathrm{pix}} = 1$.
- Two frames of different lengths give two equations for the two unknowns, solved in Eq. (7.21).

Detailed Solutions: The two frames give

$$\begin{pmatrix} 1\,\mathrm{min} & 1 \\ 3\,\mathrm{min} & 1 \end{pmatrix} \begin{pmatrix} n_{\mathrm{pix}} R_{\mathrm{dark}} \\ n_{\mathrm{pix}} \sigma_{\mathrm{RON}}^2 \end{pmatrix} = \begin{pmatrix} a \\ 2.6a \end{pmatrix}.$$

(a) **Dark variance per unit time**. By Eq. (7.21),

$$\boxed{n_{\text{pix}} R_{\text{dark}} = \frac{\sigma_2^2 - \sigma_1^2}{t_2 - t_1} = \frac{2.6a - a}{3\,\text{min} - 1\,\text{min}} = 0.8a \text{ per minute.}}$$

(b) **Read variance**. By Eq. (7.21),

$$\boxed{n_{\text{pix}} \sigma_{\text{RON}}^2 = \frac{t_2\sigma_1^2 - t_1\sigma_2^2}{t_2 - t_1} = \frac{3a - 2.6a}{2} = 0.2a.}$$

As a check, the first frame gives $0.8a \times 1 + 0.2a = a$. In the 1-minute frame the dark current gives 80 per cent of the variance, and in the 3-minute frame 92 per cent. The read variance is the same $0.2a$ in every frame, whatever its length. A 5-minute frame should show $0.8a \times 5 + 0.2a = 4.2a$, and taking one is how the straight line of Figure 7.3 would be tested.

7.7.5 Problem 5: How Many Transits

Problem Statement: A Sun-like star of magnitude 12 is watched by Kepler. Its total noise is 30 ppm over 6.5 hours, and the noise averages down like independent parts over longer times. A planet is found when the transits added together reach SNR = 7.1. Every transit is a central crossing of 13 hours.

(a) How many transits does a planet of twice the Earth's radius need?

(b) How many would an Earth need if the star did not flicker, so that only the shot noise of 10 ppm over 6.5 hours remained?

Essential Ideas:

- The depth is $(R_p/R_*)^2$, Eq. (6.243), and grows as the square of the planet's radius.
- One transit has SNR(1) equal to the depth divided by the fractional noise over the transit, Eq. (7.38). Noise that averages down like independent parts falls as one over the square root of the time.
- n_{tr} transits give $\sqrt{n_{\text{tr}}}\,\text{SNR}(1)$, Eq. (7.40).

Detailed Solutions:

(a) **Twice the Earth's radius**. The noise over 13 hours is $30\,\text{ppm} \times \sqrt{6.5/13} = 21$ ppm, as for the Earth in Section 7.5.4.

$$\text{SNR}(1) \overset{(1)}{=} \frac{2^2 \times 84\,\text{ppm}}{21\,\text{ppm}} = 16,$$

$$n_{\text{tr}} \overset{(2)}{=} \left(\frac{7.1}{\text{SNR}(1)}\right)^2 = \left(\frac{7.1}{16}\right)^2 = 0.20.$$

(1) Eq. (7.38), with the depth 4 times that of the Earth.
(2) Eq. (7.40) set equal to 7.1 and solved for n_{tr}.

$$\boxed{n_{\text{tr}} = 1 \text{ by the noise, 3 to show the period.}}$$

One transit already stands out at 16σ. Doubling the radius cuts the number of transits needed by $2^4 = 16$. Three transits are still needed to show that the dips repeat with one period.

(b) **A quiet star**. The shot noise over 13 hours is $10\,\text{ppm} \times \sqrt{6.5/13} = 7.1$ ppm.

$$\text{SNR}(1) \overset{(1)}{=} \frac{84\,\text{ppm}}{7.1\,\text{ppm}} = 12,$$

$$n_{\text{tr}} \overset{(2)}{=} \left(\frac{7.1}{12}\right)^2 = 0.36.$$

(1) Eq. (7.38).
(2) Eq. (7.40) set equal to 7.1.

$$\boxed{n_{\text{tr}} = 1 \text{ by the noise, against 4 for the real star.}}$$

The flicker of the star, not the number of photons, is what made the search for Earths take four years.

Chapter 8

The Interstellar Medium

The earlier chapters used starlight to measure positions, motions and brightness. That light travels through the **interstellar medium**, a thin mixture of gas and dust between the stars. The medium can remove light from the beam, change its colour or delay its arrival. To read a star correctly, we need to understand those effects along the whole line of sight.

The same gas is the raw material of stars. Stars form out of it, and when they die they return much of their mass to it, enriched with heavier elements.

8.1 Composition, Density and Temperature

The earlier chapters treated a star as a point of light, and read its position, distance, flux and colour from that light. They took the space between the star and the telescope to be empty. It is not, and a star seen through a dusty part of the Galaxy can be dimmed by many magnitudes.

Think of a light mist on a cool morning. Across a room nobody notices it. A hill a kilometre away is hidden by it. Each metre of mist holds very little water, but a kilometre of it holds a thousand times as much. The interstellar medium is a mist of this kind, far thinner than any mist on Earth but thousands of parsecs deep. What decides its effect on light is the total amount of matter along the line of sight, not how dense it is at any point.

Over 8 kpc toward the Galactic centre, a density of only one atom per cubic centimetre still puts 2.5×10^{22} atoms into a column 1 cm^2 wide. Sea-level air holds 2.5×10^{19} molecules per cubic centimetre, so just 10 m of air contains as many in the same cross-section. This integrated amount, the column density, explains how such thin gas can leave strong marks on distant starlight. We begin by describing what fills that column, then examine the effects of its gas and dust.

What it is. The **interstellar medium** is the matter between the stars of a galaxy. It is gas, dust, fast charged particles called cosmic rays, and a magnetic field. It holds of order ten to twenty per cent of the ordinary matter of the Milky Way. Its bulk properties are in Table 8.1.

Property	Typical value
Number density n	10^{-4}–10^{6} cm^{-3}
Temperature T	10–10^{6} K
Composition (by mass)	70% H, 28% He, $\lesssim$ 2% heavier elements
Phase split (by mass)	99% gas, 1% dust
Magnetic field strength	$\sim 10^{-10}$ T ($\sim 1\,\mu$G)

Table 8.1: Bulk properties of the interstellar medium.

How thin it is. The best laboratory vacuum, called extreme high vacuum, still holds about 10^5 molecules per cubic centimetre. Much of the interstellar medium has fewer than one atom per cubic centimetre, so it is emptier than any vacuum made on Earth. In such thin gas an excited atom can wait a long time before a collision disturbs it. Transitions too slow to be seen in laboratory gas, the so-called forbidden lines, then become the main way the gas radiates (Section 8.4).

What it is made of. The gas holds ions such as H^+, He^+ and He^{2+}, free electrons, neutral atoms, and molecules such as H_2, CO, CN, H_2O and NH_3. The dust is grains from about 1 nm to 1 μm across, made of silicates, carbon in several forms, and metals such as iron and nickel, often coated with ices of water and ammonia. Dust is only one per cent of the mass. Per kilogram, small grains block far more light than gas does (Eq. (8.13)), so dust causes almost all of the dimming (Section 8.2). Its surfaces are also where molecular hydrogen forms (Section 8.4.2).

Why it matters. There are three reasons. Every photon used in the earlier chapters is scattered, absorbed, delayed and dispersed by the medium, so the corrections for dust in Chapter 4 rest on understanding it. It holds conditions that no laboratory can make, gas thinner than any vacuum and colder than 20 K, with chemistry far from equilibrium. And it is where stars are born and where their matter returns (Chapter 10).

8.2 How the Medium Acts on Light

Chapter 4 found that dust dims a beam by the factor $e^{-\tau_\lambda}$, with optical depth $\tau_\lambda = N_d\,\sigma_{\text{ext}}(\lambda)$ (Eq. (4.33)), and that across the visible the cross-section σ_{ext} falls roughly as $1/\lambda$. That cross-section was taken from measurement. Why it depends on wavelength at all, why the sky is blue while clouds are white, and why free electrons scatter every colour alike was left open.

Hold a weight on a soft spring and move your hand up and down. Move it slowly and the weight follows the hand. Move it at the rate the weight bounces by itself and it swings wildly. Shake fast and the weight hardly moves. What decides is how fast you shake compared with the spring's own rate. Light shakes the electrons in an atom, a molecule or a grain, and each is held in place as if by a spring. A shaken charge radiates, and the light it sends off in new directions is the scattered light. One shaken spring gives the blue sky, the spectral lines and the scattering by free electrons.

For air molecules, visible light drives the charge well below its natural frequency. The scattered power then rises as the fourth power of frequency, so 450 nm blue light is scattered $(700/450)^4 = 5.9$ times more strongly than 700 nm red light. Near the natural frequency the response grows sharply; far above it the behaviour changes again. We first derive these three regimes for one charged oscillator, then use the result to understand grains and free electrons.

Extinction. Light is taken out of a beam in two ways. It can be absorbed, its energy turned into heat or internal excitation of the absorber. Or it can be scattered, sent off in a new direction. The total loss is the **extinction**,

$$\text{extinction} = \text{absorption} + \text{scattering}. \tag{8.1}$$

Absorption happens either at sharp frequencies, as lines of atoms and molecules, or over broad bands, in grains. Scattering is sorted by whether the photon keeps its energy.

8.2.1 One Oscillator, Three Regimes

The driven electron. Model a bound electron as a charge $-e$ of mass m_e held by a spring of natural angular frequency ω_0. Left alone it also loses energy, at the rate Γ, so its energy dies away as $e^{-\Gamma t}$. A light wave with electric field $E_0 \cos \omega t$ along x drives it. Newton's law is

$$m_e \ddot{x} = -m_e \omega_0^2 x - m_e \Gamma \dot{x} - eE_0 \cos \omega t. \tag{8.2}$$

To find how far it moves, write x as the real part of $\tilde{x}\, e^{-i\omega t}$ and the force as the real part of $-eE_0\, e^{-i\omega t}$.

$$\begin{aligned} -\omega^2 \tilde{x} &\overset{(1)}{=} -\omega_0^2 \tilde{x} + i\omega\Gamma\, \tilde{x} - \frac{e}{m_e} E_0 \\ \Longrightarrow \qquad \tilde{x} &\overset{(2)}{=} -\frac{e}{m_e}\, \frac{E_0}{\omega_0^2 - \omega^2 - i\Gamma\omega}. \end{aligned} \tag{8.3}$$

(1) Each time derivative becomes a factor $-i\omega$, so $\ddot{x} \to -\omega^2 \tilde{x}$ and $\dot{x} \to -i\omega\tilde{x}$. Divide by m_e.
(2) Collect the terms in $\tilde{x}$ and divide.

The cross-section. The moving electron is a dipole $-ex$. The scattering cross-section is the power it radiates divided by the power per unit area of the wave that drives it. Larmor's formula gives the radiated power of a dipole $p_0 \cos \omega t$, averaged over a cycle, as $p_0^2 \omega^4 / (12\pi\varepsilon_0 c^3)$. The wave carries $c\varepsilon_0 E_0^2/2$ per unit area, averaged over a cycle.

$$\begin{aligned} \sigma(\omega) &\overset{(1)}{=} \frac{e^2 |\tilde{x}|^2 \omega^4 / (12\pi\varepsilon_0 c^3)}{c\varepsilon_0 E_0^2/2} \\ &\overset{(2)}{=} \frac{e^4}{6\pi\varepsilon_0^2 m_e^2 c^4}\, \frac{\omega^4}{(\omega_0^2 - \omega^2)^2 + \Gamma^2\omega^2} \\ &\overset{(3)}{=} \frac{8\pi}{3} \left(\frac{e^2}{4\pi\varepsilon_0 m_e c^2} \right)^2 \frac{\omega^4}{(\omega_0^2 - \omega^2)^2 + \Gamma^2\omega^2}. \end{aligned} \tag{8.4}$$

(1) Larmor's formula with dipole amplitude $e|\tilde{x}|$, over the flux of the wave.
(2) $|\tilde{x}|^2$ from Eq. (8.3). The factor E_0^2 cancels.
(3) $6\pi\varepsilon_0^2 = \frac{3}{8\pi}(4\pi\varepsilon_0)^2$.

The constant in front is the **Thomson cross-section** σ_T, written with the **classical electron radius** $r_e = e^2/(4\pi\varepsilon_0 m_e c^2) = 2.818 \times 10^{-15}$ m. So

$$\boxed{\sigma(\omega) = \sigma_T\, \frac{\omega^4}{(\omega_0^2 - \omega^2)^2 + \Gamma^2\omega^2}.} \tag{8.5}$$

The answer depends on the frequency only through ω/ω_0 and Γ/ω_0. The damping is weak, $\Gamma \ll \omega_0$, so three regimes appear.

Slow shaking, Rayleigh scattering. For $\omega \ll \omega_0$ the denominator is ω_0^4 to leading order, and

$$\sigma \approx \sigma_T \left(\frac{\omega}{\omega_0} \right)^4 \propto \frac{1}{\lambda^4}. \tag{8.6}$$

The electrons of air molecules have natural frequencies in the ultraviolet, so visible light is in this regime. Blue light is scattered most, which makes the daytime sky blue and the setting Sun red. This is **Rayleigh scattering**.

Fast shaking, Thomson scattering. For $\omega \gg \omega_0$ the spring no longer matters, and $\sigma \to \sigma_T$. A free electron, with no spring at all ($\omega_0 = 0$), has $\sigma = \sigma_T$ at every frequency. This is the electron of Chapter 5 (Eq. (5.1)), now asked how much it radiates rather than how it bends a wave.

$$\sigma_{\mathrm{T}} = \frac{8\pi}{3}\left(\frac{e^2}{4\pi\varepsilon_0 m_e c^2}\right)^2 = \frac{8\pi}{3} r_e^2 = 6.652 \times 10^{-29}\,\mathrm{m}^2. \tag{8.7}$$

It does not depend on the wavelength, so free electrons scatter all colours alike. **Thomson scattering** is the main opacity of hot ionised gas, in the interiors and envelopes of hot stars (Chapter 11), in accretion disks, and in the early universe before the cosmic microwave background was released. It holds while the photon energy is small compared with $m_e c^2$.

Shaking at resonance, a line. Near ω_0, write $\omega = \omega_0 + \delta$ with $|\delta| \ll \omega_0$.

$$\begin{aligned} \sigma &\overset{(1)}{=} \sigma_T \frac{\omega_0^4}{(2\omega_0\delta)^2 + \Gamma^2\omega_0^2} \\ &\overset{(2)}{=} \sigma_T \frac{\omega_0^2/4}{(\omega-\omega_0)^2 + \Gamma^2/4}. \end{aligned} \tag{8.8}$$

(1) $\omega_0^2 - \omega^2 = (\omega_0 - \omega)(\omega_0 + \omega) \approx -2\omega_0\delta$, and $\omega \approx \omega_0$ elsewhere.
(2) Divide the top and bottom by $4\omega_0^2$.

The cross-section has a sharp peak of full width Γ at half its height, at the frequency where the oscillator absorbs. At the peak it is $\sigma_T(\omega_0/\Gamma)^2$, far larger than σ_T, which is why a thin gas makes strong lines. This peak is an absorption line, and its shape is found again by a second route in Section 8.3.

8.2.2 Scattering by Grains

A grain holds many electrons. How they radiate together depends on the size a of the grain compared with the wavelength. The ratio that decides is $2\pi a/\lambda$ (P6).

Grains much smaller than the wavelength. When $a \ll \lambda$ the field of the wave is the same across the whole grain. A sphere of material with relative permittivity ε_r in a uniform field becomes a dipole

$$\tilde{p} = 4\pi\varepsilon_0 a^3 \frac{\varepsilon_r - 1}{\varepsilon_r + 2} E_0,$$

a standard result of electrostatics. Here ε_r is complex when the material absorbs. All the electrons move in step, so the grain radiates as one dipole. Its scattering cross-section is

$$\begin{aligned} \sigma_{\mathrm{sca}} &\overset{(1)}{=} \frac{|\tilde{p}|^2\omega^4/(12\pi\varepsilon_0 c^3)}{c\,\varepsilon_0 E_0^2/2} \\ &\overset{(2)}{=} \frac{16\pi^2 a^6\,\omega^4}{6\pi c^4}\left|\frac{\varepsilon_r - 1}{\varepsilon_r + 2}\right|^2 \\ &\overset{(3)}{=} \frac{8\pi}{3}\left(\frac{2\pi}{\lambda}\right)^4 a^6 \left|\frac{\varepsilon_r - 1}{\varepsilon_r + 2}\right|^2. \end{aligned} \tag{8.9}$$

(1) Larmor's formula for the dipole $\tilde{p}$, over the flux of the wave, as for the electron.
(2) $|\tilde{p}|^2/E_0^2 = 16\pi^2\varepsilon_0^2 a^6 |(\varepsilon_r - 1)/(\varepsilon_r + 2)|^2$, and ε_0^2 cancels.
(3) $\omega/c = 2\pi/\lambda$.

$$\boxed{\sigma_{\text{sca}} \propto \frac{a^6}{\lambda^4}.} \tag{8.10}$$

The λ^{-4} is the same as for one molecule. The a^6 comes from radiating in step. The number of electrons grows as a^3, the total dipole grows with it, and the power goes as the square. Blue light scattered by small grains makes the blue reflection nebulae around young stars, such as the Pleiades.

The grain also absorbs. The field does work on the dipole at the rate $\frac{1}{2}\omega E_0 \operatorname{Im}\tilde{p}$, averaged over a cycle, so

$$\begin{aligned} \sigma_{\text{abs}} &\overset{(1)}{=} \frac{\frac{1}{2}\omega E_0 \operatorname{Im}\tilde{p}}{c\,\varepsilon_0 E_0^2/2} \\ &\overset{(2)}{=} \frac{2\pi}{\lambda}\, 4\pi a^3 \operatorname{Im}\frac{\varepsilon_r - 1}{\varepsilon_r + 2} \;\propto\; \frac{a^3}{\lambda}. \end{aligned} \tag{8.11}$$

(1) Power taken from the wave, over its flux.
(2) The dipole $\tilde{p}$ above, and $\omega/c = 2\pi/\lambda$.

For small grains absorption beats scattering, since $\sigma_{\text{sca}}/\sigma_{\text{abs}} \propto (a/\lambda)^3$.

Grains near the wavelength, Mie scattering. When $2\pi a/\lambda$ is near one, the field changes across the grain and its parts radiate out of step. The exact answer, found by Mie, expands the fields inside and outside the sphere in vector spherical harmonics. Over this range the cross-section of a grain grows roughly as its area times a/λ, so $\sigma_{\text{ext}} \propto a^3/\lambda$, the same form as Eq. (8.11). Interstellar grains have $a \approx 0.1\,\mu$m, so $2\pi a/\lambda = 1.1$ at 550 nm. Visible light is in the **Mie** regime, and the mix of grain sizes gives the measured extinction that falls roughly as $1/\lambda$ (Figure 4.1).

Grains much larger than the wavelength. When $a \gg \lambda$ the grain casts a sharp shadow and blocks its own area πa^2. It also bends light out of the beam by diffraction at its edge. By Babinet's principle a disc diffracts as much power as a hole of the same size lets through, another πa^2. So $\sigma_{\text{ext}} \to 2\pi a^2$, the same for every colour. Water droplets in clouds and fog are this large, which is why clouds are white or grey. Large grains in the shells around old stars behave the same way.

Tyndall scattering. Scattering by fine particles suspended in a gas or a liquid, in the range from Rayleigh to Mie, is called **Tyndall scattering**. It makes a sunbeam visible in a dusty room or a smoky forest. Dusty interstellar clouds lit by nearby stars shine in the same way.

8.2.3 Scattering That Changes the Colour

In all the cases above the photon leaves with the frequency it came with. This is **elastic** scattering. In **inelastic** scattering the photon trades energy with the scatterer and leaves with a new frequency.

Raman scattering. The photon excites or de-excites a vibration or rotation of a molecule or atom. It leaves with lower frequency if it gave energy (Stokes) and with higher frequency if it took energy from an already excited molecule (anti-Stokes). In the laboratory this identifies chemical bonds. In symbiotic stars and some planetary nebulae, ultraviolet photons of O VI at 103.2 and 103.8 nm are Raman-scattered by neutral hydrogen into broad bands at 682.5 and 708.2 nm.

Brillouin scattering. Light scatters off sound waves in a dense fluid or solid. The sound wave acts as a moving grating, so the scattered light is Doppler shifted by $\Delta\nu = \pm 2\nu_0 (c_s/c)\sin(\psi/2)$, times the refractive index of the medium, where c_s is the sound speed and ψ the scattering angle.

Compton and inverse Compton scattering. When the photon energy approaches $m_e c^2$, the electron recoils and the photon loses energy. Its wavelength grows by $\Delta\lambda = \lambda_C(1 - \cos\psi)$ (derived in Eq. (15.29)), with the Compton wavelength $\lambda_C = h/(m_e c) = 2.426 \times 10^{-12}$ m. In **inverse Compton** scattering the electron moves near the speed of light, with Lorentz factor $\gamma_e \gg 1$, and gives energy to a low-energy photon of the radio, infrared or microwave background. The photon leaves with about γ_e^2 times its energy, in X-rays or gamma rays. This lights up pulsar wind nebulae, supernova remnants and the jets of active galaxies.

Table 8.2 collects the processes.

Process	Scatterer	Dependence	Photon energy	Where seen
Rayleigh	Atoms, molecules, small grains ($a \ll \lambda$)	$\sigma \propto a^6 \lambda^{-4}$	kept	Blue sky, Pleiades reflection nebula
Mie	Dust grains ($a \sim \lambda$)	$\sigma \propto a^3 \lambda^{-1}$	kept	Interstellar extinction and reddening
Geometric	Large particles ($a \gg \lambda$)	$\sigma \to 2\pi a^2$, no colour	kept	White clouds, dust shells of old stars
Tyndall	Suspended fine particles	between Rayleigh and Mie	kept	Sunbeams, dusty clouds lit by stars
Raman	Molecular and atomic levels	at set energy steps	changed	Symbiotic stars, Raman O VI bands
Brillouin	Sound waves in dense media	shift $\propto c_s/c$	changed	Laboratory solids and liquids
Thomson	Free electrons ($h\nu \ll m_e c^2$)	σ_T, no colour	kept	Hot stars, accretion disks, early universe
Inverse Compton	Fast electrons ($\gamma_e \gg 1$)	$\nu \to \sim \gamma_e^2 \nu$	raised	Jets of active galaxies, pulsar wind nebulae

Table 8.2: Scattering processes, the scatterer, how the cross-section or the shift depends on size and wavelength, whether the photon energy changes, and where each is seen.

8.2.4 Dust Extinction and Reddening

Chapter 4 turned the cross-section into magnitudes. The extinction is $A_\lambda = 1.086\,\tau_\lambda = 1.086\,N_d\,\sigma_{\rm ext}(\lambda)$ (Eq. (4.36)). Because $\sigma_{\rm ext}$ is larger in B than in V, dust makes a star redder by the colour excess $E(B-V) = A_B - A_V$, and in the Milky Way $A_V = 3.1\,E(B-V)$ (Eq. (4.42)). The intrinsic colour and the distance modulus corrected for dust follow from Eqs. (4.40) and (4.44). What is new here is what the grains themselves tell us.

Extinction per kilogram of dust. To see why one per cent of the mass does almost all the dimming, write the extinction in terms of the mass of dust along the line of sight. Let each grain be a sphere of radius a and density $\rho_{\rm grain}$, and let Σ_d be the mass of dust in a column of cross-section $1\,\mathrm{m}^2$.

$$\begin{aligned} A_\lambda &\overset{(1)}{=} 1.086\,N_d\,\sigma_{\rm ext} \\ &\overset{(2)}{=} 1.086\,\Sigma_d\,\frac{\sigma_{\rm ext}}{\frac{4}{3}\pi a^3 \rho_{\rm grain}} \\ &\overset{(3)}{\approx} 1.086\,\Sigma_d\,\frac{3}{4a\,\rho_{\rm grain}}. \end{aligned} \tag{8.12}$$

(1) Eq. (4.36).
(2) N_d grains each of mass $\frac{4}{3}\pi a^3 \rho_{\rm grain}$ make up Σ_d.
(3) In the Mie regime the cross-section of a grain is close to its area, $\sigma_{\rm ext} \approx \pi a^2$.

$$\boxed{A_\lambda \approx 1.086\,\frac{3\,\Sigma_d}{4\,a\,\rho_{\rm grain}}.} \tag{8.13}$$

Per kilogram, a grain blocks light in proportion to $1/(a\rho_{\rm grain})$. Breaking a kilogram of matter into grains of $0.1\,\mu$m gives it an enormous total area. The same kilogram as gas atoms scatters

almost nothing at visible wavelengths, because each atom is in the Rayleigh regime, with a cross-section below σ_T by Eq. (8.6).

A check against the measured dust. Take grains of radius 0.1 μm and density 3000 kg m^{-3}, and one per cent of the mass in dust. With hydrogen making up 70 per cent of the mass, each hydrogen atom comes with $0.01\, m_{\mathrm{H}}/0.70 = 2.4\times10^{-29}$ kg of dust. Each grain has mass 1.26×10^{-17} kg, so there are 1.9×10^{-12} grains per hydrogen atom. Each blocks $\pi a^2 = 3.1\times10^{-14}$ m^2, so per hydrogen atom in the column

$$\frac{A_V}{N_{\mathrm{H}}} \approx 1.086\times1.9\times10^{-12}\times3.1\times10^{-14}\,\mathrm{m}^2 = 6.5\times10^{-26}\,\mathrm{mag\,m}^2,$$

where N_{H} is the column density of hydrogen atoms. Ultraviolet absorption lines of hydrogen toward reddened stars give $N_{\mathrm{H}}/E(B-V) = 5.8\times10^{25}\,\mathrm{m}^{-2}$ per magnitude, and with $A_V = 3.1\,E(B-V)$ that is $A_V/N_{\mathrm{H}} = 5.3\times10^{-26}\,\mathrm{mag\,m}^2$. The simple estimate is right to within 25 per cent. In the units of the field, one magnitude of A_V needs a column of 1.9×10^{21} hydrogen atoms per cm^2.

The extinction curve. Measured toward many stars, the extinction rises from the infrared through the visible into the ultraviolet. At the long-wavelength end it follows the rough $1/\lambda$ of Mie scattering. Near 217.5 nm there is a broad bump, carried by grains that contain carbon, in the form of graphite or of large molecules called polycyclic aromatic hydrocarbons. The curve differs a little from one line of sight to another, as the mix of grains does.

Dark clouds seen in the infrared. At 2.2 μm, $2\pi a/\lambda = 0.29$ for the same grains. They are close to the Rayleigh regime and block only about a tenth as much as in V. A cloud such as Barnard 68, black against the stars in visible light, is nearly transparent in the near infrared, and the stars behind it reappear, reddened.

8.2.5 Free Electrons Delay Radio Waves

Free electrons do not only scatter light. They also slow it, by an amount that depends on its frequency. In Chapter 5 the free electrons of a plasma gave the refractive index $n^2 = 1 - \nu_p^2/\nu^2$ (Eq. (5.1)), with the plasma frequency ν_p of Eq. (5.2). In interstellar gas with $n_e = 3\times10^4\,\mathrm{m}^{-3}$ (0.03 cm^{-3}), $\nu_p = 1.6$ kHz, far below any radio frequency used. The wave passes, but a pulse does not travel at c.

The speed of a pulse. A pulse moves at the group velocity $d\omega/dk$, with k the wavenumber and $\omega = 2\pi\nu$.

$$\begin{aligned}\omega^2 &\overset{(1)}{=} \frac{c^2k^2}{n^2}\\ &\overset{(2)}{=} \omega_p^2 + c^2k^2,\\ \frac{\mathrm{d}\omega}{\mathrm{d}k} &\overset{(3)}{=} \frac{c^2k}{\omega} \overset{(4)}{=} c\sqrt{1-\frac{\nu_p^2}{\nu^2}}.\end{aligned} \tag{8.14}$$

(1) The phase velocity is $\omega/k = c/n$.
(2) $n^2 = 1 - \omega_p^2/\omega^2$, with $\omega_p = 2\pi\nu_p$, multiplied by ω^2 and rearranged.
(3) Differentiate the second line, $2\omega\,d\omega = 2c^2k\,dk$.
(4) $ck/\omega = n$.

The delay. A pulse that crosses a path of length d to the source arrives after

$$t \overset{(1)}{=} \int_0^d \frac{ds}{c}\left(1 - \frac{\nu_p^2}{\nu^2}\right)^{-1/2}$$
$$\overset{(2)}{\approx} \frac{d}{c} + \frac{1}{2c\nu^2}\int_0^d \nu_p^2\, ds$$
$$\overset{(3)}{=} \frac{d}{c} + \frac{e^2}{8\pi^2\varepsilon_0 m_e c}\,\frac{1}{\nu^2}\int_0^d n_e\, ds. \tag{8.15}$$

(1) Time is $\int ds$ over the group velocity.
(2) $(1-y)^{-1/2} \approx 1 + y/2$ for $y = \nu_p^2/\nu^2 \ll 1$.
(3) $\nu_p^2 = n_e e^2/(4\pi^2\varepsilon_0 m_e)$ from Eq. (5.2).

The column of electrons $\int n_e\, ds$ is called the **dispersion measure**, and is usually quoted in $\mathrm{pc\,cm^{-3}}$. The extra delay is

$$\boxed{\Delta t = \frac{e^2}{8\pi^2\varepsilon_0 m_e c}\,\frac{\int n_e\, ds}{\nu^2} = 4.15\,\mathrm{ms} \times \frac{\int n_e\, ds}{1\,\mathrm{pc\,cm^{-3}}}\left(\frac{\nu}{1\,\mathrm{GHz}}\right)^{-2}.} \tag{8.16}$$

Low frequencies arrive later. A pulsar 1 kpc away through gas with 0.03 electrons per cm^3 has a dispersion measure of $30\,\mathrm{pc\,cm^{-3}}$. Its pulses arrive 0.78 s late at 400 MHz, and 0.58 s later than at 800 MHz. Timing a pulse at two frequencies measures the column of electrons, and with a model of the electron density it gives the distance. The column changes as the pulsar and the gas move, which is the dispersion noise of pulsar timing (Section 7.6).

8.2.6 Fast Radio Bursts

In 2007 Lorimer and colleagues searched radio data taken at the Parkes telescope in 2001 and found a single flash of radio waves, a few milliseconds long, near 1.4 GHz. It never repeated. The flash reached the top of the band first and the bottom about 0.4 s later. Many more such flashes have been found since, and they are called **fast radio bursts**.

Reading the sweep. The receiver splits the band into many narrow frequency channels and records the arrival time in each. By Eq. (8.16) the time at which frequency ν arrives is

$$t(\nu) = t_\infty + 4.15\,\mathrm{ms} \times \frac{\int n_e\, ds}{1\,\mathrm{pc\,cm^{-3}}}\left(\frac{\nu}{1\,\mathrm{GHz}}\right)^{-2}, \tag{8.17}$$

where t_∞ is the time at which a very high frequency, which the electrons hardly delay, would arrive. There are two unknowns, t_∞ and the dispersion measure. Plotted against ν^{-2}, the arrival times lie on a straight line. Its slope is 4.15 ms times the dispersion measure, and its intercept is t_∞. With many channels the fit has far more points than unknowns, so it also tests the law itself. The ν^{-2} comes from the refractive index of free electrons, Eq. (5.1), and from nothing else. A delay made at the source would have no reason to follow it. The measured times do follow it, so the sweep was made on the way, by the electrons between the source and us.

The first burst. The fit gives $375\,\mathrm{pc\,cm^{-3}}$ for the Lorimer burst. Between the top and the bottom of the band the delay is

$$t(1.2\,\mathrm{GHz}) - t(1.5\,\mathrm{GHz}) = 4.15\,\mathrm{ms} \times 375 \times \left(\frac{1}{1.2^2} - \frac{1}{1.5^2}\right) = 389\,\mathrm{ms},$$

the curve of Figure 8.1. The pulsar 1 kpc away of Section 8.2.5 has only 30 pc cm^{-3}. Models of the electrons in the Milky Way give about 25 pc cm^{-3} through the whole Galaxy in the direction of the burst. Most of the electrons that delayed it lie outside the Galaxy, so the burst came from far beyond it.

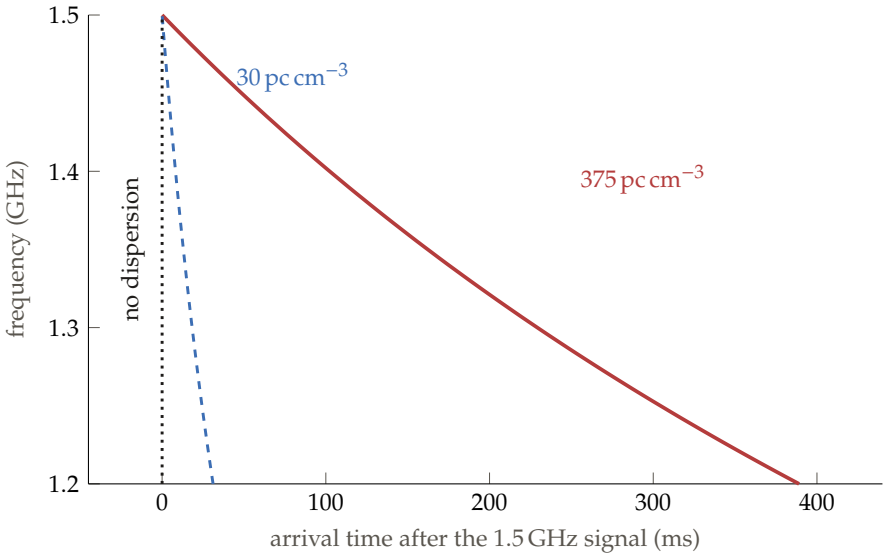


Figure 8.1: Frequency against arrival time from Eq. (8.17), with time measured from the arrival at 1.5 GHz. Solid, a dispersion measure of 375 pc cm^{-3}. Dashed, 30 pc cm^{-3}. Dotted, no dispersion, with every frequency arriving at once.

Electrons between the galaxies. To turn the dispersion measure into a distance we need the density of free electrons between the galaxies. Suppose all the ordinary matter is ionised gas spread evenly through space. Its mass density today is $\Omega_b \rho_{\text{crit},0}$ (Eq. (16.7)). A proton and a neutron each have nearly the mass m_{H}, so this is $\Omega_b \rho_{\text{crit},0}/m_{\text{H}}$ protons and neutrons per unit volume, called baryons. A fraction $1 - Y$ of them is in hydrogen, each with one electron. The rest, Y, is in helium, whose nucleus holds four baryons and brings two electrons, half an electron per baryon. The gas is neutral overall, so the electrons are counted by the charges of the nuclei (P1, conserved charge and baryon number).

$$\begin{aligned} n_{e,0} &\overset{(1)}{=} \left((1-Y)\cdot 1 + Y\cdot \tfrac{1}{2}\right)\frac{\Omega_b\,\rho_{\text{crit},0}}{m_{\text{H}}} \\ &\overset{(2)}{=} \left(1-\frac{Y}{2}\right)\frac{\Omega_b\,\rho_{\text{crit},0}}{m_{\text{H}}} \\ &\overset{(3)}{=} \left(1-\frac{Y}{2}\right)\frac{3\,\Omega_b H_0^2}{8\pi G\,m_{\text{H}}}. \end{aligned} \tag{8.18}$$

(1) Electrons per baryon times baryons per unit volume. The bracket is $1/\mu_e$ of Eq. (12.14) with no heavier elements, where the hydrogen fraction is written $1 - Y$.
(2) Multiply out the bracket.
(3) $\rho_{\text{crit},0} = 3H_0^2/(8\pi G)$, Eq. (16.7).

$$\boxed{n_{e,0} = \left(1-\frac{Y}{2}\right)\frac{3\,\Omega_b H_0^2}{8\pi G\,m_{\text{H}}} = 0.22\,\text{m}^{-3}.} \tag{8.19}$$

The numbers are those of Chapter 16. With $H_0 = 67.4\,\text{km s}^{-1}\,\text{Mpc}^{-1} = 2.184\times10^{-18}\,\text{s}^{-1}$, the critical density is $8.53\times10^{-27}\,\text{kg m}^{-3}$. With $\Omega_b = 0.049$ there are 0.250 baryons per cubic metre,

as in Problem 16.8.2. With the helium fraction $Y = 0.245$ of Section 16.4 there are 0.878 electrons per baryon, so $n_{e,0} = 0.219\,\mathrm{m}^{-3} = 2.19 \times 10^{-7}\,\mathrm{cm}^{-3}$. That is a hundred thousand times thinner than the ionised gas of the Galaxy.

Dispersion measure and redshift. For a source at small redshift, $z \ll 1$, the electron density along the path hardly differs from $n_{e,0}$, and the path is the distance d to the source. Hubble's law gives d from z (P6, a length compared with c/H_0).

$$\begin{aligned} \int n_e \, ds &\overset{(1)}{\approx} n_{e,0}\, d \\ &\overset{(2)}{=} n_{e,0}\, \frac{cz}{H_0} \\ &\overset{(3)}{=} \left(1 - \frac{Y}{2}\right) \frac{3\,c\,\Omega_b H_0}{8\pi G\, m_\mathrm{H}}\, z. \end{aligned} \tag{8.20}$$

(1) The density was higher in the past by $(1+z)^3$, which is close to 1 for $z \ll 1$.
(2) $d = v/H_0$ with $v = cz$, Eq. (16.2) and the small-distance form of Eq. (16.4).
(3) $n_{e,0}$ from Eq. (8.19).

$$\boxed{\int n_e \, ds \approx \left(1 - \frac{Y}{2}\right) \frac{3\,c\,\Omega_b H_0}{8\pi G\, m_\mathrm{H}}\, z = 975\, z\ \mathrm{pc\,cm}^{-3}.} \tag{8.21}$$

The slope is $n_{e,0} = 2.19 \times 10^{-7}\,\mathrm{cm}^{-3}$ times $c/H_0 = 4448\,\mathrm{Mpc} = 4.448 \times 10^9\,\mathrm{pc}$. It depends only on $\Omega_b H_0$ and on the helium fraction. It counts every free electron, whatever the temperature of its gas. For the Lorimer burst, the $375 - 25 = 350\,\mathrm{pc\,cm}^{-3}$ from outside the Galaxy gives $z \approx 350/975 = 0.36$. This is a rough value, uncertain both ways. The burst's own galaxy and the halo of ours add electrons, which leaves fewer for the space between and lowers z. Some of the ordinary matter sits in stars and galaxies rather than between them, which makes the slope smaller than 975 and raises z. At $z = 0.36$ the small-z form is also only rough.

At larger redshift the density was higher by $(1+z)^3$. The delay made in a piece of path at redshift z is also smaller, because the wave there had the higher frequency $\nu(1+z)$, which cuts the delay by $(1+z)^2$, and the delay is then stretched by $1+z$ on the way to us (Eq. (16.4)). Each piece of path therefore counts as $n_e\, ds/(1+z)$, and the integral of this along the path, as a function of z, is called the **Macquart relation**. In 2020 a handful of bursts whose host galaxies, and so redshifts, were known were found to follow it, with the slope expected if the ordinary matter that surveys of stars and gas had not found lies as ionised gas between the galaxies.

What makes them. Most bursts have been seen only once, but some repeat, so at least those are not one-off explosions. A flash a few milliseconds long comes from a region no larger than light crosses in that time (Eq. (15.59) with $\mathcal{D} = 1$), about 1000 km. In 2020 a burst came from SGR 1935+2154, a **magnetar** in our own Galaxy, which is a neutron star (Section 13.4) with an unusually strong magnetic field. At least some fast radio bursts come from magnetars.

8.2.7 The Gas Between the Galaxies

The space between the galaxies holds gas too, the **intergalactic medium**. Its density is of order one atom per cubic metre, against about 10^6 per cubic metre ($1\,\mathrm{cm}^{-3}$) in the interstellar medium. The mean over all space is 0.25 baryons per cubic metre (Eq. (8.19)). Gas gathered into the sheets and filaments of the cosmic web (Section 16.5) is denser than that mean, which brings it to about one per cubic metre. Its temperature runs from 10^4 to 10^8 K, hotter than most of the interstellar medium (Table 8.1). At the low end is the gas of the Lyman-α forest, near 10^4 K, held there by the heating of ultraviolet light from galaxies and quasars (Section 16.6).

Why the hot gas is hot. The hotter gas was heated by gravity. Gas falling into a group, a cluster or a filament at a few hundred km s^{-1} is stopped by shocks, which turn the energy of the bulk motion into heat, as shocks do in the interstellar medium (Section 8.6). The temperature is of order $k_B T \approx \mu m_H v^2$, the virial temperature of Eq. (10.13), as for the gas in clusters (Section 14.4). For $v = 300$ km s^{-1} and $\mu = 0.6$ this is 6.5×10^6 K. In the deep wells of clusters, with 1000 km s^{-1}, it is near 10^8 K.

Why it stays hot. Gas cools when collisions excite its atoms or deflect its electrons and the light they send out escapes. Each such loss needs two particles to meet, so the power lost per unit volume grows as n^2, as the factor $n_c\, n$ in the cooling balance of Section 8.6 shows. The heat stored per unit volume grows only as n. The cooling time, heat over power, therefore goes as $1/n$, and at a millionth of the density of the interstellar medium the gas cools a million times more slowly at the same temperature. For hot ionised gas the bremsstrahlung power of Eq. (15.16) gives a cooling time that grows as $\sqrt{T}/n$. Scaling the cluster value of 2.4×10^{18} s at $n = 10^3$ m^{-3} and 7×10^7 K to $n = 1$ m^{-3} and 10^6 K,

$$t_{\text{cool}} = 2.4 \times 10^{18}\ \text{s} \times \frac{10^3}{1} \times \sqrt{\frac{10^6}{7 \times 10^7}} = 2.9 \times 10^{20}\ \text{s} = 9 \times 10^{12}\ \text{years},$$

about 650 times the age of the universe, 1.38×10^{10} years (Eq. (16.13)). This counts bremsstrahlung only. Below about 10^7 K the lines of the few heavier atoms cool the gas faster, but not by enough to close a gap this large for gas near 10^6 K. The gas has kept most of the heat it gained when it fell in.

How it is seen. Four probes see the intergalactic medium, and each sees a different part of it (Table 8.3). The 21 cm line (Section 8.4.1) and the Lyman-α forest (Section 16.6) see only neutral hydrogen, a tiny fraction of the gas. The X-rays of bremsstrahlung (Eq. (15.16)) need gas both hot and dense enough, and the brightest is the intracluster medium (Section 14.4). The dispersion of fast radio bursts counts every free electron along the path, in proportion to n_e and whatever the temperature. Gas between about 10^5 and 10^7 K in thin filaments has almost no neutral hydrogen and is too thin to shine in X-rays. Only the dispersion measure counts it in full.

Probe	Gas it sees	Signal grows as	Where derived
21 cm emission	neutral hydrogen	column of HI	Section 8.4.1
Lyman-α forest	neutral traces in ionised gas near 10^4 K	$n_{\text{HI}} \propto n^2$	Section 16.6
Thermal X-rays	hot gas, above about 10^7 K	$n^2\sqrt{T}$	Eq. (15.16)
Burst dispersion	every free electron	n_e	Eq. (8.21)

Table 8.3: The four probes of the intergalactic medium, the gas each one sees, how its signal grows with the density n and the temperature T, and where it is derived.

8.3 Spectral Lines and Their Widths

The oscillator of Section 8.2.1 absorbs in a narrow peak near its natural frequency, and Chapter 3 showed that motion along the line of sight shifts a line (Eq. (3.25)). A real line has a width and a shape, and both carry information. The width tells the temperature of the gas, how it churns, and how dense it is. To read them we need the shape each cause produces.

Strike a bell and let it ring. It rings for a long time, and the note is pure. Touch it with a hand while it rings and the sound stops quickly, and what you hear is a dull clunk with no clear pitch.

A short ring does not have a sharp pitch. Now load many identical bells on a merry-go-round and strike them all. Those coming toward you sound higher and those going away sound lower, and together they make a blurred chord. An atom in an excited state is a bell. It radiates for about its lifetime, so its line has a width set by how long it rings. Collisions that interrupt it shorten the ring and widen the line. The atoms of a gas move at random, so the lines of different atoms are Doppler shifted by different amounts and blur together.

Hydrogen's Lyman α line joins its two lowest levels, $n = 1$ and $n = 2$. Their energy difference is $13.6\,\text{eV} \times (1 - \frac{1}{4}) = 10.2\,\text{eV}$, corresponding to 121.6 nm ultraviolet light. The upper level decays at $6.27 \times 10^8\,\text{s}^{-1}$, giving a lifetime width of 1.0×10^8 Hz. In 100 K gas, random atomic motion broadens the line over 1.8×10^{10} Hz, 176 times more. Motion shapes the central part of the profile, while the finite lifetime remains important far from line centre. We will derive those shapes separately, add the effect of collisions, and then combine them.

Formation. An atom or ion has only electronic transitions, from one arrangement of its electrons to another. A molecule can also rotate and vibrate, so its spectrum is much richer. An excited atom or molecule that drops to a lower level by itself sends out a photon, and many such photons make an **emission line**. A cool gas in front of a hotter source of continuous light takes photons out of it at the same frequencies, and makes an **absorption line**. The rates of these processes, and when each kind of line appears, are worked out in Chapter 11 (Section 11.6).

The line profile. Let ν_0 be the frequency at the centre of the line. The **line profile** $\phi(\nu)$ is the fraction of the line's photons per unit frequency, normalised so that $\int \phi(\nu)\,d\nu = 1$. Its width is measured as the full width at half the maximum.

8.3.1 Natural Broadening

The width from the uncertainty principle. Let the upper level decay at the rate Γ, so that it lives for a time $\Delta t = 1/\Gamma$. By the time–energy uncertainty relation, a state that lasts for Δt has an energy spread of about

$$\Delta E\,\Delta t \sim \hbar. \tag{8.22}$$

With $\Delta E = h\,\Delta\nu$ this is a spread in frequency,

$$\Delta\nu \sim \frac{1}{2\pi\,\Delta t} = \frac{\Gamma}{2\pi}. \tag{8.23}$$

Strongly allowed lines decay fast and are broad. Forbidden lines decay slowly and are narrow. If the lower level also decays, as for a line between two excited levels, Γ is the sum of the two rates.

The shape, from a decaying wave. The uncertainty relation gives only the width. To get the shape, treat the radiating atom as an oscillator that is excited at $t = 0$ and then rings down. Its energy decays as $e^{-\Gamma t}$, so the field it sends out decays at half that rate,

$$E(t) = E_0\,e^{-\Gamma t/2}\,e^{-i\omega_0 t} \quad (t \ge 0), \qquad E(t) = 0 \quad (t < 0).$$

The spectrum is the squared size of its Fourier transform.

$$\begin{aligned} \tilde{E}(\omega) &\overset{(1)}{=} \int_0^\infty E_0\,e^{-\Gamma t/2}\,e^{i(\omega-\omega_0)t}\,dt \\ &\overset{(2)}{=} \frac{E_0}{\Gamma/2 - i(\omega-\omega_0)}, \\ |\tilde{E}(\omega)|^2 &\overset{(3)}{=} \frac{E_0^2}{(\omega-\omega_0)^2 + \Gamma^2/4}. \end{aligned} \tag{8.24}$$

(1) The Fourier transform, $\int E(t)\, e^{i\omega t}\, dt$, over the times when the wave exists.
(2) $\int_0^\infty e^{-st} dt = 1/s$ with $s = \Gamma/2 - i(\omega - \omega_0)$, whose real part is positive.
(3) $|x + iy|^2 = x^2 + y^2$.

In terms of $\nu = \omega/2\pi$, and normalised to unit area, this is the **Lorentzian profile**,

$$\boxed{\phi_L(\nu) = \frac{\Gamma/4\pi^2}{(\nu - \nu_0)^2 + (\Gamma/4\pi)^2}.} \tag{8.25}$$

It falls to half its peak at $\nu - \nu_0 = \pm\Gamma/4\pi$, so its full width at half maximum is $\Gamma/2\pi$, the uncertainty estimate (8.23) exactly. The driven oscillator gave the same shape and width near resonance, Eq. (8.8), a second route to the same result. Far from the centre ϕ_L falls only as $(\nu - \nu_0)^{-2}$. These slow wings are the mark of an exponential decay. For Lyman α, $\Gamma = 6.27 \times 10^8\,\mathrm{s}^{-1}$ gives a width of 1.0×10^8 Hz, or 4.9×10^{-6} nm in wavelength.

8.3.2 Pressure Broadening

A collision with a neighbour disturbs the levels of an emitting atom and resets the phase of its wave at random. The wave before and after a collision no longer add in step. A collision cuts the ring short, like the hand on the bell.

The shape. Let collisions strike one atom at random times, at the mean rate $1/t_{\mathrm{col}} = n\,\sigma_{\mathrm{col}}\,\bar{v}$, where n is the density of colliders, σ_{col} the cross-section for a collision and $\bar{v}$ the mean relative speed. The chance of no collision in a time t is $e^{-t/t_{\mathrm{col}}}$, the zero-count probability of the Poisson distribution (Section 7.2). Averaged over many atoms, the part of the wave that is still in step with its start after a time t has decayed by $e^{-\Gamma t/2}$ from the lifetime and by $e^{-t/t_{\mathrm{col}}}$ from collisions. The spectrum of a steady random signal is the Fourier transform of this average, by the Wiener–Khinchin theorem. The transform is that of the previous chain with $\Gamma/2$ replaced by $\Gamma/2 + 1/t_{\mathrm{col}}$. So pressure broadening also gives a Lorentzian, Eq. (8.25) with

$$\boxed{\Gamma \;\to\; \Gamma + \frac{2}{t_{\mathrm{col}}}, \qquad \text{full width} = \frac{1}{2\pi}\left(\Gamma + \frac{2}{t_{\mathrm{col}}}\right).} \tag{8.26}$$

Denser and hotter gas has more frequent collisions and broader lines.

Where it matters. Take a collision cross-section of order $10^{-19}\,\mathrm{m}^2$, a few times the area of an atom. In interstellar gas with $n = 10^6\,\mathrm{m}^{-3}$ and $\bar{v} \approx 2\,\mathrm{km\,s}^{-1}$, $1/t_{\mathrm{col}} \approx 2 \times 10^{-10}\,\mathrm{s}^{-1}$, far below $\Gamma \approx 10^8\,\mathrm{s}^{-1}$ of a strong line. Pressure broadening does not matter there. In the Sun's photosphere, with $n \approx 10^{23}\,\mathrm{m}^{-3}$ and $\bar{v} \approx 10\,\mathrm{km\,s}^{-1}$, $1/t_{\mathrm{col}} \approx 10^8\,\mathrm{s}^{-1}$, comparable with Γ. Lines of dense dwarf stars are broader than those of thin giant atmospheres for this reason, which is how a spectrum tells a dwarf from a giant of the same temperature.

8.3.3 Doppler Broadening

The velocities of the atoms. An atom moving away from us at line-of-sight velocity v_r has its line shifted by Eq. (3.25). In gas at temperature T, the probability of an energy $\frac{1}{2}mv_r^2$ in this one direction carries the Boltzmann weight $e^{-mv_r^2/2k_BT}$, where m is the mass of the atom (P3). Normalised, this is the one-dimensional Maxwell distribution,

$$f(v_r)\, dv_r = \sqrt{\frac{m}{2\pi k_B T}}\, \exp\!\left(-\frac{mv_r^2}{2k_BT}\right) dv_r, \qquad \sigma_v \equiv \sqrt{\langle v_r^2 \rangle} = \sqrt{\frac{k_BT}{m}}. \tag{8.27}$$

It is one component of the Maxwell distribution derived by counting states in Section 9.3, which also gives $\langle v_r^2 \rangle = k_BT/m$ and the ideal gas (Eq. (9.28)).

The shape. Each atom gives a line much narrower than the spread of shifts, so the profile is the velocity distribution written in frequency. A photon from an atom with velocity v_r arrives at $\nu = \nu_0/(1 + v_r/c) \approx \nu_0(1 - v_r/c)$, so $v_r = -c(\nu - \nu_0)/\nu_0$ and $|dv_r| = (c/\nu_0)\, d\nu$.

$$\begin{aligned} \phi_D(\nu)\, d\nu &\overset{(1)}{=} f(v_r)\, |dv_r| \\ &\overset{(2)}{=} \sqrt{\frac{m}{2\pi k_B T}}\, \exp\left[-\frac{mc^2(\nu-\nu_0)^2}{2k_B T\, \nu_0^2}\right] \frac{c}{\nu_0}\, d\nu \\ &\overset{(3)}{=} \frac{1}{\sqrt{2\pi}\, \Delta\nu_D}\, \exp\left[-\frac{(\nu-\nu_0)^2}{2\,\Delta\nu_D^2}\right] d\nu. \end{aligned} \tag{8.28}$$

(1) The fraction of photons in $d\nu$ is the fraction of atoms with velocities in the matching dv_r.
(2) Eq. (8.27) with $v_r = -c(\nu - \nu_0)/\nu_0$.
(3) Define $\Delta\nu_D = \nu_0\sigma_v/c$. Then $mc^2/(k_B T\nu_0^2) = 1/\Delta\nu_D^2$ and $\sqrt{m/2\pi k_B T}\; c/\nu_0 = 1/(\sqrt{2\pi}\Delta\nu_D)$.

This is the **Gaussian profile**,

$$\boxed{\phi_D(\nu) = \frac{1}{\sqrt{2\pi}\, \Delta\nu_D}\, \exp\left[-\frac{(\nu-\nu_0)^2}{2\,\Delta\nu_D^2}\right],} \tag{8.29}$$

with the **Doppler width** $\Delta\nu_D$ as its standard deviation,

$$\frac{\Delta\nu_D}{\nu_0} = \frac{\sigma_v}{c} = \sqrt{\frac{k_B T}{mc^2}}. \tag{8.30}$$

It falls to half its peak where $(\nu - \nu_0)^2/2\Delta\nu_D^2 = \ln 2$, so its full width at half maximum is $2\sqrt{2\ln 2}\,\Delta\nu_D = 2.355\,\Delta\nu_D$. Its wings fall as $e^{-(\nu-\nu_0)^2/2\Delta\nu_D^2}$, much faster than the Lorentzian, because fast atoms are exponentially rare. The width grows as $\sqrt{T/m}$. Hydrogen at 100 K has $\sigma_v = 0.91\,\mathrm{km\,s^{-1}}$. The H$\alpha$ line at 656.28 nm from gas at 10^4 K has a full width of $21\,\mathrm{km\,s^{-1}}$, or 0.047 nm.

Turbulence and bulk motion. The gas also moves in eddies and streams. If the line-of-sight velocity of an atom is a thermal part plus an independent turbulent part with spread σ_{turb}, the variances add (Eq. (7.10)),

$$\sigma_v^2 = \frac{k_B T}{m} + \sigma_{\mathrm{turb}}^2.$$

When the turbulent part is also Gaussian, the profile stays Gaussian with this larger σ_v. The thermal part depends on the mass of the atom and the turbulent part does not. Comparing the widths of a light and a heavy atom from the same gas therefore separates the temperature from the turbulence. A steady motion of the whole cloud, such as Galactic rotation, shifts the centre ν_0 and does not widen the line.

8.3.4 The Voigt Profile

A real line has all the causes at once. Each atom has a Lorentzian line from its lifetime and collisions, centred on its own Doppler-shifted frequency. The atoms have the Gaussian spread of those centres.

The convolution. Adding up the atoms, weighted by Eq. (8.29),

$$\boxed{\phi_V(\nu) = \int_{-\infty}^{\infty} \phi_D(\nu')\,\phi_L(\nu - \nu' + \nu_0)\,d\nu',} \tag{8.31}$$

where ν' is the centre of one group of atoms and $\phi_L(\nu - \nu' + \nu_0)$ is the Lorentzian (8.25) moved to that centre. This is the **Voigt profile**. It has no closed form, but both of its limits do.

The core. In interstellar gas the Lorentzian half-width $\Gamma/4\pi$ is far smaller than $\Delta\nu_D$. Near the centre, ϕ_L is then a narrow spike of unit area, and

$$\phi_V(\nu) \approx \phi_D(\nu) \int \phi_L\,d\nu' = \phi_D(\nu).$$

The core of the line is Gaussian.

The wings. Far from the centre, $|\nu - \nu_0| \gg \Delta\nu_D$. The Gaussian is non-zero only for ν' within a few $\Delta\nu_D$ of ν_0, and across that small range $\phi_L(\nu - \nu' + \nu_0)$ hardly changes.

$$\begin{aligned}\phi_V(\nu) &\overset{(1)}{\approx} \phi_L(\nu) \int \phi_D(\nu')\,d\nu' \\ &\overset{(2)}{=} \phi_L(\nu) \overset{(3)}{\approx} \frac{\Gamma/4\pi^2}{(\nu - \nu_0)^2}.\end{aligned} \tag{8.32}$$

(1) Take ϕ_L out of the integral at $\nu' = \nu_0$.
(2) ϕ_D has unit area.
(3) Eq. (8.25) with $|\nu - \nu_0| \gg \Gamma/4\pi$.

The wings are Lorentzian. The Gaussian falls exponentially and the Lorentzian only as a power, so however small Γ is, the Lorentzian wins far enough out. For Lyman α at 100 K, the two are equal at $4.0\,\Delta\nu_D$ from the centre. Inside that the line is Gaussian, and outside it Lorentzian (Figure 8.2). When the column of gas is so large that the core absorbs all the light, the line can only grow in its wings. The wings of Lyman α in the spectra of distant quasars give the columns of the largest hydrogen clouds in this way.

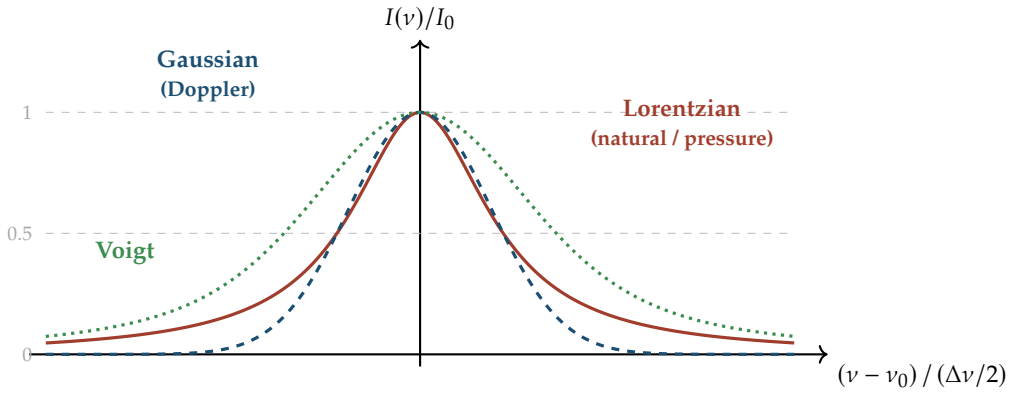


Figure 8.2: A Lorentzian profile (solid), a Gaussian profile (dashed) with the same peak intensity I_0 and the same half-width $\Delta\nu/2$, and their convolution, the Voigt profile (dotted), scaled to the same peak, against $(\nu - \nu_0)/(\Delta\nu/2)$ from -4.5 to 4.5. The Voigt profile has a half-width of 1.64 in these units. The dashed horizontal line marks half the peak.

8.4 Hydrogen in the ISM

Section 8.3 gave the width of a line from the lifetime of its upper level and from the motions of the gas. Hydrogen is most of the interstellar medium, but its strong lines are hard to use. Almost every atom sits in its ground state, and the first line up from there, Lyman α, needs 10.2 eV. At 100 K the Boltzmann factor for that is e^{-1184}, about 10^{-514}. Cold hydrogen gives no Lyman α of its own, and is seen in it only as absorption against a hot ultraviolet source behind it.

Hold two small bar magnets close together. Lined up north to south they attract and sit comfortably. Turned so that like poles face, they push apart. The two arrangements have different energies, and flipping one magnet over moves between them. For tiny magnets the energy difference is tiny. The electron and the proton in a hydrogen atom are such magnets, because both spin. Flipping the electron's spin relative to the proton's changes the energy by a few millionths of an electronvolt. The photon from that flip has a wavelength of 21 cm. Even the coldest gas has enough energy to keep both arrangements populated, so every hydrogen atom can send out this line.

The spin-flip energy is only 5.9×10^{-6} eV, equivalent to $hc/(\lambda k_B) = 0.068$ K. At 100 K, both levels are readily populated. The upper one has three states for every one in the lower, so it holds about three quarters of the atoms. Collisions move atoms between them much more often than spontaneous decay does, keeping those populations close to their thermal ratio. This makes the rare 21 cm emission useful for tracing cold neutral hydrogen. We now calculate the transition and what its brightness tells us.

Three forms. Hydrogen in the interstellar medium is neutral and atomic, written HI, ionised, written HII (the same as H^+), or molecular, H_2. Ionisation needs photons above 13.6 eV, with wavelengths below 91.2 nm. Only hot, massive stars give many of them, so HII is found around such stars. Molecular hydrogen forms on the surfaces of dust grains (Section 8.4.2).

8.4.1 The 21 cm Line

The opening of this section pictured the line as two tiny magnets, the electron and the proton, flipped from parallel to opposed. Here we work it out. We find the energy of the flip and so the wavelength, how long an atom waits before it flips on its own, how many atoms sit in the upper state in a real cloud, and why so slow a line is still easy to see. The answers rest on the energy of a magnet in a field, a balance of rates (P4) and a comparison of two times (P6).

The energy of the spin flip. Each particle's magnetic moment is set by its spin. For the electron $\vec{\mu}_e = -g_e \mu_B \vec{S}_e/\hbar$, with the Bohr magneton $\mu_B = e\hbar/2m_e$ and $g_e \approx 2$. For the proton $\vec{\mu}_p = g_p \mu_N \vec{S}_p/\hbar$, with the nuclear magneton $\mu_N = e\hbar/2m_p$ and $g_p = 5.586$. We treat the pull of the two magnets on each other as a small extra term in the energy of the atom, a perturbation. The Hamiltonian of hydrogen becomes $H = H_0 + H'$. Here H_0 is the usual electron in the electric field of the proton, and

$$H' = -\vec{\mu}_e \cdot \vec{B}_p \tag{8.33}$$

is the energy of the electron's moment in the magnetic field $\vec{B}_p$ of the proton. It is about a millionth of the binding energy, so first-order perturbation theory is enough. That says the energy of a state shifts by the average of H' over the unperturbed state, $E^{(1)} = \langle\psi|H'|\psi\rangle$, where ψ is the 1s ground state times a spin state. The 1s cloud is round, and the average of the proton's field over it leaves only the part at the proton itself, $(2\mu_0/3)\,\vec{\mu}_p\,\delta^3(\vec{r})$, a standard result of magnetostatics. So only the electron density at the proton counts, $|\psi(0)|^2 = 1/(\pi a_0^3)$, with a_0

the Bohr radius.

$$
\begin{aligned}
E^{(1)} &\overset{(1)}{=} \langle \psi | H' | \psi \rangle = - \left\langle \vec{\mu}_e \cdot \vec{B}_p \right\rangle \\
&\overset{(2)}{=} -\frac{2\mu_0}{3}\, \vec{\mu}_e \cdot \vec{\mu}_p\, |\psi(0)|^2 \\
&\overset{(3)}{=} \frac{2\mu_0}{3}\, \frac{g_e g_p \mu_B \mu_N}{\pi a_0^3}\, \frac{\vec{S}_e \cdot \vec{S}_p}{\hbar^2}.
\end{aligned}
\tag{8.34}
$$

(1) First-order perturbation theory, the average of H' over the unperturbed 1s state, that is over where the electron is.
(2) Only the contact part of the proton's field survives the average over the spherical cloud, and it acts where the electron overlaps the proton.
(3) The two moments in terms of the spins, and $|\psi(0)|^2 = 1/(\pi a_0^3)$.

The four spin states, two for the electron times two for the proton, all have the same unperturbed energy. For such states first-order theory must use the combinations in which H' has no cross terms, and since H' goes as $\vec{S}_e \cdot \vec{S}_p$ these are the states of definite total spin. The total spin $\vec{S} = \vec{S}_e + \vec{S}_p$ has $S^2 = \hbar^2 s(s+1)$, with $s = 1$ (spins parallel, three states) or $s = 0$ (spins opposed, one state). Each spin has $S_e^2 = S_p^2 = \tfrac{3}{4}\hbar^2$.

$$
\begin{aligned}
\frac{\vec{S}_e \cdot \vec{S}_p}{\hbar^2} &\overset{(1)}{=} \frac{S^2 - S_e^2 - S_p^2}{2\hbar^2} \\
&\overset{(2)}{=} \tfrac{1}{2}\big(s(s+1) - \tfrac{3}{2}\big) = \begin{cases} +\tfrac{1}{4}, & s = 1, \\ -\tfrac{3}{4}, & s = 0. \end{cases}
\end{aligned}
\tag{8.35}
$$

(1) Square $\vec{S} = \vec{S}_e + \vec{S}_p$ and solve for the cross term.
(2) $S^2 = \hbar^2 s(s+1)$ and $S_e^2 = S_p^2 = \tfrac{3}{4}\hbar^2$.

The two values differ by one. With $g_e = 2$ the gap between the levels is

$$
\boxed{E_{21} = \frac{4\mu_0\, g_p\, \mu_B \mu_N}{3\pi a_0^3} = 5.88 \times 10^{-6}\,\text{eV}, \qquad \nu_0 = \frac{E_{21}}{h} = 1421\,\text{MHz}.}
\tag{8.36}
$$

The measured frequency is 1420.406 MHz, at a wavelength of 21.1 cm. Using $g_e = 2.0023$ and the reduced mass of the electron and proton brings the formula to within 0.01 per cent of it. Written in terms of the ionisation energy of hydrogen, 13.6 eV, the gap is

$$
\frac{E_{21}}{13.6\,\text{eV}} = \frac{8}{3}\, g_p\, \alpha^2\, \frac{m_e}{m_p} = 4.3 \times 10^{-7},
$$

with α the fine-structure constant. It is small because magnetic energies carry α^2 against electric ones, and because the proton's moment is smaller than the electron's by the ratio of their masses.

How fast it decays. An oscillating electric dipole of matrix element d decays at the rate $\omega^3 |d|^2/(3\pi\varepsilon_0\hbar c^3)$. A magnetic dipole radiates the same way with $d \to \mu/c$, so

$$
A_{21} = \frac{\mu_0\, \omega^3\, |\langle 1 | \vec{\mu} | 2 \rangle|^2}{3\pi\hbar c^3},
$$

the spontaneous rate in the notation of Section 11.6, with level 1 the lower and level 2 the upper. The electron's moment is 658 times the proton's, so only the electron's spin needs to flip. Take

the upper state whose total spin along z is zero, $(\uparrow\downarrow + \downarrow\uparrow)/\sqrt{2}$, with the electron's spin written first. The z part of the electron's moment, $-2\mu_B S_{e,z}/\hbar$, gives $-\mu_B$ when the electron is up and $+\mu_B$ when it is down. Acting on the two terms it gives $-\mu_B(\uparrow\downarrow - \downarrow\uparrow)/\sqrt{2}$, which is the lower state times $-\mu_B$. So the matrix element has size μ_B, and the other two upper states give the same. Then

$$\boxed{A_{21} = \frac{\mu_0\,\omega_0^3\,\mu_B^2}{3\pi\hbar c^3} = 2.87\times10^{-15}\,\mathrm{s}^{-1}, \qquad \frac{1}{A_{21}} = 1.1\times10^{7}\ \text{years}.} \tag{8.37}$$

The rate is tiny for two reasons. The frequency is low, and the rate goes as ω^3. And the transition is magnetic, which costs a factor of about α^2 against an electric one. Lyman α decays 2×10^{23} times faster. A transition this slow is called **forbidden**. By Eq. (8.23) its natural width is $A_{21}/2\pi = 4.6\times10^{-16}$ Hz. The thermal width at 100 K, from Eq. (8.30), is 10 kHz in full. The width of the 21 cm line measures only the motions of the gas.

Level populations, a balance of rates. Collisions knock atoms between the two levels. Let one atom in the upper level be knocked down at the rate $1/t_{\mathrm{col}}$. By detailed balance (Eq. (9.39), with g_1 and g_2 the numbers of states in the lower and upper levels) an atom in the lower level is knocked up at the rate $(g_2/g_1)\,e^{-E_{21}/k_BT}/t_{\mathrm{col}}$. The upper level also decays at the rate A_{21}. Background radio light can also lift atoms, but for the 21 cm line it does so far more slowly than collisions do, so we leave it out. In a steady state the rates into and out of the upper level balance (P4). With n_1 and n_2 the densities of atoms in the two levels,

$$\begin{aligned} n_2\left(A_{21} + \frac{1}{t_{\mathrm{col}}}\right) &\overset{(1)}{=} n_1\,\frac{g_2}{g_1}\,\frac{e^{-E_{21}/k_BT}}{t_{\mathrm{col}}} \\ \Longrightarrow \qquad \frac{n_2}{n_1} &\overset{(2)}{=} \frac{g_2}{g_1}\,e^{-E_{21}/k_BT}\,\frac{1}{1+A_{21}t_{\mathrm{col}}}. \end{aligned} \tag{8.38}$$

(1) Out of the upper level by decay and by collisions, into it by collisions from below.
(2) Divide by $n_1(A_{21} + 1/t_{\mathrm{col}})$ and multiply the top and bottom by t_{col}.

$$\boxed{\frac{n_2}{n_1} = \frac{g_2}{g_1}\,e^{-E_{21}/k_BT}\,\frac{1}{1+A_{21}t_{\mathrm{col}}}.} \tag{8.39}$$

One number decides, $A_{21}t_{\mathrm{col}}$, the ratio of the decay rate to the collision rate (P6). When collisions are much faster, $A_{21}t_{\mathrm{col}} \ll 1$, the populations take the Boltzmann ratio at the gas temperature (P3). When decay is much faster, $A_{21}t_{\mathrm{col}} \gg 1$, the upper level is nearly empty and every upward collision ends in a photon. The changeover is at the **critical density**, where $A_{21}t_{\mathrm{col}} = 1$, so $n = A_{21}/(\sigma_{\mathrm{col}}\bar{v})$.

The populations of the 21 cm levels. For hydrogen at 100 K the mean relative speed of two atoms is $\sqrt{16k_BT/(\pi m_{\mathrm{H}})} = 2.0\,\mathrm{km\,s^{-1}}$. With $n = 10^6\,\mathrm{m^{-3}}$ ($1\,\mathrm{cm^{-3}}$) and a cross-section of order $10^{-19}\,\mathrm{m^2}$ for the collisions that exchange the spins, $1/t_{\mathrm{col}} = 2\times10^{-10}\,\mathrm{s^{-1}}$. An atom is knocked between the levels about once in 150 years, against once in 11 million years by decay. So $A_{21}t_{\mathrm{col}} = 1.4\times10^{-5}$, the populations follow the gas temperature, and with $g_2/g_1 = 3$ and $E_{21}/k_B = 0.068$ K,

$$\boxed{\frac{n_2}{n_1} = 3\,e^{-0.068\,\mathrm{K}/T} \approx 3, \qquad n_2 \approx \tfrac{3}{4}\,n_{\mathrm{H}}.} \tag{8.40}$$

Three quarters of the atoms are in the upper level whatever the temperature of the cloud, because every cloud is far hotter than 0.068 K.

Why the line is seen. Each atom in the upper level sends out a photon at the rate A_{21} whatever the collisions do, because they keep the upper level filled. The power per atom is tiny, $\frac{3}{4}A_{21}h\nu_0 = 2.0 \times 10^{-39}$ W, but a line of sight through the Galaxy crosses 10^{25} or more atoms per square metre, and together they make the line easy to detect with a radio telescope. A gas cell in a laboratory holds far too few atoms to show it in spontaneous emission. Instead the line is measured in the laboratory by stimulated emission in the hydrogen maser, to better than one part in 10^{12}. Because the populations do not depend on the temperature, the brightness of the line counts the atoms in the column (Problem 8.7.2).

What it maps. The rest frequency is known so well that the Doppler shift of the line, Eq. (3.25), gives the line-of-sight velocity of each cloud of HI along a sight line. Combined with the rotation of the Galaxy, the velocity places each cloud at a distance. The 21 cm line is the main tool for mapping neutral gas in the Milky Way and for measuring the rotation of other galaxies far beyond their stars.

8.4.2 Molecular Hydrogen

How it forms. Two hydrogen atoms that meet in the gas cannot stay together. Binding releases 4.5 eV, and the pair has no fast way to get rid of it, since H_2 has no electric dipole to radiate with. The atoms fly apart again within one vibration. On a grain the energy has somewhere to go. An atom that lands on the grain sticks to its surface and wanders over it, which makes it far more likely to meet a second atom. When they bind, the grain takes up the released energy, and the molecule leaves. In the thin gas, reactions without grains are far too slow to compete.

Where it survives. Ultraviolet starlight breaks H_2 apart. The molecule survives only deep in clouds, where dust absorbs that light (Section 8.2.4).

Why it is hard to see. H_2 is made of two identical atoms, so it has no permanent electric dipole moment and no allowed dipole transitions between its rotation and vibration levels. Its lowest rotational line comes from a level at $E/k_B \approx 510$ K, and is a weak quadrupole line. Cold molecular clouds, at 10 to 20 K, hardly excite it at all. Astronomers trace H_2 instead by molecules that do have a dipole. The most used is CO, whose first rotational level lies at $E/k_B = 5.5$ K and is easily excited at 10 K. The amount of H_2 follows from the strength of the CO line and a ratio calibrated by other means. Many other molecules are also found, among them OH, H_2O, CO_2, NH, CH and CH_2.

8.5 Interstellar Clouds and Nebulae

The medium holds hydrogen in three forms, with dust that dims the light, and each form has its own way of being seen. The density and temperature change smoothly over many powers of ten, so any division into kinds of cloud is rough.

Weather clouds are named by how they look, puffy or layered or wispy, and not by what they are made of, which is always water. The names are still useful, because how a cloud looks tells something about how it formed. Interstellar clouds are named the same way. The classes overlap, and the same cloud can be called by different names when seen in different ways. Each class is defined by what dominates its light, the 21 cm line, CO, the dust or the glow of hot gas.

Extinction gives one way to distinguish these clouds. A diffuse molecular cloud with A_V between 1 and 5 passes roughly 40 to 1 per cent of the visible light behind it. A dense globule with $A_V \approx 10$ passes only about one part in 10^4. The categories in the table below mark useful

ranges in a continuous spread of conditions. After identifying them, we will use a split emission line to measure how quickly a planetary nebula expands.

Clouds. Table 8.4 gives the main kinds. In **HI clouds** the hydrogen is atomic. In **diffuse and translucent molecular clouds**, with $1 < A_V < 5$, the hydrogen is still mostly atomic, with H_2 where the optical depth is higher. They are irregular in shape. **Giant molecular clouds** are large, clumpy complexes of gas and dust, and the main sites of star formation. The Horsehead Nebula is part of the giant molecular cloud complex in Orion. **Bok globules** are small, nearly round, dense clouds with $A_V \approx 10$ and $T \approx 10\,\mathrm{K}$, where stars also form. They appear to be dense cores whose surrounding gas was stripped away by the light and winds of hot massive stars.

Cloud type	T (K)	Size	Mass ($M_\odot$)	Tracer
HI clouds	~ 100	tens of pc	—	21 cm line
Diffuse molecular	15–50	several pc	3–100	CO, UV absorption
Giant molecular clouds	~ 10–20	~ 50 pc	10^5–10^6	CO
Bok globules	~ 10	below 1 pc	1–1000	Dark silhouette

Table 8.4: Approximate temperature, size, mass and tracer of the kinds of interstellar cloud. The ranges overlap between kinds.

Nebulae. *Nebula*, Latin for cloud, is an old name for anything in the sky that looks extended and diffuse rather than point-like. Some nebulae turned out to be distant galaxies, such as Andromeda. The rest are clouds in our own Galaxy, of three kinds.

- **Reflection nebulae** shine by starlight scattered from dust, such as the blue nebula around the Pleiades.
- **Dark nebulae** are clouds whose extinction is high enough to show as silhouettes against the stars behind them. Giant molecular clouds such as the Horsehead, Bok globules and diffuse molecular clouds can all appear this way. Barnard 68 is a well-known example.
- **Emission nebulae** make their own light, mostly in lines. **HII regions** are gas ionised by nearby hot stars, such as the Monkey Head Nebula, NGC 2174. **Planetary nebulae** are gas thrown off by a star late in its life, such as the Ring Nebula. **Supernova remnants** are gas thrown out by a stellar explosion, such as the Crab Nebula, left by the supernova of 1054, which is lit from inside by a pulsar.

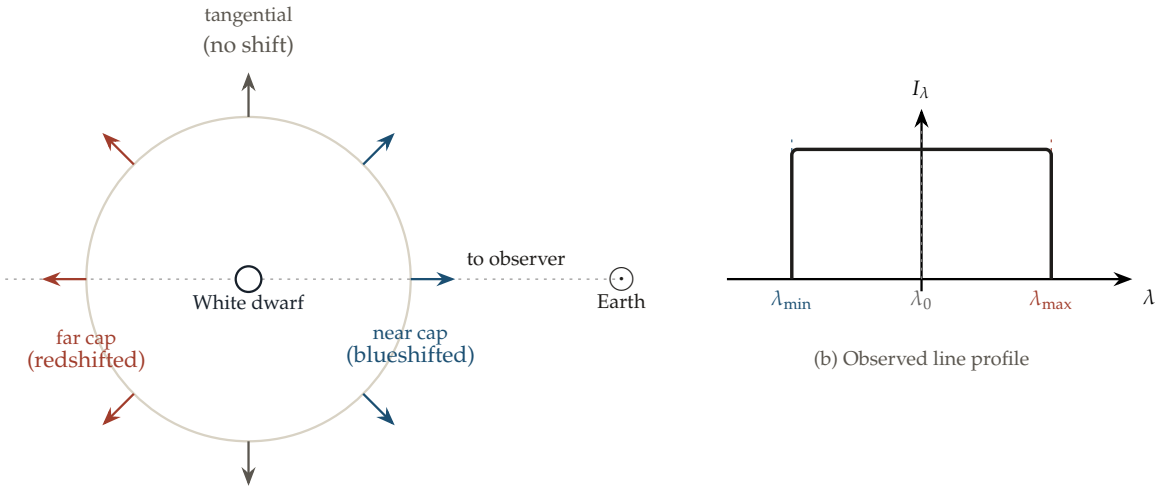
8.5.1 Expanding Planetary Nebulae and the Doppler-Split Line

A planetary nebula is a shell of gas thrown off by a star near the end of its life. The hot core left behind, a white dwarf, ionises the shell and makes it glow in emission lines. The shell expands outward, roughly as a sphere centred on the white dwarf, at a speed v_{exp} of a few tens of $\mathrm{km\,s^{-1}}$. Every point of the shell moves outward at v_{exp}, but in different directions relative to us. The near side comes toward us, the far side goes away, and the rim moves across the line of sight (Figure 8.3). We want the line profile of the whole shell and the speed it gives.

The velocity of one point. Let θ be the angle at the white dwarf between a point of the shell and the direction pointing away from us. The point moves along its radius at speed v_{exp}, and the part of that velocity along the line of sight, away from us, is

$$v_r = v_{\mathrm{exp}} \cos\theta. \tag{8.41}$$

The far cap ($\theta = 0$) has $v_r = +v_{\mathrm{exp}}$, the near cap ($\theta = \pi$) has $v_r = -v_{\mathrm{exp}}$, and the rim ($\theta = \pi/2$) has $v_r = 0$.



(a) Expanding shell, as seen from Earth

(b) Observed line profile

Figure 8.3: (a) An expanding spherical shell around a white dwarf, with Earth to the right. Arrows show the local expansion velocity, blue on the near cap, red on the far cap and grey near the rim. (b) A flat-topped emission-line profile between $\lambda_{\min}$ and $\lambda_{\max}$, with the rest wavelength λ_0 marked between them.

How much of the shell has each velocity. Take a thin shell of radius R that glows evenly and is transparent to its own line, all of it inside the telescope's aperture. The band of the shell between θ and $\theta + d\theta$ is a ring of area $2\pi R^2 \sin\theta\, d\theta$. The fraction of the shell in it is

$$\begin{aligned} \frac{dA}{A} &\overset{(1)}{=} \frac{2\pi R^2 \sin\theta\, d\theta}{4\pi R^2} \\ &\overset{(2)}{=} \frac{1}{2}\,|d(\cos\theta)| \\ &\overset{(3)}{=} \frac{|dv_r|}{2v_{\text{exp}}}. \end{aligned} \tag{8.42}$$

(1) The area of the ring over the area of the sphere.
(2) $d(\cos\theta) = -\sin\theta\, d\theta$.
(3) $v_r = v_{\text{exp}} \cos\theta$.

Each equal step of v_r between $-v_{\text{exp}}$ and $+v_{\text{exp}}$ holds an equal share of the shell. So a single sharp line becomes a flat-topped line, as in panel (b) of Figure 8.3. Its edges come from the two caps. By Eq. (3.25),

$$\lambda_{\min} = \lambda_0 \left(1 - \frac{v_{\text{exp}}}{c}\right), \qquad \lambda_{\max} = \lambda_0 \left(1 + \frac{v_{\text{exp}}}{c}\right).$$

A slit placed across the centre of the nebula sees only the two caps, and the line splits into two lines at these same wavelengths.

The expansion speed. Subtracting the two edges removes the unknown λ_0 from the difference,

$$\boxed{\lambda_{\max} - \lambda_{\min} = \frac{2\lambda_0\, v_{\text{exp}}}{c} \qquad \Longrightarrow \qquad v_{\text{exp}} = \frac{c}{2}\,\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_0}.} \tag{8.43}$$

The speed comes from the two edges of one line, without the distance of the nebula. Their mean is λ_0 shifted by any motion of the nebula as a whole, which cancels from the difference.

The thermal width of Section 8.3 rounds the edges. For hydrogen at 10^4 K it is 21 km s^{-1} in full, as large as the expansion itself, so the lines of heavier ions such as O^{2+}, with $\sigma_v = 2.3$ km s^{-1}, give sharper edges.

8.6 Heating and Cooling of the ISM

Section 8.4 balanced the rates into and out of an energy level, Eq. (8.39). That fixes how atoms share out among levels at a given temperature. It does not say what the temperature is. Gas in the medium is heated by starlight and fast particles and cooled by the light it sends out, and its temperature is wherever the two meet.

Put a pan of water on a flame with no lid. The flame heats it. Water evaporates from the surface and cools it. Evaporation rises steeply with temperature. Turn the flame up or down and the water settles at almost the same temperature, a little below boiling, because a small rise in temperature is enough to carry away a lot more heat. Interstellar gas is cooled in the same way. Collisions lift atoms and molecules into excited levels, and the photons they then send out escape and carry the energy away. The number of collisions with enough energy grows as e^{-E_{21}/k_BT}, steeply, so the temperature settles near E_{21}/k_B of the main cooling line, almost whatever the heating.

Atomic clouds cool mainly through a 157.7 μm line of C^+. Its upper level lies at $E_{21}/k_B = 91$ K, close to the roughly 100 K temperature of HI clouds. In molecular clouds, CO levels at 5.5 and 16.6 K help cool gas to 10–20 K. Cooling rises rapidly once collisions can excite the relevant level, so the temperature settles where that energy loss balances heating. We will identify the main sources on each side of that balance and find the resulting temperature.

Heating. Several processes put energy into the gas.

- **Cosmic rays**, charged particles of 10 to 10^{14} MeV accelerated in stellar flares and supernova shocks, ionise atoms and molecules. The freed electron shares its energy with the gas by collisions. Cosmic rays reach deep into dense clouds where starlight cannot.
- **Ultraviolet photons** from massive stars ionise atoms, and the freed electrons heat the gas.
- **Photoelectrons from grains**, knocked out of dust grains by ultraviolet photons, are the main heating in much of the diffuse gas.
- **X-rays** from stars ionise more deeply than ultraviolet light.
- **Dust** absorbs continuous starlight and warms.
- **Shocks** from stellar winds and supernovae turn the energy of bulk motion into heat.

Cooling. Collisions excite molecules and the fine-structure levels of atoms and ions, which then decay by sending out infrared and millimetre photons. Grains are much smaller than these wavelengths, so the photons are in the Rayleigh regime, where $\sigma \propto \lambda^{-4}$ is small (Table 8.2). They escape the cloud and carry the energy with them. Dust also cools by its own thermal emission in the far infrared.

The balance. Let each gas particle receive heat at the rate $\dot{E}_{\rm heat}$, and let a coolant with density n_c have a line of energy E_{21} that escapes. In thin gas, below the critical density ($A_{21}t_{\rm col} \gg 1$), nearly every coolant particle is in its lower level and every upward collision ends in an escaping

photon. With $1/t_{\rm col} = n\sigma_{\rm col}\bar{v}$ from Section 8.3, heat in equals heat out (P4),

$$\begin{aligned} n\,\dot{E}_{\rm heat} &\overset{(1)}{=} n_2\,A_{21}\,E_{21} \\ &\overset{(2)}{=} n_c\,\frac{g_2}{g_1}\,\frac{e^{-E_{21}/k_BT}}{t_{\rm col}}\,E_{21} \\ &\overset{(3)}{=} n_c\,n\,\sigma_{\rm col}\bar{v}\,\frac{g_2}{g_1}\,E_{21}\,e^{-E_{21}/k_BT}. \end{aligned} \tag{8.44}$$

(1) Heating per unit volume equals the energy carried off per unit volume by escaping photons.
(2) Eq. (8.39) with $A_{21}t_{\rm col} \gg 1$, so $n_2A_{21} = n_1(g_2/g_1)e^{-E_{21}/k_BT}/t_{\rm col}$, and $n_1 \approx n_c$.
(3) $1/t_{\rm col} = n\sigma_{\rm col}\bar{v}$.

Dividing by n and taking the logarithm,

$$\boxed{k_BT = \frac{E_{21}}{\ln\!\left(n_c\,\sigma_{\rm col}\,\bar{v}\,(g_2/g_1)\,E_{21}/\dot{E}_{\rm heat}\right)}.} \tag{8.45}$$

The heating and the density enter only through a logarithm. Changing the heating tenfold changes the logarithm by 2.3, so over a wide range of conditions the gas sits within a factor of a few of E_{21}/k_B of its main coolant. That is why atomic clouds, cooled by C^+, are near 100 K, and molecular clouds, cooled by CO, are near 10 to 20 K. Above the critical density the cooling per particle stops growing with n, but the Boltzmann factor still pins the temperature in the same way.

Why it matters for stars. The temperature fixes the pressure of the gas, $P = nk_BT$ (Eq. (9.28)), and the pressure is what resists gravity. A cloud that can radiate away the heat of its own compression stays cold while it shrinks, and its pressure cannot stop the collapse. Efficient cooling is therefore a condition for star formation. Whether gravity or pressure wins is worked out in Chapter 10.

8.7 Problem Set: The Interstellar Medium

8.7.1 Problem 1: Planetary Nebula Emission and Expansion

Problem Statement: The Hα line from a spherical planetary nebula, observed with the whole nebula in the aperture, spreads from a minimum wavelength of 656.24 nm to a maximum of 656.32 nm. (a) Find the rest wavelength the observation implies. (b) Estimate how fast the nebula expands, in km s^{-1}. (c) Convert the speed to AU yr^{-1}. (d) The gas is at about 10^4 K. Is the thermal width of Hα small compared with the spread of the line?

Essential Ideas:

- The near cap gives the shortest wavelength and the far cap the longest, and the profile is flat between them (Section 8.5.1).
- $v_{\rm exp} = (c/2)(\lambda_{\rm max} - \lambda_{\rm min})/\lambda_0$, Eq. (8.43).
- The thermal full width is $2.355\,\lambda_0\sigma_v/c$ with $\sigma_v = \sqrt{k_BT/m_{\rm H}}$, Eqs. (8.27) and (8.30).

Detailed Solutions:

(a) **Rest wavelength**. The caps move at equal and opposite speeds, so λ_0 is the mean of the edges.

$$\boxed{\lambda_0 = \frac{656.32 + 656.24}{2}\,\text{nm} = 656.28\,\text{nm}.}$$

This is the laboratory wavelength of Hα, so the nebula as a whole is not moving toward or away from us.

(b) **Expansion speed**.

$$\boxed{v_{\rm exp} = \frac{2.998\times10^5\,{\rm km\,s^{-1}}}{2}\times\frac{656.32-656.24}{656.28} \simeq 18.3\,{\rm km\,s^{-1}}.}$$

(c) **In astronomical units per year**. With $1\,{\rm yr} = 3.156\times10^7\,{\rm s}$ and $1\,{\rm AU} = 1.496\times10^8\,{\rm km}$,

$$\boxed{v_{\rm exp} = 18.27\,\frac{\rm km}{\rm s}\times\frac{3.156\times10^7\,{\rm s\,yr^{-1}}}{1.496\times10^8\,{\rm km\,AU^{-1}}} \simeq 3.85\,{\rm AU\,yr^{-1}}.}$$

This is a typical speed for the shell of a planetary nebula, whose observed values are 10 to $30\,{\rm km\,s^{-1}}$.

(d) **Thermal width**. $\sigma_v = \sqrt{1.381\times10^{-23}\times10^4/1.674\times10^{-27}}\,{\rm m\,s^{-1}} = 9.1\,{\rm km\,s^{-1}}$, so

$$\boxed{\text{thermal full width} = 2.355\times656.28\,{\rm nm}\times\frac{9.1}{2.998\times10^5} \simeq 0.047\,{\rm nm},}$$

more than half the spread of 0.08 nm. The thermal width is not small. The edges of the Hα line are blurred, and the measured edges should be read where the profile has fallen to half its height. Lines of heavier ions, such as [O III] at 500.7 nm, have four times smaller thermal widths and give the expansion speed more cleanly.

8.7.2 Problem 2: Counting Hydrogen with the 21 cm Line

Problem Statement: A radio telescope sees 21 cm emission from a cloud of HI with no bright source behind it. Radio astronomers express the intensity I_ν as a brightness temperature T_B, the temperature a blackbody would need to give the same I_ν. (a) Show that at this frequency $I_\nu = 2\nu^2 k_B T_B/c^2$. (b) Assume the cloud is transparent to its own line. Show that the column density of hydrogen is

$$N_{\rm H} = \frac{32\pi k_B \nu_0^2}{3hc^3 A_{21}}\int T_B\,dv,$$

where v is the line-of-sight velocity across the line. (c) Evaluate the constant in ${\rm cm^{-2}}$ per ${\rm K\,km\,s^{-1}}$. (d) The line is Gaussian with peak $T_B = 50\,{\rm K}$ and full width $10\,{\rm km\,s^{-1}}$. Find $N_{\rm H}$ and the visual extinction expected from the dust that comes with it.

Essential Ideas:

- The Planck function at $h\nu \ll k_B T$ (Eq. (4.13)).
- Three quarters of the atoms are in the upper level and each sends out photons at the rate A_{21}, evenly in all directions, Eqs. (8.37) and (8.40).
- The dust per hydrogen atom, $A_V/N_{\rm H} = 5.3\times10^{-22}\,{\rm mag\,cm^2}$ (Section 8.2.4).

Detailed Solutions:

(a) **Brightness temperature**. For $\nu_0 = 1420\,{\rm MHz}$, $h\nu_0/k_B = 0.068\,{\rm K}$, far below any T_B of interest. With $x = h\nu/k_B T_B \ll 1$, $e^x - 1 \approx x$ in the Planck function, and

$$\boxed{I_\nu = \frac{2h\nu^3}{c^2}\,\frac{1}{e^{h\nu/k_B T_B}-1} \approx \frac{2\nu^2 k_B T_B}{c^2}.}$$

(b) **Column density**. A column of cross-section $1\,\mathrm{m}^2$ holds $\frac{3}{4}N_\mathrm{H}$ atoms in the upper level. Together they send out the power $\frac{3}{4}N_\mathrm{H}A_{21}h\nu_0$, spread evenly over 4π steradians. Seen face on through a transparent cloud, the intensity summed over the line is therefore

$$\int I_\nu \, d\nu = \frac{h\nu_0 A_{21}}{4\pi}\,\frac{3}{4}N_\mathrm{H}.$$

By (a), and with $d\nu = (\nu_0/c)\,dv$ from Eq. (3.25),

$$\int I_\nu \, d\nu = \frac{2\nu_0^2 k_B}{c^2}\,\frac{\nu_0}{c}\int T_B \, dv.$$

Setting the two equal and solving,

$$\boxed{N_\mathrm{H} = \frac{32\pi k_B \nu_0^2}{3hc^3 A_{21}}\int T_B \, dv.}$$

The temperature of the gas has dropped out, because the populations of Eq. (8.40) do not depend on it. Absorption by the cloud of its own line, left out here, matters only when the cloud is not transparent, and is treated with the transfer equation of Chapter 11.

(c) **The constant**. With $\nu_0 = 1.4204\times10^9$ Hz and $A_{21} = 2.87\times10^{-15}\,\mathrm{s}^{-1}$, in SI units the constant is $1.82\times10^{19}\,\mathrm{m}^{-2}$ per $\mathrm{K\,m\,s^{-1}}$. Multiplying by 10^3 for $\mathrm{km\,s^{-1}}$ and by 10^{-4} for cm^{-2},

$$\boxed{N_\mathrm{H} = 1.82\times10^{18}\,\mathrm{cm}^{-2}\times\frac{\int T_B\,dv}{1\,\mathrm{K\,km\,s^{-1}}}.}$$

(d) **The cloud**. The area under a Gaussian is its peak times its full width times $\sqrt{\pi/4\ln 2} = 1.064$, so $\int T_B\,dv = 1.064\times50\times10 = 532\,\mathrm{K\,km\,s^{-1}}$, and

$$\boxed{N_\mathrm{H} \simeq 9.7\times10^{20}\,\mathrm{cm}^{-2}, \qquad A_V \simeq 9.7\times10^{20}\times5.3\times10^{-22} \simeq 0.5\ \mathrm{mag}.}$$

A cloud that glows this brightly in the 21 cm line dims the stars behind it by about half a magnitude. Comparing the two along many lines of sight is how the ratio of dust to gas is measured.

8.7.3 Problem 3: The Two Widths of Lyman Alpha

Problem Statement: The Lyman α line of hydrogen is at 121.567 nm, and its upper level decays at $\Gamma = 6.27\times10^8\,\mathrm{s}^{-1}$. The lower level is the ground state, which does not decay. Take interstellar gas at 100 K. (a) Find the natural full width, in Hz and in nm. (b) Find the Doppler full width in Hz. (c) Which dominates the core of the line, and by how much? (d) How far from the centre, in units of $\Delta\nu_D$, does the Lorentzian wing exceed the Gaussian?

Essential Ideas:

- The natural full width is $\Gamma/2\pi$, Eq. (8.25).
- The Doppler full width is $2.355\,\Delta\nu_D$, with $\Delta\nu_D = \nu_0\sqrt{k_BT/m_\mathrm{H}}/c$, Eq. (8.30).
- The Voigt profile is Gaussian in the core and Lorentzian in the wings (Eq. (8.31)).

Detailed Solutions:

(a) **Natural width**. In wavelength, $\Delta\lambda = \lambda_0^2\Delta\nu/c$.

$$\boxed{\frac{\Gamma}{2\pi} = \frac{6.27\times10^8}{2\pi}\,\mathrm{Hz} \simeq 1.0\times10^8\,\mathrm{Hz}, \qquad \Delta\lambda \simeq 4.9\times10^{-6}\,\mathrm{nm}.}$$

(b) **Doppler width**. $\nu_0 = c/\lambda_0 = 2.466 \times 10^{15}$ Hz and $\sigma_v = 0.908\,\mathrm{km\,s^{-1}}$, so $\Delta\nu_D = \nu_0\sigma_v/c = 7.47 \times 10^9$ Hz.

$$\boxed{2.355\,\Delta\nu_D \simeq 1.76 \times 10^{10}\ \mathrm{Hz}.}$$

(c) **The core**. The ratio of the two widths is

$$\boxed{\frac{1.76 \times 10^{10}}{1.0 \times 10^8} \simeq 176,}$$

so the core is Gaussian and set by the temperature. Even at 100 K the motions of the atoms blur the line far more than its lifetime does.

(d) **The wings**. Write $y = (\nu - \nu_0)/\Delta\nu_D$. The Gaussian is $e^{-y^2/2}/(\sqrt{2\pi}\Delta\nu_D)$ and the Lorentzian wing is $(\Gamma/4\pi^2)/(y\Delta\nu_D)^2$. They are equal when

$$\frac{e^{-y^2/2}}{\sqrt{2\pi}} = \frac{\Gamma}{4\pi^2\Delta\nu_D}\,\frac{1}{y^2} = \frac{2.12 \times 10^{-3}}{y^2},$$

which a numerical solution puts at

$$\boxed{y \simeq 4.0.}$$

Beyond four Doppler widths the line is Lorentzian. A cloud thick enough to absorb all the light out to there shows broad Lorentzian wings, and their width gives its column of hydrogen without any knowledge of its temperature.

8.7.4 Problem 4: The First Fast Radio Burst

Problem Statement: The Lorimer burst has a dispersion measure of $375\,\mathrm{pc\,cm^{-3}}$, and it lasts about 5 ms at each frequency. Models of the Milky Way give $25\,\mathrm{pc\,cm^{-3}}$ in its direction. (a) How much later does the burst arrive at 1.2 GHz than at 1.5 GHz? (b) Leaving out the burst's own galaxy, what redshift does the rest of the dispersion measure imply, and what pushes the true value up or down? (c) Find the distance from Hubble's law. (d) How large can the region that made the flash be?

Essential Ideas:

- The delay goes as ν^{-2}, Eqs. (8.16) and (8.17).
- Between the galaxies, $\int n_e\,ds \approx 975\,z\ \mathrm{pc\,cm^{-3}}$ for $z \ll 1$, Eq. (8.21).
- A source cannot change faster than light crosses it, Eq. (15.59) with $\mathcal{D} = 1$.

Detailed Solutions:

(a) **Delay across the band**. Subtracting Eq. (8.17) at the two frequencies removes t_∞.

$$\boxed{t(1.2\,\mathrm{GHz}) - t(1.5\,\mathrm{GHz}) = 4.15\,\mathrm{ms} \times 375 \times (0.694 - 0.444) \simeq 389\,\mathrm{ms}.}$$

The sweep lasts eighty times longer than the burst itself, so it is easy to see in Figure 8.1.

(b) **Redshift**. The part from outside the Galaxy is $375 - 25 = 350\,\mathrm{pc\,cm^{-3}}$, so

$$\boxed{z \approx \frac{350}{975} \simeq 0.36.}$$

The burst's own galaxy and the halo of ours add electrons, which leaves fewer for the space between, so the true z is lower. Some of the ordinary matter is in stars and galaxies, which makes the slope smaller than 975, so the true z is higher. The two work against each other, and 0.36 is only an estimate.

(c) **Distance**. With $c/H_0 = 4448\,\mathrm{Mpc}$,

$$\boxed{d \approx \frac{cz}{H_0} = 0.36 \times 4448\,\mathrm{Mpc} \simeq 1.6\,\mathrm{Gpc}.}$$

At this redshift Hubble's law in the form $v = cz$ is only rough, so the burst lies of order a gigaparsec away, far outside the Local Group.

(d) **Size**. Parts of the source more than $c\,\Delta t$ apart could not all brighten and fade within Δt.

$$\boxed{R \lesssim c\,\Delta t = 3.0 \times 10^{8}\,\mathrm{m\,s^{-1}} \times 5 \times 10^{-3}\,\mathrm{s} = 1500\,\mathrm{km}.}$$

This is far smaller than any ordinary star, which points to a compact object such as a neutron star.

Chapter 9

Six Principles

The chapters so far have used the same ideas in different settings. Conserved energy linked a star's luminosity to its flux, counting statistics set the noise of a CCD, and competing timescales decided whether an effect mattered. The formulas ahead may look more specialised, but many come from combining these familiar ideas. We collect six of them here so that, when a new result appears, we can see why it follows and when it applies.

	Principle	Statement
P1	Conservation	The mass, momentum, energy and angular momentum in a region change only by what flows across its boundary and by the forces that act on it.
P2	Gravity of a sphere	Inside a spherical body, a small blob of gas at radius r is pulled only by the mass inside r, as if it sat at the centre.
P3	Counting states	Phase space comes in cells of size h^3. Pressure, energy density and populations follow from counting the states and how full each one is.
P4	Balance of rates	In a steady state every gain is matched by a loss. In equilibrium every process is matched by its own reverse.
P5	Random walk	After N steps in random directions a particle is about $\sqrt{N}$ steps from its start, so anything that scatters often moves down its own gradient.
P6	Compare two scales	The ratio of two times or two lengths decides which physics matters. A ratio near one marks a transition.

Table 9.1: The six principles.

None of the six is new physics. Each is a familiar law put in the form that applies directly to gas, light and gravity. Each has one general form, written with as few symbols as possible, and every result built on them names the principles it combines.

9.1 Conservation (P1)

The Kepler problem of Chapter 6 was solved by two conserved quantities. Conservation of the energy and of the angular momentum turned six second-order equations into an orbit that could be written down (Section 6.1.2 onwards). A star is not two bodies. It holds about 10^{57} particles, and nobody can follow each of them.

Think of a water tank in a house. Water comes in through one pipe and leaves through another. To know how much water the tank holds a minute from now, nobody follows single molecules.

One watches the two pipes. The water inside changes by what flows in minus what flows out, and nothing that happens inside the tank can change that. A star can be treated the same way. Take any fixed region inside it. The mass in that region changes only by gas flowing across its boundary. Its momentum changes by the momentum that gas carries across the boundary and by the forces acting on the region. This is conservation written for a fixed region instead of a single particle, principle P1.

If the tank holds 200 L while water enters at 3 L/min and leaves at 5 L/min, its volume changes at $3-5=-2$ L/min and it empties after 100 minutes. We need only the flows across its boundary to predict that change. For a region of gas, the same accounting gives one equation for its mass and another for its momentum. Those equations will later describe how pressure supports a star and how its energy changes.

Mass. Let a gas have density $\rho(\vec{r},t)$ and velocity $\vec{v}(\vec{r},t)$. Take a fixed volume V with surface S, and let $d\vec{A}$ be an outward-pointing element of S. Gas crossing $d\vec{A}$ carries mass out at the rate $\rho\,\vec{v}\cdot d\vec{A}$. The mass inside therefore changes as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho\,dV &\overset{(1)}{=} -\oint_S \rho\,\vec{v}\cdot d\vec{A}\\ &\overset{(2)}{=} -\int_V \nabla\cdot(\rho\vec{v})\,dV.\end{aligned} \tag{9.1}$$

(1) Mass inside changes only by the flow across the surface (P1).
(2) The divergence theorem.

The volume is fixed, so the time derivative on the left acts only on ρ and can be taken inside the integral. Both sides are then integrals over the same V,

$$\int_V \left[\frac{\partial\rho}{\partial t}+\nabla\cdot(\rho\vec{v})\right]dV=0.$$

This holds for every choice of V, however small, so the integrand itself vanishes,

$$\boxed{\frac{\partial\rho}{\partial t}+\nabla\cdot(\rho\vec{v})=0.} \tag{9.2}$$

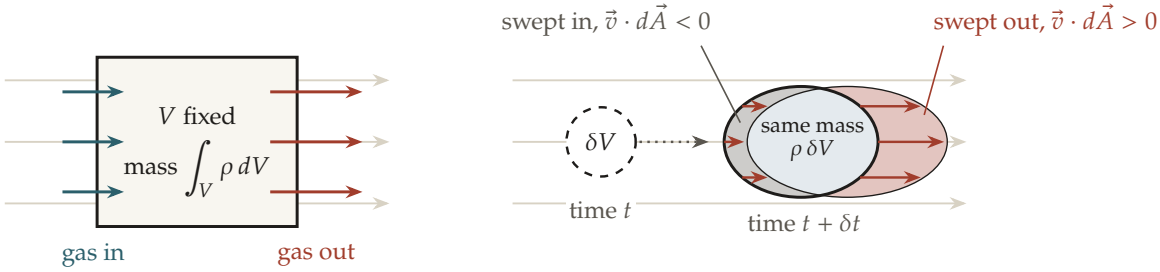
This is the **continuity equation**. It holds for any fluid and says nothing about forces.

The same law seen from a drifting blob. The box above stays still, and the gas flows through its walls. There is a second way to keep the books. Follow a small blob of gas as it drifts with the flow, like a puff of smoke carried by the wind. Its edge moves with the gas, so no gas ever crosses it, and the blob keeps its mass. What can change is its volume, and with it the density inside. Figure 9.1 shows the two views side by side.

First, the rate of change seen while riding with the blob. Call it D/Dt. For any quantity $f(\vec{r},t)$, in a short time δt the blob moves from $\vec{r}$ to $\vec{r}+\vec{v}\,\delta t$, so

$$\begin{aligned}\frac{Df}{Dt} &\overset{(1)}{=} \lim_{\delta t\to 0}\frac{f(\vec{r}+\vec{v}\,\delta t,\ t+\delta t)-f(\vec{r},t)}{\delta t}\\ &\overset{(2)}{=} \frac{\partial f}{\partial t}+(\vec{v}\cdot\nabla)f.\end{aligned} \tag{9.3}$$

(1) The blob carries us along, so the second value is read where the blob has moved to.
(2) Taylor expansion to first order in δt, $f(\vec{r}+\vec{v}\,\delta t,\ t+\delta t)=f(\vec{r},t)+\delta t\,\frac{\partial f}{\partial t}+\delta t\,(\vec{v}\cdot\nabla)f$.



(a) a fixed box

(b) a blob drifting with the gas

Figure 9.1: Two ways to count the mass of a gas that flows to the right and speeds up as it goes. (a) A box fixed in space. Gas enters through the left wall and leaves through the right wall. (b) A small blob of gas at time t (dashed) and at $t + \delta t$ (thick outline), drifting with the flow. The thin outline shows the same blob a moment later. The arrows on its edge show the velocity of the gas there. The right edge moves faster than the left edge, so the band on the right is swept out and the band on the left is swept in.

The first term is the change at a fixed point, the only change the box sees. The second is the change met by moving to a place where f is different.

Next, how the volume δV of the blob grows. Each small piece $d\vec{A}$ of its edge δS moves with the gas. In the time δt it sweeps out a thin slab of volume $\vec{v}\cdot d\vec{A}\,\delta t$, the bands of Figure 9.1(b). Where the edge runs outward this volume is positive, and where it runs inward it is negative. Adding over the whole edge,

$$\begin{aligned}\frac{D(\delta V)}{Dt} &\overset{(1)}{=} \oint_{\delta S} \vec{v}\cdot d\vec{A} \\ &\overset{(2)}{=} \int_{\delta V} \nabla\cdot\vec{v}\, dV \\ &\overset{(3)}{=} (\nabla\cdot\vec{v})\,\delta V.\end{aligned} \tag{9.4}$$

(1) The slabs swept by all pieces of the edge, added up and divided by δt.
(2) The divergence theorem.
(3) The blob is small, so $\nabla\cdot\vec{v}$ is nearly the same everywhere inside it.

So $\nabla\cdot\vec{v}$ is the rate at which a blob of gas swells, per unit of its volume.

Now the mass. It is $\rho\,\delta V$, and the product rule splits its change into a part from the density and a part from the volume,

$$\begin{aligned}0 &\overset{(1)}{=} \frac{D(\rho\,\delta V)}{Dt} \\ &\overset{(2)}{=} \delta V\,\frac{D\rho}{Dt} + \rho\,\frac{D(\delta V)}{Dt} \\ &\overset{(3)}{=} \delta V\left[\frac{D\rho}{Dt} + \rho\,\nabla\cdot\vec{v}\right].\end{aligned} \tag{9.5}$$

(1) No gas crosses the edge of the blob, so its mass never changes (P1).
(2) The product rule.
(3) The growth of the volume, from the chain above.

The bracket must vanish,

$$\boxed{\frac{D\rho}{Dt} = -\rho\,\nabla\cdot\vec{v}.} \tag{9.6}$$

A blob of gas thins out where the flow spreads ($\nabla \cdot \vec{v} > 0$) and gets denser where the flow closes in ($\nabla \cdot \vec{v} < 0$). In a collapsing cloud the gas closes in everywhere, so every blob of it gets denser.

This is the continuity equation again. Expanding D/Dt,

$$\begin{aligned} \frac{D\rho}{Dt} + \rho\,\nabla\cdot\vec{v} &\overset{(1)}{=} \frac{\partial\rho}{\partial t} + \vec{v}\cdot\nabla\rho + \rho\,\nabla\cdot\vec{v} \\ &\overset{(2)}{=} \frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\vec{v}). \end{aligned} \tag{9.7}$$

(1) Eq. (9.3) with $f = \rho$.
(2) The product rule, $\nabla\cdot(\rho\vec{v}) = \vec{v}\cdot\nabla\rho + \rho\,\nabla\cdot\vec{v}$.

So Eqs. (9.6) and (9.2) say the same thing, seen from the blob and from the box.

Where the two views differ. Both routes take the time derivative of a density times a volume, and the product rule splits it into a change of the density and a change of the volume. For the fixed box the volume never changes, so only $\partial\rho/\partial t$ is left, and the mass inside changes because gas crosses the walls (Figure 9.1a). For the blob nothing crosses the edge, so the mass stays the same, and any change of density is paid for by a change of volume (Figure 9.1b). The divergence theorem is used once in each route, at the walls of the box or at the edge of the blob. The box view, called **Eulerian**, suits equations written at fixed points in a star. The blob view, called **Lagrangian**, suits questions about what happens to a given piece of gas, and the Euler equation below is Newton's second law for one blob.

Momentum. The i-th component of the momentum inside V is $\int_V \rho v_i\,dV$. It changes in three ways. Gas crossing the surface carries momentum ρv_i at the rate $\rho v_i\,\vec{v}\cdot d\vec{A}$. The pressure P of the surrounding gas pushes inward on each surface element with the force $-P\,d\vec{A}$. Gravity pulls on every element of mass with the acceleration $\vec{g}$. Writing the surface element as dA_j and summing over repeated indices,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho v_i\,dV &\overset{(1)}{=} -\oint_S \rho v_i v_j\,dA_j - \oint_S P\,\delta_{ij}\,dA_j + \int_V \rho g_i\,dV \\ &\overset{(2)}{=} \int_V \Big[-\partial_j\big(\rho v_i v_j + P\delta_{ij}\big) + \rho g_i \Big] dV. \end{aligned} \tag{9.8}$$

(1) The three ways listed above (P1). The pressure force is written with δ_{ij} so that both surface terms have the same form.
(2) The divergence theorem, applied to each surface term.

The quantity being differentiated, $\rho v_i v_j + P\delta_{ij}$, is the flux of the i-th component of momentum in direction j. It is a 3×3 matrix, the **momentum flux**,

$$\Pi = \begin{pmatrix} \rho v_x^2 + P & \rho v_x v_y & \rho v_x v_z \\ \rho v_y v_x & \rho v_y^2 + P & \rho v_y v_z \\ \rho v_z v_x & \rho v_z v_y & \rho v_z^2 + P \end{pmatrix} = \rho\,\vec{v}\,\vec{v}^{\,\mathsf{T}} + P\,\mathbb{1}. \tag{9.9}$$

The first part is momentum carried by the bulk flow. The second part, P on the diagonal, is momentum carried across the surface by the random motion of the particles, which is what pressure is (Section 9.3). Removing the integrals as for the mass, and using Π,

$$\frac{\partial(\rho v_i)}{\partial t} + \partial_j \Pi_{ij} = \rho g_i. \tag{9.10}$$

The left side can be simplified with the continuity equation. Writing $\Pi_{ij} = \rho v_i v_j + P\delta_{ij}$,

$$\frac{\partial(\rho v_i)}{\partial t} + \partial_j(\rho v_i v_j) \overset{(1)}{=} v_i \left[\frac{\partial \rho}{\partial t} + \partial_j(\rho v_j)\right] + \rho\left[\frac{\partial v_i}{\partial t} + v_j\,\partial_j v_i\right]$$
$$\overset{(2)}{=} \rho\left[\frac{\partial v_i}{\partial t} + v_j\,\partial_j v_i\right]. \tag{9.11}$$

(1) The product rule, applied to ρv_i and to $\rho v_j \cdot v_i$.
(2) The first bracket is the continuity equation (9.2), so it vanishes.

With $\partial_j(P\delta_{ij}) = \partial_i P$, Eq. (9.10) becomes the **Euler equation**,

$$\boxed{\rho\left(\frac{\partial \vec{v}}{\partial t} + (\vec{v}\cdot\nabla)\vec{v}\right) = -\nabla P + \rho\,\vec{g}.} \tag{9.12}$$

The left side is $\rho\, D\vec{v}/Dt$ by Eq. (9.3), the mass per unit volume times the acceleration of a blob that moves with the gas. The right side is the force on it. Pressure acts only through its gradient, since a uniform pressure pushes equally on all sides of a blob. The gravity $\vec{g}$ comes from P2 and the pressure from P3.

Counting unknowns. The unknowns are ρ, the three components of $\vec{v}$, and P, five in all. The equations are continuity and the three components of Euler, four in all. One relation is missing. It is the equation of state, which gives P from the density and the temperature (P3). The temperature is then a sixth unknown, and the sixth equation is conservation of energy, written for a star in Section 12.2. This is the same count of unknowns against equations that fixed the twelve constants of the two-body problem.

Angular momentum. When gravity is the only force and it points to a fixed centre, $\vec{g} \parallel \hat{r}$, each blob of gas keeps its angular momentum per unit mass $\vec{r}\times\vec{v}$. Following a blob,

$$\frac{\mathrm{d}}{\mathrm{d}t}(\vec{r}\times\vec{v}) = \vec{v}\times\vec{v} + \vec{r}\times\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} = \vec{r}\times\vec{g} = 0,$$

exactly as for the single body of Eq. (6.17). With pressure present, neighbouring blobs can trade angular momentum through pressure forces, but the internal forces come in equal and opposite pairs, so the total angular momentum of an isolated cloud does not change. This is what turns a collapsing cloud into a disk (Section 10.6).

9.2 Gravity of a Sphere (P2)

Newton's law of gravity between two points, GMm/r^2, ran the whole of Chapter 6. A star or a gas cloud is not a point. Every bit of it pulls on every other bit, and adding all those pulls one by one looks hopeless.

Imagine a well dug straight down to the centre of the Earth, and suppose for now that the Earth has the same density everywhere. At the surface you have your usual weight. Halfway down you weigh half as much. The rock above your head forms a shell around you, and the pulls from its different sides cancel exactly. Only the ball of rock below you counts, and it is smaller. Inside a spherical star the same is true. A small blob of gas at radius r feels only the mass inside r, and feels it as if that mass sat at the centre. Everything outside r, however it is arranged in shells, pulls with zero net force. That is principle P2.

The well in numbers. Treat the Earth as a uniform sphere of density ρ. The pull at radius r comes only from the ball inside r, whose mass $m(r)$ is its volume times the density, so

$$g(r) \overset{(1)}{=} \frac{G\,m(r)}{r^2} \overset{(2)}{=} \frac{G}{r^2}\,\frac{4\pi}{3} r^3 \rho = \frac{4\pi}{3}\,G\rho\,r,$$

(1) Only the mass inside r pulls, and it pulls as if it sat at the centre.
(2) $m(r) = \frac{4}{3}\pi r^3 \rho$ for a constant density.

which falls linearly to zero at the centre. The Earth's mean density is its mass over its volume, $\rho = 5.972 \times 10^{24}\,\mathrm{kg} / (\frac{4}{3}\pi (6.371 \times 10^{6}\,\mathrm{m})^3) = 5513\,\mathrm{kg\,m^{-3}}$. Halfway down, at $r = R/2 = 3.186 \times 10^{6}\,\mathrm{m}$,

$$g = \frac{4\pi}{3} \times 6.674 \times 10^{-11} \times 5513 \times 3.186 \times 10^{6}\,\mathrm{m\,s^{-2}} = 4.91\,\mathrm{m\,s^{-2}},$$

half of the $9.82\,\mathrm{m\,s^{-2}}$ that the same formula gives at the surface.

Figure 9.2 separates the mass inside radius r from the outer shell. Only the inner mass contributes to the net pull on the blob at that radius.

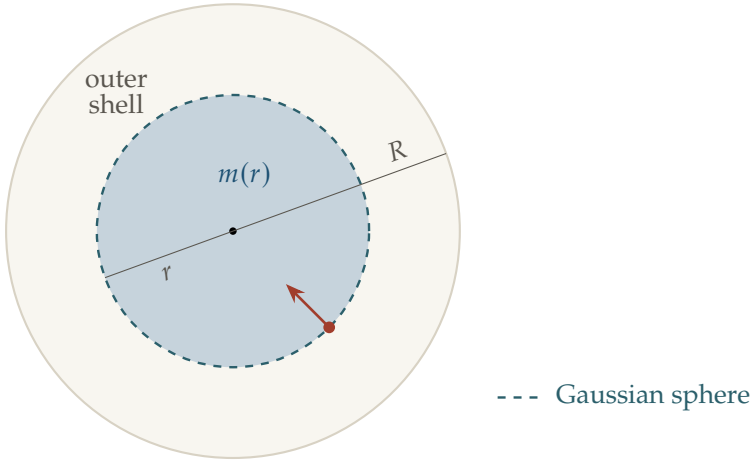


Figure 9.2: A spherical star of radius R. The dashed sphere of radius r passes through a small blob of gas (dot), which is pulled toward the centre (arrow). The shaded ball holds the mass $m(r)$. The outer shell lies between r and R.

This shell result turns a complicated spherical distribution into one number, the mass enclosed at radius r. We will repeatedly use that number to find the pull on gas inside a star and the orbital speed of matter in a galaxy.

Gauss's law for gravity. A point mass m at the origin produces $\vec{g} = -Gm\,\hat{r}/r^2$. Through a sphere of radius r around it, with $d\vec{A} = r^2\,d\Omega\,\hat{r}$,

$$\oint \vec{g} \cdot d\vec{A} \overset{(1)}{=} \oint \left(-\frac{Gm}{r^2}\right) r^2\,d\Omega \overset{(2)}{=} -4\pi G m. \tag{9.13}$$

(1) $\hat{r} \cdot \hat{r} = 1$.
(2) The r^2 cancels, and the solid angle of a whole sphere is 4π.

The radius dropped out, so the answer is the same for any closed surface around the mass. Each patch of a surface subtends some solid angle at the mass, and the flux through the patch

depends only on that solid angle. A mass outside the surface gives zero, because every line from it enters the surface once and leaves once, with equal and opposite contributions. Adding the fields of many masses therefore gives

$$\boxed{\oint_S \vec{g}\cdot d\vec{A} = -4\pi G\, m_{\text{inside}}.} \tag{9.14}$$

The field of a sphere. For a spherically symmetric body, $\vec{g}$ can only point along $\hat{r}$ and can only depend on r. Any other direction would single out a direction in space, and a sphere has none. So $\vec{g} = g(r)\,\hat{r}$. Define the mass inside radius r,

$$m(r) \equiv \int_0^r 4\pi r'^2 \rho(r')\, dr', \qquad \frac{\mathrm{d}m}{\mathrm{d}r} = 4\pi r^2 \rho. \tag{9.15}$$

Applying Eq. (9.14) to a sphere of radius r,

$$\begin{aligned} 4\pi r^2\, g(r) &\overset{(1)}{=} -4\pi G\, m(r) \\ \Longrightarrow \qquad \vec{g}(r) &= -\frac{G\, m(r)}{r^2}\,\hat{r}. \end{aligned} \tag{9.16}$$

(1) $\vec{g}\cdot d\vec{A} = g(r)\, dA$ everywhere on the sphere, and the area is $4\pi r^2$.

$$\boxed{\vec{g}(r) = -\frac{G\, m(r)}{r^2}\,\hat{r}} \tag{9.17}$$

The pull at radius r depends only on $m(r)$. How the mass outside r is arranged does not enter, as long as it is spherical.

Poisson's equation. The divergence theorem turns Eq. (9.14) into $\int_V (\nabla\cdot\vec{g} + 4\pi G\rho)\, dV = 0$ for every volume, so $\nabla\cdot\vec{g} = -4\pi G\rho$. With $\vec{g} = -\nabla\Phi$, where Φ is the gravitational potential,

$$\nabla^2\Phi = 4\pi G\rho. \tag{9.18}$$

This local form holds for any shape, spherical or not.

Gravitational energy. Build the sphere shell by shell, bringing each shell in from far away. When the ball already built has radius r and mass $m(r)$, the next shell of mass dm lands at radius r. By Eq. (9.17) the ball pulls it as a point mass $m(r)$ at the centre, and the shells not yet built do not exist. Bringing dm from infinity to r therefore releases the energy $G\, m(r)\, dm/r$. Adding all the shells gives the **gravitational energy**

$$\boxed{U = -\int_0^M \frac{G\, m}{r}\, dm.} \tag{9.19}$$

U is negative, because energy is released in building the body, and it is more negative when the same mass sits at smaller radii. Its value for a uniform sphere is found where it is first used, in Section 10.2.

9.3 Counting States (P3)

The Planck spectrum was quoted in Chapter 4 (Eq. (4.13)), and the Maxwell distribution of velocities in Chapter 8 (Eq. (8.27)). Each was taken as given. A star needs more than that. Its interior holds hot gas, a bath of light, and in some stars electrons packed so tightly that quantum mechanics stops them from coming closer. All three push. Quoting a separate formula for each hides the fact that they are the same calculation.

Think of a theatre with numbered seats. To know how many people sit in the front rows, two things are needed. One is how many seats there are. The other is how likely each seat is to be taken. Some audiences fill seats at random, some crowd the cheap seats, and some follow a strict rule of one person per seat. Particles fill quantum states in the same way. Quantum mechanics fixes how many states there are in a given range of position and momentum. Temperature, and whether particles may share a state, fix how full each state is. Counting both gives the pressure of any gas, whether of atoms, photons or electrons. That is principle P3.

The smallest example is the air in a room. At 293 K and normal pressure it has $n = 2.50 \times 10^{25}$ molecules per cubic metre, and

$$P = n k_B T = (2.50 \times 10^{25})(1.381 \times 10^{-23})(293)\,\text{Pa} = 1.01 \times 10^{5}\,\text{Pa},$$

the pressure we live in. The light in the same room, at the same temperature, pushes with $aT^4/3 = 1.9 \times 10^{-6}$ Pa (Eq. (9.30) below), about 10^{-11} of the air. In the core of the Sun the share of light rises to about 5×10^{-4} (Problem 9.7.2), and in the most massive stars light carries most of the pressure.

We count states in the combined space of position and momentum, with one quantum state per cell of volume h^3 for each spin direction. How particles fill those cells determines how much momentum they carry across a surface and hence their pressure. Changing the occupancy rule will take us from ordinary gas to photons and then to degenerate electrons.

How many states. Put one particle in a cube of side b and volume $V = b^3$, with periodic boundary conditions. Its wavefunction $e^{i\vec{p}\cdot\vec{r}/\hbar}$ must repeat after a distance b in each direction, so $p_x b/\hbar = 2\pi$ times an integer, and the same for y and z,

$$p_x = \frac{h}{b}\,q_x, \qquad p_y = \frac{h}{b}\,q_y, \qquad p_z = \frac{h}{b}\,q_z, \qquad q_x, q_y, q_z \in \mathbb{Z}. \tag{9.20}$$

The allowed momenta are the points of a grid with spacing h/b. The number of grid points in a small box of momentum space is then

$$\begin{aligned} dq_x\, dq_y\, dq_z &\overset{(1)}{=} \frac{b}{h}\,dp_x\,\frac{b}{h}\,dp_y\,\frac{b}{h}\,dp_z \\ &\overset{(2)}{=} \frac{V\,d^3p}{h^3}. \end{aligned} \tag{9.21}$$

(1) One grid point per interval h/b in each direction.
(2) $b^3 = V$ and $d^3p = dp_x\,dp_y\,dp_z$.

$$\boxed{\text{number of states in } d^3x\,d^3p = \frac{d^3x\,d^3p}{h^3} \quad \text{(per spin state)}} \tag{9.22}$$

This does not depend on what the particle is. Every particle gets one state per h^3 of phase space for each spin direction.

How full. Let $f(p)$ be the mean number of particles in one state of momentum p. Statistical mechanics gives it in the three cases needed.

- Particles far apart compared with their quantum wavelength (an ordinary gas) occupy states rarely, with $f \propto e^{-p^2/2mk_BT}$.
- Photons can pile into one state without limit, and $f = 1/(e^{pc/k_BT} - 1)$.
- Electrons allow at most one particle per state, $f \le 1$. At low temperature every state up to some momentum is full and every state above it is empty. This case is worked out in Section 12.3.

Combining f with Eq. (9.22), the number of particles per unit volume with momentum in d^3p is

$$dn = (\text{spin states})\,\frac{f(p)\,d^3p}{h^3}, \qquad n = \int dn. \tag{9.23}$$

Pressure as momentum flux. A particle with momentum $\vec{p}$ moves with velocity $\vec{v}(p)$ in the same direction. Across a surface it carries the momentum component p_i at the rate v_j per unit area in direction j, so the particles together carry the momentum flux $\int p_i v_j\,dn$. This is the random-motion part of the momentum flux Π of Eq. (9.9), seen from a frame in which the gas is at rest. When f depends only on the size of $\vec{p}$, as in all three cases above,

$$\begin{pmatrix} \int p_x v_x\,dn & \int p_x v_y\,dn & \int p_x v_z\,dn \\ \int p_y v_x\,dn & \int p_y v_y\,dn & \int p_y v_z\,dn \\ \int p_z v_x\,dn & \int p_z v_y\,dn & \int p_z v_z\,dn \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} P & 0 & 0 \\ 0 & P & 0 \\ 0 & 0 & P \end{pmatrix}. \tag{9.24}$$

(1) Off the diagonal, the integrand $p_x v_y$ changes sign when $p_y \to -p_y$ while dn does not, so the integral vanishes. On the diagonal the three integrals are equal because no direction is special. Their common value is the pressure.

Adding the three diagonal entries gives $3P = \int \vec{p}\cdot\vec{v}\,dn$, and since $\vec{v} \parallel \vec{p}$,

$$\boxed{P = \frac{1}{3}\int p\,v(p)\,dn.} \tag{9.25}$$

The factor $\frac{1}{3}$ is the average of $\cos^2\theta$ over all directions. Only the distribution of momenta enters. The temperature, the quantum rules and the kind of particle enter only through dn.

Pressure and energy density. The kinetic energy per unit volume is $u = \int(\text{kinetic energy})\,dn$. In two limits the integrand of Eq. (9.25) is a fixed multiple of the kinetic energy.

- Slow particles have $v = p/m$ and kinetic energy $p^2/2m$, so $pv = p^2/m = 2\times$ kinetic energy, and $P = \frac{2}{3}u$.
- Particles at the speed of light (photons, or electrons with $pc \gg mc^2$) have $v = c$ and energy pc, so $pv = pc =$ energy, and $P = \frac{1}{3}u$.

Writing both as one formula with an exponent γ,

$$\boxed{P = (\gamma - 1)\,u, \qquad \gamma = \tfrac{5}{3}\ (\text{slow particles}), \qquad \gamma = \tfrac{4}{3}\ (\text{particles at speed } c).} \tag{9.26}$$

This depends only on how a particle's energy grows with its momentum, not on f. It holds for an ordinary gas, for light and for degenerate electrons alike. The difference between $\gamma = \frac{5}{3}$ and $\gamma = \frac{4}{3}$ decides whether a star is firmly bound or barely bound (Section 10.2).

The ideal gas. For an ordinary gas, $f \propto e^{-p^2/2mk_BT}$. Normalising dn so that $\int dn = n$ and writing $d^3p = 4\pi p^2 dp$, Eq. (9.25) gives

$$\begin{aligned} P &\overset{(1)}{=} \frac{1}{3}\,\frac{n\int_0^\infty (p^2/m)\, e^{-p^2/2mk_BT}\, 4\pi p^2\, dp}{\int_0^\infty e^{-p^2/2mk_BT}\, 4\pi p^2\, dp} \\ &\overset{(2)}{=} \frac{1}{3}\, n\, \frac{3mk_BT}{m} \\ &= nk_BT. \end{aligned} \tag{9.27}$$

(1) Eq. (9.25) with $v = p/m$. Dividing by the same integral without p^2/m normalises dn to n.

(2) The Gaussian moments $\int_0^\infty p^4 e^{-p^2/2mk_BT} dp = 3mk_BT \int_0^\infty p^2 e^{-p^2/2mk_BT} dp$, which follows from one integration by parts.

$$\boxed{P = nk_BT, \qquad u = \tfrac{3}{2}nk_BT.} \tag{9.28}$$

The second form follows from $u = \frac{3}{2}P$ (Eq. (9.26) with $\gamma = \frac{5}{3}$). For N particles the total kinetic energy is $K = \frac{3}{2}Nk_BT$. The one-dimensional distribution of Eq. (8.27) is one component of this one.

Light. Photons have two polarisations, energy pc and $f = 1/(e^{pc/k_BT} - 1)$. Their energy density is

$$\begin{aligned} u &\overset{(1)}{=} \int_0^\infty pc\, \frac{2}{h^3}\, \frac{4\pi p^2\, dp}{e^{pc/k_BT} - 1} \\ &\overset{(2)}{=} \frac{8\pi (k_BT)^4}{h^3c^3} \int_0^\infty \frac{x^3\, dx}{e^x - 1} \\ &\overset{(3)}{=} \frac{8\pi^5 k_B^4}{15\, h^3c^3}\, T^4. \end{aligned} \tag{9.29}$$

(1) Eq. (9.23) with two spin states, times the energy pc of each photon.

(2) The substitution $x = pc/k_BT$.

(3) $\int_0^\infty x^3 dx/(e^x - 1) = \pi^4/15$, the same integral that gave the Stefan–Boltzmann constant in Eq. (4.25).

The constant in front is the **radiation constant**. Comparing with Eq. (4.25), it is $a = 4\sigma_{\mathrm{SB}}/c = 7.566 \times 10^{-16}\,\mathrm{J\,m^{-3}\,K^{-4}}$. With Eq. (9.26) at $\gamma = \frac{4}{3}$,

$$\boxed{u = aT^4, \qquad P = \tfrac{1}{3}aT^4.} \tag{9.30}$$

The pressure of light depends on the temperature alone, not on any density.

The Planck spectrum. Keeping the photon energy $h\nu = pc$ instead of integrating over it gives the energy per unit volume and per unit frequency. With $p = h\nu/c$ and $dp = h\, d\nu/c$ in the first line above,

$$u_\nu\, d\nu = \frac{8\pi h\nu^3}{c^3}\, \frac{d\nu}{e^{h\nu/k_BT} - 1}.$$

This light moves at c and is spread evenly over all 4π steradians of direction, so the energy crossing a unit area per unit time, per unit frequency and per steradian is $c\, u_\nu/4\pi$,

$$\boxed{B_\nu(T) = \frac{c}{4\pi}\, u_\nu = \frac{2h\nu^3}{c^2}\, \frac{1}{e^{h\nu/k_BT} - 1}.} \tag{9.31}$$

This is the Planck function used in Chapter 4, now derived by counting states. The quantity it describes, the intensity of light in one direction, is developed in Chapter 11.

The mechanism does not matter. Nothing in this derivation asked how the light was made. It used only the number of photon states, Eq. (9.22), and how full each state is at temperature T. So light that has come into thermal equilibrium with matter at temperature T has the Planck spectrum, whatever produced it. Electrons braking in collisions with ions, atoms jumping between energy levels and gamma rays from nuclear fusion all end in the same $B_\nu(T)$, once the light has been absorbed and re-emitted often enough to share the temperature of the matter. This is detailed balance at work (P4). In equilibrium each process that makes a photon of frequency ν is matched by its own reverse, which absorbs one, and that balance fixes the ratio of emission to absorption at $B_\nu(T)$ for every process separately. This is Kirchhoff's law (Section 11.3).

The Sun shows this. Fusion in its core releases gamma rays with energies near a million electron volts, yet the light that leaves its surface is close to a blackbody at 5772 K. On the way out the energy is absorbed and re-emitted about 5×10^{23} times (Problem 11.9.3), and at every depth it takes on the local temperature, so nothing of the fusion spectrum survives. The microwave background is another case. It was made by many processes in the hot early universe and is today the most perfect blackbody known (Section 16.3).

What the mechanism does decide is how fast equilibrium is reached, and whether it is reached at all. The matter must be thick enough that light is absorbed many times before it escapes, an optical depth much larger than one (Section 11.2). Thin matter shows its mechanism instead. A nebula glows in the spectral lines of its own atoms, and the fast electrons of a radio galaxy give a synchrotron power law (Chapter 15). Neither is a blackbody, because in neither does the light come into equilibrium with the matter.

9.4 Balance of Rates (P4)

The temperature of interstellar gas was set by matching heating to cooling (Section 8.6). That balance says where the gas ends up. It does not say how long it takes to get there, or how the many processes inside the gas share the work.

Think of two rooms joined by a door, with people walking through it both ways. After a while the number in each room stops changing. People have not stopped walking. As many walk one way as the other. If one room is more pleasant, people leave it more slowly, and it ends up holding more of them. Atoms in a gas move between energy levels in the same way, kicked up and down by collisions. In a steady state the rate into each level equals the rate out. In full equilibrium each process is balanced by its own reverse. That is principle P4.

Suppose ninety people move between two rooms. Each person in A leaves once a minute on average, while each person in B leaves twice a minute. Equal flows require $1 \times N_A = 2 \times N_B$, so the steady populations are $N_A = 60$ and $N_B = 30$. The people keep moving even though the numbers stop changing. The rooms approach that split in about 20 s (Eq. (9.38)). In a gas, the same distinction between steady populations and ongoing transitions will let us derive both equilibrium ratios and the time needed to reach them.

Steady state. Let N be a number of particles, supplied at a constant rate S per unit time and each lost at the rate $1/\tau$, so that it survives for a time τ on average. Then

$$\frac{\mathrm{d}N}{\mathrm{d}t} = S - \frac{N}{\tau}. \tag{9.32}$$

Starting from $N = 0$ at $t = 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(N\, e^{t/\tau}\right) \overset{(1)}{=} S\, e^{t/\tau}$$
$$\Longrightarrow \qquad N(t) \overset{(2)}{=} S\tau\left(1 - e^{-t/\tau}\right). \tag{9.33}$$

(1) Multiply the equation by $e^{t/\tau}$. The left side becomes a total derivative.
(2) Integrate from 0 to t and divide by $e^{t/\tau}$.

$$\boxed{N \to S\,\tau \quad \text{after a few } \tau.} \tag{9.34}$$

The steady amount is the supply rate times the lifetime. How the loss happens does not matter, only its rate.

Detailed balance. Take two states A and B, with energies E_A and E_B. A particle in A jumps to B at the rate k_{AB}, and one in B jumps back at k_{BA}. The populations change as one matrix equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} N_A \\ N_B \end{pmatrix} = \begin{pmatrix} -k_{AB} & k_{BA} \\ k_{AB} & -k_{BA} \end{pmatrix}\begin{pmatrix} N_A \\ N_B \end{pmatrix}. \tag{9.35}$$

The solutions are set by the eigenvalues λ of the matrix,

$$0 \overset{(1)}{=} (-k_{AB} - \lambda)(-k_{BA} - \lambda) - k_{AB}k_{BA}$$
$$\overset{(2)}{=} \lambda\,(\lambda + k_{AB} + k_{BA}). \tag{9.36}$$

(1) The determinant of the matrix minus λ times the unit matrix.
(2) Multiply out. The products $k_{AB}k_{BA}$ cancel.

The eigenvalue $\lambda = 0$ belongs to the state that does not change, the equilibrium. Its eigenvector satisfies $-k_{AB}N_A + k_{BA}N_B = 0$, so

$$\frac{N_B}{N_A} = \frac{k_{AB}}{k_{BA}}. \tag{9.37}$$

The other eigenvalue, $-(k_{AB} + k_{BA})$, is the rate at which any departure from equilibrium dies away,

$$\text{relaxation time} = \frac{1}{k_{AB} + k_{BA}}. \tag{9.38}$$

For the two rooms, $1/(1 + 2)$ minutes, which is 20 s.

Now add P3. In thermal equilibrium at temperature T, the two populations also follow the Boltzmann factor, $N_B/N_A = e^{-(E_B - E_A)/k_BT}$. Both statements hold at once, so

$$\boxed{\frac{k_{AB}}{k_{BA}} = e^{-(E_B - E_A)/k_BT}.} \tag{9.39}$$

The ratio of an upward rate to the downward rate is fixed by the temperature alone, whatever the physics of the collisions. The rates themselves belong to single collisions, which do not know whether the populations are in equilibrium. So the relation still holds when the populations are far from equilibrium, as long as the colliding particles have a Maxwell distribution at temperature T. That is what makes it useful. (When a level contains several states, each side is multiplied by its number of states.)

9.5 Random Walk (P5)

In the earlier chapters light travelled in straight lines, from a star to a telescope.

Inside a star it cannot. A photon is absorbed after a short distance and re-emitted in a new direction, again and again. Treating its path as a straight line gives an escape time that is wrong by many orders of magnitude.

Think of a bottle of perfume opened in the corner of a still room. Each molecule flies a tiny distance, hits an air molecule and goes off in a random direction. Each one moves at hundreds of metres per second, yet in perfectly still air the scent would take days to cross the room (Problem 9.7.4). We smell it within minutes only because air currents carry it.

Anything that scatters many times moves this way, by a random walk. After N steps it is not N steps from its start but only about $\sqrt{N}$ steps. That is principle P5.

In numbers, 100 steps of 1 m in random directions leave a walker about 10 m from the start, and 10 000 steps about 100 m. Figure 9.3 shows 500 such walks of 100 steps. Over 100 000 walks the root-mean-square distance comes out as 9.98 steps.

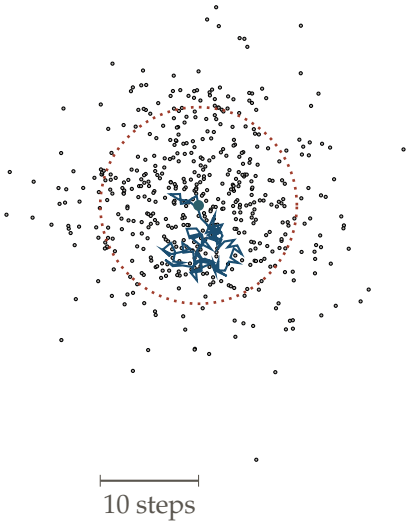


Figure 9.3: End points of 500 random walks of 100 unit steps each in the plane (grey dots), all starting at the centre (dark dot). One of the walks is drawn in full (blue). The dotted circle has a radius of 10 steps.

The $\sqrt{N}$ rule explains why repeated scattering slows the escape of light from a star and the movement of particles through gas. When many walkers start in a region of high concentration, more leave it than enter, giving a net flux down the concentration gradient. We will derive that flux after establishing the distance reached in N steps.

The $\sqrt{N}$ law. Let a particle take N steps $\vec{s}_1, \dots, \vec{s}_N$, each of length s and each in a random direction independent of the others. Its displacement is $\vec{R} = \sum_i \vec{s}_i$. Averaging over many walks,

$$\begin{aligned} \langle R^2 \rangle &\overset{(1)}{=} \sum_{i=1}^{N} \sum_{j=1}^{N} \langle \vec{s}_i \cdot \vec{s}_j \rangle \\ &\overset{(2)}{=} \sum_{i=1}^{N} \langle s_i^2 \rangle + \sum_{i \neq j} \langle \vec{s}_i \cdot \vec{s}_j \rangle \\ &\overset{(3)}{=} N s^2. \end{aligned} \tag{9.40}$$

(1) $R^2 = \vec{R} \cdot \vec{R}$, with each $\vec{R}$ written as its sum of steps.
(2) Split the double sum into the terms with $i = j$ and the rest.
(3) Each step has length s. Two different steps point in independent random directions, so their dot product is as often negative as positive and averages to zero. This is the same step that makes the variances of independent noise sources add (Eq. (7.10)).

$$\boxed{R_{\text{rms}} = \sqrt{N}\, s.} \tag{9.41}$$

Real free paths vary in length. For the usual exponential spread of free paths with mean s, the average of s_i^2 is $2s^2$, which changes only the factor in front, not the $\sqrt{N}$.

To cover a distance D the walk needs $N = (D/s)^2$ steps. At speed v each step takes s/v, so the time is

$$t = N \frac{s}{v} = \frac{D^2}{s\, v}. \tag{9.42}$$

The time grows as D^2. Twice the distance takes four times as long. Compared with flying straight across, the walk is slower by D/s, the number of free paths across the distance.

Flux down a gradient. Let the number density $n(x)$ of the walkers change along x, and look at the particles crossing the plane at x. A particle crossing it at an angle θ to the x-axis last scattered, on average, one free path back along its flight, at $x - s\cos\theta$. It therefore brings the density found there. Particles fly in all directions equally, a fraction $d\Omega/4\pi$ of them within a solid angle $d\Omega$, so the net number crossing per unit area per unit time is

$$\begin{aligned} J &\overset{(1)}{=} \int \frac{d\Omega}{4\pi}\, v \cos\theta\; n(x - s\cos\theta) \\ &\overset{(2)}{=} \int \frac{d\Omega}{4\pi}\, v \cos\theta \left[n(x) - s\cos\theta\, \frac{\mathrm{d}n}{\mathrm{d}x} \right] \\ &\overset{(3)}{=} -v\, s\, \frac{\mathrm{d}n}{\mathrm{d}x} \int \frac{d\Omega}{4\pi} \cos^2\theta \\ &\overset{(4)}{=} -\frac{1}{3}\, v\, s\, \frac{\mathrm{d}n}{\mathrm{d}x}. \end{aligned} \tag{9.43}$$

(1) Particles moving at angle θ cross the plane at the rate $v\cos\theta$ per unit density, positive for those moving toward $+x$.
(2) A Taylor expansion of n over one free path.
(3) $\int \cos\theta\, d\Omega = 0$, so the uniform part $n(x)$ carries no net flux.
(4) The average of $\cos^2\theta$ over all directions is $\frac{1}{3}$, as in the pressure integral (9.25).

In three dimensions the same holds along each axis,

$$\boxed{\vec{J} = -\frac{1}{3}\, v\, s\, \nabla n.} \tag{9.44}$$

Particles flow from where there are many to where there are few. No force is needed, only random motion and a difference in numbers. The flow is larger when the steps are longer. For photons inside a star, with the energy density aT^4 in place of n, $v = c$ and the mean free path in place of s, this becomes the flow of energy out of the star (Section 11.4).

9.6 Compare Two Scales (P6)

Chapter 5 decided whether two stars are resolved by comparing their separation with λ/D (Eq. (5.46)). Chapter 8 chose between Rayleigh, Mie and geometric scattering by comparing the size of a grain with the wavelength (Table 8.2). Each time, a single ratio decided which physics mattered. A star has many competing effects at once, and most of them are negligible in any given place. Knowing which ones can be dropped is most of the work.

Think of crossing a road. If you can cross before the car reaches you, you are safe, however fast the car is. If the car arrives first, you are not. Only the ratio of the two times matters, and when it is close to one, small changes decide. In a gas cloud, sound tries to smooth out a squeeze while gravity tries to deepen it. Whichever acts faster wins. Every regime in astrophysics is chosen in this way. That is principle P6.

Gravity and density give a simple example. The constant G has units $\mathrm{m^3\,kg^{-1}\,s^{-2}}$ and density ρ has units $\mathrm{kg\,m^{-3}}$, so $G\rho$ has units $\mathrm{s^{-2}}$. A time made from these quantities must therefore scale as $(G\rho)^{-1/2}$. At the Earth's mean density, $5514\,\mathrm{kg\,m^{-3}}$, that scale is 27 minutes. The units reveal the dependence, but not the numerical factor. We can find that factor in a concrete gravitational problem and then compare its time with other timescales.

A time from units, checked against Kepler. The orbit of a satellite that skims the surface of a uniform sphere of density ρ and radius a gives the same time. Kepler's third law, Eq. (6.50), sets the orbital angular speed to $\sqrt{GM/a^3}$, so the period is

$$\begin{aligned} \text{period} &\overset{(1)}{=} 2\pi\sqrt{\frac{a^3}{GM}} \\ &\overset{(2)}{=} 2\pi\sqrt{\frac{a^3}{G\,\frac{4}{3}\pi\rho\, a^3}} \\ &= \sqrt{\frac{3\pi}{G\rho}}. \end{aligned} \tag{9.45}$$

(1) Eq. (6.50), one full turn at the angular speed $\sqrt{GM/a^3}$.
(2) The mass of a uniform sphere, $M = \frac{4}{3}\pi\rho a^3$.

The radius has dropped out. The period depends on the density alone, and it is $(G\rho)^{-1/2}$ times the number $\sqrt{3\pi} \approx 3.1$. For the Earth it is 84 minutes, the period of a satellite skimming the surface. Real satellites, 400 km up, take about 92 minutes. The units had already said that only the number in front could be missing.

Reading a ratio. A ratio far from one means that one of the two effects can be dropped, and the physics simplifies. A ratio near one means that both matter, and something changes there, often an instability. The ratios that decide the regime in stars and gas are listed in Table 9.2, with the place where each is used.

Ratio	Small	Large	Where
size / mean free path of light	transparent	opaque	Ch. 11
sound-crossing time / free-fall time	stable	collapses	Sec. 10.4
free-fall time / Kelvin–Helmholtz time	slow evolution	–	Sec. 10.5
particle spacing / quantum wavelength	degenerate	classical	Sec. 12.3
pressure of light / pressure of gas	gas holds the star	light holds the star	Ch. 12
luminosity / Eddington luminosity	bound	blown apart	Sec. 11.5
GM/Rc^2	Newtonian gravity	general relativity	Ch. 13

Table 9.2: Ratios that decide the regime, and where each is used.

9.7 Problem Set: Six Principles

9.7.1 Problem 1: Weight in a Well

Problem Statement: Model the Earth as a uniform sphere of radius $R = 6371\,\mathrm{km}$ with surface gravity $9.81\,\mathrm{m\,s^{-2}}$.

(a) Find the gravity at a depth d below the surface, and at a height h above it.

(b) At what height above the surface is gravity as weak as it is 1000 km below the surface?

Essential Ideas:

- Inside, only the mass within radius r pulls (Eq. (9.17)), and for uniform density it is $M(r/R)^3$.
- Outside, the whole mass pulls as a point at the centre.

Detailed Solutions:

(a) **Inside and outside**. With $g_0 = GM/R^2$,

$$\boxed{g(R-d) = g_0\left(1 - \frac{d}{R}\right), \qquad g(R+h) = \frac{g_0}{(1+h/R)^2}.}$$

(b) **Equal gravity**. At $d = 1000\,\mathrm{km}$, $g = g_0(1 - 1000/6371) = 0.843\,g_0$. Setting $(1+h/R)^{-2} = 0.843$ gives $1 + h/R = 1.089$.

$$\boxed{h \simeq 0.089\,R \simeq 570\,\mathrm{km}.}$$

Gravity falls faster going up than going down. Going down, the loss of mass below is partly made up by coming closer to the centre. Going up, nothing makes up for the greater distance.

9.7.2 Problem 2: Light Against Gas

Problem Statement: Compare the pressure of light with the pressure of gas.

(a) In a room at 293 K with $n = 2.50 \times 10^{25}\,\mathrm{m^{-3}}$.

(b) At the centre of the Sun, where $\rho = 1.5 \times 10^5\,\mathrm{kg\,m^{-3}}$, $T = 1.57 \times 10^7\,\mathrm{K}$, and the ionised gas has a mean mass per particle of $0.62\,m_{\mathrm{H}}$.

(c) How does the ratio scale with T and n?

Essential Ideas:

- Gas pressure nk_BT (Eq. (9.28)), light pressure $aT^4/3$ (Eq. (9.30)).
- The ratio is a P6 ratio, $aT^3/(3nk_B)$.

Detailed Solutions:

(a) **Room**. $P_{\mathrm{gas}} = 1.01 \times 10^5\,\mathrm{Pa}$ and $P_{\mathrm{light}} = 7.566 \times 10^{-16} \times 293^4/3 = 1.86 \times 10^{-6}\,\mathrm{Pa}$.

$$\boxed{P_{\mathrm{light}}/P_{\mathrm{gas}} \simeq 1.8 \times 10^{-11}.}$$

(b) **Centre of the Sun**. $n = \rho/(0.62\,m_{\mathrm{H}}) = 1.45 \times 10^{32}\,\mathrm{m^{-3}}$, so $P_{\mathrm{gas}} = nk_BT = 3.1 \times 10^{16}\,\mathrm{Pa}$. $P_{\mathrm{light}} = aT^4/3 = 1.5 \times 10^{13}\,\mathrm{Pa}$.

$$\boxed{P_{\mathrm{light}}/P_{\mathrm{gas}} \simeq 5 \times 10^{-4}.}$$

(c) **Scaling**. $P_{\mathrm{light}}/P_{\mathrm{gas}} = aT^3/(3nk_B)$ grows as T^3/n. Hot, thin gas is held up by light. This is why light matters most in the most massive stars, which are hotter inside but not denser.

9.7.3 Problem 3: A Two-Level Atom

Problem Statement: Collisions move atoms between a ground level A and an excited level B that lies 0.010 eV higher. Both levels are single states.

(a) In gas at 100 K, what is N_B/N_A in equilibrium?

(b) If the downward rate is $k_{BA} = 10^{-3}\,\mathrm{s^{-1}}$, what is the upward rate, and how fast is equilibrium reached?

Essential Ideas:

- Detailed balance, Eq. (9.39), fixes the ratio of the two rates.
- The relaxation time is $1/(k_{AB} + k_{BA})$, Eq. (9.38).

Detailed Solutions:

(a) **Populations**. The gap in temperature units is $0.010\,\mathrm{eV}/k_B = 116\,\mathrm{K}$.

$$\boxed{N_B/N_A = e^{-116/100} = 0.31.}$$

(b) **Rates**. $k_{AB} = 0.31\,k_{BA} = 3.1 \times 10^{-4}\,\mathrm{s^{-1}}$, and the relaxation time is $1/(1.31 \times 10^{-3}\,\mathrm{s^{-1}})$.

$$\boxed{k_{AB} = 3.1 \times 10^{-4}\,\mathrm{s^{-1}}, \qquad \text{relaxation time} \simeq 760\,\mathrm{s}.}$$

9.7.4 Problem 4: Perfume in Still Air

Problem Statement: Air molecules move at about $v = 400\,\mathrm{m\,s^{-1}}$ with a mean free path $s = 6.8 \times 10^{-8}\,\mathrm{m}$.

(a) How long does a perfume molecule take to cross a 3 m room by diffusion alone?

(b) How long would it take flying straight?

Essential Ideas:

- The random-walk time $D^2/(sv)$, Eq. (9.42).
- The ratio of the two times is D/s, a P6 ratio.

Detailed Solutions:

(a) **Diffusion**. $t = (3)^2/(6.8 \times 10^{-8} \times 400)\,\mathrm{s} = 3.3 \times 10^5\,\mathrm{s}$.

$$\boxed{t \simeq 4 \text{ days}.}$$

(b) **Straight flight**. $3/400\,\mathrm{s} = 7.5\,\mathrm{ms}$. The ratio of the two times is $D/s = 4.4 \times 10^7$. In a real room, air currents carry the scent across in minutes, so the bulk motion of the air beats diffusion by far. Inside stars the same competition between carrying energy by motion and by diffusion decides where a star convects (Section 12.2).

9.7.5 Problem 5: Gravity's Clock

Problem Statement: Evaluate $(G\rho)^{-1/2}$ for the Sun (mean density $1410\,\mathrm{kg\,m^{-3}}$) and for a white dwarf (mean density $10^9\,\mathrm{kg\,m^{-3}}$). What does each number say about how quickly the object can respond to a disturbance?

Essential Ideas:

- $(G\rho)^{-1/2}$ is the only time made from G and ρ (Section 9.6).

Detailed Solutions:

$$\boxed{\text{Sun: } (G\rho)^{-1/2} = 3.3 \times 10^3\,\mathrm{s} \simeq 54\,\mathrm{min}, \qquad \text{white dwarf: } 3.9\,\mathrm{s}.}$$

A star responds to a change in its own balance of forces within times of this order. The Sun readjusts within an hour, a white dwarf within seconds. Both are far shorter than the times over which their heat and fuel change, which is why both can be treated as being in balance at every moment (Section 10.5).

Chapter 10

Self-Gravitating Gas

A molecular cloud can span tens of parsecs and remain near 10 K (Table 8.4). Some clouds hold that shape for millions of years, while others break into dense cores and form stars. The difference begins with a contest between gravity, which pulls gas inward, and pressure, which resists compression. Once a cloud starts shrinking, its rotation speeds up (ω increases) and changes the shape of the collapse.

The source of the opposing pressure depends on the body. In a star it is a sum,

$$P = P_{\text{gas}} + P_{\text{light}} + P_{\text{degenerate}} + \cdots ,$$

with each term following from the available states of matter or light (P3, Section 9.3). Gas pressure dominates much of the Sun, radiation pressure matters more in massive stars (Problem 9.7.2), and degenerate electrons support white dwarfs. Planets rely mainly on compressed liquid and solid matter. Whichever source applies, gravity and pressure must be compared before we can tell whether the body holds its shape.

That comparison also reveals a less familiar result. When a self-gravitating body loses energy and contracts, it can become hotter.

10.1 Hydrostatic Equilibrium

A star can remain nearly unchanged for billions of years only if each layer supports the weight above it. The Euler equation describes the motion that would follow from an unbalanced force (P1, Eq. (9.12)). For a nearly stationary, spherical star, that force must vanish. The shell theorem supplies the inward pull of gravity (P2, Eq. (9.17)), leaving the pressure gradient to oppose it.

The bottom book in a pile carries the weight of every book above it. Water pressure grows with depth for the same reason. In a star, each inner layer carries the gas outside it, so pressure increases inward. Across a thin layer, the difference between the pressure below and above must equal the layer's weight. This is momentum conservation for gas at rest (P1) in a spherical gravitational field (P2).

For water with $\rho = 1000\,\text{kg}\,\text{m}^{-3}$ and $g = 9.8\,\text{m}\,\text{s}^{-2}$, pressure rises by $\rho g = 9800\,\text{Pa}$ per metre, or one atmosphere every 10.3 m. The same relation applies to a thin slab of stellar gas: the stronger push on its lower face holds up its weight. Once we know how gravity varies, that balance gives the pressure profile. It also gives the height of an atmosphere, the thickness of a gas disk, and the mass needed to hold hot gas in a galaxy cluster (Problem 10.8.2).

The equation. Set $\vec{v} = 0$ in the Euler equation (9.12) and use the gravity of a sphere (9.17). Only the radial component survives,

$$\begin{aligned} 0 &\overset{(1)}{=} -\nabla P + \rho\,\vec{g} \\ &\overset{(2)}{=} -\frac{\mathrm{d}P}{\mathrm{d}r}\,\hat{r} - \rho\,\frac{G\,m(r)}{r^2}\,\hat{r}. \end{aligned} \tag{10.1}$$

(1) Eq. (9.12) with $\vec{v} = 0$ everywhere and at all times.
(2) Spherical symmetry, so P depends only on r, and $\vec{g}$ from Eq. (9.17).

$$\boxed{\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{G\,m(r)\,\rho}{r^2}.} \tag{10.2}$$

This is **hydrostatic equilibrium**. It describes any star whose layers are nearly at rest. The equation sets the required pressure gradient; the equation of state tells us what produces that pressure.

How large is the central pressure? Estimate the pressure gradient as $-P_c/R$ and use M/R^3 for a typical density. Hydrostatic balance then gives $P_c \sim GM^2/R^4$, or 1.1×10^{15} Pa for the Sun. Detailed models give 2.5×10^{16} Pa (Chapter 12). The Sun's mass is concentrated toward the centre, where gravity and pressure are stronger than this average estimate assumes.

An atmosphere. Near the surface of a planet, $m(r)$ and r hardly change, so Gm/r^2 is the constant surface gravity g. If the air is an ideal gas at one temperature T, with mean mass μm_{H} per particle, Eq. (9.28) gives $\rho = P\,\mu m_{\mathrm{H}}/(k_B T)$. With height z in place of r,

$$\begin{aligned} \frac{\mathrm{d}P}{\mathrm{d}z} &\overset{(1)}{=} -\rho g \\ &\overset{(2)}{=} -\frac{\mu m_{\mathrm{H}}\,g}{k_B T}\,P, \end{aligned} \tag{10.3}$$

(1) Eq. (10.2) with $Gm/r^2 = g$ held constant.
(2) The ideal-gas law solved for ρ.

so the pressure falls exponentially,

$$P(z) = P(0)\,e^{-z/H}, \qquad H = \frac{k_B T}{\mu m_{\mathrm{H}}\,g}. \tag{10.4}$$

The **scale height** H is the height over which the pressure falls by a factor e. For the Earth's air ($T = 288$ K, mean molecular mass 28.97 atomic mass units) it is 8.4 km. A hotter or lighter gas reaches higher.

Planets, and why their interiors are layered. Planets obey Eq. (10.2) too, though rock, ice, and liquid provide much of their pressure. Deeper layers carry more weight, so pressure rises inward. Temperature usually rises as well. Each depth therefore has a different position (P, T) on a material's phase diagram. Crossing a phase boundary changes the stable form of the material. Rock can take denser mineral structures, while hydrogen in giant planets becomes metallic. Hydrostatic balance sets the pressure profile; the phase diagram sets the layers.

10.2 The Virial Theorem

Hydrostatic equilibrium gives a pressure profile only after the density and equation of state are known. Sometimes we need less detail. A single relation between a star's gravity, temperature, and total energy can tell us what happens when it loses heat.

A satellite that loses energy to drag sinks to a lower orbit, where it moves faster. At 400 km altitude the station moves at $7.67\,\mathrm{km\,s^{-1}}$; at 300 km it would move at $7.73\,\mathrm{km\,s^{-1}}$. Gravity makes lower-energy orbits faster. In a star, the gas particles also move within a gravitational well, and their average kinetic energy sets the temperature. Adding hydrostatic balance over the whole star gives the virial theorem (P1 and P2), which we read as a temperature relation using P3.

For a circular orbit, $mv^2/r = GMm/r^2$, so $K = \frac{1}{2}mv^2 = GMm/(2r)$ and $U = -GMm/r$. Thus $2K + U = 0$ and $E = K + U = -K$, as in Eq. (6.34). A lower total energy means a higher kinetic energy. The same balance applies, on average, to a self-gravitating gas and to collections of stars or galaxies. Zwicky used galaxy speeds in the Coma cluster to infer more mass than the visible galaxies supplied, an early clue to dark matter. We can reach the virial relation from hydrostatic balance, moving gas, or point-mass gravity.

Route A, from hydrostatic equilibrium. Multiply Eq. (10.2) by $4\pi r^3$ and integrate from the centre to the surface R. On the left,

$$\int_0^R 4\pi r^3 \frac{\mathrm{d}P}{\mathrm{d}r}\,dr \overset{(1)}{=} \Big[4\pi r^3 P\Big]_0^R - 3\int_0^R 4\pi r^2 P\,dr \overset{(2)}{=} -3\int P\,dV. \tag{10.5}$$

(1) Integration by parts.
(2) At the centre $r^3 = 0$, and at the surface $P = 0$. The volume element is $dV = 4\pi r^2 dr$.

On the right,

$$\int_0^R 4\pi r^3 \left(-\frac{Gm\rho}{r^2}\right) dr \overset{(1)}{=} -\int_0^M \frac{Gm}{r}\,dm \overset{(2)}{=} U. \tag{10.6}$$

(1) $4\pi r^2 \rho\,dr = dm$, Eq. (9.15).
(2) The gravitational energy, Eq. (9.19).

Setting the two equal,

$$\boxed{3\int P\,dV + U = 0.} \tag{10.7}$$

Only gravity and pressure enter, so this relation does not depend on the equation of state. Weighting hydrostatic balance by radius removes the local detail and keeps the star's overall balance.

Route B, for a gas that moves. Route A assumed the gas is at rest. To see what happens when it is not, follow the second moment of the mass, $I \equiv \int r^2\,dm$, where each dm moves with the

gas. Differentiating twice,

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} &\overset{(1)}{=} \int v^2\, dm + \int \vec{r}\cdot\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}\, dm \\ &\overset{(2)}{=} 2K_{\text{bulk}} + \int \vec{r}\cdot\big(-\nabla P + \rho\,\vec{g}\big)\, dV \\ &\overset{(3)}{=} 2K_{\text{bulk}} + 3\int P\, dV + U. \end{aligned} \tag{10.8}$$

(1) $\mathrm{d}(r^2)/\mathrm{d}t = 2\vec{r}\cdot\vec{v}$, then once more. The mass elements do not change (P1).

(2) $K_{\text{bulk}} = \frac{1}{2}\int v^2\, dm$ is the kinetic energy of bulk motion. The Euler equation (9.12) gives $\rho\ \mathrm{d}\vec{v}/\mathrm{d}t$.

(3) $-\int \vec{r}\cdot\nabla P\, dV = 3\int P\, dV$ by integration by parts, as in Route A. For a sphere, $\vec{r}\cdot\vec{g} = -Gm/r$ by Eq. (9.17), so the gravity term is U as in Route A. (For any shape the same result follows by writing U as a sum over pairs, Route C.)

For a gas of slow particles, $\gamma = \frac{5}{3}$ in Eq. (9.26), so $\int P\, dV = \frac{2}{3}K_{\text{th}}$, where K_{th} is the kinetic energy of the random thermal motion. With the total kinetic energy $K = K_{\text{bulk}} + K_{\text{th}}$,

$$\boxed{\frac{1}{2}\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} = 2K + U.} \tag{10.9}$$

For a cloud whose size stays steady on average, $\langle \ddot{I} \rangle = 0$ and $2K + U = 0$. This is the static result again. If $2K < |U|$, then $\ddot{I} < 0$ and contraction accelerates. If $2K > |U|$, the cloud tends to expand.

Route C, for point masses over a long time. The stars of a cluster and the galaxies of a group are separate bodies, not a fluid, and their system need not be round or steady. Take N point masses m_i at $\vec{r}_i$ with velocities $\vec{v}_i$. Let $\vec{f}_{ij}$ be the force on body i from body j. Forces come in equal and opposite pairs (P1), $\vec{f}_{ji} = -\vec{f}_{ij}$, and for gravity

$$\vec{f}_{ij} = -\frac{Gm_i m_j\,(\vec{r}_i - \vec{r}_j)}{|\vec{r}_i - \vec{r}_j|^3}.$$

The second moment of the mass is now the sum $I = \sum_i m_i r_i^2$. Differentiating it twice,

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}I}{\mathrm{d}t} &\overset{(1)}{=} \sum_i m_i\,\vec{r}_i\cdot\vec{v}_i, \\ \frac{1}{2}\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} &\overset{(2)}{=} \sum_i m_i v_i^2 + \sum_i \vec{r}_i\cdot m_i\frac{\mathrm{d}\vec{v}_i}{\mathrm{d}t} \\ &\overset{(3)}{=} 2K + \sum_i\sum_{j\neq i}\vec{r}_i\cdot\vec{f}_{ij} \\ &\overset{(4)}{=} 2K + \frac{1}{2}\sum_i\sum_{j\neq i}(\vec{r}_i - \vec{r}_j)\cdot\vec{f}_{ij} \\ &\overset{(5)}{=} 2K - \frac{1}{2}\sum_i\sum_{j\neq i}\frac{Gm_i m_j}{|\vec{r}_i - \vec{r}_j|} \overset{(6)}{=} 2K + U. \end{aligned} \tag{10.10}$$

(1) $\mathrm{d}(r_i^2)/\mathrm{d}t = 2\vec{r}_i\cdot\vec{v}_i$.

(2) Differentiate the product once more.
(3) $K = \frac{1}{2}\sum_i m_i v_i^2$ is the kinetic energy of all the motion. Newton's second law, $m_i \, \mathrm{d}\vec{v}_i/\mathrm{d}t = \sum_{j\neq i} \vec{f}_{ij}$.
(4) Renaming $i \leftrightarrow j$ does not change the double sum, and with $\vec{f}_{ji} = -\vec{f}_{ij}$ it turns $\vec{r}_i \cdot \vec{f}_{ij}$ into $-\vec{r}_j \cdot \vec{f}_{ij}$. The sum is the average of its two forms.
(5) The gravitational force above, with $(\vec{r}_i - \vec{r}_j)\cdot(\vec{r}_i - \vec{r}_j) = |\vec{r}_i - \vec{r}_j|^2$.
(6) The double sum counts each pair twice, and the $\frac{1}{2}$ counts it once. The energy of point masses is $U = -\sum_{\text{pairs}} G m_i m_j/|\vec{r}_i - \vec{r}_j|$, the sum form of Eq. (9.19).

This is Eq. (10.9) again, now for any shape. Instead of asking for balance at every moment, average it over a long time t,

$$\left\langle \frac{1}{2}\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} \right\rangle \overset{(1)}{=} \frac{1}{t}\int_0^t \frac{1}{2}\frac{\mathrm{d}^2 I}{\mathrm{d}t'^2}\, dt' \overset{(2)}{=} \frac{1}{t}\left[\sum_i m_i \vec{r}_i \cdot \vec{v}_i\right]_0^t \overset{(3)}{\longrightarrow} 0. \tag{10.11}$$

(1) The time average.
(2) The integral of a derivative, with the first line of the chain above.
(3) In a bound system the bodies stay in a finite region with finite speeds, so the bracket stays finite. Divided by t it goes to zero as $t \to \infty$.

$$\boxed{2\langle K\rangle + \langle U\rangle = 0.} \tag{10.12}$$

The point masses need not form a sphere or balance forces at every instant. If they remain bound, their long-time averages obey the same relation. This form weighs star clusters and galaxy clusters (Chapter 14).

Reading it, gravity sets the temperature. Let the gas be ideal with mean mass μm_{H} per particle, where μ is the **mean molecular weight**, so that the number density is $n = \rho/(\mu m_{\mathrm{H}})$. Then

$$\int P\, dV = \int n k_B T\, dV = \frac{k_B}{\mu m_{\mathrm{H}}}\int T\, dm \equiv \frac{M k_B \bar{T}}{\mu m_{\mathrm{H}}},$$

which defines $\bar{T}$, the temperature averaged over the mass. With Eq. (10.7),

$$\boxed{k_B \bar{T} = \frac{\mu m_{\mathrm{H}}\,|U|}{3M}.} \tag{10.13}$$

The depth of the gravitational well per unit mass sets the average temperature. The equation contains no term for how the star produces energy.

A uniform sphere. To put numbers in, take a sphere of uniform density. The mass inside r is $m = M(r/R)^3$, so $dm = 3Mr^2 dr/R^3$, and Eq. (9.19) gives

$$\begin{aligned} U &\overset{(1)}{=} -\int_0^R \frac{G\,M(r/R)^3}{r}\,\frac{3Mr^2}{R^3}\,dr \\ &\overset{(2)}{=} -\frac{3GM^2}{R^6}\int_0^R r^4\, dr \\ &= -\frac{3}{5}\,\frac{GM^2}{R}. \end{aligned} \tag{10.14}$$

(1) Eq. (9.19) with $m = M(r/R)^3$ and $dm = 3Mr^2 dr/R^3$.

(2) Collect the powers of r and R.

Equation (10.13) then gives $k_B\bar{T} = GM\mu m_{\mathrm{H}}/(5R)$. For the Sun, taking $\mu = 0.6$ gives $\bar{T} = 2.8 \times 10^6$ K. Its centre is hotter, 1.57×10^7 K, because the real Sun is denser toward the middle. The virial theorem predicts a million-degree interior without using nuclear physics. At such temperatures hydrogen fusion can proceed (Section 12.3).

How firmly a star is bound. Let the gas have $P = (\gamma - 1)u$, Eq. (9.26). Its internal energy is $\int u\,dV = \int P\,dV/(\gamma - 1)$, and the total energy is

$$\begin{aligned} E &\overset{(1)}{=} \frac{1}{\gamma - 1}\int P\,dV + U \\ &\overset{(2)}{=} -\frac{U}{3(\gamma - 1)} + U \\ &= \frac{3\gamma - 4}{3(\gamma - 1)}\,U. \end{aligned} \tag{10.15}$$

(1) Internal energy plus gravitational energy, for a star at rest.
(2) The virial theorem (10.7), $\int P\,dV = -U/3$.

$$\boxed{E = \frac{3\gamma - 4}{3(\gamma - 1)}\,U.} \tag{10.16}$$

For slow particles, $\gamma = \frac{5}{3}$ and $E = U/2 < 0$. The star is bound, and taking it apart costs half its gravitational energy. As $\gamma \to \frac{4}{3}$, $E \to 0$. The star is barely bound, and a small push can make it collapse or fly apart. For $\gamma < \frac{4}{3}$ no static balance exists at all. This single line explains three later results. Stars held up mainly by light are fragile (Section 12.5). White dwarfs whose electrons move near the speed of light have a maximum mass (Section 13.2). And a collapsing protostar collapses a second time while its molecules are being broken apart (Section 10.6).

10.3 Negative Heat Capacity

For a star supported by slow particles, the virial theorem gives $E = -K$ (Eq. (10.16) with $\gamma = \frac{5}{3}$). Losing total energy therefore raises its kinetic energy and temperature. This seems at odds with the familiar rule that an object cools when it loses heat. The difference lies in the star's self-gravity.

A cup of tea cools toward room temperature because the room acts as a heat bath. The usual proof that heat capacity is positive assumes such a bath can set an object's temperature. A star radiates into space instead. No surrounding reservoir holds its temperature fixed, and gravity ties its internal motion to its size. When radiation removes energy, the star can contract into a deeper gravitational well and heat up. The heat-bath proof does not describe that sequence.

For the Sun, $\mu = 0.6$ gives $N = M/(\mu m_{\mathrm{H}}) = 2.0 \times 10^{57}$ particles and $E = -K = -\frac{3}{2}Nk_B\bar{T}$. If fusion stopped but the Sun shone at its present luminosity for a million years, it would lose $\Delta E = -1.2 \times 10^{40}$ J. Its mean temperature would rise by 3×10^5 K as it contracted. A graph of temperature against total energy therefore slopes downward: lower energy means higher temperature. A similar effect appears in collapsing star clusters and in black holes losing mass to Hawking radiation.

The heat capacity of a star. Along a sequence of stars in virial balance, each with a slightly different energy, $E = -\frac{3}{2}Nk_B\bar{T}$. So

$$\boxed{C \equiv \frac{\mathrm{d}E}{\mathrm{d}\bar{T}} = -\frac{3}{2}Nk_B < 0.} \tag{10.17}$$

Losing energy by radiation, $dE < 0$, raises the temperature, $d\bar{T} = dE/C > 0$. Of the gravitational energy released by contraction, half goes into heating the gas and half is radiated away, since $\Delta K = -\Delta U/2$ and $\Delta E = \Delta U/2$.

From one equilibrium to the next. A ball of gas in balance has its energy fixed by its size. For a uniform sphere of slow particles, Eqs. (10.16), (10.14) and (10.13) give

$$E = \frac{U}{2} = -\frac{3GM^2}{10R}, \qquad k_B\bar{T} = \frac{\mu m_{\mathrm{H}}\, GM}{5R}.$$

Both depend on R alone. A smaller ball in balance has a lower energy and a higher temperature. So the equilibria of a given mass form a ladder. Each rung is a ball in balance, and each lower rung is smaller, hotter and more tightly bound. To step down from one rung to the next, from R_1 to a smaller R_2, the gas must change its energy by

$$\begin{aligned} \Delta E &\overset{(1)}{=} \tfrac{1}{2}\,\Delta U < 0, \\ \Delta K &\overset{(2)}{=} \Delta E - \Delta U \overset{(3)}{=} -\tfrac{1}{2}\,\Delta U = -\Delta E > 0. \end{aligned} \tag{10.18}$$

(1) $E = U/2$ on both rungs, and U is lower on the smaller one.
(2) $E = K + U$ on both rungs.
(3) The first line.

The cloud must radiate $-\Delta E = \frac{1}{2}|\Delta U|$ to reach the smaller equilibrium. The other half of the released gravitational energy remains as increased kinetic energy. Radiation and heating thus occur together during contraction. If gravity alone powers the light, $\mathrm{d}E/\mathrm{d}t = -L$ gives the heating rate

$$\frac{\mathrm{d}\bar{T}}{\mathrm{d}t} = -\frac{L}{C} = \frac{2L}{3Nk_B} > 0,$$

the 3×10^5 K per million years of the Sun above.

The physical way down the ladder is contraction, and a real cloud takes it as follows.

- A molecular cloud starts near a rung, close to balance, held up by its thermal pressure, by turbulence, whose random gas motions push like an extra pressure, and by magnetic fields threaded through the gas.
- When gravity wins, by the Jeans criterion of Section 10.4, the cloud leaves its rung and contracts, and its gravitational energy falls, $\Delta U < 0$.
- The released energy goes first into the infall itself, and as the gas is squeezed, into heat. The warm gas radiates in the lines of its molecules and atoms and from its dust. In a thin cloud this light escapes.
- If the cooling keeps up with the compression, the heat leaves as fast as it is made. The pressure never builds up, the temperature stays near 10 K, and the cloud does not stop on any rung. It falls freely (Section 10.5) and breaks into pieces (Section 10.4). During the fall it is not in balance at all.

- The fall ends when the centre becomes opaque to its own cooling light. The heat now stays in, the pressure rises, and the gas settles on a new rung, the first hydrostatic core of Section 10.6. It has a lower energy and a higher temperature than the cloud it came from, and the difference has left as light.
- From then on the steps are slow. The core, and later the protostar, is in balance at every moment and moves down the ladder on the Kelvin–Helmholtz time (Section 10.5), radiating and heating, until its centre is hot enough to burn hydrogen (Section 12.3).

Once a hydrostatic core forms, it can move slowly through a sequence of smaller equilibria. Each step requires radiation to carry energy away, while the core grows hotter.

Why the usual rule says $C \geq 0$. The usual rule belongs to the canonical ensemble, a system in contact with a heat bath at temperature T. With $\beta = 1/k_B T$ its partition function is $Z(\beta) = \sum_s e^{-\beta E_s}$, summed over its states s. The mean energy and its spread follow from Z,

$$\langle E \rangle \overset{(1)}{=} -\frac{\partial \ln Z}{\partial \beta},$$
$$\left\langle (\Delta E)^2 \right\rangle = \left\langle E^2 \right\rangle - \langle E \rangle^2 \overset{(2)}{=} \frac{\partial^2 \ln Z}{\partial \beta^2} \overset{(3)}{=} -\frac{\partial \langle E \rangle}{\partial \beta} \overset{(4)}{=} k_B T^2 \, \frac{\partial \langle E \rangle}{\partial T}. \tag{10.19}$$

(1) Differentiate $\ln Z$ once. Each state enters with weight $e^{-\beta E_s}/Z$.
(2) Differentiate once more. The second derivative of $\ln Z$ is the mean of E^2 minus the square of the mean of E.
(3) The first line.
(4) $\beta = 1/k_B T$, so $\partial/\partial\beta = -k_B T^2 \, \partial/\partial T$.

$$\boxed{C_{\text{can}} = \frac{\left\langle (\Delta E)^2 \right\rangle}{k_B T^2} \geq 0.} \tag{10.20}$$

A spread in energy cannot be negative, so in contact with a bath the heat capacity cannot be negative either. This is a statement about the canonical ensemble only.

The isolated system. An isolated star has a fixed energy and belongs to the microcanonical ensemble, where the temperature is defined from the entropy $S(E)$ by $1/T = \mathrm{d}S/\mathrm{d}E$. Comparing neighbouring equilibria with slightly different E,

$$\frac{\mathrm{d}}{\mathrm{d}E}\left(\frac{1}{T}\right) \overset{(1)}{=} \frac{\mathrm{d}^2 S}{\mathrm{d}E^2}$$
$$\Longrightarrow \quad -\frac{1}{T^2}\frac{\mathrm{d}T}{\mathrm{d}E} = \frac{\mathrm{d}^2 S}{\mathrm{d}E^2}$$
$$\Longrightarrow \quad C_{\text{micro}} = \frac{\mathrm{d}E}{\mathrm{d}T} \overset{(2)}{=} -\frac{1}{T^2 \, \mathrm{d}^2 S/\mathrm{d}E^2}. \tag{10.21}$$

(1) Differentiate the definition of T once more.
(2) Invert $\mathrm{d}T/\mathrm{d}E$.

Nothing here forbids $C < 0$. It only requires $\mathrm{d}^2 S/\mathrm{d}E^2 > 0$, an entropy that curves upward. For short-range forces, a large system can be split into parts whose energies add, and then the two ensembles agree and $S(E)$ curves downward. Gravity is long-range. The energy of a star is not the sum of the energies of its halves, because the halves attract each other. So the two ensembles need not agree, and a self-gravitating body can have $C < 0$ as an isolated system even though no system with $C < 0$ can sit in steady contact with a bath.

That last point is itself physical. Put a star in a bath slightly hotter than itself. Heat flows in, and with $C < 0$ the star cools, so even more heat flows in. Put it in a slightly cooler bath and it heats as it loses heat. Either way it runs away from the bath's temperature. A self-gravitating body cannot come into steady equilibrium with a heat bath.

What it does for a star. Suppose the nuclear burning in a core runs a little too fast. The extra energy expands the core, and by Eq. (10.17) the core cools, which slows the burning. If the burning runs too slow, the core contracts and heats until the burning catches up. Negative heat capacity makes a star its own thermostat (P4 on the energy, Section 9.4). When fuel runs out in the core, the same rule makes the core contract and heat until the next fuel ignites. Where the pressure stops depending on temperature, in degenerate matter, the thermostat breaks, and burning can run away (Section 12.3).

10.4 The Jeans Criterion

The virial theorem compares gravity with motion for a whole cloud. It shows that $2K < |U|$ favours contraction, but it does not tell us which regions of the cloud will collapse on their own. That requires comparing gravity with the pressure response at each size.

A small traffic jam can dissolve as drivers adjust their spacing. A long jam may grow faster than that adjustment can travel through the line of cars. Gas faces a similar race. A slightly denser patch draws in more gas by gravity, while its extra pressure sends a restoring signal outward at the speed of sound. Sound can smooth a small patch before gravity gathers much more gas. A large patch grows before sound crosses it. The dividing scale is the Jeans length, found by comparing sound-crossing and collapse times (P6) using conservation (P1) and gravity (P2).

Consider a dense core of molecular gas at 10 K, with $n = 10^{10}\,\mathrm{m}^{-3}$ ($10^4\,\mathrm{cm}^{-3}$) and $\mu = 2.33$. Its sound speed is $c_s = \sqrt{k_B T/(\mu m_\mathrm{H})} = 188\,\mathrm{m\,s}^{-1}$, while its free-fall time is $t_\mathrm{ff} = 3.4 \times 10^5$ years (Section 10.5). Sound can cross only $c_s t_\mathrm{ff} = 0.065\,\mathrm{pc}$ in that time. Pressure can respond across a smaller region before it falls, but not across a much larger one. Figure 10.1 shows the same distinction for waves: short disturbances oscillate as sound, while long ones grow under gravity.

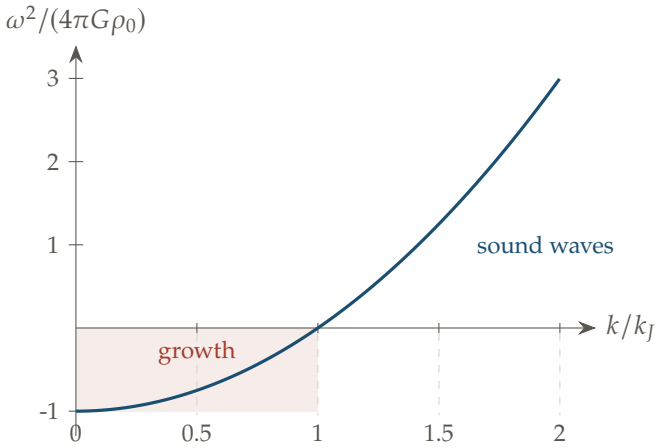


Figure 10.1: The squared frequency ω^2 of a small wave in a uniform self-gravitating gas against its wavenumber k, in units of $4\pi G\rho_0$ and of k_J. Below $k = k_J$ (shaded) $\omega^2 < 0$.

The dividing scale between pressure support and collapse will give us both a critical length and a critical mass. Related instabilities form clumps in rotating disks and, against cosmic expansion, help grow galaxies and clusters from the early universe's small density variations.

By energy. Take a uniform sphere of mass M, radius R and temperature T. Its thermal energy is $K = \frac{3}{2}Nk_BT = \frac{3}{2}Mk_BT/(\mu m_{\rm H})$ (Eq. (9.28)), and its gravitational energy is Eq. (10.14). The marginal case of Eq. (10.9), $2K = |U|$, gives

$$3\,\frac{Mk_BT}{\mu m_{\rm H}} = \frac{3}{5}\,\frac{GM^2}{R}$$
$$\Longrightarrow \qquad R = \frac{GM\mu m_{\rm H}}{5k_BT}. \tag{10.22}$$

Writing $M = \frac{4}{3}\pi R^3\rho$ and eliminating R or M gives the **Jeans mass** and **Jeans length**,

$$\boxed{M_J = \left(\frac{5k_BT}{G\mu m_{\rm H}}\right)^{3/2}\left(\frac{3}{4\pi\rho}\right)^{1/2}, \qquad R_J = \left(\frac{15\,k_BT}{4\pi G\mu m_{\rm H}\rho}\right)^{1/2}.} \tag{10.23}$$

A cloud with $M > M_J$ cannot be held up by its thermal pressure. For the dense core above, $M_J = 5.4\,M_\odot$ and $R_J = 0.13\,$pc. Hot, thin gas is stable. Cold, dense gas is not, which is why stars form in cold molecular clouds rather than in the warm diffuse gas.

By waves. Take an infinite gas at rest with uniform density ρ_0 and one temperature, so that a small change of density ρ_1 changes the pressure by $c_s^2\rho_1$. Disturb it slightly, $\rho = \rho_0 + \rho_1$, $\vec{v} = \vec{v}_1$, $\Phi = \Phi_0 + \Phi_1$. To first order in the small quantities, continuity (9.2), Euler (9.12) and Poisson (9.18) read

$$\frac{\partial\rho_1}{\partial t} + \rho_0\nabla\cdot\vec{v}_1 = 0, \qquad \rho_0\frac{\partial\vec{v}_1}{\partial t} = -c_s^2\nabla\rho_1 - \rho_0\nabla\Phi_1, \qquad \nabla^2\Phi_1 = 4\pi G\rho_1. \tag{10.24}$$

(Strictly, a uniform ρ_0 cannot be at rest, because Poisson's equation then needs a Φ_0 that grows without bound. Jeans set this aside, and the result is still right for regions much smaller than the cloud.) For a wave $\propto e^{i(kx-\omega t)}$ along x, with $\vec{v}_1 = v_1\hat{x}$, each derivative becomes a factor, $\partial_t \to -i\omega$ and $\partial_x \to ik$. The three equations become one matrix equation for the three amplitudes,

$$\begin{pmatrix} -i\omega & i\rho_0 k & 0 \\ ic_s^2 k & -i\omega\rho_0 & i\rho_0 k \\ 4\pi G & 0 & k^2 \end{pmatrix}\begin{pmatrix}\rho_1 \\ v_1 \\ \Phi_1\end{pmatrix} = 0. \tag{10.25}$$

A wave exists only if the determinant vanishes,

$$0 \overset{(1)}{=} -i\omega\big(-i\omega\rho_0 k^2\big) - i\rho_0 k\big(ic_s^2k^3 - i\rho_0 k\cdot 4\pi G\big)$$
$$\overset{(2)}{=} \rho_0 k^2\big(c_s^2k^2 - 4\pi G\rho_0 - \omega^2\big). \tag{10.26}$$

(1) Expand along the first row.
(2) Multiply out, using $i^2 = -1$, and collect.

$$\boxed{\omega^2 = c_s^2k^2 - 4\pi G\rho_0.} \tag{10.27}$$

Short waves, with large k, have $\omega^2 > 0$ and oscillate as sound. Long waves with k below $k_J = \sqrt{4\pi G\rho_0}/c_s$ have $\omega^2 < 0$, so ω is imaginary and the disturbance grows exponentially instead of oscillating. The critical wavelength is

$$\lambda_J = \frac{2\pi}{k_J} = c_s\sqrt{\frac{\pi}{G\rho_0}}. \tag{10.28}$$

It scales with T and ρ exactly as R_J does, and differs only in the number in front, because one route treats a finite sphere and the other an infinite medium. Both say $\lambda_J \approx c_s\,t_{\rm ff}$. With $t_{\rm ff}$ from Eq. (10.32) the ratio is $\sqrt{32/3} = 3.3$.

Why stars form in groups. A giant molecular cloud holds 10^5 to $10^6\,M_\odot$ (Table 8.4), many thousands of Jeans masses. It does not collapse into one star. While it collapses it stays thin enough to radiate away the heat of compression, so its temperature stays near 10 K. By Eq. (10.23), $M_J \propto T^{3/2}\rho^{-1/2}$ then falls as the density rises, so parts that were stable become unstable on their own. The cloud breaks into clumps, the clumps into cores, and each core into stars. This is why most stars are born in clusters and in pairs, although many clusters later drift apart.

10.5 Free Fall and the Three Clocks

Once a region passes the Jeans limit, gravity draws it inward. The next question is how long the fall takes if pressure cannot slow it.

Imagine a hole drilled straight through a uniform Earth. A ball dropped into it reaches the centre after 21 minutes, whatever the hole's length. The time depends only on the Earth's density, as the units of $G\rho$ demand (Section 9.6). A cloud whose pressure can no longer resist gravity falls in on itself in the same way. Each shell falls toward the centre pulled by the mass inside it. The time to reach the centre depends only on the density, so every shell of a uniform cloud arrives at the same moment. This is conservation of energy (P1) for each shell, with the gravity of a sphere (P2).

The free-fall time depends on density. At the Earth's mean density it is 15 minutes; at the Sun's, 30 minutes; and for a dense molecular-cloud core, 3.4×10^5 years. Figure 10.2 shows that a collapsing cloud shrinks slowly at first and then rapidly as gravity strengthens.

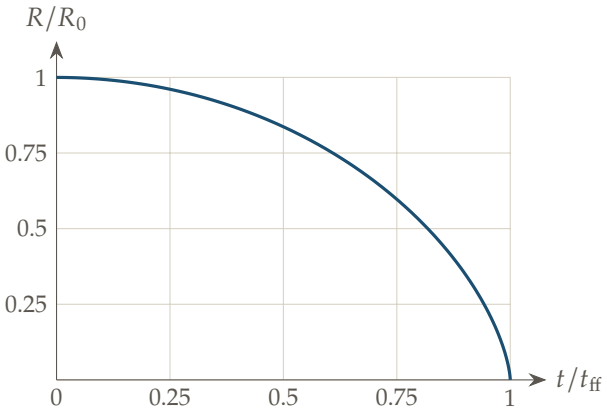


Figure 10.2: The radius of a shell in a uniform, pressure-free sphere that starts at rest, against time in units of the free-fall time.

We can calculate that time by treating a shell of the cloud as a radial Kepler orbit with no angular momentum (Chapter 6). The result also sets the initial collapse time of a massive stellar core and the scale of collapse in a forming galaxy halo.

The fall of one shell. Take a uniform cloud of density ρ_0 at rest, with no pressure. Follow a shell that starts at radius R_0, with mass $M = \frac{4}{3}\pi R_0^3 \rho_0$ inside it. While the shells do not cross, M stays the same, and by Eq. (9.17)

$$\frac{\mathrm{d}^2 R}{\mathrm{d}t^2} = -\frac{GM}{R^2}. \tag{10.29}$$

Multiplying by $\mathrm{d}R/\mathrm{d}t$ and integrating from rest at R_0 gives the energy of the shell per unit mass,

$$\frac{1}{2}\left(\frac{\mathrm{d}R}{\mathrm{d}t}\right)^2 - \frac{GM}{R} = -\frac{GM}{R_0}$$
$$\Longrightarrow \quad \left(\frac{\mathrm{d}R}{\mathrm{d}t}\right)^2 = \frac{2GM}{R_0}\left(\frac{R_0}{R} - 1\right). \tag{10.30}$$

The substitution $R = R_0 \cos^2\xi$, with ξ running from 0 at the start to $\pi/2$ at the centre, turns the bracket into $\tan^2\xi$,

$$\left(-2R_0 \cos\xi \sin\xi \,\frac{\mathrm{d}\xi}{\mathrm{d}t}\right)^2 \overset{(1)}{=} \frac{2GM}{R_0}\,\frac{\sin^2\xi}{\cos^2\xi}$$
$$\Longrightarrow \quad \cos^2\xi \,\frac{\mathrm{d}\xi}{\mathrm{d}t} \overset{(2)}{=} \sqrt{\frac{GM}{2R_0^3}}$$
$$\Longrightarrow \quad \xi + \sin\xi\cos\xi \overset{(3)}{=} \sqrt{\frac{2GM}{R_0^3}}\; t. \tag{10.31}$$

(1) $\mathrm{d}R/\mathrm{d}t = -2R_0\cos\xi\sin\xi\;\mathrm{d}\xi/\mathrm{d}t$, and $R_0/R - 1 = 1/\cos^2\xi - 1 = \tan^2\xi$.
(2) Divide both sides by $4R_0^2\sin^2\xi$ and take the positive root, since ξ grows.
(3) $\int_0^{\xi} 2\cos^2\xi'\,d\xi' = \xi + \sin\xi\cos\xi$, with $\xi = 0$ at $t = 0$.

The shell reaches the centre at $\xi = \pi/2$. With $M = \frac{4}{3}\pi R_0^3\rho_0$, $\sqrt{2GM/R_0^3} = \sqrt{8\pi G\rho_0/3}$, and

$$\boxed{t_{\mathrm{ff}} = \frac{\pi/2}{\sqrt{8\pi G\rho_0/3}} = \sqrt{\frac{3\pi}{32\,G\rho_0}}.} \tag{10.32}$$

R_0 has dropped out. Every shell of a uniform cloud reaches the centre at the same time, so shells never cross, which is what the calculation assumed. The time is $(G\rho_0)^{-1/2}$ times a number, as P6 said it must be.

Homologous collapse. More is true. The last line of the chain ties ξ to t through $\sqrt{8\pi G\rho_0/3}$ alone, the same for every shell. So $R/R_0 = \cos^2\xi$ is the same function of time for every shell. A shell that started at half the cloud's radius stays at half its radius at every moment. The cloud keeps its shape and only shrinks in scale. This is called **homologous collapse**, the same word as the homology of stars in Section 12.6, one shape rescaled. Two things follow. Every shell keeps its mass while every volume shrinks by $\cos^6\xi$, so the density stays uniform, $\rho = \rho_0/\cos^6\xi$. And the speed of a shell, $\mathrm{d}R/\mathrm{d}t = R_0\;\mathrm{d}(\cos^2\xi)/\mathrm{d}t$, is proportional to its radius at each moment, as in a uniform shrinking. A real core is denser at its centre. With $t_{\mathrm{ff}} \propto \rho^{-1/2}$ its inner gas arrives first and its outer gas later, so its collapse is not homologous. A small dense centre forms while most of the gas is still falling in. That centre becomes the first core and then the protostar (Section 10.6), and the rest of the gas reaches it later (Section 10.7).

The same result as a Kepler orbit. A shell falling from rest is a body on a Kepler orbit of zero angular momentum. As the angular momentum shrinks to zero at fixed energy, the ellipse of Chapter 6 flattens into a straight segment from R_0 to the centre, with semi-major axis $R_0/2$. The fall from one end to the other is half an orbital period. By Eq. (6.50),

$$t_{\mathrm{ff}} = \frac{1}{2}\cdot 2\pi\sqrt{\frac{(R_0/2)^3}{GM}} = \pi\sqrt{\frac{R_0^3}{8GM}} = \sqrt{\frac{3\pi}{32\,G\rho_0}},$$

the same as Eq. (10.32).

The three clocks of a star. A star changes on three very different times.

- The **free-fall** or dynamical time, Eq. (10.32). It is how fast the star restores hydrostatic equilibrium after a disturbance. For the Sun's mean density, 1410 kg m^{-3}, it is 30 minutes.
- The **Kelvin–Helmholtz** time, $t_{\mathrm{KH}} = |U|/L$. It is how long the star could shine on its gravitational energy alone, by contracting (Section 10.3). With $|U| \sim GM^2/R$, the Sun gives $t_{\mathrm{KH}} \approx 3 \times 10^7$ years.
- The **nuclear** time. Fusing hydrogen into helium releases 0.7% of the rest mass, and about a tenth of the Sun's hydrogen is burned on the main sequence, so $t_{\mathrm{nuc}} \approx 0.007 \times 0.1 \times Mc^2/L \approx 10^{10}$ years.

Figure 10.3 places them on one axis. Each is separated from the next by many orders of magnitude.

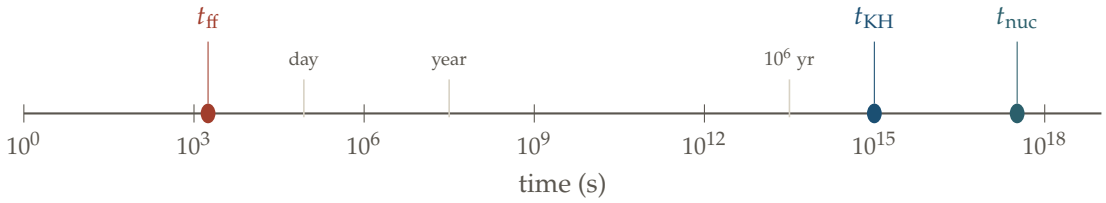


Figure 10.3: The free-fall, Kelvin–Helmholtz and nuclear times of the Sun on a logarithmic axis, with a day, a year and a million years for comparison.

This ordering is a P6 statement, and it is why a star can be described as a sequence of hydrostatic models. A star restores its balance within an hour, while its heat content changes over tens of millions of years and its composition over billions. At every moment it is in hydrostatic equilibrium, to great accuracy, and it moves slowly from one equilibrium to the next as it loses energy or burns fuel. A collapsing cloud is different. During free fall, t_{ff} is the only clock, the gas is not in balance at any moment, and no equilibrium description holds.

10.6 Rotation and the End of Collapse

The free-fall solution assumed that pressure and rotation never halt the gas. A real cloud violates both assumptions as it shrinks. Its weak initial spin grows, and eventually the centre becomes opaque enough to retain the heat of compression.

Think of a figure skater spinning with arms held out. When the skater pulls the arms in, the spin becomes much faster. Nothing pushed the skater. The angular momentum stayed the same while the size shrank. A collapsing cloud does the same (P1 for angular momentum, Section 9.1). Its spin grows as it shrinks, and the outward pull of rotation grows faster than gravity. Along the axis of rotation nothing resists, so the cloud keeps falling there. Across the axis the spin stops it. The cloud flattens into a disk.

As the core shrinks, rotation matters more. The ratio of centrifugal acceleration to gravity grows as $1/R$ (derived below), so shrinking from 0.1 pc to the Sun's size raises it by 4×10^6. A small initial spin can therefore stop gas from falling directly onto the centre. Instead, the gas flattens into a disk and must shed angular momentum before it can move farther inward.

The rotation ratio. Let a parcel at radius R circle the axis with speed v_{rot}, so its angular momentum per unit mass is Rv_{rot}, which is conserved. The ratio of the rotational pull to gravity

is

$$\frac{v_{\mathrm{rot}}^2/R}{GM/R^2} \overset{(1)}{=} \frac{v_{\mathrm{rot}}^2 R}{GM} \overset{(2)}{=} \frac{(Rv_{\mathrm{rot}})^2}{GM}\,\frac{1}{R}. \tag{10.33}$$

(1) Rotational acceleration v_{rot}^2/R divided by the gravity of the mass M inside.
(2) Group the conserved angular momentum per unit mass.

With Rv_{rot} and M fixed, the ratio grows as $1/R$. Shrinking R by some factor raises the ratio by the same factor.

Where the spin comes from. A cloud as a whole may hardly turn at all. Its gas is not still, though. Molecular clouds are turbulent. The widths of their lines show random motions of a few kilometres per second, far faster than the speed of sound in gas at 10 K, about $0.2\,\mathrm{km\,s^{-1}}$. Turbulence makes an irregular field of velocities, with eddies and shears on every scale. Take any small region and measure the motion of its gas about the region's own centre of mass. Some of the gas goes round one way more than the other, so the region has a nonzero angular momentum, even where the cloud as a whole has none. When the cloud breaks into cores (Section 10.4), each core keeps the angular momentum of the gas it gathered (P1). By the ratio above, its spin matters more and more as it shrinks, so each core flattens into its own disk, turning about its own axis. This is why young stars are so often found with disks around them, even in clouds that do not rotate as a whole.

The end of the isothermal collapse. While the gas can radiate away the heat of compression, its temperature stays nearly constant and the collapse goes on. Once the centre becomes dense enough to be opaque to its own cooling radiation, the heat stays in. The pressure now rises faster than the density, close to an adiabat with $\gamma = \frac{7}{5}$ for molecular hydrogen, above $\frac{4}{3}$. By Eq. (10.16) a static balance is again possible. The centre stops falling and forms the **first hydrostatic core**, a few hundredths of a solar mass, which grows by catching the gas still falling in.

The core heats as it grows. Near 2000 K the hydrogen molecules start to break apart. Breaking them takes energy that would otherwise raise the pressure, so during this phase the gas behaves as if $\gamma < \frac{4}{3}$. By Eq. (10.16) no static balance exists, and the core collapses a second time, until the molecules are gone and the gas, now atomic and soon ionised, can stop it again. That second core is the protostar.

Three kinds of equilibrium. The word equilibrium has been used in three senses, and they differ.

- **Hydrostatic equilibrium** is mechanical. The forces on each parcel balance, Eq. (10.2). A star is in it to great accuracy.
- **Global thermodynamic equilibrium** means no net flow of heat anywhere, and in Newtonian physics a single temperature throughout. A star is never in it. Its temperature falls from the centre outward, and that is why energy flows out. (In general relativity the condition is a little different, Section 13.3.)
- **Local thermodynamic equilibrium** means that each small parcel is close to equilibrium inside itself, with one temperature for its gas and its light, even though the temperature changes from parcel to parcel. This holds deep inside stars, and it is what lets the gas pressure and the light of a parcel be written with the formulas of P3. Chapters 11 and 12 rely on it.

A star can be in excellent hydrostatic equilibrium while far from global thermodynamic equilibrium, and for most of its interior local thermodynamic equilibrium sits between the two.

10.7 From Protostar to Star

The second collapse leaves a protostar inside a spinning, dusty core (Section 10.6). It will contract and heat toward hydrogen burning as its total energy falls (Section 10.3), but much of its mass has yet to arrive. The remaining gas carries angular momentum, so it settles into a disk rather than falling directly onto the star. That disk controls both the star's growth and the material left for planets.

Stir the water in a bath and pull the plug. The water does not drop straight into the hole. It circles round it, faster near the middle, and it drains only as fast as it loses its swirl, by rubbing against the bottom and against the slower water further out. Gas falling onto a protostar does the same. It settles into a disk in which each ring circles the star, and it reaches the star only by handing its angular momentum to gas further out or by throwing it away in a jet (P1). A star is built through a disk, and what is left of the disk becomes planets.

For example, gas starting 0.05 pc from the centre of a one-solar-mass core with sideways speed $0.05\,\mathrm{km\,s^{-1}}$ settles into a ring about 300 AU from the star (derived below). To reach a protostar of three solar radii, it must lose 99.3 per cent of its angular momentum. Figure 10.4 shows the resulting disk inside its dusty envelope, with jets escaping along the rotation axis.

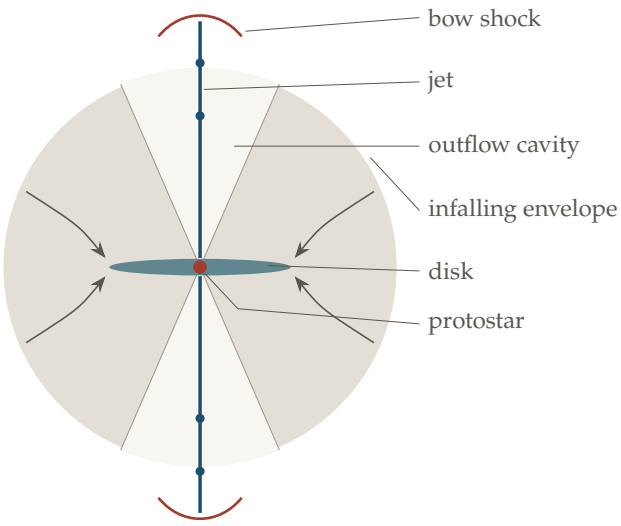


Figure 10.4: Schematic side view of an embedded protostar, not to scale. The envelope is grey, the outflow cavities are light, the disk is dark, and the jet runs along the axis to a bow shock at each end. The curved arrows mark gas falling onto the disk.

The need to shed angular momentum is common to accretion disks around white dwarfs, black holes and active galactic nuclei (Section 15.5). Here we follow it from the falling cloud to the protostar, then ask how the young star clears its surroundings and what remains for planets to form.

What starts a collapse. A cloud close to balance sits on a rung of the ladder. It collapses on its own once its mass exceeds the Jeans mass. A cloud just below the limit needs a push. At a fixed temperature $M_J \propto \rho^{-1/2}$ (Eq. (10.23)), so squeezing the gas to a higher density lowers the Jeans mass. A cloud of a third of its Jeans mass becomes unstable once it is squeezed nine times denser. The squeeze comes from outside.

- The blast wave of a supernova sweeps up the gas around it.
- A cloud that enters a spiral arm of the Galaxy (Chapter 14) is squeezed as gas piles up in the arm.
- Two clouds collide.
- A hot, massive star ionises the gas around it (Chapter 8). The ionised gas at 10^4 K is a thousand times hotter than the 10 K molecular gas next to it, so at a similar density it pushes thousands of times harder. The edge of the ionised gas, the ionisation front, pushes a shock ahead of it into the cold gas.

Each push is a shock, a sudden jump in density that moves through the gas.

How hard a shock squeezes. In the frame that moves with the shock, gas of density ρ_a flows in at speed v_a and flows out, squeezed, with density ρ_b and speed v_b. Cold molecular gas radiates the heat of the shock almost at once, so it leaves at its old temperature, and on both sides the pressure is $c_s^2\rho$. The mass and the momentum flowing in each second equal those flowing out (P1),

$$\rho_a v_a = \rho_b v_b, \qquad \rho_a v_a^2 + c_s^2\rho_a = \rho_b v_b^2 + c_s^2\rho_b.$$

Eliminating v_b,

$$\begin{aligned} c_s^2(\rho_b - \rho_a) &\overset{(1)}{=} \rho_a v_a^2 - \rho_b v_b^2 \\ &\overset{(2)}{=} \rho_a v_a\,(v_a - v_b) \\ &\overset{(3)}{=} \rho_a v_a^2\,\frac{\rho_b - \rho_a}{\rho_b}. \end{aligned} \tag{10.34}$$

(1) The momentum balance, rearranged.
(2) $\rho_b v_b = \rho_a v_a$, from the mass balance.
(3) $v_b = v_a\rho_a/\rho_b$, from the mass balance again.

A shock has $\rho_b \neq \rho_a$, so dividing by $\rho_b - \rho_a$ leaves $c_s^2\rho_b = \rho_a v_a^2$,

$$\boxed{\frac{\rho_b}{\rho_a} = \left(\frac{v_a}{c_s}\right)^2.} \tag{10.35}$$

The squeeze grows as the square of the shock's speed in units of the sound speed. In gas at 10 K, with $c_s = 0.19\,\mathrm{km\,s^{-1}}$, a shock of $0.56\,\mathrm{km\,s^{-1}}$ already squeezes the gas nine times, enough to tip the cloud above. A shock of $10\,\mathrm{km\,s^{-1}}$ would squeeze it 2800 times if nothing else resisted. Magnetic fields in the gas resist and lower this, but even weak shocks are strong triggers. Not every collapse needs one. A cloud that cools and condenses on its own crosses the Jeans limit without help.

How fast a protostar grows. A core of about one Jeans mass falls in about one free-fall time, so mass reaches the centre at a rate $\dot{M}$ of about M_J/t_{ff}. With $c_s^2 = k_B T/(\mu m_{\mathrm{H}})$, Eqs. (10.23) and (10.32) give

$$\begin{aligned} \dot{M} \approx \frac{M_J}{t_{\mathrm{ff}}} &\overset{(1)}{=} \left(\frac{5c_s^2}{G}\right)^{3/2}\left(\frac{3}{4\pi\rho}\right)^{1/2}\left(\frac{32\,G\rho}{3\pi}\right)^{1/2} \\ &\overset{(2)}{=} 5^{3/2}\,\frac{\sqrt{8}}{\pi}\,\frac{c_s^3}{G} = 10.1\,\frac{c_s^3}{G}. \end{aligned} \tag{10.36}$$

(1) M_J from Eq. (10.23) with $k_B T/(\mu m_{\mathrm{H}}) = c_s^2$, and $1/t_{\mathrm{ff}}$ from Eq. (10.32).

(2) The density cancels, $G^{-3/2}G^{1/2} = G^{-1}$, and $\frac{3}{4\pi} \cdot \frac{32}{3\pi} = \frac{8}{\pi^2}$.

$$\boxed{\dot{M} \approx 10\,\frac{c_s^3}{G}.} \tag{10.37}$$

The density has dropped out. How fast a star grows depends only on the temperature of the gas it is made from. Gas at 10 K gives $1.6 \times 10^{-5}\,M_\odot$ per year. The number in front is rough. The outer gas starts less dense and arrives later than the mean free-fall time says, and detailed models of a collapsing core give a rate near $c_s^3/G = 1.6 \times 10^{-6}\,M_\odot$ per year. At that rate a star of half a solar mass is assembled in 3×10^5 years, about one free-fall time of the core.

The light of a protostar. Gas that falls from far away onto the surface of the star releases GM/R per kilogram, so a protostar shines with the accretion luminosity $L = GM\dot{M}/R$ of Eq. (15.45). For $M = 0.5\,M_\odot$, $R = 3\,R_\odot$ and $\dot{M} = 1.6 \times 10^{-6}\,M_\odot$ per year, $L = 8\,L_\odot$. Not all of this light leaves from the star. Half of the energy of the falling gas is released in the disk, as orbital energy lost on the way in, and the other half where the gas meets the surface (Section 15.5). To this the star adds the light of its own slow contraction down the ladder.

Its inside is set by the virial theorem. For ionised gas with $\mu = 0.6$, Eq. (10.13) with the uniform sphere gives

$$\bar{T} = \frac{GM\mu m_{\mathrm{H}}}{5k_B R} = 2.8 \times 10^6\,\mathrm{K}\,\frac{M/M_\odot}{R/R_\odot}.$$

A young protostar of $0.1\,M_\odot$ and $5\,R_\odot$ is at 5.5×10^4 K inside, and one of $1\,M_\odot$ and $2\,R_\odot$ at 1.4×10^6 K. The interior is at tens of thousands to about a million kelvin, below the several million kelvin at which hydrogen burns fast enough to stop the contraction (Section 12.6). For a while the burning of its small store of deuterium slows the contraction (Chapter 12). The star itself has a surface at 2000 to 4000 K. By Wien's law (4.19), a surface at 3000 K shines most at 0.97 μm, in the near infrared, with some visible light.

Seen through dust. The protostar sits inside its envelope, behind tens of magnitudes of visual extinction. Dust blocks the near infrared about ten times less than the visible (Section 8.2.4). Behind $A_V = 50$ the visible light is dimmed by 10^{-20} and is lost. At 2.2 μm the extinction is about 5 magnitudes, and one per cent of the light gets through. The light the dust absorbs warms it to tens of kelvin, and the dust shines in turn. At 30 K its light peaks near 100 μm, by the same Wien's law. So a protostar is seen in the infrared, by its own near-infrared light through the thinnest parts of the envelope and by the far-infrared glow of the warmed dust. The near-infrared camera of the James Webb Space Telescope imaged the protostar in the dark cloud L1527 in Taurus this way. Its light escapes through two cavities, above and below a disk seen edge on.

Where the infalling gas lands. A parcel of gas keeps its angular momentum per unit mass Rv_{rot} as it falls. By Eq. (10.33) its rotational pull grows against gravity as $1/R$, and the fall across the axis stops where the two are equal. There the parcel circles the star. Setting the ratio to one at a radius r_{disk},

$$\frac{(Rv_{\mathrm{rot}})^2}{GM}\,\frac{1}{r_{\mathrm{disk}}} = 1 \quad \Longrightarrow \quad \boxed{r_{\mathrm{disk}} = \frac{(Rv_{\mathrm{rot}})^2}{GM}.} \tag{10.38}$$

The landing radius grows as the square of the angular momentum the parcel carries, and it depends on nothing else but the mass it circles. For $R = 0.05$ pc, $v_{\mathrm{rot}} = 0.05\,\mathrm{km\,s^{-1}}$ and $M = M_\odot$, $r_{\mathrm{disk}} = 300$ AU, three per cent of the starting radius. Gas from further in carries less angular momentum and lands closer, so the disk fills from the inside out. Disks around young stars are observed at tens to hundreds of AU.

A Keplerian disk. When the disk holds much less mass than the star, each ring circles the star as a planet would. By the third law (6.50), with the radius r of the ring in place of a, the ring turns at the angular rate $\sqrt{GM/r^3}$, moves at $\sqrt{GM/r}$ and carries the angular momentum $\sqrt{GMr}$ per unit mass. A disk whose rings obey this is called **Keplerian**, and the disks of protostars nearly are. The angular momentum per unit mass grows outward as $\sqrt{r}$. Gas at 300 AU that ends on the surface of a star of $3\,R_\odot$ keeps the fraction $\sqrt{3\,R_\odot/300\text{ AU}} = 0.007$ of its angular momentum. The other 99.3 per cent must go somewhere.

Why rings move in. Take two rings of the disk, of masses m_1 and m_2, at radii $r_1 < r_2$. Friction between them trades angular momentum, but the total stays the same (P1). The friction turns orbital energy into heat, which is radiated, so the orbital energy can only fall. Each ring has the orbital energy $-GMm_1/(2r_1)$ or $-GMm_2/(2r_2)$ of a circular orbit, Eq. (6.34) with $a = r$. For small changes dr_1 and dr_2,

$$
\begin{aligned}
0 = d\Big(m_1\sqrt{GMr_1} + m_2\sqrt{GMr_2}\Big) &\overset{(1)}{=} \sqrt{GM}\left(\frac{m_1\,dr_1}{2\sqrt{r_1}} + \frac{m_2\,dr_2}{2\sqrt{r_2}}\right) \quad\Longrightarrow\quad m_2\,dr_2 = -m_1\,dr_1\sqrt{\frac{r_2}{r_1}},\\
dE = d\left(-\frac{GMm_1}{2r_1} - \frac{GMm_2}{2r_2}\right) &\overset{(2)}{=} \frac{GM}{2}\left(\frac{m_1\,dr_1}{r_1^2} + \frac{m_2\,dr_2}{r_2^2}\right)\\
&\overset{(3)}{=} \frac{GMm_1\,dr_1}{2}\left(\frac{1}{r_1^2} - \frac{1}{\sqrt{r_1}\,r_2^{3/2}}\right)\\
&= \frac{GMm_1\,dr_1}{2\sqrt{r_1}}\left(\frac{1}{r_1^{3/2}} - \frac{1}{r_2^{3/2}}\right).
\end{aligned}
\tag{10.39}
$$

(1) The total angular momentum is fixed. Differentiate each $\sqrt{r}$.
(2) Differentiate each $1/r$.
(3) Replace $m_2\,dr_2$ by the first line.

The last bracket is positive, since $r_1 < r_2$. So $dE < 0$ needs $dr_1 < 0$, and then $dr_2 > 0$,

$$
\boxed{dE < 0 \quad\Longleftrightarrow\quad \text{the inner ring moves in and the outer ring moves out.}}
$$

Friction lowers the energy by moving mass in and angular momentum out. A disk spreads. Most of its mass drains onto the star, and a small part moves out carrying the angular momentum. The bookkeeping for a steady disk, with the torque that one ring exerts on the next, is Eq. (15.46).

What does the rubbing. The ordinary viscosity of the gas is far too weak. It would take longer than the age of the universe to drain a disk. Three things act much more strongly.

- **Turbulence.** Eddies carry gas, and its momentum, between rings far further than single molecules do. Magnetic fields wound up by the shear of the disk keep the gas stirred.
- **Magnetic torques.** Field lines that join one ring to another, or the disk to the star or the envelope, pull like stretched bands and pass angular momentum along.
- **Tidal torques.** Spiral arms in a massive disk, a companion star, or planets forming in the disk pull on the gas and take angular momentum from it.

Jets and outflows. Gas can also leave, and take angular momentum with it. Magnetic field lines anchored in the disk turn with it. Gas lifted off the disk slides out along them like a bead on a spinning wire, and it is made to turn with the disk until it is well away from it. Gas from a

ring at r that turns at the ring's rate $\sqrt{GM/r^3}$ out to a distance r_{lever} from the axis carries, per kilogram,

$$\sqrt{\frac{GM}{r^3}}\, r_{\text{lever}}^2 = \sqrt{GMr}\left(\frac{r_{\text{lever}}}{r}\right)^2,$$

which is $(r_{\text{lever}}/r)^2$ times the angular momentum it had in the ring. For $r_{\text{lever}} = 3r$ the factor is nine. So a jet that carries away a tenth of the infalling mass can carry away all the angular momentum that the rest must lose. This is why a disk that feeds a star also drives a jet. The gas leaves along the axis, where the disk and the envelope do not block it, in two opposite jets. Launched near the star, it leaves at about the escape speed from there, $\sqrt{2GM/R} = 250\,\text{km}\,\text{s}^{-1}$ for $0.5\,M_\odot$ and $3\,R_\odot$. Slower, wider outflows of swept-up gas surround the jets. The infall is uneven, so the outflow comes in bursts, and the knots along a jet record them. Some young stars brighten a hundredfold within a year as their disk dumps gas onto them, and stay bright for decades. FU Orionis, in 1937, was the first seen.

Herbig–Haro objects. Where a jet runs into the gas of the cloud, it drives a shock. The shock turns the energy of bulk motion into heat, $\frac{3}{2}k_BT \approx \frac{1}{2}\mu m_{\text{H}}v^2$, so a shock of $50\,\text{km}\,\text{s}^{-1}$ in gas with $\mu = 0.6$ heats it to about 6×10^4 K. The hot gas cools by shining in lines of hydrogen, oxygen and sulphur, and makes the glowing patches called **Herbig–Haro objects**. In HH 30, in the Taurus cloud 450 light years away, the disk is seen edge on as a dark lane, with a jet leaving it on both sides.

Clearing the birth cloud. The envelope goes in three ways. Accretion uses it up. Outflows sweep out the cavities along the axis and push gas away. Radiation pushes on the dust. Dust is far more opaque than free electrons. From the dust column of Section 8.2.4, one magnitude of A_V per 1.9×10^{25} hydrogen atoms per square metre, with hydrogen 70 per cent of the mass,

$$\kappa \approx \frac{1}{1.086 \times 1.9 \times 10^{25}\,\text{m}^{-2} \times m_{\text{H}}/0.70} \approx 20\,\text{m}^2\,\text{kg}^{-1}$$

in visible light, about 500 times the opacity of free electrons. By Eq. (11.27) the Eddington luminosity for dusty gas is then about $60\,L_\odot$ per solar mass. A young star of ten solar masses shines with thousands of $L_\odot$, mostly in visible and ultraviolet light, and its light blows the dusty envelope away. A young star of one solar mass shines with a few $L_\odot$, and its outflows do most of the clearing.

The disk survives this for three reasons. It is close to the star and tightly bound. A kilogram at 30 AU is bound by GM/r, about 340 times more firmly than a kilogram of the envelope at 0.05 pc. The outflows leave along the axis, away from the plane of the disk. And the disk is thin, so it catches little of the starlight, and it is opaque, so only its surface feels the push.

The names of young stars. Strictly, a **protostar** is a young star still hidden in its envelope. Once the envelope is gone, the star is seen in visible light, but it is still contracting slowly down the ladder on the Kelvin–Helmholtz time (Section 10.5). Such stars of low mass, up to about two solar masses, are **T Tauri stars**. Those of intermediate mass, about two to eight solar masses, are **Herbig Ae/Be stars**. Protostars and these stars together are **young stellar objects**. They vary in brightness. The accretion changes from week to week, spots on the surface turn with the star, clumps of dust in the disk pass in front of it and change its extinction, and some of them pulsate.

Planets form in what is left. In the disk the dust grains collide and stick. They grow into pebbles, rocks and bodies kilometres across, and these gather into planets. In a massive disk

a dense region can also collapse under its own gravity, the disk form of the Jeans criterion (Section 10.4). A growing planet pulls on the gas near its orbit by its tide and pushes it aside, opening a gap. The ALMA image of the young star HL Tau, taken in the millimetre light of its dust, shows bright rings separated by dark gaps of this kind. The gas of a disk lasts a few million years, much less than the tens of millions of years a Sun-like star takes to contract to the main sequence. Giant planets must gather their gas in that time, so they form while the star is still contracting. The rocky planets grow in the same disk more slowly. The Earth took a few tens of millions of years to form, about as long as the Sun took to reach the main sequence. As the star nears the main sequence, its light and wind blow away the gas and fine dust that remain, and only the planets and the larger solid bodies are left.

10.8 Problem Set: Self-Gravitating Gas

10.8.1 Problem 1: Scale Heights of Planetary Atmospheres

Problem Statement: Find the scale height of (a) the Earth's atmosphere ($T = 288\,\mathrm{K}$, mean molecular mass $28.97\,\mathrm{u}$, $g = 9.81\,\mathrm{m\,s^{-2}}$) and (b) the atmosphere of Mars ($T = 210\,\mathrm{K}$, mostly CO_2 with mean molecular mass $43.3\,\mathrm{u}$, $g = 3.71\,\mathrm{m\,s^{-2}}$). (c) Why is the scale height of Mars larger, although its air is heavier and colder?

Essential Ideas:

- Hydrostatic equilibrium with an ideal gas at one temperature gives $H = k_B T/(\mu m_{\mathrm{H}} g)$, Eq. (10.4).

Detailed Solutions:

(a) **Earth**. $H = (1.381 \times 10^{-23} \times 288)/(28.97 \times 1.661 \times 10^{-27} \times 9.81)\,\mathrm{m}$.

$$\boxed{H_\oplus \simeq 8.4\,\mathrm{km}.}$$

(b) **Mars**.

$$\boxed{H_{\mathrm{Mars}} \simeq 10.9\,\mathrm{km}.}$$

(c) **Comparison**. $H \propto T/(\mu g)$. The heavier, colder air of Mars lowers H by a factor $(210/288)(28.97/43.3) = 0.49$, but its weaker gravity raises it by $9.81/3.71 = 2.6$, which wins.

10.8.2 Problem 2: The Temperature of a Galaxy Cluster

Problem Statement: The galaxies of a rich cluster move with a line-of-sight velocity dispersion $\sigma = 1000\,\mathrm{km\,s^{-1}}$. The cluster also holds hot ionised gas ($\mu = 0.6$) in hydrostatic equilibrium in the same gravitational well. Estimate the temperature of that gas. In which band of light does it shine?

Essential Ideas:

- The galaxies and the gas feel the same well, so by the virial theorem their particles have the same kinetic energy per unit mass on average.
- For galaxies, the kinetic energy per unit mass is $\frac{3}{2}\sigma^2$ (three directions). For gas, it is $\frac{3}{2}k_B T/(\mu m_{\mathrm{H}})$.

Detailed Solutions: Setting the two equal, $k_B T = \mu m_{\mathrm{H}} \sigma^2$, so

$$\boxed{T = \frac{0.6 \times 1.673 \times 10^{-27} \times (10^6)^2}{1.381 \times 10^{-23}}\,\mathrm{K} \simeq 7 \times 10^7\,\mathrm{K}.}$$

Gas at this temperature has $k_B T \approx 6\,\mathrm{keV}$ and shines in X-rays, as clusters are observed to. Measuring this temperature and the density profile of the gas, and using Eq. (10.2), gives the mass of the cluster, which comes out far larger than the mass in stars and gas. This is the X-ray version of Zwicky's argument for dark matter.

10.8.3 Problem 3: Jeans Mass of a Dense Core

Problem Statement: For a molecular core with $\mu = 2.33$, find the Jeans mass (a) at 10 K and $n = 10^{10}\,\mathrm{m}^{-3}$, (b) at 20 K and the same density, and (c) at 10 K and $n = 10^{12}\,\mathrm{m}^{-3}$. What does (c) say about a core that collapses while staying at 10 K?

Essential Ideas:

- Eq. (10.23), with $M_J \propto T^{3/2}\rho^{-1/2}$.

Detailed Solutions:

(a) **Reference core**. $\rho = n\mu m_{\mathrm{H}} = 3.9 \times 10^{-17}\,\mathrm{kg\,m}^{-3}$ in Eq. (10.23).

$$\boxed{M_J \simeq 5.4\,M_\odot.}$$

(b) **Twice as warm**. M_J grows by $2^{3/2} = 2.8$.

$$\boxed{M_J \simeq 15\,M_\odot.}$$

(c) **A hundred times denser**. M_J falls by $\sqrt{100} = 10$.

$$\boxed{M_J \simeq 0.54\,M_\odot.}$$

A core that collapses while staying cold has a Jeans mass that keeps falling, so its parts become unstable on their own and it fragments.

10.8.4 Problem 4: The Age of the Sun Before Fusion Was Known

Problem Statement: In the nineteenth century, Kelvin and Helmholtz assumed the Sun shines by slowly contracting. Estimate how long it could have shone at its present luminosity $L_\odot = 3.83 \times 10^{26}\,\mathrm{W}$. Compare with the age of the Earth from geology and radioactivity, about 4.5×10^9 years.

Essential Ideas:

- $t_{\mathrm{KH}} = |U|/L$ with $|U| \sim GM^2/R$.
- By the virial theorem only half of the released energy is radiated, which changes only the factor in front.

Detailed Solutions:

$$\boxed{t_{\mathrm{KH}} \sim \frac{GM_\odot^2}{R_\odot L_\odot} = \frac{6.674 \times 10^{-11} \times (1.989 \times 10^{30})^2}{6.957 \times 10^8 \times 3.83 \times 10^{26}}\,\mathrm{s} \simeq 3 \times 10^7\ \text{years}.}$$

This is about a hundred times shorter than the age of the Earth. Gravity alone cannot have powered the Sun for that long, which pointed to a much larger store of energy, later found in the nucleus.

10.8.5 Problem 5: A Star That Heats as It Cools

Problem Statement: A young star of one solar mass, made of ionised gas with $\mu = 0.6$, contracts toward the main sequence. It shines with $L = L_\odot$ for 10^6 years without any fusion. By how much does its mean temperature change? Does it rise or fall?

Essential Ideas:

- The heat capacity of a star in virial balance is $C = -\frac{3}{2}Nk_B$, Eq. (10.17).
- The energy lost is $L\times$ time.

Detailed Solutions: $N = M/(\mu m_{\rm H}) = 2.0\times10^{57}$, so $C = -4.1\times10^{34}\,{\rm J\,K^{-1}}$. The energy lost is $\Delta E = -3.83\times10^{26}\times3.16\times10^{13}\,{\rm J} = -1.2\times10^{40}\,{\rm J}$.

$$\boxed{\Delta\bar{T} = \frac{\Delta E}{C} \simeq +3\times10^5\,{\rm K}.}$$

The temperature rises. The star loses energy and gets hotter, and it keeps contracting and heating until its centre is hot enough for hydrogen to fuse.

10.8.6 Problem 6: How a Protostar Grows

Problem Statement: A core of molecular gas with $\mu = 2.33$ collapses onto a protostar. (a) Estimate the accretion rate c_s^3/G for gas at 10 K and at 20 K. (b) A protostar of $0.3\,M_\odot$ and radius $2.5\,R_\odot$ grows at the rate for 20 K. Find its accretion luminosity. (c) Gas starts 0.02 pc from the axis and moves sideways at $0.05\,{\rm km\,s^{-1}}$. Where does it settle around the $0.3\,M_\odot$ star?

Essential Ideas:

- $\dot{M} \propto c_s^3/G$, Eq. (10.37), with $c_s = \sqrt{k_BT/(\mu m_{\rm H})} \propto T^{1/2}$.
- $L = GM\dot{M}/R$, Eq. (15.45).
- $r_{\rm disk} = (Rv_{\rm rot})^2/(GM)$, Eq. (10.38).

Detailed Solutions:

(a) **Accretion rate**. At 10 K, $c_s = 188\,{\rm m\,s^{-1}}$ and $c_s^3/G = 1.0\times10^{17}\,{\rm kg\,s^{-1}}$. Twice the temperature gives $2^{3/2} = 2.8$ times the rate.

$$\boxed{\dot{M} \simeq 1.6\times10^{-6}\text{ and }4.5\times10^{-6}\,M_\odot\,{\rm yr^{-1}}.}$$

(b) **Luminosity**. $\dot{M} = 2.8\times10^{17}\,{\rm kg\,s^{-1}}$, so $L = 6.674\times10^{-11}\times0.3\times1.989\times10^{30}\times2.8\times10^{17}/(2.5\times6.957\times10^8)\,{\rm W}$.

$$\boxed{L \simeq 6.5\times10^{27}\,{\rm W} \simeq 17\,L_\odot.}$$

(c) **Landing radius**. $Rv_{\rm rot} = 6.2\times10^{14}\,{\rm m}\times50\,{\rm m\,s^{-1}} = 3.1\times10^{16}\,{\rm m^2\,s^{-1}}$ and $GM = 4.0\times10^{19}\,{\rm m^3\,s^{-2}}$.

$$\boxed{r_{\rm disk} \simeq 2.4\times10^{13}\,{\rm m} \simeq 160\,{\rm AU}.}$$

The gas lands in a disk the size of the outer Solar System and must lose almost all of its angular momentum before it reaches the star.

Chapter 11

Radiation in Matter

Most of what we know about stars comes from their light. Before reaching a telescope, that light passes through the star's own gas and sometimes through interstellar matter (Chapter 8). Along the way, matter can absorb some of it and emit more.

Tracking those gains and losses along one ray gives the transfer equation (P1 and P4). In thin gas it explains bright and dark lines. In thick gas repeated absorption and emission make the light nearly thermal, while a small imbalance lets energy leak outward by a random walk (P5).

11.1 Specific Intensity

Flux measures the net light crossing an area and falls with distance as in Eq. (4.2) (Chapter 4). Inside a star, rays travel in many directions. Their opposing flows can nearly cancel in the flux even while the radiation field remains strong. To track what matter does to each ray, we need a directional measure of light.

Look at an evenly lit wall from one metre away, then from five. It covers less of your view at the greater distance, but each small patch appears just as bright. The Sun behaves the same way: from Mars it covers less sky than from Earth, while each patch of its surface keeps the same brightness. Distance reduces the total flux because the source subtends a smaller solid angle. The light carried along one ray is measured by **specific intensity**, energy per area, time, frequency, and solid angle. In empty space, energy conservation keeps that intensity constant (P1).

The Sun's angular radius is 0.266° from Earth and 0.175° from Mars, which is 1.52 times farther away. Its solid angle and total flux fall by $1.52^2 = 2.3$, while the intensity of a surface patch stays the same. Think of a narrow cone of rays leaving that patch: the cone widens, but its energy spreads over area and solid angle together. This conservation also explains why lensing preserves surface brightness. In a static space, a galaxy's surface brightness would likewise be independent of distance; departures from that rule reveal cosmic expansion.

Definition. Light of frequency between ν and $\nu + d\nu$ crosses an area dA in time dt, travelling within a solid angle $d\Omega$ about a direction that makes an angle θ with the normal of dA. The energy it carries defines the **specific intensity** I_ν,

$$dE = I_\nu \cos\theta \, dA \, dt \, d\nu \, d\Omega. \tag{11.1}$$

The factor $\cos\theta$ is the area dA as seen from the direction of travel.

Constant along a ray. Take two small areas dA_1 and dA_2, a distance d apart in empty space, each facing the other. The light leaving dA_1 toward dA_2 is the light arriving at dA_2 from

dA_1. Seen from dA_1, the area dA_2 subtends $d\Omega_1 = dA_2/d^2$, and seen from dA_2, dA_1 subtends $d\Omega_2 = dA_1/d^2$. Equating the two energies (P1),

$$\begin{aligned} & I_{\nu,1}\, dA_1\, d\Omega_1 = I_{\nu,2}\, dA_2\, d\Omega_2 \\ \Longrightarrow \quad & I_{\nu,1}\, \frac{dA_1\, dA_2}{d^2} \overset{(1)}{=} I_{\nu,2}\, \frac{dA_2\, dA_1}{d^2} \\ \Longrightarrow \quad & I_{\nu,1} = I_{\nu,2}. \end{aligned} \tag{11.2}$$

(1) The two solid angles written with the areas they subtend.

$$\boxed{I_\nu \text{ is constant along a ray in empty space.}} \tag{11.3}$$

The inverse-square fall of the flux is entirely the shrinking of the solid angle that a source covers.

Flux, energy density and pressure. Three quantities used throughout are sums of I_ν over direction. Take one beam, the light inside a small solid angle $d\Omega$ at angle θ to the normal of a unit area. Seen from along the beam the area shrinks to $\cos\theta$, so the beam carries energy across it at the rate $dF_\nu = I_\nu \cos\theta\, d\Omega$. The same light moves at speed c, so it fills space with energy density $I_\nu\, d\Omega/c$. Each photon carries momentum equal to its energy over c, pointing along the beam, and only the part $\cos\theta$ of it points across the area. The momentum the beam carries across the area per second, which is the pressure it exerts, is therefore

$$dP_\nu \overset{(1)}{=} \frac{\cos\theta}{c}\, dF_\nu \overset{(2)}{=} \frac{1}{c}\, I_\nu \cos^2\theta\, d\Omega,$$

(1) Momentum is energy over c, and only the part $\cos\theta$ pushes across the area.
(2) $dF_\nu = I_\nu \cos\theta\, d\Omega$.

so one $\cos\theta$ counts how much light crosses the area and the other how much of its push points across it. Adding all the beams gives

$$F_\nu = \int I_\nu \cos\theta\, d\Omega, \qquad u_\nu = \frac{1}{c}\int I_\nu\, d\Omega, \qquad P_\nu = \frac{1}{c}\int I_\nu \cos^2\theta\, d\Omega. \tag{11.4}$$

The flux F_ν is the net energy crossing a unit area, u_ν the energy density and P_ν the pressure. For light that is the same in every direction,

$$F_\nu = 0, \qquad u_\nu = \frac{4\pi}{c} I_\nu, \qquad P_\nu = \frac{u_\nu}{3},$$

the same $\frac{1}{3}$ as in the pressure integral (9.25), from the average of $\cos^2\theta$. With $I_\nu = B_\nu$ this is the relation $B_\nu = c\, u_\nu/4\pi$ of Eq. (9.31).

A surface that sends intensity B_ν outward into every direction of the outer half-space gives the flux $F_\nu = \pi B_\nu$ (Eq. (4.20)), and integrated over frequency the surface flux $\sigma_{\mathrm{SB}} T^4$ of a blackbody, Eq. (4.24). In the language of intensity, the factor π is the first moment of Eq. (11.4) taken over the outward half of the sphere.

11.2 The Transfer Equation

Extinction removes light from a ray by the factor $e^{-\tau}$ (Eq. (4.36)). But matter can add light too. A hot cloud emits its own glow, and a stellar atmosphere both absorbs light from below and emits into the same ray. We therefore need an equation that tracks both changes together.

On a foggy street, a distant lamp fades behind the fog while the fog scatters light into your view. Through a thin patch, the lamp still dominates. Through enough fog, you mainly see the

fog's own glow. Along a ray, matter removes a fraction of the passing light and adds its own emission. The transfer equation balances these changes (P1 and P4). Once enough matter lies in the way, the emerging light reflects the nearby matter rather than the distant lamp.

The slab's optical depth decides how much of its own glow replaces the entering light. For incident intensity I_0 and source function S, a slab with $\tau = 1$ transmits $e^{-1} = 0.37$ of I_0 and adds $0.63\,S$. At $\tau = 5$, it transmits only 0.7% and adds 99.3% of S. Figure 11.1 shows this approach to the slab's own glow as it becomes optically thick.

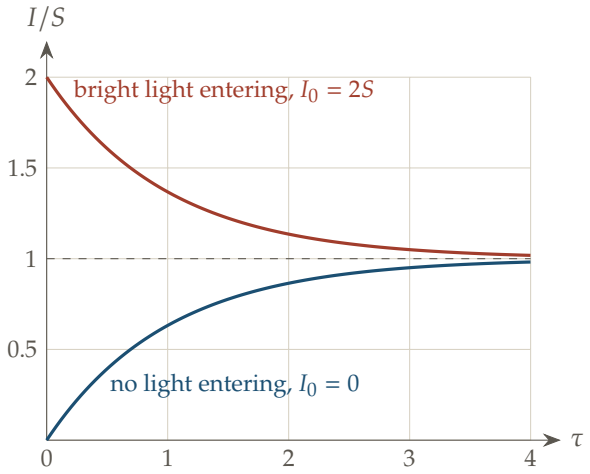


Figure 11.1: The intensity leaving a uniform slab against its optical depth τ, in units of the source function S, for entering intensity $I_0 = 2S$ (upper curve) and $I_0 = 0$ (lower curve).

We now derive the transfer equation that produces this result. It applies to cold gas seen in radio light, hot clusters seen in X-rays and the Lyman-α absorption of distant quasars (Section 8.3).

Absorption by a thin slab. Let a beam cross a thin slab of matter, of thickness ds along the beam and area A across it. The slab holds absorbers, atoms, ions, electrons or dust grains, with number density n. Each one blocks light of frequency ν over an area σ_ν, its **cross-section**, which atomic physics supplies (the Thomson cross-section of Eq. (8.7) is one). The slab is thin enough that the absorbers do not hide behind one another, so their areas simply add. Then

$$\begin{aligned} \text{absorbers in the slab} &\overset{(1)}{=} nA\,ds, \\ \text{area they block} &\overset{(2)}{=} \sigma_\nu\, nA\,ds, \\ \text{fraction of the beam absorbed} &\overset{(3)}{=} \frac{\sigma_\nu\, nA\,ds}{A} = n\sigma_\nu\,ds, \\ dI_\nu\big|_{\text{abs}} &\overset{(4)}{=} -n\sigma_\nu\, I_\nu\,ds. \end{aligned} \tag{11.5}$$

(1) Number density times the volume $A\,ds$ of the slab.
(2) Each absorber blocks σ_ν.
(3) A photon crossing the area A at a random point is caught if it lands on a blocked part.
(4) Each part of the beam loses the same fraction, so the intensity drops by that fraction of itself.

The fraction absorbed is proportional to ds because the number of absorbers met is. Its inverse, $1/(n\sigma_\nu)$, is the mean free path, the s of the random walk (Section 9.5). Light covers $ds = c\,dt$, so the same result says that a photon is absorbed at the rate $n\sigma_\nu c$ per unit time, the collision rate of a particle moving at c through targets of area σ_ν. Astronomers count absorbing area per unit

mass of matter rather than per unit volume, because the mass of a star or a cloud is what is known. Dividing by the density ρ defines the **opacity**,

$$\boxed{\kappa_\nu \equiv \frac{n\sigma_\nu}{\rho}, \qquad dI_\nu\big|_{\text{abs}} = -\kappa_\nu \rho\, I_\nu\, ds.} \tag{11.6}$$

ρ/n is the mass of matter that comes with each absorber, so κ_ν is the cross-section divided by that mass. Ionised hydrogen has one electron for each mass m_{H}, so its electrons give $\kappa = \sigma_T/m_{\text{H}}$ (Section 11.5). The opacity collects all the microscopic physics, how many absorbers there are and how large each one looks, into one number per unit mass.

Emission by a thin slab. Let j_ν be the energy the matter emits per unit volume, time, frequency and solid angle, so that the slab sends

$$dE_{\text{em}} = j_\nu\,(A\,ds)\,dt\,d\nu\,d\Omega$$

into the direction of the beam. By the definition of intensity, Eq. (11.1) with $\cos\theta = 1$ for an area facing the beam, the energy $dE = dI_\nu\,A\,dt\,d\nu\,d\Omega$ is carried by an extra intensity dI_ν. Equating the two,

$$dI_\nu\big|_{\text{em}} = j_\nu\,ds.$$

The equation. Along a path element ds, matter of density ρ removes the fraction $\kappa_\nu\rho\,ds$ of the intensity, where κ_ν is the opacity of Eq. (11.6), the absorbing area per unit mass. It also adds its own emission. Let j_ν be the energy it emits per unit volume, time, frequency and solid angle. Adding the two changes (P1, P4),

$$\boxed{\frac{\mathrm{d}I_\nu}{\mathrm{d}s} = -\kappa_\nu\rho\, I_\nu + j_\nu.} \tag{11.7}$$

Scattering removes light from a ray just as absorption does, so it counts in κ_ν, and the light it sends into the ray counts in j_ν.

Optical depth and source function. Two combinations make the equation simpler. The **optical depth** along the ray, $d\tau_\nu = \kappa_\nu\rho\,ds$, is the path length counted in mean free paths $1/(\kappa_\nu\rho)$, the same τ as in Chapter 4. The **source function** $S_\nu = j_\nu/(\kappa_\nu\rho)$ is the ratio of emission to absorption. Dividing Eq. (11.7) by $\kappa_\nu\rho$,

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}\tau_\nu} = -I_\nu + S_\nu. \tag{11.8}$$

The intensity always moves toward the source function. If $I_\nu > S_\nu$ it falls, and if $I_\nu < S_\nu$ it rises.

Formal solution. Multiply Eq. (11.8) by e^{τ_ν} and integrate from 0, where the ray enters with $I_\nu(0)$, to τ_ν,

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\tau_\nu}\left(I_\nu\, e^{\tau_\nu}\right) \overset{(1)}{=} S_\nu\, e^{\tau_\nu} \\ \Longrightarrow\quad & I_\nu(\tau_\nu) \overset{(2)}{=} I_\nu(0)\, e^{-\tau_\nu} + \int_0^{\tau_\nu} S_\nu(\tau')\, e^{-(\tau_\nu-\tau')}\, d\tau'. \end{aligned} \tag{11.9}$$

(1) The product rule. The terms in I_ν combine into a total derivative.
(2) Integrate, then divide by e^{τ_ν}.

The entering light is dimmed by $e^{-\tau_\nu}$. Each layer adds its emission, dimmed by the matter between it and the end of the ray. For a uniform slab, S_ν is constant and

$$\boxed{I_\nu = I_\nu(0)\, e^{-\tau_\nu} + S_\nu\big(1 - e^{-\tau_\nu}\big).} \tag{11.10}$$

Two limits follow, chosen by the ratio of the size of the slab to the mean free path (P6).

- **Optically thin**, $\tau_\nu \ll 1$. Then $I_\nu \approx I_\nu(0) + S_\nu\tau_\nu$. The slab adds emission in proportion to the amount of matter. Where κ_ν is large, at the frequency of a line, the emission is large, so a thin glowing gas shows emission lines.
- **Optically thick**, $\tau_\nu \gg 1$. Then $I_\nu \to S_\nu$, whatever entered. We see only the last mean free path or so of the matter.

With no emission, $S_\nu = 0$, this is the extinction law of Eq. (4.34).

11.3 Kirchhoff's Law and Local Thermodynamic Equilibrium

The Planck spectrum describes light in equilibrium with matter at one temperature (Eq. (9.31)). A star has no single temperature: its outer layers are cooler, and light escapes through them. Why, then, is its spectrum close to a blackbody, and why do dark lines interrupt it?

Look into a furnace. A poker left in the coals glows with exactly the colour of the coals around it, and its outline nearly disappears. Everything in the furnace glows alike, whatever it is made of. A material that absorbs well also glows well, in just the right proportion, because otherwise the poker would end up hotter or colder than the coals. The same rule holds for each small parcel of a star. Matter that absorbs strongly at some frequency also emits strongly there, and the ratio of the two is fixed by its temperature. This is detailed balance (P4) for absorption and emission, with the counted spectrum of light (P3).

An atmosphere lets us see to different depths at different frequencies. At most frequencies, the continuum comes from a deeper, hotter layer. At a strong spectral line, greater opacity hides that layer and exposes cooler gas above it. For example, if the continuum at 500 nm comes from gas at 6000 K and the line centre from gas at 4500 K, the line intensity is $B_\nu(4500\,\mathrm{K})/B_\nu(6000\,\mathrm{K}) = 0.20$ of the continuum. The line appears dark because we see cooler gas there. This connection between absorption, emission and temperature also explains the Sun's Fraunhofer lines and the darkening of its edge.

Kirchhoff's law. Inside a closed cavity in full thermodynamic equilibrium at temperature T, the light is the same everywhere and in every direction, and by P3 it is Planck light, $I_\nu = B_\nu(T)$. Nothing changes along a ray, so Eq. (11.7) gives

$$\begin{aligned} 0 &\overset{(1)}{=} -\kappa_\nu\rho\, B_\nu(T) + j_\nu \\ \Longrightarrow \qquad S_\nu &\overset{(2)}{=} \frac{j_\nu}{\kappa_\nu\rho} = B_\nu(T). \end{aligned} \tag{11.11}$$

(1) $\mathrm{d}I_\nu/\mathrm{d}s = 0$ with $I_\nu = B_\nu$. Every process is balanced by its reverse (P4).
(2) The definition of S_ν.

$$\boxed{j_\nu = \kappa_\nu\rho\, B_\nu(T).} \tag{11.12}$$

Emission and absorption are tied together by the Planck function of the matter's temperature.

Local thermodynamic equilibrium. The emission and absorption of a parcel depend on the parcel itself, on its atoms and how they are excited, not on the light passing through. Where collisions keep each parcel close to equilibrium inside itself, the local thermodynamic equilibrium of Section 10.6, its atoms are excited as in equilibrium at the local temperature. Eq. (11.12) then holds point by point, even though the light passing through is not Planck light,

$$S_\nu = B_\nu\big(T(\vec{r})\big) \qquad \text{(local thermodynamic equilibrium).} \tag{11.13}$$

A thick layer at one temperature then emits Planck light, as a blackbody does, without being a closed box.

The transfer equation in local equilibrium. Put Eq. (11.12), now at the local temperature, into the transfer equation (11.7),

$$\boxed{\frac{\mathrm{d}I_\nu}{\mathrm{d}s} = -\kappa_\nu\rho\left(I_\nu - B_\nu\big(T(\vec{r})\big)\right).} \tag{11.14}$$

Read it as a relaxation. Over each mean free path $1/(\kappa_\nu\rho)$ the light is pulled toward the Planck light of the gas it is passing through. If the temperature were the same everywhere, the light would settle at B_ν and stay there, the same in every direction, carrying no net flux. That is global thermodynamic equilibrium (Section 10.6). In a star the temperature changes from place to place, so the target $B_\nu(T)$ moves along every ray, and the light never quite catches up. At each point it still carries a trace of the gas it has just come from, which was hotter or cooler than the gas here. Deep inside, that trace is tiny, but it is the whole flux of the star (Section 11.4).

Absorption lines. In an atmosphere the temperature falls outward. By Eq. (11.10) we see light from about one mean free path deep, $\tau_\nu \approx 1$ along our line of sight (more exactly $\tau_\nu \approx \frac{2}{3}$, Section 11.8). At the frequency of a line, κ_ν is large, so $\tau_\nu = 1$ is reached higher up, where the gas is cooler and B_ν is smaller. The line is darker than the continuum on either side. If the temperature rose outward, as it does in the solar chromosphere, the same argument would give emission lines. The edge of the Sun looks darker than its centre for the same reason. Looking at the edge, the line of sight runs slantwise and reaches $\tau = 1$ in higher, cooler layers.

11.4 Diffusion of Light

Deep inside a star, light is absorbed and re-emitted many times before it reaches the surface. It is therefore nearly Planck light at the local temperature, as the thick limit of Eq. (11.10) predicts. Exactly Planck light would be equally bright in every direction and carry no net flux (Section 11.1). The star shines because the outward light is just a little brighter than the inward light.

Think again of perfume in still air (Problem 9.7.4). Each molecule wanders at random, but a little more perfume drifts away from the bottle than toward it, because there is more perfume near the bottle. A net flow appears with no push in any direction, only a gradient. Photons in a star do the same. Each is absorbed and re-emitted after a short free path, in a random direction. There is slightly more light where the gas is hotter, so a little more light drifts outward than inward. This is a random walk (P5), with the energy density of light aT^4 in place of the number of perfume molecules.

The difference between a straight flight and a random walk is enormous. Near the Sun's centre, a photon's free path is of order a millimetre to a centimetre. To cross a radius of 7×10^8 m by random walking takes $R^2/(sc)$ (Eq. (9.42)): 5×10^3 years for $s = 1$ cm or 5×10^4 years for $s = 1$ mm. A straight flight would take 2.3 seconds. Figure 9.3 shows why so many short steps are needed. We can derive the small outward flux produced by this walk in three ways.

Route A, the random walk. The diffusion flux of Eq. (9.44) applies to anything that walks at random. For light, the walkers are photons and the quantity that flows is energy. With energy density aT^4 (Eq. (9.30)) in place of n, speed c, and step length $1/(\kappa\rho)$,

$$F = -\frac{1}{3}\, c\, \frac{1}{\kappa\rho}\, \frac{\mathrm{d}(aT^4)}{\mathrm{d}r} = -\frac{c}{3\kappa\rho}\, \frac{\mathrm{d}(aT^4)}{\mathrm{d}r}. \tag{11.15}$$

Route B, the momentum of light. Multiply the transfer equation (11.7) by $\cos\theta/c$ and integrate over all directions. Along a radial coordinate, $\mathrm{d}/\mathrm{d}s = \cos\theta\ \mathrm{d}/\mathrm{d}r$ for the part of the change that matters deep inside a star, where curvature over one free path is negligible. Then

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}r}\, \frac{1}{c} \int I_\nu \cos^2\theta\, d\Omega &\overset{(1)}{=} -\frac{\kappa_\nu \rho}{c} \int I_\nu \cos\theta\, d\Omega + \frac{j_\nu}{c} \int \cos\theta\, d\Omega \\ \Longrightarrow \qquad \frac{\mathrm{d}P_\nu}{\mathrm{d}r} &\overset{(2)}{=} -\frac{\kappa_\nu \rho}{c}\, F_\nu. \end{aligned} \tag{11.16}$$

(1) Eq. (11.7) with $\mathrm{d}/\mathrm{d}s = \cos\theta\ \mathrm{d}/\mathrm{d}r$, times $\cos\theta/c$, integrated over $d\Omega$.
(2) The moments of Eq. (11.4). Emission is the same in all directions, so $\int \cos\theta\, d\Omega = 0$ removes it.

This is momentum conservation (P1) for light. The gradient of the pressure of light equals the momentum per unit volume and time that the light gives to the matter. Deep inside, $P = aT^4/3$, and with κ a suitable average over frequency,

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{aT^4}{3}\right) = -\frac{\kappa\rho}{c}\, F, \tag{11.17}$$

which is Eq. (11.15) again. (The right average weights $1/\kappa_\nu$ by $\partial B_\nu/\partial T$, so the frequencies where matter is most transparent carry most of the flux. It is called the Rosseland mean, and Route C derives it.)

Route C, the light that lags behind. Routes A and B both assumed that the light deep inside is Planck light plus a small part that carries the flux. The transfer equation in local equilibrium, Eq. (11.14), gives that small part directly, and shows where it comes from. Take a point at radius r and a ray through it at angle θ to the outward radial direction. Measure optical depth backward along the ray from the point, $\tau' = \kappa_\nu\rho \times$ (distance back). The ray has come from very deep, so the light that entered it far back is dimmed to nothing, and the formal solution (11.9) gives

$$I_\nu(r,\theta) = \int_0^\infty B_\nu\big(\text{at } \tau' \text{ back}\big)\, e^{-\tau'}\, d\tau'. \tag{11.18}$$

The light at a point is the Planck light of the gas behind it along the ray, each layer dimmed by the matter in between. The weight $e^{-\tau'}$ means that most of it was emitted within about one mean free path. That place lies at radius $r - (\tau'/\kappa_\nu\rho)\cos\theta$, since the ray climbs $\cos\theta$ in radius per unit length. Over one free path the temperature changes very little, so expand to first order,

$$\begin{aligned} I_\nu(r,\theta) &\overset{(1)}{=} \int_0^\infty \left[B_\nu(r) - \frac{\tau'\cos\theta}{\kappa_\nu\rho}\, \frac{\mathrm{d}B_\nu}{\mathrm{d}r}\right] e^{-\tau'}\, d\tau' \\ &\overset{(2)}{=} B_\nu\big(T(r)\big) - \frac{\cos\theta}{\kappa_\nu\rho}\, \frac{\mathrm{d}B_\nu}{\mathrm{d}r}. \end{aligned} \tag{11.19}$$

(1) Taylor expansion of B_ν about r, with $\kappa_\nu\rho$ taken as constant over a free path.
(2) $\int_0^\infty e^{-\tau'} d\tau' = 1$ and $\int_0^\infty \tau' e^{-\tau'} d\tau' = 1$.

$$\boxed{I_\nu = B_\nu(T) + \delta I_\nu, \qquad \delta I_\nu = -\frac{\cos\theta}{\kappa_\nu \rho}\,\frac{\partial B_\nu}{\partial T}\,\frac{\mathrm{d}T}{\mathrm{d}r}.} \tag{11.20}$$

δI_ν is the emission of neighbouring gas at a different temperature, seen from here. A ray heading outward, $\cos\theta > 0$, was emitted about one free path deeper, where the gas is hotter, so it is a little brighter than the local B_ν. A ray heading inward comes from the cooler gas above and is a little dimmer. The same result follows by putting $I_\nu = B_\nu + \delta I_\nu$ into Eq. (11.14) and dropping $\mathrm{d}\delta I_\nu/\mathrm{d}s$, which is smaller by one more factor of the free path over the distance on which the temperature changes. This is the lag of Section 11.3 in numbers.

The departure from equilibrium is very small. $\delta I_\nu / B_\nu$ is about the free path divided by the distance over which T changes by itself. For the Sun that is roughly 1 cm over 7×10^8 m, about 10^{-11}. Each parcel is in local equilibrium at its own temperature, and the light is Planck light to eleven figures. If the temperature were the same everywhere, $\mathrm{d}T/\mathrm{d}r = 0$, the light would be exactly Planck light and would carry nothing. A star shines because it is not in global thermodynamic equilibrium.

Now the flux. Only δI_ν contributes,

$$\begin{aligned} F_\nu &\overset{(1)}{=} \int I_\nu \cos\theta\, d\Omega \overset{(2)}{=} B_\nu \int \cos\theta\, d\Omega - \frac{1}{\kappa_\nu \rho}\,\frac{\partial B_\nu}{\partial T}\,\frac{\mathrm{d}T}{\mathrm{d}r}\int \cos^2\theta\, d\Omega \\ &\overset{(3)}{=} -\frac{4\pi}{3\kappa_\nu \rho}\,\frac{\partial B_\nu}{\partial T}\,\frac{\mathrm{d}T}{\mathrm{d}r}. \end{aligned} \tag{11.21}$$

(1) The flux moment of Eq. (11.4).
(2) Eq. (11.20).
(3) $\int \cos\theta\, d\Omega = 0$, the isotropic part carries nothing, and $\int \cos^2\theta\, d\Omega = 4\pi/3$, the $\frac{1}{3}$ of the pressure integral (9.25).

Each frequency carries a flux in proportion to $1/\kappa_\nu$, so light leaks out mostly through the windows where matter is most transparent. Integrating over frequency defines the **Rosseland mean** opacity κ_R,

$$\boxed{\frac{1}{\kappa_R} \equiv \frac{\displaystyle\int_0^\infty \frac{1}{\kappa_\nu}\,\frac{\partial B_\nu}{\partial T}\,d\nu}{\displaystyle\int_0^\infty \frac{\partial B_\nu}{\partial T}\,d\nu}.} \tag{11.22}$$

It is a harmonic mean, an average of $1/\kappa_\nu$, because transparent windows matter most, and it is weighted by $\partial B_\nu/\partial T$, the frequencies whose Planck light changes most from one layer to the next. The denominator follows from $B_\nu = c\,u_\nu/4\pi$ and $u = aT^4$ (Eq. (9.31)),

$$\int_0^\infty B_\nu\, d\nu = \frac{acT^4}{4\pi} \quad\Longrightarrow\quad \int_0^\infty \frac{\partial B_\nu}{\partial T}\, d\nu = \frac{acT^3}{\pi}.$$

Then

$$\begin{aligned} F &\overset{(1)}{=} -\frac{4\pi}{3\rho}\,\frac{\mathrm{d}T}{\mathrm{d}r}\int_0^\infty \frac{1}{\kappa_\nu}\,\frac{\partial B_\nu}{\partial T}\,d\nu \\ &\overset{(2)}{=} -\frac{4\pi}{3\kappa_R\rho}\,\frac{acT^3}{\pi}\,\frac{\mathrm{d}T}{\mathrm{d}r} = -\frac{4acT^3}{3\kappa_R\rho}\,\frac{\mathrm{d}T}{\mathrm{d}r} = -\frac{c}{3\kappa_R\rho}\,\frac{\mathrm{d}(aT^4)}{\mathrm{d}r}. \end{aligned} \tag{11.23}$$

(1) Integrate F_ν over ν. $\mathrm{d}T/\mathrm{d}r$ is the same at every frequency.
(2) Eq. (11.22) and the integral above.

This is Eq. (11.15) of Route A, with $\kappa = \kappa_R$. The three routes agree, and Route C shows what each assumed. The isotropic part B_ν of the light holds the energy density aT^4 and the pressure $aT^4/3$ that Route B used. The small part δI_ν is the flux.

The transport equation. Write $F = L(r)/(4\pi r^2)$, where $L(r)$ is the luminosity through the sphere of radius r, $\mathrm{d}(aT^4)/\mathrm{d}r = 4aT^3\ \mathrm{d}T/\mathrm{d}r$, and κ for the Rosseland mean κ_R. Then

$$\boxed{\frac{\mathrm{d}T}{\mathrm{d}r} = -\frac{3\,\kappa\rho\,L(r)}{16\pi ac\,r^2T^3}.} \tag{11.24}$$

The temperature gradient needed to carry a luminosity grows with the opacity. Opaque matter needs a steep fall of temperature to push the same energy through. This is the radiative transport equation of a star (Section 12.2).

Read with Route C, the equation runs from the flux to the temperature. The luminosity a layer must pass on sets the flux. The flux sets how unequal the light going up and down must be, δI_ν. And δI_ν is only produced by a difference of temperature over a free path, so the flux fixes how steeply T falls. Where $L > 0$, $\mathrm{d}T/\mathrm{d}r < 0$, and the temperature falls outward.

The equation does not give $T(r)$ by itself. The opacity, the density and $L(r)$ all change with radius, so it has to be solved together with hydrostatic equilibrium, the mass and energy equations and the equation of state. Chapter 12 does this, writing it per unit mass as Eq. (12.9). Its shape shows in a simple case. Over a range of radii where κ, ρ and L hardly change, use $T^3\ \mathrm{d}T/\mathrm{d}r = \frac{1}{4}\ \mathrm{d}T^4/\mathrm{d}r$ and integrate from some radius r_0,

$$\begin{aligned} \frac{\mathrm{d}T^4}{\mathrm{d}r} &\overset{(1)}{=} -\frac{3\kappa\rho L}{4\pi ac\,r^2} \\ \Longrightarrow \qquad T^4(r) &\overset{(2)}{=} T^4(r_0) + \frac{3\kappa\rho L}{4\pi ac}\left(\frac{1}{r} - \frac{1}{r_0}\right). \end{aligned} \tag{11.25}$$

(1) Eq. (11.24) times $4T^3$.
(2) $\int_{r_0}^{r} dr'/r'^2 = 1/r_0 - 1/r$.

For $r > r_0$ the bracket is negative and T^4 falls. It falls fastest where the opacity is largest, since there the light needs the largest difference of temperature over each free path to carry the same flux. This is only an illustration over a short range. The grey atmosphere of Section 11.8 follows the same balance all the way out to where the light escapes.

11.5 The Eddington Luminosity

The outward flux that carries a star's energy also transfers momentum to its gas, as Route B showed. Increasing the luminosity strengthens this outward push. Eventually it can balance the inward pull of gravity.

Think of standing in a strong wind. A breeze does nothing, a gale pushes you over. Somewhere between, the push of the wind equals the grip your weight gives you on the ground. The matter at the surface of a star is held down by gravity and pushed out by the momentum of the light streaming past it (P1 for light, P2 for gravity). Both forces fall off as $1/r^2$, so their ratio depends only on the luminosity and the mass. Above a certain luminosity, the Eddington luminosity, no star can hold on to its outer layers.

For ionised hydrogen around one solar mass, the limiting luminosity is 1.3×10^{31} W, or about 3.3×10^4 times the Sun's luminosity. The Sun is well below this limit. The same balance matters for gas falling onto a black hole: as the gas radiates, its light pushes against the incoming material. A quasar shining at 10^{39} W therefore needs at least 10^8 solar masses if its luminosity is below this limit.

The limit. From Route B, light of flux F gives the matter a force per unit volume $\kappa\rho F/c$, so a force per unit mass $\kappa F/c$. Gravity pulls with GM/r^2 per unit mass. With $F = L/(4\pi r^2)$ the two balance when

$$\frac{\kappa}{c}\frac{L}{4\pi r^2} = \frac{GM}{r^2}$$
$$\Longrightarrow \qquad L = \frac{4\pi GMc}{\kappa}. \tag{11.26}$$

$$\boxed{L_{\mathrm{Edd}} = \frac{4\pi GMc}{\kappa}.} \tag{11.27}$$

The radius has dropped out. In hot ionised gas the opacity is scattering by free electrons, with the Thomson cross-section σ_T of Eq. (8.7) per electron. For pure hydrogen there is one electron per mass m_{H}, so by Eq. (11.6) $\kappa = \sigma_T/m_{\mathrm{H}} = 0.040\,\mathrm{m^2\,kg^{-1}}$ and

$$L_{\mathrm{Edd}} = 1.3\times10^{31}\,\mathrm{W}\left(\frac{M}{M_\odot}\right) = 3.3\times10^{4}\,L_\odot\left(\frac{M}{M_\odot}\right).$$

Any larger opacity lowers the limit.

11.6 Einstein Coefficients

Kirchhoff's law fixed the ratio of emission to absorption for matter in local thermodynamic equilibrium (Eq. (11.12)). It did not say what an atom actually does, or what happens when collisions are too rare to keep its levels in equilibrium, as in thin interstellar gas.

Think of a ball resting at the bottom of a step. A kick can lift it onto the step, and from the step it can roll back down by itself. A kick can also knock it down sooner than it would have rolled. If the ball spends time on and off the step in the proportions the temperature demands, the three kinds of move must be related. An atom with two levels makes three kinds of transition. Light can lift it, it can drop by itself, and light can push it down sooner (stimulated emission). Detailed balance (P4) with Planck light and Boltzmann populations (P3) fixes how the three rates are related. Since the rates belong to the atom alone, the relations hold everywhere, in equilibrium or not.

The hydrogen 21 cm line shows why the competing rates matter. Its upper level drops spontaneously at only $2.9\times10^{-15}\,\mathrm{s^{-1}}$ (Eq. (8.37)), about once in 11 million years. In interstellar gas, collisions change the level much more often, so the populations usually follow the gas temperature (Section 8.4). We will describe absorption and both ways an excited atom can emit with three coefficients, then use equilibrium to relate them.

Three rates. Let level 1 (lower) and level 2 (upper) be single states separated by $h\nu$. In light of energy density u_ν at the line frequency,

- absorption lifts an atom from 1 to 2 at the rate $B_{12}\,u_\nu$ per atom,
- spontaneous emission drops it from 2 to 1 at the rate A_{21} per atom,
- stimulated emission drops it from 2 to 1 at the rate $B_{21}\,u_\nu$ per atom.

These are the **Einstein coefficients**. They belong to the atom.

Relations. In thermodynamic equilibrium at temperature T, the upward and downward flows balance (P4), the populations follow the Boltzmann factor, $N_2/N_1 = e^{-h\nu/k_BT}$, and the

light is Planck light (P3),

$$N_1 B_{12}\, u_\nu \overset{(1)}{=} N_2\big(A_{21} + B_{21}\, u_\nu\big)$$
$$\Longrightarrow \qquad u_\nu \overset{(2)}{=} \frac{A_{21}}{B_{12}\, e^{h\nu/k_B T} - B_{21}}. \tag{11.28}$$

(1) Detailed balance between level 1 and level 2.
(2) Solve for u_ν and use $N_1/N_2 = e^{h\nu/k_B T}$.

This must equal the Planck energy density, $u_\nu = (8\pi h\nu^3/c^3)/(e^{h\nu/k_B T} - 1)$ from Section 9.3, at every temperature. That is possible only if

$$\boxed{B_{12} = B_{21}, \qquad A_{21} = \frac{8\pi h\nu^3}{c^3}\, B_{21}.} \tag{11.29}$$

One coefficient fixes the other two. The relations hold for the atom in any surroundings, since the coefficients do not depend on whether the gas is in equilibrium. Stimulated emission is required by them. Without it no choice of coefficients could reproduce the Planck spectrum. (For levels with several states, each B is multiplied by the number of states of its starting level.)

When local equilibrium fails. The populations follow the gas temperature when collisions move atoms between the levels much faster than spontaneous emission does. The ratio of the collision rate to A_{21} decides (P6). In dense stellar atmospheres collisions win, and local thermodynamic equilibrium holds. In thin nebulae, light and spontaneous decay set the populations instead, and lines must be computed from the rates themselves. If some process fills the upper level faster than the lower one, $N_2 > N_1$, the net absorption $(N_1 - N_2)B_{12}u_\nu$ turns negative and the gas amplifies light passing through it. Such masers are seen in water and hydroxyl molecules in star-forming clouds.

11.7 Spectral Classes

A star shows dark lines where cooler gas lies above hotter gas (Section 11.3), and the strength of a line is set by how many atoms sit in its lower level (Section 11.6). Real spectra differ far more than their compositions do. The Sun shows thousands of lines of iron and calcium and only weak lines of hydrogen. Vega shows strong hydrogen lines and few others. Both stars are about three quarters hydrogen by mass. Something other than composition decides which lines show.

Think of a shop that sells umbrellas to people walking past. On a dry day nobody needs one. In a storm nobody walks past. The shop sells most on a day of light rain, because a sale needs two things at once, and the weather that brings one takes away the other. Transitions from higher energy levels to $n = 2$ in hydrogen produce visible spectral lines, and these form the Balmer series. In a star's spectrum they appear in absorption, the same jumps upward from $n = 2$, so they need atoms already in that level. Heat lifts atoms into it, and more heat strips the electron off altogether. So the hydrogen lines are strongest at one temperature in between, near 10^4 K. The pattern of lines in a star measures its temperature, not its composition.

The rise and fall is large enough to see in the numbers. At the Sun's 5772 K, only 5 hydrogen atoms in 10^9 are in level 2. At 9870 K, the share is 9 in a million. By 2×10^4 K, only 3 atoms in 10^5 still hold their electron, and the share in level 2 has fallen to 3 in 10^7. Figure 11.2 shows the resulting peak. We will use the same competition between excitation and ionisation to read the sequence of stellar spectra.

The sequence. Stars are sorted by the lines they show into the classes O, B, A, F, G, K and M, from hot to cool. The letters are left over from an older alphabetical order by the strength of the hydrogen lines, which was later rearranged by temperature. Each class is split into subtypes 0 to 9, with 0 the hottest, so a G2 star is a little cooler than a G0 star. Table 11.1 gives the classes. The Sun is G2.

Class	T_{eff} (K)	Strongest lines	Example
O	above 30 000	ionised helium (He II), hydrogen weak	–
B	10 000–30 000	neutral helium (He I), hydrogen growing	Rigel
A	7500–10 000	hydrogen, strongest at A0	Vega
F	6000–7500	hydrogen weaker, ionised calcium (Ca II H and K) and other ionised metals growing	Procyon
G	5200–6000	Ca II H and K strong, neutral metals growing	Sun
K	3700–5200	neutral metals, first molecular bands	Arcturus
M	below 3700	molecules, titanium oxide (TiO) bands	Betelgeuse

Table 11.1: The spectral classes, their effective temperatures and their strongest absorption lines.

Each entry follows from the competition of the opening. Ionised helium needs 24.6 eV to remove the first electron, so it exists only in the hottest stars, and its lines mark class O. Neutral helium absorbs only from levels about 21 eV above its ground state, so it needs B-star temperatures to fill them. Calcium loses its first electron at 6.1 eV and its second at 11.9 eV. Its H and K lines, at 396.8 and 393.4 nm, start from the ground state of Ca^+, so they need calcium ionised once but not twice, which happens in the F, G and K stars. Neutral metals survive only in cooler gas. Molecules such as TiO are broken apart by collisions above about 4000 K, so their bands appear only in M stars.

Why a line tracks a population. By Eq. (11.6), the opacity at the frequency of a line is $n\sigma_\nu/\rho$, where n is now the density of atoms that can absorb that line, the atoms in its lower level. A larger n makes the gas opaque at the line higher up, where it is cooler, so the line is darker (Section 11.3). Comparing the same line from star to star therefore compares the density of atoms in its lower level.

The hydrogen lines start from level 2. A hydrogen atom has levels of energy $E_n = -13.6\,\text{eV}/n^2$, where $n = 1, 2, 3, \ldots$ is the principal quantum number. A line from level 2 to level n has photon energy $13.6\,\text{eV}\,(1/4 - 1/n^2)$. For $n = 3$ that is 1.889 eV, a wavelength of 656 nm, which is red light. For $n = 4$ it is 486 nm, and the series ends at 365 nm. These are the **Balmer lines**, and they lie in visible light. Lines from level 1 (the Lyman series, starting at 121.6 nm) lie in the ultraviolet. So the hydrogen lines of an optical spectrum measure n_2, the density of atoms in level 2.

The share of hydrogen in level 2 (P3 + P4). Two kinds of balance set n_2. Within the neutral atom, collisions share atoms among the levels by the Boltzmann factor. Between the atom and its ion, ionisation is balanced by recombination, which is the Saha equation. We want the share of all hydrogen, neutral and ionised, that sits in level 2. Write n_1 and n_2 for the densities of atoms in levels 1 and 2, $n_{\text{HI}} = n_1 + n_2$ for the neutral atoms, n_p for the protons, and $n_{\text{H,tot}} = n_{\text{HI}} + n_p$. Level n holds $g_n = 2n^2$ states (two spin directions times n^2 orbital states), so $g_1 = 2$ and $g_2 = 8$. Levels 3 and above hold about 10^{-5} of the neutral atoms at 10^4 K and a few per cent at 2×10^4 K, so we keep only levels 1 and 2. The electron density n_e is taken as given. In the cooler atmospheres most free electrons come from metals such as sodium, magnesium

and iron, which lose an electron far more easily than hydrogen, so n_e is not n_p. The ideal-gas law gives $n_e = P_e/k_BT$, where P_e is the pressure of the electrons alone. Then

$$\begin{aligned}
\frac{n_2}{n_1} &\overset{(1)}{=} \frac{g_2}{g_1}\, e^{-(E_2-E_1)/k_BT} \overset{(2)}{=} 4\, e^{-10.2\,\mathrm{eV}/k_BT}, \\
\frac{n_p}{n_1} &\overset{(3)}{=} \frac{1}{n_e}\left(\frac{2\pi m_e k_B T}{h^2}\right)^{3/2} e^{-13.6\,\mathrm{eV}/k_BT} \overset{(4)}{=} \frac{k_BT}{P_e}\left(\frac{2\pi m_e k_B T}{h^2}\right)^{3/2} e^{-13.6\,\mathrm{eV}/k_BT}, \\
\frac{n_2}{n_{\mathrm{H,tot}}} &\overset{(5)}{=} \frac{n_2}{n_1+n_2+n_p} \overset{(6)}{=} \frac{n_2/n_1}{1+n_2/n_1+n_p/n_1}.
\end{aligned} \tag{11.30}$$

(1) The Boltzmann factor for two levels, with each side counted by its number of states (Eq. (9.39) and the remark after it).
(2) $g_2/g_1 = 8/2$ and $E_2 - E_1 = 13.6\,\mathrm{eV}\,(1 - \frac{1}{4}) = 10.2\,\mathrm{eV}$.
(3) The Saha equation (16.16) divided by n_e. Its n_H counts atoms in the ground state, which is n_1 here, and its binding energy is 13.6 eV.
(4) $n_e = P_e/k_BT$.
(5) $n_{\mathrm{H,tot}} = n_1 + n_2 + n_p$, with only levels 1 and 2 kept.
(6) Divide the top and the bottom by n_1.

Read as a product, the last line is the Boltzmann share within the neutral atoms times the neutral share of all hydrogen,

$$\boxed{\frac{n_2}{n_{\mathrm{H,tot}}} = \underbrace{\frac{n_2/n_1}{1+n_2/n_1}}_{n_2/n_{\mathrm{HI}}} \times \underbrace{\frac{1}{1+\dfrac{n_p/n_1}{1+n_2/n_1}}}_{n_{\mathrm{HI}}/n_{\mathrm{H,tot}}},} \tag{11.31}$$

with n_2/n_1 and n_p/n_1 from the first two lines of the chain. With the electron pressure given, n_p/n_1 follows from the temperature alone. The first factor rises with T for as long as k_BT is well below 10.2 eV. The second factor is close to one while hydrogen is neutral and falls toward zero once it ionises. The product rises and then falls.

Numbers. The constants are $(2\pi m_e k_B/h^2)^{3/2} = 2.415\times10^{21}\,\mathrm{m^{-3}\,K^{-3/2}}$, $13.6\,\mathrm{eV}/k_B = 1.578\times10^5\,\mathrm{K}$ and $10.2\,\mathrm{eV}/k_B = 1.184\times10^5\,\mathrm{K}$. We take $P_e = 20\,\mathrm{N\,m^{-2}}$, a value typical of model atmospheres of stars near 10^4 K. Then

$$\frac{n_2}{n_1} = 4\, e^{-1.184\times10^5\,\mathrm{K}/T}, \qquad \frac{n_p}{n_1} = 1.667\times10^{-3}\left(\frac{T}{1\,\mathrm{K}}\right)^{5/2} e^{-1.578\times10^5\,\mathrm{K}/T}.$$

At the Sun's 5772 K, $n_2/n_1 = 5.0\times10^{-9}$ and $n_p/n_1 = 5.6\times10^{-6}$. Hydrogen is neutral, and almost none of it is in level 2. Figure 11.2 shows Eq. (11.31) from 4000 to 25 000 K.

Where the peak lies. Near the peak n_2/n_1 is about 10^{-5}, so it can be dropped next to 1, and $n_2/n_{\mathrm{H,tot}} \approx (n_2/n_1)/(1+n_p/n_1)$. The peak is where its logarithm stops changing with T,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}T}\ln\frac{n_2}{n_{\mathrm{H,tot}}} &\overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}T}\ln\frac{n_2}{n_1} - \frac{n_p/n_1}{1+n_p/n_1}\frac{\mathrm{d}}{\mathrm{d}T}\ln\frac{n_p}{n_1} \\
&\overset{(2)}{=} \frac{1}{T}\frac{10.2\,\mathrm{eV}}{k_BT} - \frac{n_p/n_1}{1+n_p/n_1}\frac{1}{T}\left(\frac{5}{2}+\frac{13.6\,\mathrm{eV}}{k_BT}\right) \overset{(3)}{=} 0.
\end{aligned} \tag{11.32}$$

(1) The logarithm of a ratio, and $\mathrm{d}\ln(1+y)/\mathrm{d}T = \big(y/(1+y)\big)\,\mathrm{d}\ln y/\mathrm{d}T$ with $y = n_p/n_1$.

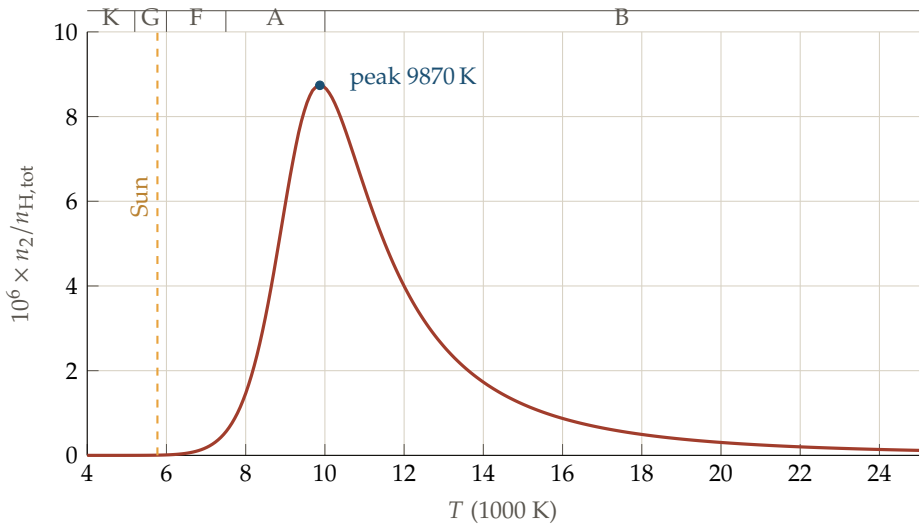


Figure 11.2: The share of all hydrogen in level $n = 2$ against temperature, Eq. (11.31), for an electron pressure of $20\,\mathrm{N\,m^{-2}}$ and levels 1 and 2 only. The dot marks the peak. The dashed line marks the Sun's effective temperature. The letters along the top mark the spectral classes.

(2) $\ln(n_2/n_1) = \ln 4 - 10.2\,\mathrm{eV}/k_BT$, and $\ln(n_p/n_1) = \text{const} + \frac{5}{2}\ln T - 13.6\,\mathrm{eV}/k_BT$.
(3) The condition for the peak.

$$\boxed{\frac{n_p}{n_1 + n_p} = \frac{10.2\,\mathrm{eV}/k_BT}{\frac{5}{2} + 13.6\,\mathrm{eV}/k_BT} \quad \text{at the peak.}} \tag{11.33}$$

The left side is the ionised share of the hydrogen. Since $13.6\,\mathrm{eV}/k_BT$ is much larger than $\frac{5}{2}$, the right side is a little below $10.2/13.6 = \frac{3}{4}$. The Balmer lines are strongest when about two thirds of the hydrogen has already lost its electron. Bisection on Eq. (11.33) with $P_e = 20\,\mathrm{N\,m^{-2}}$,

$$\begin{aligned} T_{\mathrm{peak}} &\overset{(1)}{=} 9873\,\mathrm{K}, \\ \frac{n_p}{n_1 + n_p} &\overset{(2)}{=} \frac{11.99}{2.5 + 15.98} = 0.649, \\ \frac{n_2}{n_{\mathrm{H,tot}}} &\overset{(3)}{=} 8.74 \times 10^{-6}. \end{aligned} \tag{11.34}$$

(1) The difference of the two sides is -0.420 at 9000 K and $+0.053$ at 10 000 K. Sixty halvings of that interval end at 9873 K. Bisecting instead on the slope of the full Eq. (11.31), without dropping n_2/n_1, gives the same 9873 K.
(2) At 9873 K, $10.2\,\mathrm{eV}/k_BT = 11.99$, $13.6\,\mathrm{eV}/k_BT = 15.98$ and $n_p/n_1 = 1.846$, so both sides are 0.649. There $n_2/n_1 = 2.5 \times 10^{-5}$, so dropping it was safe.
(3) Eq. (11.31) at 9873 K, the dot in Figure 11.2. The curve is above half this value from 8750 to 11 850 K.

This is the temperature of the A0 stars, where the Balmer lines are strongest. The electron pressure enters only through $n_p/n_1 \propto 1/P_e$, inside a steep exponential, so the peak moves little. The same bisection gives 8790 K for $P_e = 2\,\mathrm{N\,m^{-2}}$ and 11 230 K for $200\,\mathrm{N\,m^{-2}}$.

At the peak $k_BT = 0.85\,\mathrm{eV}$, only a sixteenth of the 13.6 eV needed to ionise. Hydrogen ionises at so low a temperature because a free electron has far more states to choose from than a bound one (P6). At 9873 K the factor $(2\pi m_e k_B T/h^2)^{3/2}/n_e$ is 1.6×10^7, so a Boltzmann factor $e^{-13.6\,\mathrm{eV}/k_BT} = 1.1 \times 10^{-7}$ is already enough to free most electrons. Recombination in the early

universe is the same balance, with a billion photons per atom in place of the free electron's many states (Section 16.3).

Luminosity classes. Stars of one spectral type can differ in size a hundredfold. Their spectra differ too, through the pressure in the atmosphere. Hydrostatic equilibrium (P1 + P2), Eq. (10.2), with the surface gravity g held constant over the thin atmosphere and with optical depth measured inward, $d\tau = -\kappa\rho\, dr$, gives

$$\frac{\mathrm{d}P}{\mathrm{d}\tau} \overset{(1)}{=} \frac{\mathrm{d}P}{\mathrm{d}r}\,\frac{\mathrm{d}r}{\mathrm{d}\tau} \overset{(2)}{=} (-\rho g)\left(-\frac{1}{\kappa\rho}\right) = \frac{g}{\kappa}$$

$$\Longrightarrow \qquad P(\tau) \overset{(3)}{=} \frac{g}{\kappa}\,\tau \overset{(4)}{=} \frac{2g}{3\kappa} \quad \text{at the photosphere.} \tag{11.35}$$

(1) The chain rule.
(2) Eq. (10.2) with $Gm/r^2 = g$, and $d\tau = -\kappa\rho\, dr$.
(3) Integrate from the top, where $P = 0$ and $\tau = 0$, with κ taken as constant.
(4) The photosphere lies at $\tau = \frac{2}{3}$ (Eq. (11.43)).

At a given temperature the opacity is roughly the same, so the pressure of the layer we see scales with $g = GM/R^2$. The Sun has $g = 274\,\mathrm{m\,s^{-2}}$. A giant of one solar mass and ten solar radii has $g = 2.7\,\mathrm{m\,s^{-2}}$, a hundred times less, and a supergiant less again. Lower pressure changes the spectrum in two ways.

- **More ionisation.** At lower pressure n_e is lower, and by the Saha equation $n_p/n_1 \propto 1/n_e$, so the gas is more ionised at the same temperature. A giant reaches a given spectral type at a lower temperature than a dwarf. For the Balmer lines, the peak moves from 9870 K to 8790 K when P_e falls tenfold.
- **Narrower lines.** Collisions between atoms broaden lines in proportion to the density, Eq. (8.26). In the thin atmosphere of a giant they are rarer, so its lines are narrower than a dwarf's of the same type.

The **luminosity class**, read from these signs of the surface gravity, is written as a Roman numeral after the type. Ia and Ib are supergiants (bright and less bright), II bright giants, III giants, IV subgiants and V main-sequence stars, also called dwarfs. The Sun is G2V. Type and class together place a star on the HR diagram, Figure 12.8, the type along the temperature axis and the class across the luminosity axis. That is what lets a spectrum alone give a star's absolute magnitude, and so its distance (Section 3.7.2).

11.7.1 Planet Atmospheres in Transit

A planet crossing its star blocks the fraction $(R_p/R_*)^2$ of the star's light, Eq. (6.243), with R_p and R_* the radii of planet and star. Around the dark disk of the planet lies a thin ring of air. Starlight grazing the planet crosses that ring on its way to us. At the wavelength of a line of some molecule in the air, the ring is opaque higher up, just as a star's atmosphere is (Section 11.3), so the planet looks larger and the transit is a little deeper. Measuring the depth wavelength by wavelength gives the planet's absorption spectrum. This is **transmission spectroscopy**.

The grazing ray. The calculation combines hydrostatic equilibrium (P1 + P2) with the dimming of a ray by absorption (P1, P4), and asks where the optical depth along the ray reaches one (P6). Take an isothermal atmosphere. By Eq. (10.4) its pressure falls as $e^{-z/H}$ with height z, and at one temperature the density falls the same way, $\rho(z) = \rho(0)\, e^{-z/H}$, with the scale height $H = k_B T/(\mu m_{\mathrm{H}} g)$. A ray passes the planet at height z above the radius R_p. Let x be the distance

along the ray from its closest point. The ray is then at distance $\sqrt{(R_p+z)^2+x^2}$ from the centre, and at height

$$\sqrt{(R_p+z)^2+x^2}-R_p \overset{(1)}{=} (R_p+z)\left(1+\frac{x^2}{2(R_p+z)^2}\right)-R_p \overset{(2)}{=} z+\frac{x^2}{2R_p}. \tag{11.36}$$

(1) $\sqrt{1+y}\approx 1+y/2$ for $x \ll R_p$, which holds where the air is dense enough to matter.
(2) $z \ll R_p$ in the second term.

The optical depth along the whole ray is the opacity times the mass crossed (Eq. (11.6)),

$$\tau_\nu \overset{(1)}{=} \int_{-\infty}^{\infty} \kappa_\nu\, \rho(0)\, e^{-z/H}\, e^{-x^2/(2R_pH)}\, dx \overset{(2)}{=} \kappa_\nu\, \rho(z)\sqrt{2\pi R_p H}. \tag{11.37}$$

(1) $d\tau_\nu = \kappa_\nu \rho\, dx$, with ρ at the height just found and κ_ν the same at all heights.
(2) The Gaussian integral $\int_{-\infty}^{\infty} e^{-x^2/(2s^2)}\, dx = \sqrt{2\pi}\, s$, with $s^2 = R_pH$, and $\rho(0)e^{-z/H} = \rho(z)$.

The grazing ray crosses $\sqrt{2\pi R_p/H}$ times more air than a ray going straight down from height z, which crosses $\rho(z)H$ per unit area.

How much larger. The planet blocks the rays with τ_ν above about one. So it looks opaque up to the height where $\tau_\nu = 1$,

$$z_\nu = H \ln\!\left(\kappa_\nu\, \rho(0)\sqrt{2\pi R_p H}\right).$$

Compare a wavelength in a line, with opacity κ_{line}, and one just beside it, with κ_{cont}. Everything but the opacity cancels,

$$\boxed{\Delta z = z_{\text{line}} - z_{\text{cont}} = H \ln \frac{\kappa_{\text{line}}}{\kappa_{\text{cont}}}.} \tag{11.38}$$

The planet grows by a few scale heights. A line a hundred times more opaque than its surroundings lifts the edge by $H \ln 100 = 4.6\,H$. A thousand times only reaches $6.9\,H$, since the opacity enters through its logarithm. By Eq. (6.243) with $R_p + \Delta z$ in place of R_p,

$$\text{depth in the line} - \text{depth beside it} \overset{(1)}{=} \frac{(R_p+\Delta z)^2 - R_p^2}{R_*^2} = \frac{2R_p\,\Delta z + \Delta z^2}{R_*^2} \overset{(2)}{=} \frac{2R_p\,H}{R_*^2} \ln \frac{\kappa_{\text{line}}}{\kappa_{\text{cont}}}. \tag{11.39}$$

(1) Eq. (6.243) at the two radii.
(2) $\Delta z \ll R_p$, so Δz^2 is dropped, and Eq. (11.38).

The signal is the area of a ring of width Δz around the planet, over the area of the star.

A hot Jupiter and the Earth. Take a hot Jupiter with $T = 1200\,\text{K}$, $\mu = 2.3$ for hydrogen and helium, $g = 25\,\text{m}\,\text{s}^{-2}$, and Jupiter's radius $R_p = 7.15\times10^7\,\text{m}$, in front of a star like the Sun,

$R_* = 6.96 \times 10^8$ m. Then

$$H = \frac{(1.381\times10^{-23})(1200)}{2.3\,(1.674\times10^{-27})(25)}\ \text{m} = 172\,\text{km},$$
$$\frac{2R_pH}{R_*^2} = \frac{2\,(7.15\times10^7)(1.72\times10^5)}{(6.96\times10^8)^2} = 5.1\times10^{-5}\ \text{per scale height},$$
$$\text{signal} \overset{(1)}{=} 5.1\times10^{-5}\times4.6 = 2.3\times10^{-4}. \tag{11.40}$$

(1) A line a hundred times more opaque than its surroundings, Eq. (11.39). The edge rises by 790 km.

The transit itself is $(R_p/R_*)^2 = 1.06\times10^{-2}$ deep, so the line deepens it by about two per cent. The air at the limb is thin, but the grazing ray crosses $\sqrt{2\pi R_p/H} = 51$ times more of it than a vertical one, which is why the ring shows at all. For the Earth in front of the Sun, $H = 8.4$ km (Eq. (10.4)) and $R_p = 6.37\times10^6$ m, and the same line gives $2.2\times10^{-7}\times4.6 = 1.0\times10^{-6}$, against a transit depth of 8.4×10^{-5}.

The signal is proportional to $H = k_BT/(\mu m_{\text{H}}g)$. Hot planets with light air and weak gravity show the largest features, and that is why hot Jupiters were studied first. The size of the features measures H, and so μ. Air made of hydrogen and helium, $\mu = 2.3$, gives features eight times larger than air made of water, $\mu = 18$, at the same temperature and gravity. A layer of cloud adds opacity at every wavelength and flattens the features, since only the ratio $\kappa_{\text{line}}/\kappa_{\text{cont}}$ enters.

Sodium was the first gas found this way, in 2002 with the Hubble Space Telescope, in the hot Jupiter HD 209458 b. Water has since been found in many hot Jupiters, and in 2022 the James Webb Space Telescope found carbon dioxide in WASP-39 b. A signal of 10^{-6} for an Earth around a Sun-like star is still out of reach. Around a small M dwarf, R_* is about ten times smaller and the signal a hundred times larger (Problem 11.9.6). No sign of life has been found in any planet's air.

11.8 The Grey Atmosphere

The effective temperature is defined by $L = 4\pi R^2\sigma_{\text{SB}}T_{\text{eff}}^4$ (Eq. (4.26)), as though a star radiated from one surface at one temperature. In reality, its temperature falls continuously outward, and the escaping light comes from a range of depths. We can still ask where in that atmosphere the local temperature equals T_{eff}.

Think of looking into a pool of murky water. You see down to some depth where the water stops being clear, and no farther. The bottom you see is not a surface in the water but the depth at which the murk has added up to enough. A star's photosphere is such a depth. It lies at a fixed optical depth, and the temperature there is the effective temperature. The calculation uses momentum conservation for light (P1), the steady flow of energy through the atmosphere (P4), and Planck light at the local temperature (P3).

The model below places that depth at $\tau = \frac{2}{3}$. At the top, where $\tau = 0$, the temperature is $2^{-1/4} = 0.84$ times as high. For the Sun, this gives 4850 K at the top and 5772 K at the photosphere. Figure 11.3 shows the gradual rise in temperature with depth.

The same calculation, with the opacity's dependence on frequency added, is the starting point of every model stellar atmosphere, and of planetary atmospheres warmed from below. It gives the temperature through the atmosphere.

Assumptions. Take the opacity the same at all frequencies (grey), so frequency can be integrated out. Measure the optical depth τ inward from the top. The atmosphere is thin

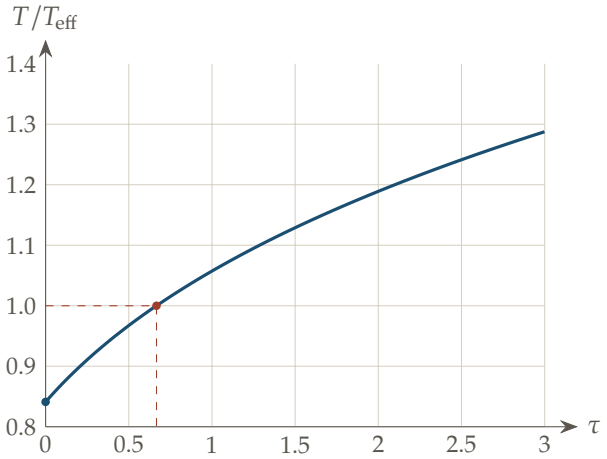


Figure 11.3: The temperature of a grey atmosphere against optical depth, in units of the effective temperature. The point at $\tau = \frac{2}{3}$ marks $T = T_{\text{eff}}$, and the point at $\tau = 0$ marks $T = 2^{-1/4} T_{\text{eff}}$.

compared with the star, so it is treated as flat, and no energy is made in it, so the flux F is the same at every depth (P1, P4).

Temperature against depth. The momentum equation of Route B, with $d\tau = -\kappa\rho\, dr$, reads $\mathrm{d}P_{\text{light}}/\mathrm{d}\tau = F/c$. With F constant,

$$P_{\text{light}}(\tau) = \frac{F}{c}\,\tau + P_{\text{light}}(0). \tag{11.41}$$

Two approximations close it. Deep down the light is nearly the same in all directions, so $P_{\text{light}} = u/3$, and Eddington took this to hold at all depths. At the top, light only leaves and none comes in. If the outgoing light is spread evenly over the outer half of directions with intensity I^+, then $F = \pi I^+$ and $u = 2\pi I^+/c$ from Eq. (11.4), so $u(0) = 2F/c$. Then

$$\begin{aligned} \frac{u}{3} &\overset{(1)}{=} \frac{F}{c}\,\tau + \frac{2F}{3c} \\ \Longrightarrow \qquad aT^4 &\overset{(2)}{=} \frac{3F}{c}\left(\tau + \frac{2}{3}\right) \\ &\overset{(3)}{=} \frac{3}{4}\,aT_{\text{eff}}^4\left(\tau + \frac{2}{3}\right). \end{aligned} \tag{11.42}$$

(1) $P_{\text{light}} = u/3$ at all depths and $P_{\text{light}}(0) = u(0)/3 = 2F/(3c)$.
(2) Local thermodynamic equilibrium, $u = aT^4$.
(3) $F = \sigma_{\text{SB}} T_{\text{eff}}^4$ and $\sigma_{\text{SB}} = ac/4$ (Section 9.3).

$$\boxed{T^4 = \frac{3}{4}\,T_{\text{eff}}^4\left(\tau + \frac{2}{3}\right).} \tag{11.43}$$

The temperature equals T_{eff} at $\tau = \frac{2}{3}$. That depth is the **photosphere**, the layer the effective temperature describes. Above it the temperature levels off at $2^{-1/4} T_{\text{eff}}$, and below it the temperature keeps rising, as $\tau^{1/4}$ deep down, where the diffusion of Section 11.4 takes over.

11.9 Problem Set: Radiation in Matter

11.9.1 Problem 1: The Sun Seen From Mars

Problem Statement: Mars is 1.52 times farther from the Sun than the Earth. Compare (a) the flux of sunlight and (b) the specific intensity of the solar disk seen from the two planets. (c) Why does a camera on Mars need a longer exposure for a picture of the whole landscape, but not for a picture of the Sun's disk filling the same number of pixels?

Essential Ideas:

- The flux is defined at each point, as the power crossing unit area there. The planets are tiny next to their distances from the Sun, so each one counts as a point at distance d. The Sun's power spreads over a sphere of radius d through that point, so $F = L_\odot/4\pi d^2$ falls as $1/d^2$ (Eq. (4.2)). Each planet then receives one flux, and the two are compared through $F_{\rm Mars}/F_\oplus = (d_\oplus/d_{\rm Mars})^2$.
- The intensity is constant along a ray (Eq. (11.3)).

Detailed Solutions:

(a) **Flux**. $F_{\rm Mars}/F_\oplus = 1/1.52^2$.

$$\boxed{F_{\rm Mars} = 0.43\,F_\oplus.}$$

(b) **Intensity**. Unchanged.

$$\boxed{I_{\rm Mars} = I_\oplus.}$$

(c) **Exposures**. A landscape is lit by the total flux, which is lower on Mars. A pixel looking at the Sun's disk collects intensity times the solid angle of one pixel, and both are the same.

11.9.2 Problem 2: How Opaque Is a Dark Cloud?

Problem Statement: A dark cloud dims background starlight in the visual band by $A_V = 10$ magnitudes, and the cloud's own glow in that band is negligible. (a) What is its optical depth? (b) What fraction of the starlight gets through? (c) At 2.2 μm the extinction is about 0.11 of the visual value. What fraction gets through there?

Essential Ideas:

- With $S_\nu = 0$, Eq. (11.10) gives $I = I_0 e^{-\tau}$, and $A = 2.5\log_{10}(I_0/I) = 1.086\,\tau$ (Eq. (4.36)).

Detailed Solutions:

(a) **Optical depth**. $\tau_V = 10/1.086$.

$$\boxed{\tau_V \simeq 9.2.}$$

(b) **Visual transmission**. $e^{-9.2}$.

$$\boxed{I/I_0 \simeq 1.0\times10^{-4}.}$$

(c) **Infrared**. $\tau = 0.11\times9.2 = 1.0$, so $e^{-1.0} = 0.36$.

$$\boxed{I/I_0 \simeq 0.36.}$$

The cloud is opaque in visible light and nearly transparent in the near infrared, which is why young stars inside such clouds are found with infrared telescopes.

11.9.3 Problem 3: How Long Light Takes to Leave the Sun

Problem Statement: Take the mean free path of a photon inside the Sun to be $s = 1\,\mathrm{mm}$ throughout. Estimate the time for energy made at the centre to random-walk to the surface, $R_\odot = 6.96 \times 10^8\,\mathrm{m}$. How many free paths is that?

Essential Ideas:

- The random-walk time $R^2/(sc)$, Eq. (9.42), and $N = (R/s)^2$ steps.

Detailed Solutions:

$$\boxed{t = \frac{(6.96\times10^8)^2}{10^{-3}\times 3.00\times10^8}\,\mathrm{s} \simeq 1.6\times10^{12}\,\mathrm{s} \simeq 5\times10^4\ \text{years}, \qquad N \simeq 5\times10^{23}.}$$

The light we see today left the core tens of thousands of years ago, as energy. The individual photons were absorbed and re-emitted along the way, about 5×10^{23} times.

11.9.4 Problem 4: Eddington Limits

Problem Statement: Using electron scattering in ionised hydrogen, $\kappa = 0.040\,\mathrm{m^2\,kg^{-1}}$, find the Eddington luminosity of (a) a star of $100\,M_\odot$ and (b) a black hole of $10^8\,M_\odot$. (c) A quasar emits $10^{40}\,\mathrm{W}$. What is the least mass that can power it?

Essential Ideas:

- $L_{\mathrm{Edd}} = 4\pi GMc/\kappa = 1.3\times10^{31}\,\mathrm{W}\,(M/M_\odot)$, Eq. (11.27).

Detailed Solutions:

(a) **Massive star**.

$$\boxed{L_{\mathrm{Edd}} \simeq 1.3\times10^{33}\,\mathrm{W} \simeq 3\times10^6\,L_\odot.}$$

The most massive stars shine at a good fraction of this, so they lose mass in strong winds.

(b) **Black hole**.

$$\boxed{L_{\mathrm{Edd}} \simeq 1.3\times10^{39}\,\mathrm{W}.}$$

(c) **Quasar**. $M \geq L\kappa/(4\pi Gc) = 10^{40}/(1.3\times10^{31})\,M_\odot$.

$$\boxed{M \gtrsim 8\times10^8\,M_\odot.}$$

11.9.5 Problem 5: The Surface of a Grey Star

Problem Statement: In the grey atmosphere of Eq. (11.43) for a star with $T_{\mathrm{eff}} = 5772\,\mathrm{K}$, find the temperature (a) at the top, (b) at $\tau = \frac{2}{3}$, and (c) at $\tau = 10$.

Essential Ideas:

- $T = T_{\mathrm{eff}}\,[\frac{3}{4}(\tau + \frac{2}{3})]^{1/4}$.

Detailed Solutions:

(a) **Top**. $T = 2^{-1/4}T_{\mathrm{eff}}$.

$$\boxed{T(0) \simeq 4850\,\mathrm{K}.}$$

(b) **Photosphere**.

$$\boxed{T(\tfrac{2}{3}) = T_{\mathrm{eff}} = 5772\,\mathrm{K}.}$$

(c) **Deeper**. $[\frac{3}{4}\times10.67]^{1/4} = 1.68$.

$$\boxed{T(10) \simeq 9700\,\mathrm{K}.}$$

11.9.6 Problem 6: Seeing a Planet's Air

Problem Statement: A planet of radius 1.27 times Jupiter's ($R_p = 9.08 \times 10^7$ m) transits a star like the Sun ($R_* = 6.96 \times 10^8$ m). Its air is hydrogen and helium ($\mu = 2.3$) at $T = 1100$ K, and its surface gravity is only $g = 5\,\mathrm{m\,s^{-2}}$. (a) Find the scale height. (b) Find the extra transit depth in a molecular band a hundred times more opaque than its surroundings, and compare it with the transit depth. (c) Find the same signal for an Earth ($R_p = 6.37 \times 10^6$ m, $H = 8.4$ km) in front of an M dwarf of 0.12 solar radii.

Essential Ideas:

- The scale height $H = k_B T/(\mu m_\mathrm{H} g)$, Eq. (10.4).
- The extra depth $(2R_p H/R_*^2)\ln(\kappa_\mathrm{line}/\kappa_\mathrm{cont})$, Eq. (11.39).

Detailed Solutions:

(a) **Scale height**. $H = (1.381 \times 10^{-23})(1100)/\big(2.3\,(1.674 \times 10^{-27})(5)\big)$ m.

$$\boxed{H \simeq 790\,\mathrm{km}.}$$

It is about five times the 172 km of the hot Jupiter of Section 11.7.1, mostly because the gravity is five times weaker.

(b) **Signal**. $2R_pH/R_*^2 = 2\,(9.08\times10^7)(7.89\times10^5)/(6.96\times10^8)^2 = 2.96\times10^{-4}$, times $\ln 100 = 4.6$.

$$\boxed{\text{extra depth} \simeq 1.4 \times 10^{-3}.}$$

The transit depth is $(R_p/R_*)^2 = 1.70 \times 10^{-2}$, so the band deepens it by 8%. Planets with weak gravity and light air are the easiest to read.

(c) **Earth and M dwarf**. $R_* = 0.12 \times 6.96 \times 10^8$ m $= 8.35 \times 10^7$ m, and $2R_pH/R_*^2 = 2\,(6.37 \times 10^6)(8.4 \times 10^3)/(8.35 \times 10^7)^2 = 1.53 \times 10^{-5}$, times 4.6.

$$\boxed{\text{extra depth} \simeq 7 \times 10^{-5}.}$$

The same planet in front of the Sun gives 1.0×10^{-6}. The star's area is $1/0.12^2 = 69$ times smaller, and the signal is that much larger. This is why the search for the air of rocky planets starts with small, cool stars.

Chapter 12

The Equations of a Star

A star is held together by gravity, supported by pressure, heated by nuclear burning and cooled by the light escaping its surface. To describe it, we need to know how its radius, pressure, temperature and luminosity change from the centre outward. Four equations connect those quantities shell by shell. The gas's pressure, opacity and burning rate supply the material properties needed to solve them. Together with conditions at the centre and surface, these are the equations solved by stellar-evolution codes.

What makes a body a star is where its light comes from. A **star** is a ball of gas held together by its own gravity that shines for a long time by nuclear fusion in its core, hydrogen for most of its life and heavier fuels later. Other bodies shine for other reasons. Planets, moons, asteroids and comets reflect the light of a star. Protostars and brown dwarfs shine by slow contraction, the Kelvin–Helmholtz source of Section 10.5, helped for a while by burning deuterium (Eq. (12.20)). White dwarfs shine by cooling, giving away heat they already have. Neutron stars shine by losing spin, by gas falling onto them or by the slow decay of their magnetic field, and black holes only by the gas falling in (Section 15.5).

12.1 The Unknowns

Chapter 10 described a star at each radius r, with $m(r)$ the mass inside. With radius as the coordinate, the temperature is a function of radius and time,

$$T = T(r, t).$$

But the gas inside a real star does not stay at one radius. When the star expands, the gas moves outward, and a fixed r no longer holds the same gas. A thermometer held at fixed r then reads two changes mixed together, the gas there heating or cooling, and hotter or cooler gas sliding past it. So we need a label that stays with the gas.

Think of the rings of a tree. As the tree grows, every ring moves outward, but the ring laid down in a given year stays the same ring. Counting rings from the centre labels the wood, whatever the size of the trunk. Inside a star the mass coordinate m plays the part of the ring count. The shell of gas with mass m inside it keeps that label while the star contracts, expands or burns, because gas never crosses from one shell to the next. Written with this label,

$$T = T(m, t),$$

the temperature follows the same shell of gas as the star evolves. The time derivative at fixed m is then the change that this one shell feels,

$$\left.\frac{\partial T}{\partial t}\right|_m \overset{(1)}{=} \left.\frac{\partial T}{\partial t}\right|_r + v \left.\frac{\partial T}{\partial r}\right|_t , \tag{12.1}$$

(1) Eq. (9.3) in one dimension, with the blob a whole shell and $v = \partial r/\partial t|_m$ the speed at which the shell moves outward.

The first term on the right is the thermometer at fixed r, and the second is the gas sliding past it. Writing everything with m removes the second term from sight. The composition of a shell changes only by its own burning, so everything that happens to a shell happens at fixed m.

The Sun's radius has grown by about 10% since it began to burn hydrogen. Its shell enclosing half the mass has moved outward too, but it is still labelled $m = M_\odot/2$. Following shells in this way is the Lagrangian description of a fluid; watching fixed positions is the Eulerian description. We will use m as our coordinate, from 0 at the centre to M at the surface. At each m we seek four quantities: the radius $r(m)$, pressure $P(m)$, temperature $T(m)$ and luminosity $L(m)$ passing through that shell.

12.2 The Four Equations

We already have equations for hydrostatic balance (Eq. (10.2)) and for the flow of light (Eq. (11.24)). Both use radius as the coordinate. To follow the mass shells introduced above, we must rewrite them in terms of m. We also need to track the radius of each shell and the energy added to the luminosity as we move outward.

Think of a household budget kept month by month. One line records the money coming in. Another records the money going out. The balance at the end of each month follows from the balance before plus the difference. Nothing else is needed, and nothing can be left out. A star keeps four such accounts, shell by shell, for space, momentum, energy and the flow of heat. Each one says how a quantity changes from one shell to the next. Mass and energy are conservation (P1). Momentum is conservation with the gravity of a sphere (P1 and P2). The flow of heat is a random walk of light (P5), or, where that is too slow, the bulk motion of the gas itself.

In the Sun, burning builds nearly all of its 3.83×10^{26} W luminosity within the inner quarter of the radius. The outer shells then pass that energy on without adding much. This is why luminosity must be followed from shell to shell alongside radius, pressure and temperature. The same equations can describe other self-gravitating bodies when we supply the appropriate source of energy and form of pressure.

Mass. Inverting the definition of $m(r)$, Eq. (9.15),

$$\boxed{\frac{\mathrm{d}r}{\mathrm{d}m} = \frac{1}{4\pi r^2 \rho}.} \tag{12.2}$$

Momentum. Hydrostatic equilibrium (10.2), divided by $\mathrm{d}m/\mathrm{d}r = 4\pi r^2\rho$,

$$\begin{aligned} \frac{\mathrm{d}P}{\mathrm{d}m} &\overset{(1)}{=} \frac{\mathrm{d}P}{\mathrm{d}r}\,\frac{\mathrm{d}r}{\mathrm{d}m} \\ &\overset{(2)}{=} -\frac{Gm\rho}{r^2}\,\frac{1}{4\pi r^2\rho} \\ &= -\frac{Gm}{4\pi r^4}. \end{aligned} \tag{12.3}$$

(1) The chain rule.
(2) Eq. (10.2) and Eq. (12.2).

$$\boxed{\frac{\mathrm{d}P}{\mathrm{d}m} = -\frac{Gm}{4\pi r^4}.} \tag{12.4}$$

The density has dropped out. The pressure falls outward at a rate set by gravity alone.

The same balance follows from Newton's second law for the shell between m and $m + dm$, and that route also says what happens when the balance fails. Gravity pulls the shell inward with force $Gm\,dm/r^2$, from the mass m inside it. The pressure $P(m)$ below pushes its inner face, of area $4\pi r^2$, outward, and the pressure $P(m + dm)$ above pushes its outer face inward. The two faces have the same area to first order in dm. Then

$$\begin{aligned} dm\,\frac{\partial^2 r}{\partial t^2} &\overset{(1)}{=} -\frac{Gm\,dm}{r^2} + 4\pi r^2\big[P(m) - P(m+dm)\big] \\ &\overset{(2)}{=} -\frac{Gm\,dm}{r^2} - 4\pi r^2\,\frac{\partial P}{\partial m}\,dm \\ \Longrightarrow \qquad \frac{\partial^2 r}{\partial t^2} &\overset{(3)}{=} -\frac{Gm}{r^2} - 4\pi r^2\,\frac{\partial P}{\partial m}. \end{aligned} \tag{12.5}$$

(1) Mass times acceleration equals the sum of the forces, gravity inward and the net push of the pressure on the two faces.
(2) To first order, $P(m + dm) - P(m) = (\partial P/\partial m)\,dm$.
(3) Divide by dm.

The minus sign in front of $4\pi r^2\,\partial P/\partial m$ says that the pressure pushes the shell toward the side where the pressure is lower. In a star the pressure falls outward, $\partial P/\partial m < 0$, so this term is positive and pushes the shell outward, against gravity. The time derivative is taken at fixed m, so $\partial^2 r/\partial t^2$ is the acceleration of the shell itself (Eq. (12.1)). When the shell does not accelerate, the left side is zero and Eq. (12.5) is Eq. (12.4). When the pressure is too weak to hold the shell, the gravity term wins, the shell falls inward, and the star collapses on the free-fall time of Section 10.5.

Energy. Take the shell between m and $m + dm$. Nuclear burning releases energy in it at the rate $\varepsilon\,dm$, where ε is the energy released per unit mass and time. Light carries $L(m)$ in through its inner face and $L(m + dm)$ out through its outer face. Whatever is left heats the shell. With s the entropy per unit mass, the heat added to the shell per unit time is $T\,(\mathrm{d}s/\mathrm{d}t)\,dm$, so by conservation of energy (P1)

$$\begin{aligned} T\,\frac{\mathrm{d}s}{\mathrm{d}t}\,dm &\overset{(1)}{=} \varepsilon\,dm - \big[L(m+dm) - L(m)\big] \\ &\overset{(2)}{=} \left(\varepsilon - \frac{\mathrm{d}L}{\mathrm{d}m}\right) dm. \end{aligned} \tag{12.6}$$

(1) Heat gained by the shell equals energy released inside it minus the net energy carried out.
(2) To first order, $L(m + dm) - L(m) = (\mathrm{d}L/\mathrm{d}m)\,dm$.

The heat on the left can also be counted with the first law of thermodynamics, for the same shell. The shell has volume $dV = dm/\rho$ and internal energy $u\,dm$, where u is the internal energy per unit mass (not per unit volume as in Eq. (9.26)). Heat added to the shell either raises its internal energy or is spent as work $P\,dV$ when the shell expands and pushes its neighbours away. Per

unit time,

$$\begin{aligned} T\,\frac{\mathrm{d}s}{\mathrm{d}t}\,dm &\overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\big(u\,dm\big) + P\,\frac{\mathrm{d}}{\mathrm{d}t}\big(dV\big) \\ &\overset{(2)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\big(u\,dm\big) + P\,\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{dm}{\rho}\Big) \\ &\overset{(3)}{=} \Big[\frac{\mathrm{d}u}{\mathrm{d}t} + P\,\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{1}{\rho}\Big)\Big]\,dm. \end{aligned} \tag{12.7}$$

(1) The first law for the shell. The heat gained equals the rise of the internal energy plus the work done.
(2) $dV = dm/\rho$.
(3) dm does not change in time, since the shell keeps its gas (Eq. (12.1)), so it comes out of both derivatives.

So $T\ \mathrm{d}s/\mathrm{d}t = \mathrm{d}u/\mathrm{d}t + P\ \mathrm{d}(1/\rho)/\mathrm{d}t$, with every time derivative taken at fixed m. A shell that contracts has $\mathrm{d}(1/\rho)/\mathrm{d}t < 0$, so its neighbours do work on it and heat it.

$$\boxed{\frac{\mathrm{d}L}{\mathrm{d}m} = \varepsilon - T\,\frac{\mathrm{d}s}{\mathrm{d}t}.} \tag{12.8}$$

In a star that is not changing, $\mathrm{d}s/\mathrm{d}t = 0$ and the luminosity grows outward exactly by the burning in each shell. In a star that contracts, as before hydrogen ignites, the entropy of each shell falls and $-T\ \mathrm{d}s/\mathrm{d}t > 0$ supplies luminosity from gravity. This term is the Kelvin–Helmholtz contraction of Section 10.5, written for each shell. (Neutrinos that escape directly add a loss term to ε.)

Heat flow by light. The radiative transport equation (11.24), divided by $4\pi r^2\rho$,

$$\boxed{\frac{\mathrm{d}T}{\mathrm{d}m} = -\frac{3\kappa L}{64\pi^2 a c\, r^4 T^3} \qquad \text{(radiative).}} \tag{12.9}$$

Heat flow by convection. Where light alone would need too steep a fall of temperature, the gas starts to move and carries heat itself. To find where, follow a blob of gas lifted a little, from pressure P to the lower pressure $P + dP$ above it ($dP < 0$). It rises too fast to exchange heat with its surroundings but slowly enough to match their pressure, since sound crosses it quickly (P6). So it expands along an adiabat, $P \propto \rho^\gamma$, and for an ideal gas its temperature changes as

$$\left(\frac{dT}{T}\right)_{\text{blob}} = \frac{\gamma-1}{\gamma}\,\frac{dP}{P}.$$

At the same pressure, a hotter blob is less dense than its surroundings (ideal gas, Eq. (9.28)) and keeps rising. The blob ends up hotter than its surroundings if its temperature falls less than theirs,

$$\begin{aligned} &\frac{\gamma-1}{\gamma}\,\frac{dP}{P} > \left(\frac{\mathrm{d}\ln T}{\mathrm{d}\ln P}\right)_{\text{star}}\frac{dP}{P} \\ \Longrightarrow\quad &\left(\frac{\mathrm{d}\ln T}{\mathrm{d}\ln P}\right)_{\text{star}} \overset{(1)}{>} \frac{\gamma-1}{\gamma}. \end{aligned} \tag{12.10}$$

(1) Divide by dP/P, which is negative, so the inequality turns over.

$$\boxed{\text{convection where} \quad \left(\frac{\mathrm{d}\ln T}{\mathrm{d}\ln P}\right)_{\text{radiative}} > \frac{\gamma-1}{\gamma} \left(= \tfrac{2}{5} \text{ for } \gamma = \tfrac{5}{3}\right).} \tag{12.11}$$

This is the **Schwarzschild criterion**. The left side is computed from Eq. (12.9) and Eq. (12.4). Two things make it large, and so make a region convect. A high opacity needs a steep temperature fall to push the light through, as in the cool outer third of the Sun. A large luminosity from a small mass needs one too, as in the cores of massive stars, whose steeply temperature-sensitive burning (Section 12.3) packs the energy release into a small centre. Once convection starts it is so efficient that it holds the temperature gradient very close to the adiabatic value, so in a convective region

$$\frac{\mathrm{d}\ln T}{\mathrm{d}\ln P} \approx \frac{\gamma-1}{\gamma} \qquad \text{(convective)} \tag{12.12}$$

replaces Eq. (12.9). Codes use mixing-length theory, a model of how far a blob rises before it mixes, to compute the small excess over the adiabatic gradient. The same criterion sets the fall of temperature with height in the Earth's lower atmosphere (Problem 12.8.5).

The system, and the count. Collect the four unknowns in one vector. The four equations then form one first-order system in m,

$$\frac{\mathrm{d}}{\mathrm{d}m}\begin{pmatrix} r \\ P \\ L \\ T \end{pmatrix} = \begin{pmatrix} 1/(4\pi r^2\rho) \\ -Gm/(4\pi r^4) \\ \varepsilon - T\ \mathrm{d}s/\mathrm{d}t \\ -3\kappa L/(64\pi^2 ac\, r^4 T^3) \text{ or Eq. (12.12)} \end{pmatrix}. \tag{12.13}$$

Four unknown functions, four first-order equations. The right side needs three properties of the gas, its density $\rho(P,T)$, its opacity $\kappa(\rho,T)$ and its burning rate $\varepsilon(\rho,T)$, each depending also on the composition at that m. They are worked out in Section 12.3. Four first-order equations need four boundary conditions, found in Section 12.4.

12.3 Pressure, Opacity and Burning

The four structure equations tell us how a star changes from shell to shell, but they need three inputs from the material in each shell. Pressure determines how the gas resists gravity. Opacity determines how readily light crosses it (Section 11.4). The burning rate determines how much energy it adds. All three depend on the gas's composition and state; at the densities of white dwarfs, even the form of the pressure changes.

Think of a crowded lift. When there is room, people move about freely, and the more energetic they are, the harder they push on the walls. When the lift is packed, nobody can move whatever their mood, and the push on the walls comes only from being squeezed. The push no longer depends on how energetic people feel. Electrons in very dense gas are the packed lift. Each quantum state holds at most one electron, so once the low states are full the electrons are forced into states of high momentum, and they push however cold the gas is. Counting states (P3) gives this degeneracy pressure, the opacity follows from counting absorbers (P3) along a random walk of light (P5), and the burning rate from counting fast particles (P3) that tunnel through the electric barrier of the nucleus.

The contrast appears in the ratio of Fermi energy to thermal energy $k_B T$. At the Sun's centre, the Fermi energy is about half as large, so degeneracy is mild. At a white dwarf's centre, with density $10^9\,\mathrm{kg\,m^{-3}}$ and temperature 10^7 K, it is 190 times larger (Problem 12.8.2). Figure 12.1 shows where ordinary gas, radiation and degenerate electrons each dominate the pressure.

We now supply these three material properties to the structure equations, beginning with composition and the pressure it produces.

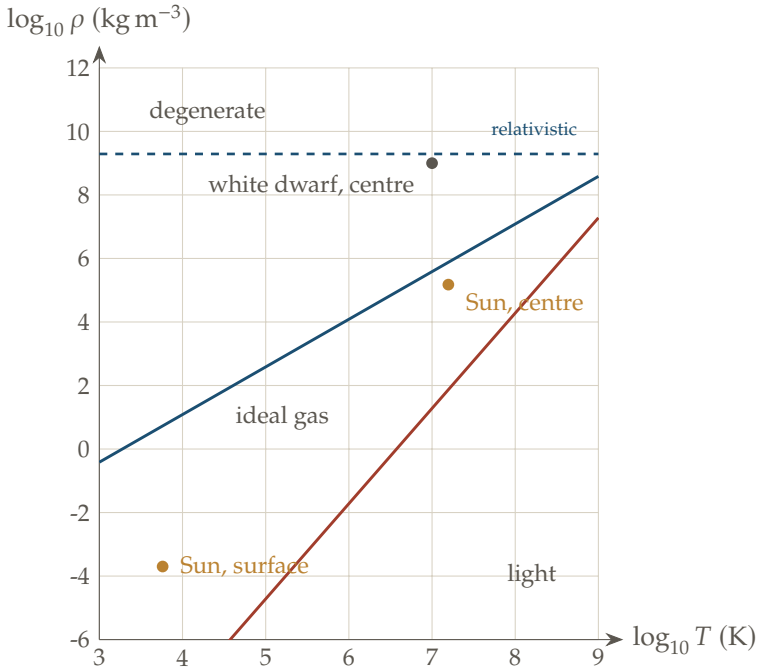


Figure 12.1: Density against temperature, on logarithmic axes. The red line marks equal pressure of light and of an ideal gas ($\mu = 0.62$). The blue line marks a Fermi energy equal to k_BT ($\mu_e = 2$). The dashed line marks the density at which the electrons' Fermi momentum reaches $m_e c$. Points mark the centre and surface of the Sun and the centre of a white dwarf.

Composition and mean molecular weight. Let X, Y and Z be the mass fractions of hydrogen, helium and everything heavier, with $X + Y + Z = 1$. In a fully ionised gas, a mass m_H of hydrogen gives two particles (a proton and an electron), a mass $4m_\mathrm{H}$ of helium gives three (a nucleus and two electrons), and a heavy atom of mass Am_H gives about $A/2$ particles (it has about $A/2$ electrons, and the one nucleus can be neglected). The number of particles per mass m_H is therefore

$$\boxed{\frac{1}{\mu} = 2X + \frac{3}{4}Y + \frac{1}{2}Z, \qquad \frac{1}{\mu_e} = X + \frac{1}{2}Y + \frac{1}{2}Z = \frac{1+X}{2},} \tag{12.14}$$

where μ is the mean molecular weight of Section 10.2, and μ_e counts the mass per electron, so that $n_e = \rho/(\mu_e m_\mathrm{H})$. For the Sun's composition, $X = 0.70$, $Y = 0.28$, $Z = 0.02$, this gives $\mu = 0.617$ and $\mu_e = 1.18$. Pure helium gives $\mu = \frac{4}{3}$ and $\mu_e = 2$, as do carbon and oxygen to good accuracy.

Degenerate electrons. At high density and low temperature every electron state up to some momentum p_F, the **Fermi momentum**, is full and every state above it is empty. With two spin states, Eq. (9.23) gives the electron density

$$n_e = \int_0^{p_F} \frac{2}{h^3}\, 4\pi p^2\, dp = \frac{8\pi}{3h^3}\, p_F^3, \qquad p_F = h\left(\frac{3n_e}{8\pi}\right)^{1/3}. \tag{12.15}$$

The pressure integral (9.25) in the two limits of Section 9.3 gives

$$P_{\text{slow}} \overset{(1)}{=} \frac{1}{3}\int_0^{p_F} p\,\frac{p}{m_e}\,\frac{2}{h^3}4\pi p^2\,dp \overset{(2)}{=} \frac{8\pi p_F^5}{15\,m_e h^3} \overset{(3)}{=} \frac{h^2}{20\,m_e}\left(\frac{3}{\pi}\right)^{2/3} n_e^{5/3},$$
$$P_{\text{fast}} \overset{(4)}{=} \frac{1}{3}\int_0^{p_F} p\,c\,\frac{2}{h^3}4\pi p^2\,dp \overset{(5)}{=} \frac{2\pi c\,p_F^4}{3h^3} \overset{(6)}{=} \frac{hc}{8}\left(\frac{3}{\pi}\right)^{1/3} n_e^{4/3}. \tag{12.16}$$

(1) Slow electrons, $v = p/m_e$.
(2) $\int_0^{p_F} p^4\,dp = p_F^5/5$.
(3) p_F from Eq. (12.15).
(4) Electrons at the speed of light, $v = c$.
(5) $\int_0^{p_F} p^3\,dp = p_F^4/4$.
(6) p_F from Eq. (12.15).

With $n_e = \rho/(\mu_e m_{\text{H}})$ both are power laws of the density alone, $P \propto \rho^{5/3}$ for slow electrons and $P \propto \rho^{4/3}$ for fast ones, the two values of γ from Eq. (9.26). The temperature does not appear. Heating degenerate gas does not raise its pressure, so it does not expand and cool, and the thermostat of Section 10.3 fails. Burning that starts in degenerate gas runs away, as in the helium flash of low-mass giants and in the thermonuclear explosion of a white dwarf. The ions stay classical, because at the same momentum their much larger mass gives them a far smaller kinetic energy.

Still gas pressure. Degeneracy pressure is not a new kind of force. It is the same gas pressure as $P = nk_BT$, the momentum carried across a surface by moving particles, and it comes from the same pressure integral (9.25). Only the occupation number $f(p)$ of the momentum states differs. In a thin gas the states are mostly empty, $f \ll 1$, and the particles get their momenta from heat, so $P = nk_BT$. In a dense gas of fermions every state is full up to p_F, and the particles get their momenta from crowding, because the Pauli principle allows only one in each state. In both cases the pressure is about the density times the energy of one particle. That energy is k_BT in the thin gas and the Fermi energy $E_F = p_F^2/2m_e \propto n_e^{2/3}$ in the dense one, and $P_{\text{slow}} = \frac{2}{5}n_eE_F$. This is where the extra factor $n_e^{2/3}$ of $P \propto n_e^{5/3}$ comes from. Figure 12.2 shows the two occupations, the integrands of the one pressure integral, and the pressure of one gas at a fixed temperature as its density grows, computed with the occupation number that holds at any density, the Fermi–Dirac form. The curve follows nk_BT while $E_F \ll k_BT$, turns near $E_F = k_BT$, where $P = 1.13\,nk_BT$, and follows the degenerate law beyond. Nothing new switches on along the way. The name degenerate marks the state of the gas, with all its low states full, not a different pressure.

When degeneracy matters. Degeneracy matters when the Fermi energy $p_F^2/2m_e$ is comparable with k_BT, or equivalently when the spacing between electrons is comparable with their quantum wavelength (P6). Setting $p_F^2/2m_e = k_BT$ gives the blue line of Figure 12.1, $\rho \propto T^{3/2}$. The electrons move near the speed of light when p_F approaches m_ec, which by Eq. (12.15) with $\mu_e = 2$ happens at $\rho = 2.0 \times 10^9\,\text{kg}\,\text{m}^{-3}$, the dashed line. The pressure of light exceeds that of the gas where $aT^4/3 > \rho k_BT/(\mu m_{\text{H}})$, below the red line, at low density and high temperature.

Opacity. In hot, fully ionised gas the electrons scatter light with the Thomson cross-section, so $\kappa = \sigma_T/(\mu_e m_{\text{H}}) = 0.020(1 + X)\,\text{m}^2\,\text{kg}^{-1}$, independent of density and temperature. At lower temperatures, electrons bound to ions or passing close to them absorb light. Their cross-section falls with frequency roughly as ν^{-3}, the light that carries the flux has $h\nu \sim k_BT$ (the Rosseland weighting of Section 11.4), and the number of absorbers per unit mass grows with the density.

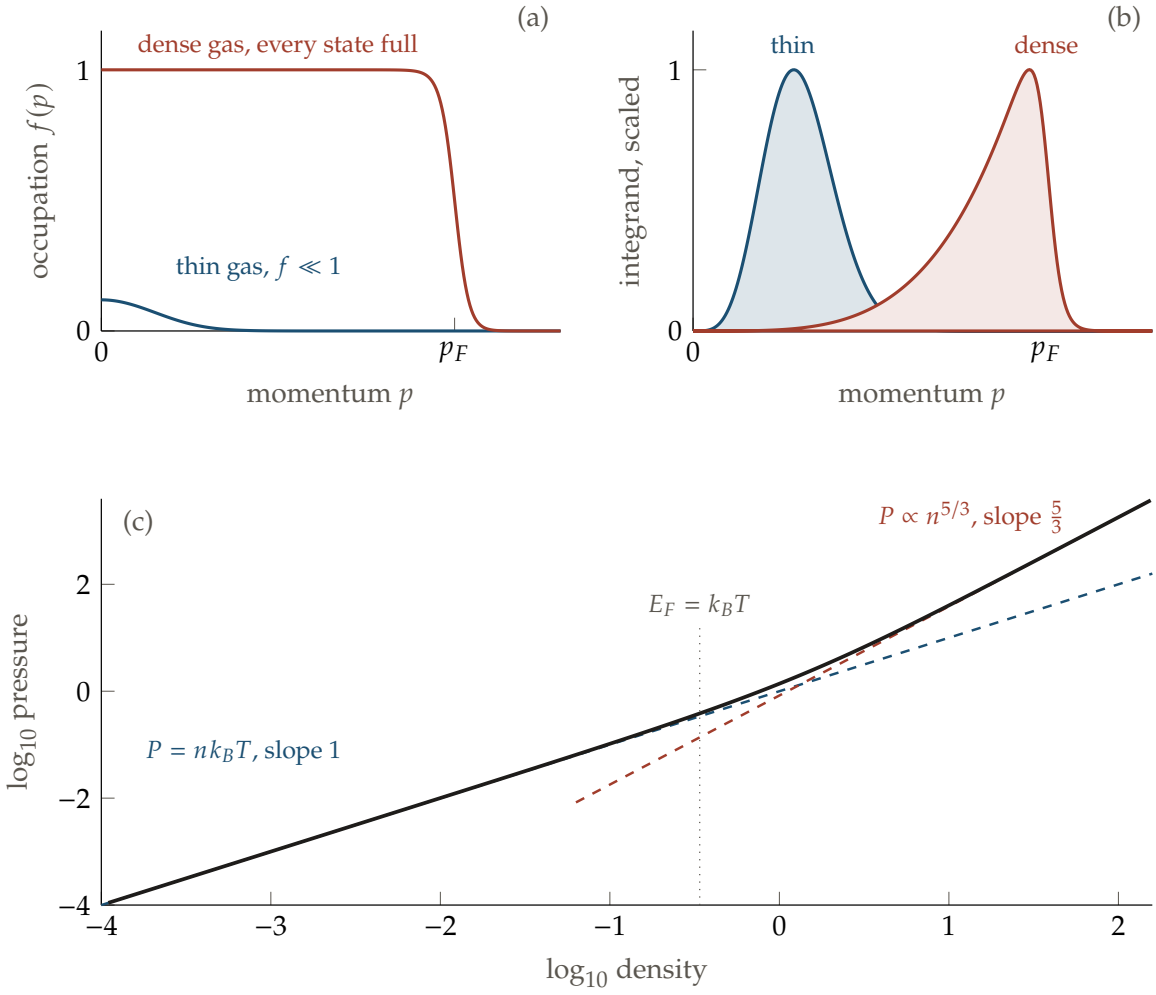


Figure 12.2: A thin and a dense gas of fermions at the same temperature. (a) The occupation number of the momentum states. (b) The integrand $p\,v\,f(p)\,p^2$ of the pressure integral, each scaled to its own peak. The area under each curve is proportional to the pressure. (c) The pressure against the density at fixed temperature, in units of $8\pi(2mk_BT)^{3/2}k_BT/h^3$ and $8\pi(2mk_BT)^{3/2}/h^3$, computed with the Fermi–Dirac occupation (black), with the thin-gas law $P = nk_BT$ and the degenerate law $P = \frac{2}{5}nE_F$ (dashed). The dotted line marks $E_F = k_BT$.

Together these give Kramers' law, $\kappa \propto \rho\,T^{-7/2}$, whose coefficient comes from atomic physics (the free–free part is derived in Eq. (15.18)). Kramers opacity is what makes the outer layers of the Sun opaque enough to convect.

The burning rate and the Gamow peak. Two protons at the Sun's centre, at $k_BT = 1.35\,\mathrm{keV}$, approach each other against an electric barrier of about 1 MeV. Classically they never touch. Quantum mechanically they tunnel, with a probability $e^{-\sqrt{E_G/E}}$ at relative energy E, where the **Gamow energy** $E_G = 2m_rc^2(\pi\alpha Z_1Z_2)^2$ depends on their charges Z_1, Z_2 and their reduced mass m_r, and α is the fine-structure constant. For two protons, $E_G = 493\,\mathrm{keV}$. The number of pairs at energy E falls as the Maxwell tail e^{-E/k_BT} (P3). The reaction rate is dominated by the energies where the product of the two is largest (Figure 12.3). Setting the derivative of its exponent to

zero,

$$0 \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}E}\left(-\frac{E}{k_B T} - \sqrt{\frac{E_G}{E}}\right) \overset{(2)}{=} -\frac{1}{k_B T} + \frac{1}{2}\,\frac{E_G^{1/2}}{E^{3/2}}$$

$$\Longrightarrow \qquad E_0 \overset{(3)}{=} \left(\frac{E_G\,(k_B T)^2}{4}\right)^{1/3}. \tag{12.17}$$

(1) The maximum of $e^{-E/k_B T} e^{-\sqrt{E_G/E}}$ is where its exponent is largest.
(2) $\mathrm{d}/\mathrm{d}E\ E^{-1/2} = -\frac{1}{2}E^{-3/2}$.
(3) Solve for E.

$$\boxed{E_0 = \left(\frac{E_G\,(k_B T)^2}{4}\right)^{1/3}}. \tag{12.18}$$

For the Sun's centre, $E_0 = 6.1\,\mathrm{keV} = 4.5\,k_B T$. Only the rare protons far out on the Maxwell tail react. Figure 12.3(b) shows how rare. At the peak the Maxwell factor is 10^{-2} and the tunnelling factor about 10^{-4}, so their product is only 1.4×10^{-6}. A star holds so many protons that this tiny share is enough, which is why a temperature of millions of kelvin, set by gravity through the virial theorem, is enough to burn.

How steeply burning depends on temperature. Near its peak the product is close to a Gaussian of width proportional to $(E_0 k_B T)^{1/2}$, and its height is $e^{-3E_0/k_B T}$ (the two terms of the exponent at E_0 add to $3E_0/k_B T$). The rate per pair, averaged over the Maxwell distribution with its normalisation $(k_B T)^{-3/2}$, is therefore

$$\text{rate} \propto (k_B T)^{-3/2}\,(E_0 k_B T)^{1/2}\,e^{-3E_0/k_B T} \propto T^{-2/3}\,e^{-3E_0/k_B T},$$

using $E_0 \propto T^{2/3}$. Its logarithmic slope is

$$\frac{\mathrm{d}\ln(\text{rate})}{\mathrm{d}\ln T} \overset{(1)}{=} -\frac{2}{3} - \frac{\mathrm{d}}{\mathrm{d}\ln T}\left(\frac{3E_0}{k_B T}\right)$$

$$\overset{(2)}{=} -\frac{2}{3} + \frac{E_0}{k_B T}. \tag{12.19}$$

(1) The logarithm of the rate above.
(2) $3E_0/k_B T \propto T^{2/3}/T = T^{-1/3}$, so its logarithmic derivative is $-\frac{1}{3} \times 3E_0/k_B T$.

For the proton–proton chain at the Sun's centre this is 3.8, so its rate goes as about T^4. For the CNO cycle, whose slowest step is a proton meeting nitrogen ($Z_2 = 7$, $E_G = 45\,\mathrm{MeV}$), it is about 20. A one percent rise of temperature raises the CNO rate by twenty percent. This is why the thermostat of Section 10.3 holds the temperature of a burning core so tightly, why stars more massive than the Sun, whose hotter cores run on CNO, have convective cores, and why the temperature at which a fuel ignites is so sharply defined. Each fusion of four protons into helium releases 26.7 MeV, 0.7% of their rest mass, the figure used for the nuclear time in Section 10.5.

Two ways to burn hydrogen. The proton–proton chain and the CNO cycle do the same job. Both turn four protons into one helium nucleus. The pp chain builds the helium directly, proton by proton. The CNO cycle uses a carbon nucleus as a catalyst. It adds four protons to it one at a time, with two beta decays on the way through nitrogen and oxygen, and the last step throws out a helium nucleus and gives the carbon back. The start and the end are the same in both,

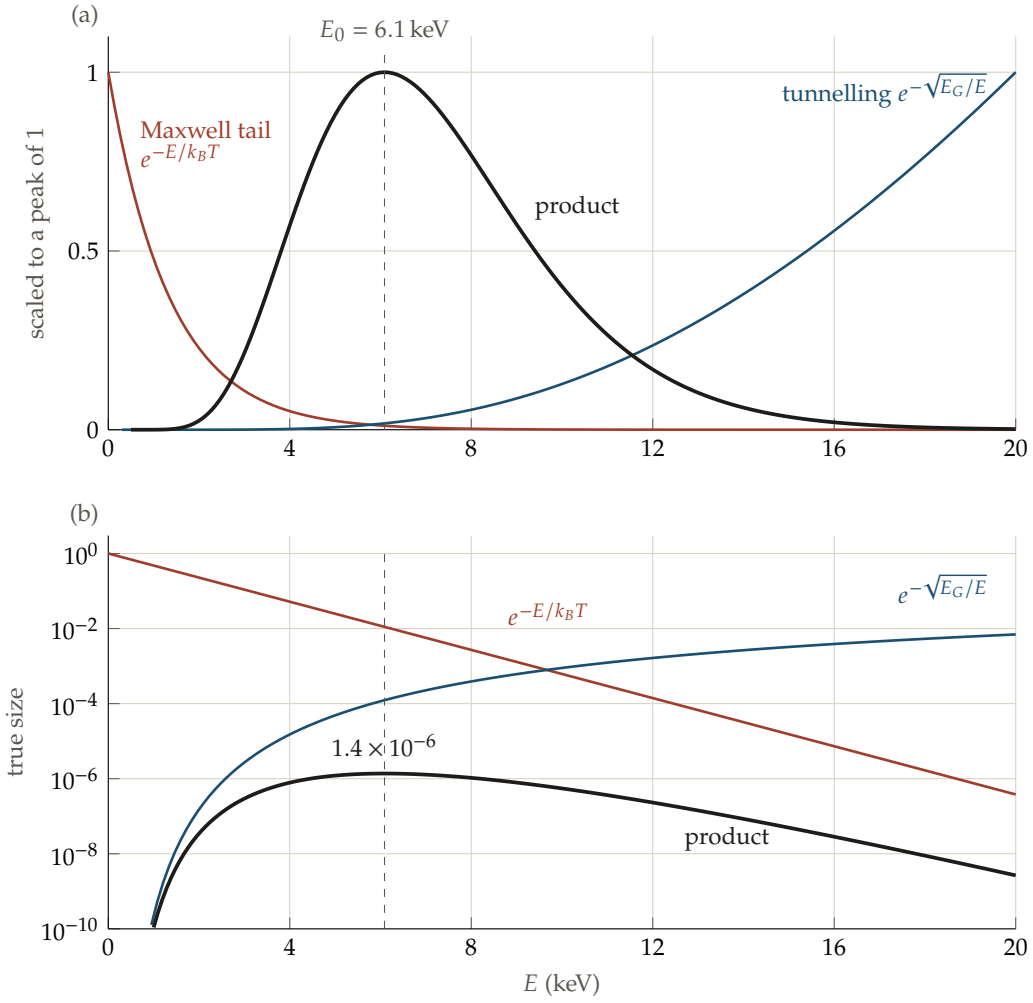


Figure 12.3: Two protons at the Sun's centre ($k_BT = 1.35\,\text{keV}$, $E_G = 493\,\text{keV}$). The Maxwell tail e^{-E/k_BT}, the tunnelling probability $e^{-\sqrt{E_G/E}}$, and their product. (a) Each curve scaled to a maximum of one within the range shown. (b) The same three curves at their true sizes, on a log scale. The dashed line marks the Gamow peak E_0.

four protons and one helium nucleus, so both release the same 26.7 MeV, the difference of their masses. Only the share carried off by neutrinos differs, about 2% in the pp chain and 6% in the CNO cycle, so slightly less of the energy stays in the star as heat when it burns by CNO. What differs greatly is how fast each goes at a given temperature (Figure 12.4).

As a protostar contracts, its centre heats (Section 10.3), and the pp chain switches on first. It needs only protons, and its barrier, with $Z_1Z_2 = 1$, is the lowest there is. The CNO cycle must push a proton into nitrogen, against a barrier with $E_G = 45\,\text{MeV}$ instead of 0.49 MeV, and its catalysts are rare, about 1.4% of the mass. At low temperature it is slower by many powers of ten. But its rate climbs as T^{20} against T^4 for the pp chain, so it catches up. The two curves cross at $17.7 \times 10^6\,\text{K}$. Below that the pp chain makes most of the energy, and above it the CNO cycle takes over. Both rates are proportional to the density, since each step is a meeting of two particles, so the density cancels in their ratio. The crossover depends only on the temperature and on the share of carbon, nitrogen and oxygen. The pressure matters only through the temperature it goes with, which gravity sets through the virial theorem (Eq. (10.13)).

Which cycle a star runs on is decided by where its contraction stops. The centre keeps contracting and heating until burning makes as much energy as the star loses. The Sun stopped

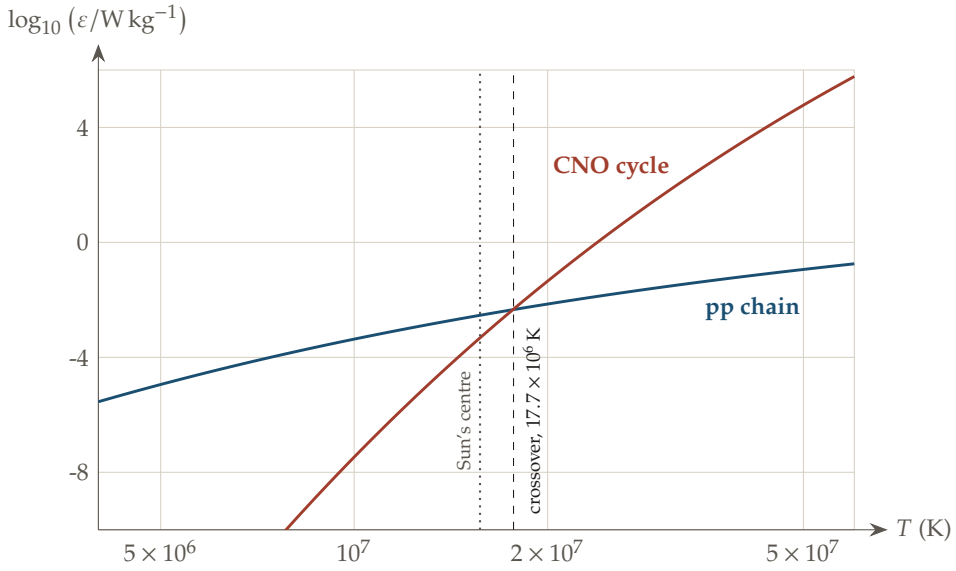


Figure 12.4: Energy released per unit mass and time by the pp chain and the CNO cycle against temperature, for $X = 0.70$, a mass fraction 0.014 of carbon, nitrogen and oxygen, and density $10^5\,\mathrm{kg\,m^{-3}}$. The dashed line marks where the two are equal, and the dotted line the temperature at the Sun's centre.

at 15.7×10^6 K, below the crossover. At that temperature, with the composition of Figure 12.4, the CNO cycle runs at 0.16 of the pp rate, and over the whole core it makes only about 1% of the Sun's light. A heavier star loses energy faster and must reach a hotter centre before its burning catches up. Above about $1.3\,M_\odot$ the centre passes the crossover and the CNO cycle takes over. Because its rate rises so steeply, the burning crowds into the very centre, and the light cannot carry such a concentrated output, which is why these stars have convective cores. The Sun's own centre slowly heats as its hydrogen is used up, so its CNO share grows with age.

Deuterium, the first fuel. Before either way of burning hydrogen starts, a contracting protostar burns the deuterium it was born with,

$$^2\mathrm{H} + {}^1\mathrm{H} \;\longrightarrow\; {}^3\mathrm{He} + \gamma. \tag{12.20}$$

This is the second step of the pp chain. Its barrier is no lower than that of two protons, since both nuclei carry one charge, $Z_1 Z_2 = 1$. With the reduced mass $m_r = \frac{2}{3} m_p$ in place of $\frac{1}{2} m_p$, the Gamow energy is even a little higher, $E_G = \frac{4}{3} \times 493\,\mathrm{keV} = 657\,\mathrm{keV}$. What differs is what happens once the nuclei touch. In the first step of the pp chain, two protons can only stick if one of them turns into a neutron while they touch, which needs the weak force, and this almost never happens. A proton and a deuteron stick and give off a photon, through the electromagnetic force. Per close encounter this is about 5×10^{17} times more likely, the ratio of the two reaction strengths at low energy. So deuterium burns at a far lower temperature, at about 10^6 K, where the tunnelling penalty is harsh. There $k_B T = 0.086\,\mathrm{keV}$, and Eq. (12.18) gives

$$E_0 = \left(\frac{657 \times 0.086^2}{4} \right)^{1/3} \mathrm{keV} = 1.07\,\mathrm{keV}, \qquad e^{-3E_0/k_B T} = e^{-37.2} = 7 \times 10^{-17},$$

against $e^{-13.5} = 1.4 \times 10^{-6}$ for two protons at the Sun's centre. The penalty is 10^{10} times worse, and the factor 5×10^{17} more than pays for it. The rate goes as $T^{E_0/k_B T - 2/3} = T^{12}$, so deuterium ignites sharply, as hydrogen does.

The store is small. There are about 2×10^{-5} deuterium nuclei per hydrogen nucleus, and each reaction releases 5.5 MeV. With $X = 0.70$ this is

$$0.70 \times \frac{2 \times 10^{-5}}{m_{\mathrm{H}}} \times 5.5\,\mathrm{MeV} = 7.4 \times 10^{9}\,\mathrm{J\,kg^{-1}},$$

or 7.3×10^{39} J for a protostar of $0.5\,M_\odot$. That star, at $3\,R_\odot$ and $8\,L_\odot$ (Section 10.7), holds a gravitational energy of order $GM^2/R = 3.2 \times 10^{40}$ J. Deuterium adds about a quarter of that and pays for its light for 8×10^4 years. While it burns, the protostar stops contracting. Then it resumes its slide down the ladder of Section 10.3 until hydrogen ignites.

A body that never gets hot enough for deuterium never burns anything. Detailed models put this line at about 13 Jupiter masses ($0.012\,M_\odot$), and it is the line between a planet and a brown dwarf. A **brown dwarf**, from about 13 to about 80 Jupiter masses, burns its deuterium once and never ignites hydrogen (Section 12.6). A planet burns nothing.

12.4 Boundary Conditions and How Stars Are Computed

The system (12.13) has four unknowns and four first-order equations, now closed by the properties of the gas. Four first-order equations need four conditions to fix a solution, and here they are split between the two ends. Two are known at the centre and two at the surface, and neither end alone fixes the star.

Think of building a bridge from both banks. Each team starts from its own bank with a guess for the height and slope of its half. The halves meet in the middle only if the guesses were right, and the gap between them says how to correct the guesses. A star is computed the same way. Two conditions hold at the centre, two at the surface, and the solution must satisfy all four. Integrating from both ends and closing the gap in the middle, or adjusting the whole profile at once until every equation and every condition holds, gives the star. The method is conservation, applied numerically.

A modern model may divide a star into about a thousand shells and solve for roughly four thousand values at each time step. It repeats that calculation as burning changes the composition. First, though, we need to specify what the solution must satisfy at the centre and surface.

Four boundary conditions. At the centre there is no radius and no luminosity yet,

$$r = 0, \qquad L = 0 \qquad \text{at } m = 0. \tag{12.21}$$

At the surface the pressure has dropped to almost nothing and the temperature is that of the photosphere. With the grey atmosphere of Section 11.8,

$$P \approx 0 \ \text{(more exactly, the pressure at } \tau = \tfrac{2}{3}), \qquad L = 4\pi R^2 \sigma_{\mathrm{SB}} T^4 \qquad \text{at } m = M. \tag{12.22}$$

The second surface condition says that the temperature at $m = M$ is the effective temperature. Two conditions at each end make this a **two-point boundary value problem**.

Shooting. Guess the central pressure P_c and temperature T_c and integrate outward from the centre. Guess the radius R and luminosity L and integrate inward from the surface. Stop both at a fitting mass m_f in between. The four differences between the two solutions there, Δr, ΔP, ΔL and ΔT, are functions of the four guesses. Newton's method corrects all four guesses at once,

$$\begin{pmatrix} \delta P_c \\ \delta T_c \\ \delta R \\ \delta L \end{pmatrix} = -\left[\frac{\partial(\Delta r, \Delta P, \Delta L, \Delta T)}{\partial(P_c, T_c, R, L)} \right]^{-1} \begin{pmatrix} \Delta r \\ \Delta P \\ \Delta L \\ \Delta T \end{pmatrix}, \tag{12.23}$$

where the 4×4 matrix of derivatives is found by repeating the integrations with each guess changed slightly. Near the centre the integration starts from the series $r \approx (3m/4\pi\rho_c)^{1/3}$, since the equations themselves are singular at $r = 0$.

Relaxation. Divide the star into K shells and write the four unknowns at every shell boundary, $4K$ numbers in all. Replace each differential equation by its difference between neighbouring boundaries, $y_{k+1} - y_k = \Delta m\, f\big(\tfrac{1}{2}(y_k + y_{k+1})\big)$ for the vector y and right side f of Eq. (12.13). This gives $4(K-1)$ equations, and the four boundary conditions make $4K$. Starting from a guessed profile, Newton's method corrects all $4K$ numbers together. Each equation involves only two neighbouring shells, so the matrix of derivatives is zero except for small blocks along its diagonal, and it can be solved in a number of steps proportional to K. This is the Henyey method, used by modern codes such as MESA.

Evolution in time. Burning changes the composition at each m, slowly, at rates set by ε. A star is evolved by alternating two steps. Advance the composition of every shell over a time step, then solve the four equations again for a new model in equilibrium. The ordering of the three clocks of Section 10.5 is what allows this, since the star regains hydrostatic equilibrium far faster than its composition changes. Given its mass and the composition at every m, the four equations with their boundary conditions fix the structure of a star. This is the Vogt–Russell statement, true for ordinary stars, though some sets of conditions admit more than one solution.

12.5 Polytropes and the Isothermal Sphere

The full structure equations usually require a computer. To see what controls a star's shape, we can set aside its luminosity and use only the mass and hydrostatic equations. These two equations become a closed problem when pressure is specified as a function of density.

Think of a soft pillow and a firm one, each stacked with books. The firmness of the filling, how much its push grows when squeezed, decides how much the pillow flattens under a given weight. Nothing else about the pillow matters. In a star the same is true when $P = \mathcal{K}\rho^{\gamma}$ for some constant $\mathcal{K}$. Then the stiffness γ alone decides how the star's mass is spread out, and the mass equation with hydrostatic equilibrium (P1 and P2), closed by a pressure from counting states (P3), give one equation for the density. This happens for degenerate gas, where Eq. (12.16) has exactly this form, and for gas that is convective throughout, where $P \propto \rho^{\gamma}$ along the adiabat.

For example, a $\gamma = \frac{4}{3}$ model scaled to the Sun's mass and radius gives central density $7.64 \times 10^{4}\,\mathrm{kg\,m^{-3}}$, pressure $1.25 \times 10^{16}\,\mathrm{Pa}$ and temperature $1.22 \times 10^{7}\,\mathrm{K}$. Detailed solar models give $1.5 \times 10^{5}\,\mathrm{kg\,m^{-3}}$, $2.5 \times 10^{16}\,\mathrm{Pa}$ and $1.57 \times 10^{7}\,\mathrm{K}$. Figure 12.5 shows how changing γ changes the density profile: stiffer gas spreads mass more evenly, while softer gas concentrates it toward the centre.

The same models describe white dwarfs and the cores of giant planets, fully convective low-mass stars, and, in the limit $\gamma \to 1$, the isothermal sphere that models globular clusters and the dark matter halos of galaxies. They give one equation for the density, its solutions, the mass–radius relation, and two classic applications.

The Lane–Emden equation. Divide hydrostatic equilibrium (10.2) by ρ, multiply by r^2 and differentiate. With $\mathrm{d}m/\mathrm{d}r = 4\pi r^2 \rho$,

$$\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r^2}{\rho}\,\frac{\mathrm{d}P}{\mathrm{d}r}\right) = -4\pi G\rho. \tag{12.24}$$

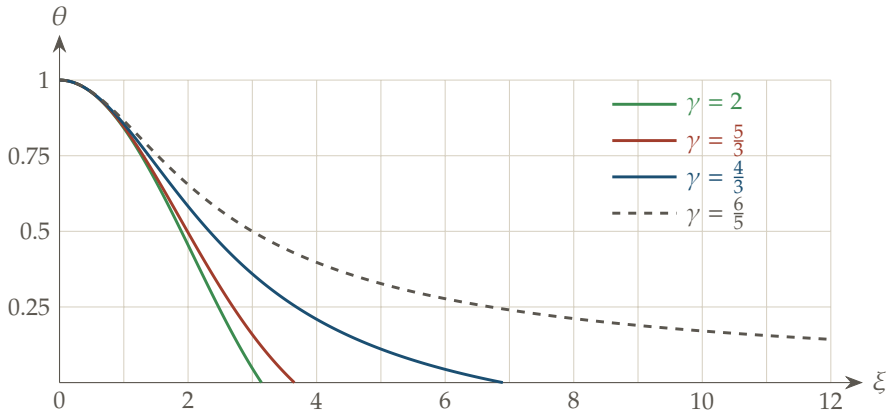


Figure 12.5: Solutions $\theta(\xi)$ of the Lane–Emden equation for $\gamma = 2$, $\frac{5}{3}$, $\frac{4}{3}$ and $\frac{6}{5}$. The density is $\rho_c \theta^{1/(\gamma-1)}$.

Write the density as $\rho = \rho_c\, \theta^{1/(\gamma-1)}$, with ρ_c the central density and $\theta = 1$ at the centre. Then $P = \mathcal{K}\rho_c^{\gamma}\theta^{\gamma/(\gamma-1)}$, and

$$\frac{1}{\rho}\frac{\mathrm{d}P}{\mathrm{d}r} \overset{(1)}{=} \frac{1}{\rho_c\theta^{1/(\gamma-1)}}\,\mathcal{K}\rho_c^{\gamma}\,\frac{\gamma}{\gamma-1}\,\theta^{1/(\gamma-1)}\,\frac{\mathrm{d}\theta}{\mathrm{d}r} = \frac{\gamma\,\mathcal{K}\rho_c^{\gamma-1}}{\gamma-1}\,\frac{\mathrm{d}\theta}{\mathrm{d}r}. \tag{12.25}$$

(1) The chain rule on $\theta^{\gamma/(\gamma-1)}$, whose exponent minus one is $1/(\gamma-1)$.

Putting this into Eq. (12.24) and measuring radius in the unit

$$r_0^2 = \frac{\gamma\,\mathcal{K}\,\rho_c^{\gamma-2}}{4\pi G(\gamma-1)}, \qquad r = r_0\,\xi, \tag{12.26}$$

all the constants collect into r_0 and

$$\boxed{\frac{1}{\xi^2}\frac{\mathrm{d}}{\mathrm{d}\xi}\left(\xi^2\frac{\mathrm{d}\theta}{\mathrm{d}\xi}\right) = -\theta^{1/(\gamma-1)}, \qquad \theta(0) = 1, \quad \theta'(0) = 0.} \tag{12.27}$$

This is the **Lane–Emden equation**. The second condition says that the density is smooth at the centre. The surface of the star is the first zero ξ_1 of θ, where density and pressure vanish.

Solutions. Three values of γ give solutions in closed form, which can be checked by putting them into Eq. (12.27).

- $\gamma \to \infty$, an incompressible star of uniform density, $\theta = 1 - \xi^2/6$, $\xi_1 = \sqrt{6}$.
- $\gamma = 2$, $\theta = \sin\xi/\xi$, $\xi_1 = \pi$.
- $\gamma = \frac{6}{5}$, $\theta = (1 + \xi^2/3)^{-1/2}$, which never reaches zero. The radius is infinite, yet the mass is finite.

The two cases that matter most are solved numerically. For $\gamma = \frac{5}{3}$, $\xi_1 = 3.6538$ and the central density is 5.99 times the mean density. For $\gamma = \frac{4}{3}$, $\xi_1 = 6.8968$ and the central density is 54.18 times the mean.

Mass and radius. The radius is $R = r_0\xi_1$. The mass follows from Eq. (12.27) itself,

$$M \overset{(1)}{=} \int_0^R 4\pi r^2 \rho \, dr \overset{(2)}{=} 4\pi r_0^3 \rho_c \int_0^{\xi_1} \xi^2 \theta^{1/(\gamma-1)} \, d\xi$$
$$\overset{(3)}{=} 4\pi r_0^3 \rho_c \Big[-\xi^2 \frac{d\theta}{d\xi} \Big]_{\xi_1}. \tag{12.28}$$

(1) The mass equation.
(2) $r = r_0\xi$ and $\rho = \rho_c \theta^{1/(\gamma-1)}$.
(3) Eq. (12.27) says $\xi^2\theta^{1/(\gamma-1)} = -\,\mathrm{d}/\mathrm{d}\xi\,(\xi^2\,\mathrm{d}\theta/\mathrm{d}\xi\,)$, and the term vanishes at $\xi = 0$.

For a fixed $\mathcal{K}$, Eq. (12.26) gives $r_0 \propto \rho_c^{(\gamma-2)/2}$, so $R \propto \rho_c^{(\gamma-2)/2}$ and $M \propto r_0^3\rho_c \propto \rho_c^{(3\gamma-4)/2}$. Eliminating ρ_c,

$$\boxed{M \propto R^{(3\gamma-4)/(\gamma-2)}.} \tag{12.29}$$

For $\gamma = \frac{5}{3}$, $M \propto R^{-3}$. A heavier star of this kind is smaller, as white dwarfs are (Section 13.2). For $\gamma = \frac{4}{3}$ the exponent vanishes. Only one mass is possible, whatever the radius. The factor $3\gamma - 4$ is the one that made a star barely bound in Eq. (10.16).

The isothermal sphere. As $\gamma \to 1$ the gas has one temperature and $P = c_s^2\rho$. Writing $\rho = \rho_c e^{-\psi}$ and $r_0^2 = c_s^2/(4\pi G\rho_c)$, Eq. (12.24) becomes

$$\frac{1}{\xi^2}\frac{d}{d\xi}\left(\xi^2 \frac{d\psi}{d\xi}\right) = e^{-\psi}, \qquad \psi(0) = 0, \quad \psi'(0) = 0. \tag{12.30}$$

The density never reaches zero, so an isothermal sphere has no surface. Far out it approaches a power law. Trying $\rho = A/r^2$ in hydrostatic equilibrium, with $m = 4\pi A r$,

$$c_s^2 \frac{d}{dr}\left(\frac{A}{r^2}\right) = -\frac{G\,(4\pi A r)}{r^2}\,\frac{A}{r^2}$$
$$\Longrightarrow \quad -\frac{2c_s^2 A}{r^3} = -\frac{4\pi G A^2}{r^3} \quad \Longrightarrow \quad A = \frac{c_s^2}{2\pi G}. \tag{12.31}$$

So $\rho = c_s^2/(2\pi G r^2)$, which is $\rho/\rho_c = 2/\xi^2$. The numerical solution approaches this law (Figure 12.6). At $\xi = 10^4$ its density is 0.991 of $2\rho_c/\xi^2$. The mass inside r grows as $m = 2c_s^2 r/G$, so a body circling at radius r has the speed

$$v_{\text{circ}}^2 = \frac{Gm}{r} = 2c_s^2, \tag{12.32}$$

the same at every radius. Rotation curves of spiral galaxies are observed to be flat far beyond their visible stars. An isothermal halo of unseen matter is the simplest model that gives this, which is part of the case for dark matter.

Eddington's standard model, and the most massive stars. Divide the momentum equation for light (Route B of Section 11.4, $\mathrm{d}P_{\text{light}}/\mathrm{d}r = -\kappa\rho L/(4\pi r^2 c)$) by hydrostatic equilibrium (10.2),

$$\frac{dP_{\text{light}}}{dP} = \frac{\kappa L(r)}{4\pi c G\, m(r)}. \tag{12.33}$$

Eddington assumed that $\kappa L(r)/m(r)$ is about the same throughout the star. Then P_{light} is a fixed fraction of the total pressure everywhere. Call the gas fraction β, so $P_{\text{gas}} = \beta P$ and $P_{\text{light}} = (1-\beta)P$, with

$$1 - \beta = \frac{\kappa L}{4\pi c G M} = \frac{L}{L_{\text{Edd}}}, \tag{12.34}$$

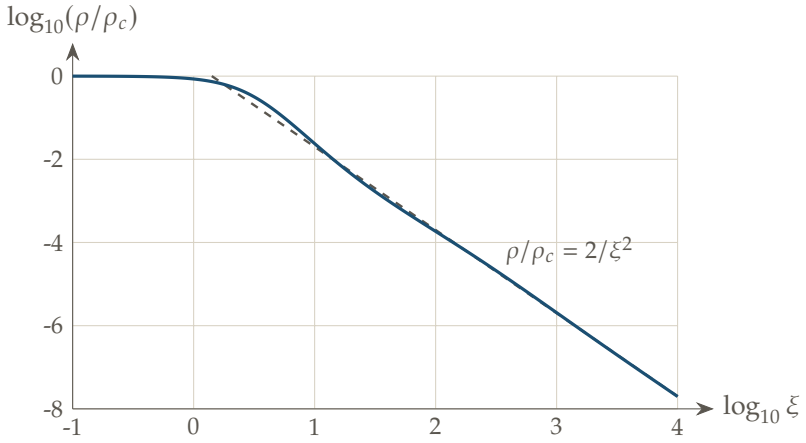


Figure 12.6: The density of the isothermal sphere against radius, on logarithmic axes (solid), and the power law $\rho/\rho_c = 2/\xi^2$ (dashed).

by Eq. (11.27). Eliminating T between $P_{\text{gas}} = \rho k_B T/(\mu m_{\text{H}})$ and $P_{\text{light}} = aT^4/3$,

$$
\begin{aligned}
T &\overset{(1)}{=} \frac{\mu m_{\text{H}}\,\beta P}{k_B \rho} \\
\Longrightarrow \quad (1-\beta)P &\overset{(2)}{=} \frac{a}{3}\left(\frac{\mu m_{\text{H}}\,\beta P}{k_B \rho}\right)^4 \\
\Longrightarrow \quad P &= \left[\frac{3(1-\beta)}{a\,\beta^4}\left(\frac{k_B}{\mu m_{\text{H}}}\right)^4\right]^{1/3} \rho^{4/3}.
\end{aligned}
\tag{12.35}
$$

(1) The ideal-gas law for the gas share of the pressure.
(2) The pressure of light, Eq. (9.30).

So the star is a polytrope with $\gamma = \frac{4}{3}$, whose single mass is fixed by $\mathcal{K}$, and so by β. Turned around, the mass fixes β through Eddington's quartic equation. The script gives $\beta = 0.9995$ for the Sun, 0.961 at $10\,M_\odot$, 0.558 at $100\,M_\odot$ and 0.204 at $1000\,M_\odot$. In massive stars light carries a growing share of the pressure, Eq. (12.34) puts them close to the Eddington luminosity, and the effective γ of their gas and light together falls toward $\frac{4}{3}$, where Eq. (10.16) leaves them barely bound. Stars much above about a hundred solar masses blow off their outer layers or become unstable, which sets the upper end of the masses of stars. The profiles of the solar model are drawn in Figure 12.7.

12.6 Scaling Laws and the HR Diagram

Polytropes show how pressure shapes a star, but leave out the energy flowing through it. For luminosity we return to the full equations. We can first replace each quantity by a typical value to see how luminosity changes with mass, without solving a separate model for every star. The same estimates help explain stellar lifetimes and the main sequence in the HR diagram.

Think of model ships. A model built at half size, with everything in proportion, behaves in a way that can be predicted from the scale alone, without testing each new size from scratch. Stars of different mass are close to scaled copies of each other in the same way. If every quantity in the four equations is replaced by its typical size, each equation becomes a relation between powers of M and R. This is the comparison of scales (P6) applied to the equations themselves.

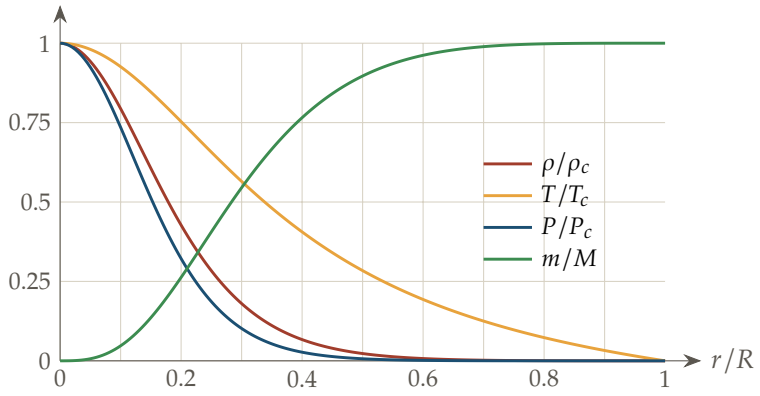


Figure 12.7: Eddington's standard model ($\gamma = \frac{4}{3}$) scaled to the Sun. Density, temperature, pressure and enclosed mass against radius, each relative to its central or total value.

With constant opacity, the estimate below gives $L \propto M^3$; near the Sun, the observed relation is closer to $L \propto M^{3.5}$. A more massive star therefore has more fuel but spends it much faster, giving a lifetime that falls roughly as $M^{-2.5}$. A ten-solar-mass star lives about 3×10^7 years, compared with 10^{10} years for the Sun (Problem 12.8.4). Figure 12.8 shows how luminosity and temperature place these stars on the main sequence.

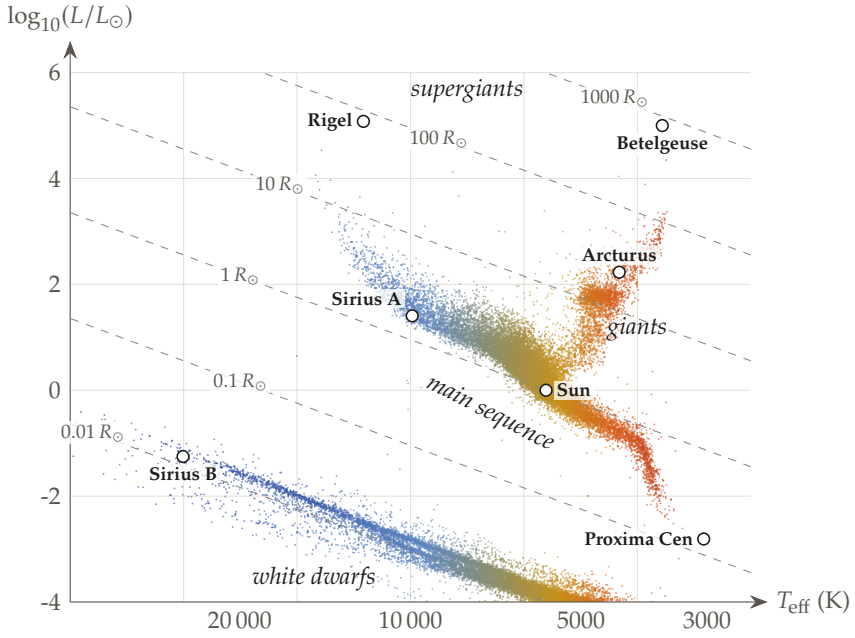


Figure 12.8: Luminosity against effective temperature, with temperature increasing to the left, for 20 298 Hipparcos stars with parallaxes better than 10% (all within 220 pc) and 11 990 Gaia white dwarfs within 100 pc, coloured by temperature. Dashed lines join points of equal radius, labelled in solar radii. Circles mark seven well-known stars.

Homology. Replace $\mathrm{d}/\mathrm{d}m$ by $1/M$, r by R and every other quantity by its typical size. The momentum equation (12.4) and the ideal gas give

$$P \sim \frac{GM^2}{R^4}, \qquad T \sim \frac{\mu m_{\mathrm{H}} P}{k_B \rho} \sim \frac{\mu m_{\mathrm{H}}}{k_B} \frac{GM^2/R^4}{M/R^3} = \frac{\mu m_{\mathrm{H}}\, GM}{k_B R}, \tag{12.36}$$

which is the virial temperature again. The radiative equation (12.9) gives

$$\begin{aligned} L &\overset{(1)}{=} \frac{64\pi^2 ac\, R^4 T^4}{3\kappa M} \\ &\overset{(2)}{=} \frac{64\pi^2 ac\, R^4}{3\kappa M}\left(\frac{\mu m_{\mathrm{H}}\, GM}{k_B R}\right)^4 \\ &\propto \frac{\mu^4 M^3}{\kappa}. \end{aligned} \tag{12.37}$$

(1) $\mathrm{d}T/\mathrm{d}m \sim -T/M$ in Eq. (12.9), solved for L.
(2) The temperature above.

$$\boxed{L \propto \frac{\mu^4 M^3}{\kappa}.} \tag{12.38}$$

The radius has dropped out, and so has the burning. A star's luminosity is set by how fast light can leak out of it, not by how fast it burns. The burning adjusts, through the thermostat, to supply what leaks out. With electron-scattering opacity, as in massive stars, $L \propto M^3$. Kramers opacity, which grows at the lower temperatures of smaller stars, steepens this, and near the Sun $L \propto M^{3.5}$ is observed. The fuel scales as M, so the lifetime scales as $M/L \propto M^{-2.5}$.

The smallest stars. A contracting ball of gas heats up (Section 10.3), but as its density rises its electrons become degenerate and start to push without heating. Add the electrons' Fermi energy to the virial balance of a uniform sphere of ionised hydrogen, with $N = M/m_{\mathrm{H}}$ protons and as many electrons. The protons carry the thermal energy, $\frac{3}{2}Nk_B\bar{T}$, and the electrons their degenerate energy, $\frac{3}{5}NE_F$ with $E_F = (\hbar^2/2m_e)(9\pi N/4R^3)^{2/3}$. The virial theorem, $2\times$ kinetic $= |U| = \frac{3}{5}GM^2/R$, gives

$$k_B\bar{T} = \frac{GMm_{\mathrm{H}}}{5R} - \frac{2}{5}E_F = \frac{\mathcal{A}}{R} - \frac{\mathcal{B}}{R^2}, \tag{12.39}$$

with $\mathcal{A} = GMm_{\mathrm{H}}/5$ and $\mathcal{B} = (\hbar^2/5m_e)(9\pi/4)^{2/3}N^{2/3}$. As R shrinks the temperature first rises and then falls, with a maximum at $R = 2\mathcal{B}/\mathcal{A}$,

$$k_B\bar{T}_{\max} = \frac{\mathcal{A}^2}{4\mathcal{B}} \propto M^{4/3}. \tag{12.40}$$

Hydrogen burns fast enough to stop the contraction only above a few million kelvin. Setting $\bar{T}_{\max} = 3\times10^6\,\mathrm{K}$ gives $M \approx 0.16\,M_\odot$. The centre of a star is about twice as hot as its mean, which lowers this to near the $0.08\,M_\odot$ found by detailed models. Lighter bodies never ignite hydrogen. They are brown dwarfs, and they cool forever at a radius near the peak, about 1.5 times Jupiter's. Those above about 13 Jupiter masses burn their deuterium on the way (Eq. (12.20)).

The range of stars. Table 12.1 lists how far the properties of stars spread. Below $0.08\,M_\odot$ degeneracy stops the heating before hydrogen ignites, by the virial balance with degenerate electrons. Above about $150\,M_\odot$ the luminosity nears the Eddington limit (11.27), light carries most of the pressure (Eq. (12.34)), and the star blows off its outer layers. The other ranges follow from the mass. The luminosity runs from the $M^{3.5}$ law at the bottom to the Eddington luminosity at the top, $4\pi GMc/\kappa = 5.8\times10^6\,L_\odot$ for $150\,M_\odot$ with $\kappa = 0.034\,\mathrm{m^2\,kg^{-1}}$. The lifetime has a floor for the same reason. A star near the Eddington limit has $L \propto M$, so the nuclear time $0.007\times0.1\times Mc^2/L$ of Section 10.5 stops falling with mass,

$$0.007\times0.1\times\frac{Mc^2}{4\pi GMc/\kappa} = \frac{0.0007\,c\,\kappa}{4\pi G} = 2.7\times10^5\ \text{years},$$

whatever the mass. The most massive stars burn a larger share of their hydrogen and shine somewhat below the limit, so they live a few million years. The $M^{-2.5}$ law of Problem 12.8.4 misses this floor and gives too short a life above a few tens of solar masses.

Property	smallest	largest
Mass	$0.08\,M_\odot$	about $150\,M_\odot$
Radius	$0.1\,R_\odot$	$1700\,R_\odot$
Surface temperature	2600 K	about 40 000 K
Luminosity	$3\times10^{-4}\,L_\odot$	10^6 to $10^7\,L_\odot$
Lifetime	a few 10^6 years	over 10^{12} years

Table 12.1: Ranges of the masses, radii, surface temperatures, luminosities and main-sequence lifetimes of stars.

The HR diagram. The Stefan–Boltzmann law, $L = 4\pi R^2 \sigma_{\mathrm{SB}} T_{\mathrm{eff}}^4$ (Eq. (4.26)), gives

$$\log L = 2\log R + 4\log T_{\mathrm{eff}} + \text{constant},$$

so in a diagram of $\log L$ against $\log T_{\mathrm{eff}}$, stars of equal radius lie on straight lines (Figure 12.8). Stars burning hydrogen in their cores form the main sequence, running from cool dim dwarfs like Proxima Centauri to hot bright stars like Sirius A, ordered by mass through Eq. (12.38). Giants and supergiants like Arcturus and Betelgeuse lie above it at large radius, and white dwarfs like Sirius B below it at about a hundredth of the Sun's radius.

Each point in the figure is a real star. Its distance comes from its parallax (Section 3.1), and its brightness and distance give its luminosity. Its colour gives its temperature, and a small correction adds the light that falls outside the visual band. For the white dwarfs the catalogue gives the temperature, the surface gravity g and the mass, so the radius is $R = \sqrt{GM/g}$ and the luminosity follows from Stefan–Boltzmann. The stars do not fill the diagram. Most sit on the main sequence, because a star spends about nine tenths of its life burning hydrogen in its core. The giant branch is thinner, since the giant stage is short. Hot bright stars are rare among the points, because massive stars are few and short-lived, and few lie within 220 pc. The dimmest red dwarfs, such as Proxima Centauri, were too faint for Hipparcos to measure well, so the main sequence stops above it in the figure.

A star leaves the main sequence when its core runs out of hydrogen. The core, now helium, has no burning to support it. It loses energy and, by Section 10.3, contracts and heats. Hydrogen burns on in a shell around it, the shell's output rises as the core shrinks and heats, and the extra luminosity swells the envelope to a hundred times its former radius. The star moves to the upper right of the diagram and becomes a red giant. When the core reaches about 10^8 K, helium ignites.

The age of a cluster. The stars of a cluster formed from one cloud at about the same time, so they have one age and one composition and differ only in mass. The heavier a star, the sooner it leaves the main sequence. After a time t, every star whose main-sequence lifetime is shorter than t has left, and the main sequence of the cluster ends at a **turn-off**, where the lifetime equals the age. The luminosity at the turn-off is what is measured, so we write the age in terms of it.

With the lifetime $10^{10}\,\mathrm{yr}\,(M/M_\odot)^{-2.5}$ (Eq. (12.38) and Problem 12.8.4),

$$t \overset{(1)}{=} 10^{10}\,\mathrm{yr}\left(\frac{M}{M_\odot}\right)^{-2.5} \overset{(2)}{=} 10^{10}\,\mathrm{yr}\left[\left(\frac{L}{L_\odot}\right)^{1/3.5}\right]^{-2.5} = 10^{10}\,\mathrm{yr}\left(\frac{L}{L_\odot}\right)^{-5/7}. \tag{12.41}$$

(1) The turn-off star has just used up its core hydrogen, so its lifetime is the age.
(2) $L \propto M^{3.5}$ near the Sun, inverted to $M/M_\odot = (L/L_\odot)^{1/3.5}$.

$$\boxed{t \approx 10^{10}\,\mathrm{yr}\left(\frac{L_{\text{turn-off}}}{L_\odot}\right)^{-5/7}.} \tag{12.42}$$

A cluster whose main sequence ends at $10\,L_\odot$ has its turn-off at $10^{1/3.5} = 1.9\,M_\odot$ and is $10^{10} \times 10^{-5/7} = 1.9 \times 10^9$ years old. The distance needed to turn the turn-off brightness into a luminosity comes from the same diagram, by sliding the cluster's main sequence onto that of nearby stars (Chapter 3). The oldest clusters, the globular clusters of the Galaxy's halo, turn off just below the Sun's mass. There the $M^{3.5}$ law is too rough, and detailed models of the turn-off give ages of about 1.2×10^{10} years, a lower limit on the age of the universe. As a worked example, the turn-off of M67 in Figure 12.9 is at $M_G = 3.0$. With M_G of the Sun taken as 4.67 and no bolometric correction, this is $4.7\,L_\odot$, and Eq. (12.42) gives 3.3×10^9 years, in rough agreement with the four billion years usually quoted for M67.

Blue stragglers. Many old clusters hold a few stars on the main sequence above the turn-off. By Eq. (12.42) such stars should have left long ago, so they did not start with their present mass. Each gained mass late, either by taking gas from a companion in a close binary or by merging with another star in the crowded centre of the cluster. Fresh hydrogen is mixed into its core, and it burns on as a heavier, younger-looking star. These are the **blue stragglers**, named because they are bluer than the turn-off and seem to lag behind the evolution of the other stars.

In a colour–magnitude diagram a blue straggler is a star that lies on the line of the main sequence but above the turn-off, so it is brighter and bluer than the turn-off stars. Figure 12.9 shows M67, with the main sequence of a young cluster (the Pleiades) as the dashed line above the turn-off. A box picks the stragglers out. They are members brighter than the turn-off by more than 0.5 magnitudes and bluer than the turn-off colour, but not more than 2.5 magnitudes above it, which leaves out the subgiant branch and the red clump. The box finds 5 stars. Their mean M_G is 1.4, against 3.0 at the turn-off, so each is on average 4.4 times as luminous as a turn-off star. Eq. (12.38), $L \propto M^{3.5}$, turns this into mass. The turn-off stars have $4.7\,L_\odot$ and so $1.6\,M_\odot$, and the stragglers have $2.4\,M_\odot$. A star of that mass lives about 1.2×10^9 years by the lifetime $10^{10}\,\mathrm{yr}\,(M/M_\odot)^{-2.5}$, much less than the 3.3×10^9 years of the cluster. Stars this massive should be gone, and they are there only because they gained mass.

12.7 Pulsating Stars

Hydrostatic balance does not mean a star cannot move. If its surface is displaced, gravity and pressure can make it oscillate on roughly the free-fall timescale (Section 10.5). Such motion should die away unless some process replaces the energy lost each cycle. In δ Cephei, the surface moves in and out while the star brightens and fades every 5.37 days. What keeps the motion going?

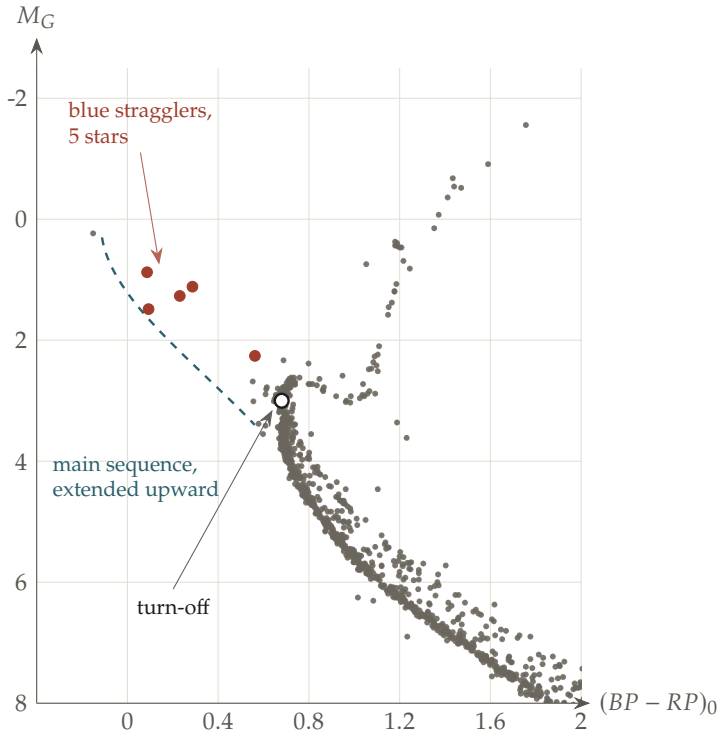


Figure 12.9: Absolute magnitude M_G against colour $(BP - RP)_0$ for the 1040 Gaia DR3 members of M67 (grey), with the 5 selected blue stragglers in red and the turn-off marked by the circle. The dashed line is the main sequence of the Pleiades, a cubic fit, continued above the turn-off.

Think of the engine of a car. The fuel burns when the piston is at the top and the gas is squeezed hardest. The hot gas then pushes harder on the way out than it resisted on the way in, and each cycle leaves some work over to turn the wheels. If the cylinder instead lost heat at the top, through a leak to cold walls, the gas would push less on the way out than on the way in, and the piston would stop. The layers of a star move in and out like the piston. In most layers a squeeze heats the gas and lowers its opacity, so heat leaks out at the top of the cycle and the motion dies away. In a layer where helium is partly ionised, the squeeze goes into ionising helium instead of heating it, the opacity rises, and the layer holds back the light at the top of the cycle. That layer is the engine. It is called the κ mechanism, after the symbol for opacity.

In numbers, Kramers' law gives $\kappa \propto \rho T^{-7/2}$. In gas that heats as $P \propto \rho^{5/3}$ when squeezed, a ten per cent rise of density lowers the opacity by twelve per cent. In a layer where ionisation holds the gas to $P \propto \rho^{1.1}$, the same squeeze raises the opacity by six per cent. The picture is Figure 12.10, the cycle of the layer as a loop of four stages. The period comes out as $(G\bar{\rho})^{-1/2}$ times a number, where $\bar{\rho}$ is the mean density. It is the same clock as free fall and as the Jeans time, because in a ball held by its own gravity every mechanical time is set by $G\bar{\rho}$ (P6). The engine is conservation of energy over a closed cycle (P1), the work $\oint P\,dV$ of any heat engine. The same valve drives the RR Lyrae stars of old clusters and the δ Scuti stars near the main sequence. We find the period, the valve, the growth of the swing, the band of the HR diagram where the valve works, and why brighter Cepheids pulsate more slowly.

The period. To find how fast a star swings, we take the simplest model that keeps gravity and pressure, a one-zone star (P1 and P2). A thin shell of mass Δm sits at radius R on gas of pressure P that fills the sphere inside it, and the mass inside is M. Newton's second law for the

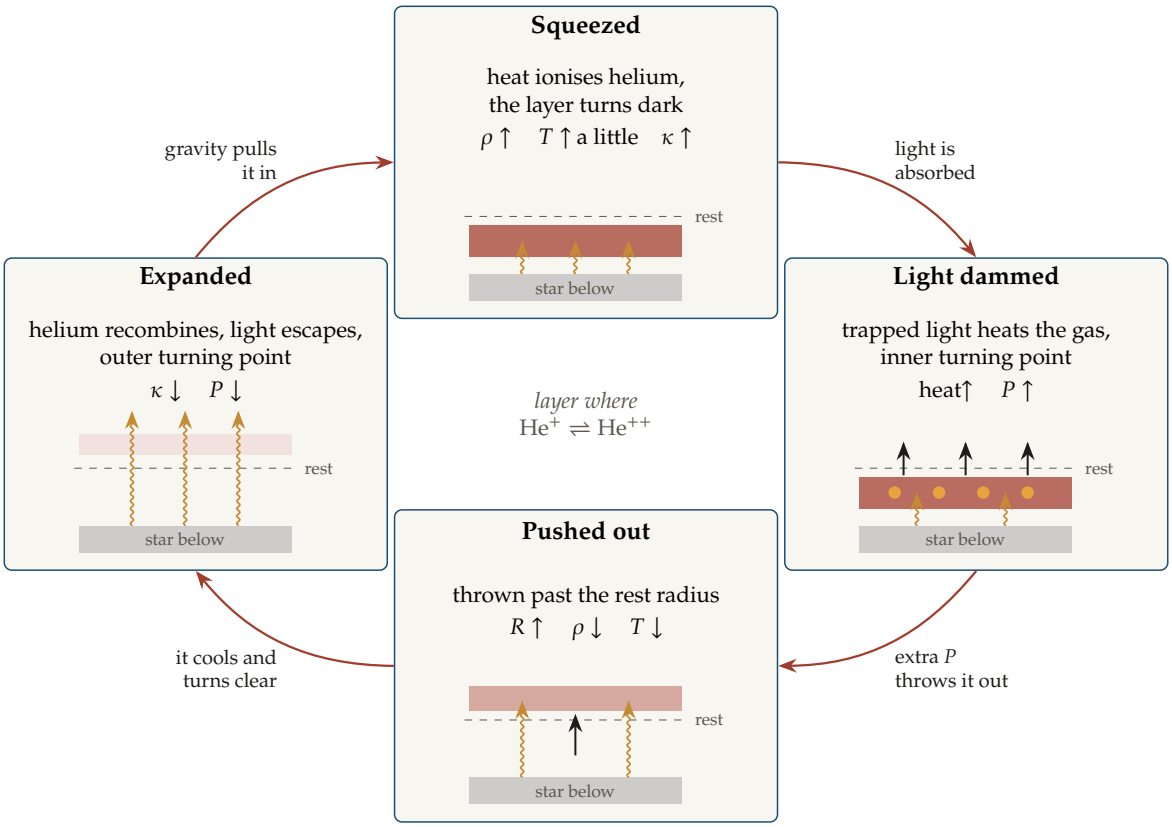


Figure 12.10: One cycle of a layer of partly ionised helium, as four stages joined by arrows in the order they happen. In each picture the grey bar is the star below, the dashed line is the rest radius, and the coloured band is the helium layer, drawn below the line when squeezed, above it when stretched, and darker for a higher density. Wavy arrows are light rising from below, black arrows show which way the layer moves or pushes, the labels on the cycle arrows give what happens between the stages, and the symbols give what rises (↑) and what falls (↓).

shell is

$$\Delta m\, \ddot{R} = -\frac{GM\Delta m}{R^2} + 4\pi R^2 P. \tag{12.43}$$

At rest, $R = R_0$, $P = P_0$ and $4\pi R_0^2 P_0 = GM\Delta m/R_0^2$. Now let the shell move a little, $R = R_0 + \delta R$, so that the pressure changes to $P_0 + \delta P$. The gas inside is squeezed too fast to exchange heat, so $P \propto \rho^\gamma \propto R^{-3\gamma}$ and $\delta P = -3\gamma P_0\, \delta R/R_0$. To first order,

$$\begin{aligned}
\Delta m\, \delta\ddot{R} &\overset{(1)}{=} \frac{2GM\Delta m}{R_0^3}\,\delta R + 8\pi R_0 P_0\,\delta R + 4\pi R_0^2\,\delta P \\
&\overset{(2)}{=} \frac{2GM\Delta m}{R_0^3}\,\delta R + \frac{2GM\Delta m}{R_0^3}\,\delta R - 3\gamma\,\frac{GM\Delta m}{R_0^3}\,\delta R \\
&= -(3\gamma - 4)\,\frac{GM\Delta m}{R_0^3}\,\delta R.
\end{aligned} \tag{12.44}$$

(1) Expand Eq. (12.43) to first order, with $1/R^2 \approx 1/R_0^2 - 2\,\delta R/R_0^3$ and $R^2 \approx R_0^2 + 2R_0\,\delta R$. The terms of zeroth order cancel by the condition at rest.

(2) The condition at rest gives $4\pi R_0 P_0 = GM\Delta m/R_0^3$, and $\delta P = -3\gamma P_0\,\delta R/R_0$.

This is a spring, $\delta\ddot{R} = -\omega^2 \delta R$, with

$$\omega^2 = (3\gamma - 4)\,\frac{GM}{R_0^3} = (3\gamma - 4)\,\frac{4\pi}{3}\,G\bar{\rho}, \tag{12.45}$$

where $\bar{\rho} = M/(\frac{4}{3}\pi R_0^3)$ is the mean density. The period $P_{\text{puls}} = 2\pi/\omega$ is

$$\boxed{P_{\text{puls}} = \sqrt{\frac{3\pi}{(3\gamma - 4)\,G\bar{\rho}}}.} \tag{12.46}$$

The period depends on the mean density alone, like the free-fall time (10.32). For $\gamma = \frac{5}{3}$ the two differ by a fixed factor, $P_{\text{puls}}/t_{\text{ff}} = \sqrt{32} = 5.7$. The factor $3\gamma - 4$ is the one that set the binding of a star in Eq. (10.16). As $\gamma \to \frac{4}{3}$ the star loses its spring and its binding together.

For the Sun's mean density, $1410\,\text{kg}\,\text{m}^{-3}$, with $\gamma = \frac{5}{3}$, Eq. (12.46) gives 2.8 hours. A giant of $5\,M_\odot$ and $50\,R_\odot$ has a mean density $5/50^3$ of the Sun's, so its period is $\sqrt{50^3/5} = 158$ times longer, 18 days. The one-zone star spreads its mass evenly. A Cepheid has most of its mass in a small dense core, and pulsation models of such stars give periods about a third of the one-zone value for the same mean density, as do Cepheids whose masses and radii are measured. The number in front depends on how the mass is spread. The scaling $P_{\text{puls}} \propto \bar{\rho}^{-1/2}$ does not, and the period–luminosity relation needs only the scaling.

Heat changes the pressure. The period treated the gas as if it gained no heat. Now let the gas inside the shell gain heat at the rate $L_{\text{in}} - L_{\text{out}}$, the light entering from below minus the light leaving through the top. Its internal energy is $PV/(\gamma - 1)$ (Eq. (9.26)), with $V = \frac{4}{3}\pi R^3$. By conservation of energy (P1),

$$\begin{aligned} &\frac{1}{\gamma - 1}\,\frac{\mathrm{d}(PV)}{\mathrm{d}t} = L_{\text{in}} - L_{\text{out}} - P\,\frac{\mathrm{d}V}{\mathrm{d}t} \\ \Longrightarrow\quad &V\frac{\mathrm{d}P}{\mathrm{d}t} + P\frac{\mathrm{d}V}{\mathrm{d}t} \overset{(1)}{=} (\gamma - 1)(L_{\text{in}} - L_{\text{out}}) - (\gamma - 1)\,P\frac{\mathrm{d}V}{\mathrm{d}t} \\ \Longrightarrow\quad &\frac{\mathrm{d}P}{\mathrm{d}t} \overset{(2)}{=} -\gamma\,\frac{P}{V}\,\frac{\mathrm{d}V}{\mathrm{d}t} + \frac{\gamma - 1}{V}\,(L_{\text{in}} - L_{\text{out}}). \end{aligned} \tag{12.47}$$

(1) Multiply by $\gamma - 1$ and expand $\mathrm{d}(PV)/\mathrm{d}t$ by the product rule.
(2) Collect the two terms in $P\ \mathrm{d}V/\mathrm{d}t$ and divide by V.

The first term is the change without heat, which integrates to $P \propto V^{-\gamma}$ and was used for the period. The second is the extra pressure of the heat the layer gains. Over the small swing, with $V \approx V_0 = \frac{4}{3}\pi R_0^3$, it adds

$$\delta P_{\text{heat}} = \frac{\gamma - 1}{V_0}\int (L_{\text{in}} - L_{\text{out}})\,dt. \tag{12.48}$$

The valve. The light reaching the layer from below comes from deep in the star, where nothing changes during a cycle. The light leaving depends on the layer itself. Across a layer of mass Δm, Eq. (12.9) with $\mathrm{d}T/\mathrm{d}m \sim -T/\Delta m$ gives $L_{\text{out}} \propto r^4 T^4/\kappa$. So a layer that becomes hotter or more transparent when squeezed lets more light out and loses heat, and a layer that becomes more opaque without heating much holds the light back and gains heat. Which one happens depends on the opacity. Putting the squeezed gas's temperature, $T \propto \rho^{\gamma-1}$ (from $P \propto \rho^\gamma$ and the ideal gas), into Kramers' law,

$$\begin{aligned} \kappa &\overset{(1)}{\propto} \rho\,T^{-7/2} \\ &\overset{(2)}{\propto} \rho \cdot \rho^{-7(\gamma-1)/2} = \rho^{(9-7\gamma)/2}. \end{aligned} \tag{12.49}$$

(1) Kramers' law, Section 12.3.
(2) $T \propto \rho^{\gamma-1}$ for gas squeezed without exchanging heat.

$$\boxed{\text{a squeeze raises the opacity only if} \quad \gamma < \tfrac{9}{7}.} \tag{12.50}$$

Ordinary ionised gas has $\gamma = \frac{5}{3}$, so $\kappa \propto \rho^{-4/3}$. A squeezed layer becomes both hotter and more transparent, and both let more light out. It loses heat at the top of the cycle, like the leaking cylinder. Every layer of ordinary gas damps the swing.

A layer where an element is partly ionised behaves differently. Helium loses its second electron, $\mathrm{He}^{+} \to \mathrm{He}^{++}$, with a binding energy of 54.4 eV. The weights of the states give the same factor of one as for hydrogen, so the Saha equation (16.16) holds with $B = 54.4\,\mathrm{eV}$. Setting $n_{\mathrm{He}^{++}} = n_{\mathrm{He}^{+}}$ in it and solving for T gives the temperature at which helium is half ionised. For electron densities from 10^{20} to $10^{24}\,\mathrm{m}^{-3}$ it is 3.4×10^4 to $6.0 \times 10^4\,\mathrm{K}$, with B/k_BT from 19 to 10.5. So the zone lies near $4 \times 10^4\,\mathrm{K}$ in any star, whatever the density there. A squeeze ionises this gas, and the reason takes two steps. First, the squeeze heats the gas. The layer above falls in and does work on the gas below, and a pulsation is too quick for heat to leak away, so the work $-P\,dV$ all stays in the gas. In ordinary gas it all goes into the motion of the particles. A particle that bounces off a wall moving toward it comes back faster, and on average that is the rise of the temperature, $T \propto \rho^{\gamma-1}$, which with $\gamma = \frac{5}{3}$ is 6.6% for a 10% squeeze. Second, a small warming shifts the helium balance a lot. Taking the logarithm of the Saha equation for helium and letting T and n_e change a little,

$$\ln \frac{n_{\mathrm{He}^{++}}}{n_{\mathrm{He}^{+}}} \overset{(1)}{=} \tfrac{3}{2} \ln T - \frac{B}{k_B T} - \ln n_e + \text{constant}$$

$$\Longrightarrow \quad d \ln \frac{n_{\mathrm{He}^{++}}}{n_{\mathrm{He}^{+}}} \overset{(2)}{=} \left(\frac{3}{2} + \frac{B}{k_B T} \right) d \ln T - d \ln n_e . \tag{12.51}$$

(1) The logarithm of Eq. (16.16) with $n_{\mathrm{He}^{++}}$, $n_{\mathrm{He}^{+}}$ in place of n_p, n_{H} and $B = 54.4\,\mathrm{eV}$.
(2) Change each term a little. $d(-B/k_BT) = (B/k_BT)\,d \ln T$.

In this zone B/k_BT is 10.5 to 19, so a 1% rise of the temperature raises the share of He^{++} by 12 to 20%, while a 1% rise of the density lowers it by only 1%. A denser gas on its own would recombine more, but the warming wins, and the squeeze pushes the balance toward ionisation. Most of the work done on the gas then goes into removing electrons, 54.4 eV each, instead of into heat. The temperature rises little, which is what a γ close to one means. Detailed models give γ of about 1.1 to 1.2 there, below $\frac{9}{7}$. The squeezed zone becomes more opaque and barely hotter, so it holds back the light at the top of the cycle. This is the valve. The zone of hydrogen ionisation, near $10^4\,\mathrm{K}$ by the same equation with $B = 13.6\,\mathrm{eV}$, works the same way, but it lies closer to the surface. A low γ also makes $(\gamma - 1)/\gamma$ small, so by the Schwarzschild criterion (12.11) these zones tend to convect, which sets the cool edge of the instability strip.

One cycle. Figure 12.10 follows the layer around one swing. Its rest radius is the one place where the pressure below balances gravity. Anywhere else the net force points back toward it, outward when the layer is squeezed and inward when it is stretched, so the equilibrium is stable and the layer swings about it like a pendulum. Any small disturbance starts a tiny swing. At the outer turning point the pressure below is too weak to hold the layer, and gravity pulls it in. It speeds up, passes the rest radius still moving, and squeezes the gas below. The squeeze heats the gas, the heat goes mostly into ionising helium, and the gas comes out denser but hardly hotter, so its opacity rises and it absorbs the light rising from below. The absorbed light heats the gas further, and the gas pressure climbs above what the squeeze alone would give. At the inner turning point this extra pressure stops the fall and throws the layer outward, past the rest radius. As the layer expands it cools, helium recombines, the gas turns clear, the

trapped heat escapes, and the pressure falls below normal. Gravity takes over again at the outer turning point. The gas takes in heat while squeezed and gives it out while stretched, so each swing is pushed outward a little harder than it is pulled inward, and the swing grows. A contracting star (Section 10.3) loses energy for good and passes slowly from one equilibrium to the next. The pulsating layer swings about one equilibrium, and the energy it takes from gravity on the way in, as motion and then as heat and ionisation, it hands back on the way out.

Growth of the swing. To see what the valve does to the swing, write it as $L_{\text{in}} - L_{\text{out}} = -\mathcal{H}\,\delta R$. A valve has $\mathcal{H} > 0$, gaining heat while squeezed ($\delta R < 0$). A leaking layer has $\mathcal{H} < 0$. Over one swing, $\delta R = A\cos\omega t$ with amplitude A, so

$$\int (L_{\text{in}} - L_{\text{out}})\,dt \overset{(1)}{=} -\mathcal{H}\,\frac{A\sin\omega t}{\omega} \overset{(2)}{=} \frac{\mathcal{H}}{\omega^2}\,\delta\dot{R}. \tag{12.52}$$

(1) $\int \cos\omega t\,dt = \sin(\omega t)/\omega$.
(2) $\delta\dot{R} = -A\omega\sin\omega t$.

By Eq. (12.48) the pressure of the trapped heat is $\delta P_{\text{heat}} = (\gamma-1)\mathcal{H}\,\delta\dot{R}/(V_0\omega^2)$. It is largest while the layer moves out. Its force on the shell is $4\pi R_0^2\,\delta P_{\text{heat}}$, with $4\pi R_0^2/V_0 = 3/R_0$. Adding it to the spring, the swing obeys

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\delta R\\ \delta\dot{R}\end{pmatrix} = \begin{pmatrix}0 & 1\\ -\omega^2 & 2/t_{\text{grow}}\end{pmatrix}\begin{pmatrix}\delta R\\ \delta\dot{R}\end{pmatrix}, \qquad \frac{2}{t_{\text{grow}}} = \frac{3(\gamma-1)\mathcal{H}}{\Delta m\,R_0\,\omega^2}. \tag{12.53}$$

The energy of the swing is $E_{\text{osc}} = \frac{1}{2}\Delta m\,(\delta\dot{R}^2 + \omega^2\delta R^2)$. Its rate of change is

$$\frac{\mathrm{d}E_{\text{osc}}}{\mathrm{d}t} \overset{(1)}{=} \Delta m\,\delta\dot{R}\left(\delta\ddot{R} + \omega^2\,\delta R\right) \overset{(2)}{=} \Delta m\,\delta\dot{R}\,\frac{2\,\delta\dot{R}}{t_{\text{grow}}} = \frac{2\Delta m}{t_{\text{grow}}}\,\delta\dot{R}^2. \tag{12.54}$$

(1) Differentiate E_{osc}.
(2) The second row of Eq. (12.53), $\delta\ddot{R} = -\omega^2\delta R + 2\,\delta\dot{R}/t_{\text{grow}}$.

Over a cycle the kinetic and stored parts of E_{osc} are equal on average, so the average of $\Delta m\,\delta\dot{R}^2$ is E_{osc}, and $\mathrm{d}E_{\text{osc}}/\mathrm{d}t = 2E_{\text{osc}}/t_{\text{grow}}$. If t_{grow} is long compared with a period, as assumed in writing $\delta R = A\cos\omega t$,

$$\boxed{E_{\text{osc}} \propto e^{2t/t_{\text{grow}}}, \qquad A \propto e^{t/t_{\text{grow}}}.} \tag{12.55}$$

The swing grows only if the layer gains heat while squeezed ($\mathcal{H} > 0$). The pressure of that heat then peaks a quarter of a cycle later, while the layer moves out, and pushes it on its way. A leaking layer has $\mathcal{H} < 0$, so $t_{\text{grow}} < 0$, and the swing decays at the same kind of rate. A star pulsates when the valve zone gains more energy per cycle than all the other layers lose.

The same answer follows from the first law over a closed cycle. The internal energy returns to its starting value, so the work the gas does is $\oint P\,dV = \oint T\,ds$, with s the entropy per unit mass as in Eq. (12.8). The valve takes heat in ($ds > 0$) near the squeeze, when T is highest, and gives it out ($ds < 0$) after the expansion, when T is lowest. So $\oint T\,ds > 0$ and the gas does net work on its surroundings each cycle, which is the car engine. A leaking layer gives heat out at the hot end, and $\oint T\,ds < 0$.

Where the valve works. The helium zone always lies near 4×10^4 K, but how deep it lies depends on the surface temperature. The grey atmosphere (11.43), $T^4 = \frac{3}{4}T_{\text{eff}}^4(\tau + \frac{2}{3})$, places it at an optical depth

$$\tau_{\text{zone}} \approx \frac{4}{3}\left(\frac{4 \times 10^4\,\text{K}}{T_{\text{eff}}}\right)^4,$$

which is about 1100 for $T_{\text{eff}} = 7500$ K and 3700 for 5500 K. A hotter star has its zone higher up, under less gas.

The valve works only at the right depth, and the test compares two times (P6). The gas above the zone holds heat, about $\frac{3}{2}k_BT/(\mu m_{\text{H}})$ per kilogram. The star's light carries that heat away in a time equal to the heat held divided by L. If this time is much shorter than a period, the zone lies so close to the surface that the little gas there cannot hold the trapped heat through a cycle, and too little mass takes part to drive the star. This sets the hot edge. If this time is much longer than a period, the zone lies deep, the heat it traps in one cycle is tiny next to the heat it holds, and the pressure hardly changes. Only when the two times are about equal does the valve drive the star.

In cool stars a second effect takes over. Their outer layers convect, and the ionisation zones themselves tend to convect, by the Schwarzschild criterion with low γ. Moving gas carries heat across the zone whatever its opacity, so holding back the light no longer holds back the heat, and the valve leaks. This sets the cool edge.

Between the two edges lies the **instability strip**, a narrow band of the HR diagram, about a thousand kelvin wide, that leans toward cooler temperatures at higher luminosity (Figure 12.11). Stars cross it as they evolve. Stars of about 4 to 10 $M_\odot$, burning helium after the main sequence, cross it as Cepheids. Low-mass helium-burning stars of old clusters sit in it as RR Lyrae stars, and stars near the main sequence as δ Scuti stars. The Sun lies on its cool side and does not pulsate this way.

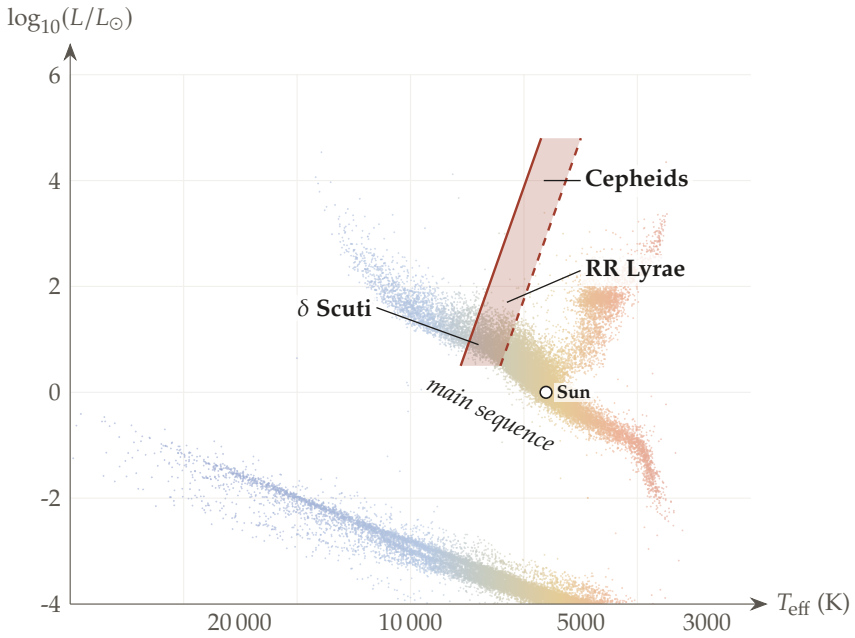


Figure 12.11: The stars of Figure 12.8, shown faint, with a schematic instability strip. The solid line is its hot edge and the dashed line its cool edge. Labels mark where Cepheids, RR Lyrae stars and δ Scuti stars lie in the strip.

Period and luminosity. All Cepheids sit in the strip, so all have nearly the same surface temperature. Brighter ones must then be bigger, and bigger ones are less dense and swing more slowly. To make this quantitative, combine the period (12.46) with the Stefan–Boltzmann law $L = 4\pi R^2 \sigma_{\rm SB} T_{\rm eff}^4$ and a mass–luminosity relation,

$$\begin{aligned} P_{\rm puls} &\overset{(1)}{\propto} \bar{\rho}^{-1/2} \overset{(2)}{\propto} R^{3/2} M^{-1/2} \\ &\overset{(3)}{\propto} \left(L^{1/2} T_{\rm eff}^{-2}\right)^{3/2} M^{-1/2} = L^{3/4}\, T_{\rm eff}^{-3}\, M^{-1/2} \\ &\overset{(4)}{\propto} L^{3/4}\, T_{\rm eff}^{-3}\, L^{-1/7} = L^{0.61}\, T_{\rm eff}^{-3}. \end{aligned} \tag{12.56}$$

(1) Eq. (12.46), at fixed γ.
(2) $\bar{\rho} \propto M/R^3$.
(3) $R \propto L^{1/2} T_{\rm eff}^{-2}$, from the Stefan–Boltzmann law.
(4) $L \propto M^{3.5}$, so $M^{-1/2} \propto L^{-1/7}$. Cepheids have left the main sequence and are brighter for their mass than Eq. (12.38) says, but their luminosity still rises steeply with mass. An exponent of 3 or 4 instead of 3.5 changes $\frac{1}{7}$ only to $\frac{1}{6}$ or $\frac{1}{8}$.

If the strip were exactly vertical, $T_{\rm eff}$ would be the same for all Cepheids and

$$\boxed{L \propto P_{\rm puls}^{1/0.61} = P_{\rm puls}^{1.6}.} \tag{12.57}$$

A brighter Cepheid is bigger and less dense, so it pulsates more slowly. This is the physical root of Leavitt's period–luminosity relation. The strip leans. In Figure 12.11, $T_{\rm eff} \propto L^{-0.033}$, which lowers the temperature by 14% across the hundredfold range of Cepheid luminosities. The factor $T_{\rm eff}^{-3}$ then becomes $L^{0.10}$, $P_{\rm puls} \propto L^{0.71}$, and $L \propto P_{\rm puls}^{1.4}$. The observed relation in visual light (Chapter 3) brightens by 2.43 magnitudes per tenfold rise of period, so $L \propto P_{\rm puls}^{2.43/2.5} = P_{\rm puls}^{0.97}$. The scaling gets the sign right and a power between one and two, against the observed power of about one. The exact slope needs pulsation models and the share of the light that falls in the visual band. Once calibrated, the relation turns a measured period into a luminosity, and with the measured flux into a distance.

12.8 Problem Set: The Equations of a Star

12.8.1 Problem 1: Mean Molecular Weights

Problem Statement: Find μ and μ_e for fully ionised gas of (a) solar composition, $X = 0.70$, $Y = 0.28$, $Z = 0.02$, (b) pure helium, and (c) pure carbon.

Essential Ideas:

- Count particles and electrons per mass $m_{\rm H}$, Eq. (12.14).
- Carbon, $A = 12$, has a nucleus and six electrons, seven particles.

Detailed Solutions:

(a) **Solar**. $1/\mu = 1.40 + 0.21 + 0.01 = 1.62$ and $1/\mu_e = 0.85$.

$$\boxed{\mu = 0.62, \qquad \mu_e = 1.18.}$$

(b) **Helium**. $1/\mu = \frac{3}{4}$ and $1/\mu_e = \frac{1}{2}$.

$$\boxed{\mu = \tfrac{4}{3}, \qquad \mu_e = 2.}$$

(c) **Carbon**. Seven particles and six electrons per $12\,m_{\mathrm{H}}$.

$$\boxed{\mu = \tfrac{12}{7} = 1.71, \qquad \mu_e = 2.}$$

12.8.2 Problem 2: Is It Degenerate?

Problem Statement: Compare the electrons' Fermi energy with $k_B T$ (a) at the Sun's centre ($\rho = 1.5 \times 10^5\,\mathrm{kg\,m^{-3}}$, $T = 1.57 \times 10^7\,\mathrm{K}$, $\mu_e = 1.18$) and (b) at the centre of a white dwarf ($\rho = 10^9\,\mathrm{kg\,m^{-3}}$, $T = 10^7\,\mathrm{K}$, $\mu_e = 2$). (c) In case (b), are the electrons relativistic?

Essential Ideas:

- $n_e = \rho/(\mu_e m_{\mathrm{H}})$, p_F from Eq. (12.15), $E_F = p_F^2/2m_e$ for slow electrons.

Detailed Solutions:

(a) **Sun**.

$$\boxed{E_F/k_B T \simeq 0.5.}$$

Mildly degenerate. The ideal-gas law is accurate to some percent there.

(b) **White dwarf**.

$$\boxed{E_F/k_B T \simeq 190.}$$

Strongly degenerate. The temperature barely matters for the pressure.

(c) **Relativistic?** $p_F/m_e c = 0.80$. The electrons move at a large fraction of c, and neither limit of Eq. (12.16) is exact.

12.8.3 Problem 3: The Gamow Peak

Problem Statement: At the Sun's centre, $k_B T = 1.35\,\mathrm{keV}$. (a) Find the Gamow peak E_0 for two protons ($E_G = 493\,\mathrm{keV}$) and the temperature exponent of the rate. (b) Repeat for a proton and a nitrogen nucleus ($E_G = 45\,\mathrm{MeV}$).

Essential Ideas:

- $E_0 = (E_G (k_B T)^2/4)^{1/3}$, Eq. (12.18), and the exponent $E_0/k_B T - \tfrac{2}{3}$.

Detailed Solutions:

(a) **Two protons**. $E_0 = (493 \times 1.35^2/4)^{1/3}\,\mathrm{keV}$.

$$\boxed{E_0 \simeq 6.1\,\mathrm{keV} = 4.5\,k_B T, \qquad \text{rate} \propto T^{3.8}.}$$

(b) **Proton and nitrogen**.

$$\boxed{E_0 \simeq 27\,\mathrm{keV} = 20\,k_B T, \qquad \text{rate} \propto T^{19.6}.}$$

12.8.4 Problem 4: Lifetimes

Problem Statement: Take $L \propto M^{3.5}$ and a main-sequence lifetime of 10^{10} years for the Sun. Estimate the lifetimes of stars of 0.5, 10 and $50\,M_\odot$.

Essential Ideas:

- Lifetime $\propto M/L \propto M^{-2.5}$.

Detailed Solutions:

$$\boxed{0.5\,M_\odot:\ 6 \times 10^{10}\ \mathrm{yr}, \qquad 10\,M_\odot:\ 3 \times 10^{7}\ \mathrm{yr}, \qquad 50\,M_\odot:\ 6 \times 10^{5}\ \mathrm{yr}.}$$

Every star of half a solar mass ever formed is still on the main sequence, since the universe is only 1.4×10^{10} years old. Massive stars live and die while the cloud they formed in is still nearby.

12.8.5 Problem 5: The Temperature Fall of the Earth's Air

Problem Statement: The Earth's lower atmosphere is stirred by convection. Use the convective gradient, Eq. (12.12), with $\gamma = \frac{7}{5}$ for air (mean molecular mass 28.97 u) and hydrostatic equilibrium near the ground to find how fast the temperature falls with height.

Essential Ideas:

- $\mathrm{d}\ln T/\mathrm{d}\ln P = (\gamma-1)/\gamma$, and $\mathrm{d}P/\mathrm{d}z = -\rho g$ with $\rho = P\mu m_{\mathrm{H}}/(k_B T)$.

Detailed Solutions:

$$\frac{\mathrm{d}T}{\mathrm{d}z} \overset{(1)}{=} \frac{T}{P}\,\frac{\gamma-1}{\gamma}\,\frac{\mathrm{d}P}{\mathrm{d}z} \overset{(2)}{=} -\frac{\gamma-1}{\gamma}\,\frac{\mu m_{\mathrm{H}}\, g}{k_B}.$$

(1) The convective gradient.
(2) Hydrostatic equilibrium with the ideal gas, $\mathrm{d}P/\mathrm{d}z = -P\mu m_{\mathrm{H}} g/(k_B T)$.

$$\boxed{\frac{\mathrm{d}T}{\mathrm{d}z} = -\frac{2}{7}\times\frac{28.97\times1.661\times10^{-27}\times9.81}{1.381\times10^{-23}}\,\mathrm{K\,m^{-1}} \simeq -9.8\,\mathrm{K\,km^{-1}}.}$$

This is the dry adiabatic lapse rate of meteorology. The same criterion that decides where the Sun convects decides how cold mountain tops are.

12.8.6 Problem 6: Cepheid Periods

Problem Statement: (a) A Cepheid of fixed mass swells from 40 to 50 $R_\odot$ as it evolves across the instability strip. By what factor does its period change? (b) Two Cepheids have the same surface temperature, and one is ten times as luminous as the other. Using Eq. (12.46) and $L \propto M^{3.5}$, find the ratio of their periods. (c) Compare with the observed visual relation, $L \propto P_{\mathrm{puls}}^{0.97}$.

Essential Ideas:

- $P_{\mathrm{puls}} \propto \bar{\rho}^{-1/2} \propto R^{3/2}M^{-1/2}$, Eq. (12.46).
- At fixed T_{eff}, $P_{\mathrm{puls}} \propto L^{0.61}$ (Section 12.7).

Detailed Solutions:

(a) **Swelling**. At fixed mass, $P_{\mathrm{puls}} \propto R^{3/2}$, and $(50/40)^{3/2} = 1.25^{1.5}$.

$$\boxed{P_{\mathrm{puls}} \text{ grows by a factor } 1.40.}$$

(b) **Ten times brighter**. $P_{\mathrm{puls}} \propto L^{3/4-1/7} = L^{0.61}$, and $10^{0.61}$.

$$\boxed{\text{period ratio} \simeq 4.0.}$$

(c) **Observed**. $P_{\mathrm{puls}} \propto L^{1/0.97}$, and $10^{1/0.97}$.

$$\boxed{\text{period ratio} \simeq 11.}$$

The observed periods spread more than the estimate at fixed temperature. Part of the difference is that the brighter Cepheid is also cooler, so larger and less dense than a vertical strip allows. The lean of the strip raises the estimate only to $10^{0.71} = 5.1$, and the rest needs pulsation models and the share of the light in the visual band. The scaling explains why the relation exists, and its calibration on Cepheids of known distance makes it a measuring tool.

Chapter 13

Degenerate and Compact Stars

When a star exhausts its nuclear fuel, its core loses the heat source that helped support it. The core contracts and heats as it releases gravitational energy (Section 10.3). If electron degeneracy pressure stops the contraction, the remnant is a white dwarf. A heavier core may collapse further until neutron matter supports it as a neutron star. If no stable neutron star can form, the collapse continues toward a black hole.

Two different effects enter this story. In a heavy white dwarf, electrons move near the speed of light. Their pressure then grows less rapidly under compression, approaching $\gamma = \frac{4}{3}$; this sets a mass limit even with Newtonian gravity. In a neutron star, gravity itself must be described by general relativity, which changes the structure equations. We will keep these two effects separate and quote the needed relativistic equations where they arise.

13.1 Compactness

Chapter 12 treated gravity with Newton's law. That works when a star's gravitational field is weak, but a neutron star packs a large mass into a small radius. We need a measure of when the Newtonian approximation begins to fail.

Think of a ball thrown up from the surface of a small asteroid. A gentle throw sends it away for good. On the Earth the same throw brings it straight back. How hard gravity grips a body at the surface depends on how much mass is packed into how small a radius. The grip of gravity, measured against the speed of light, is the ratio $GM/(Rc^2)$, the **compactness**. It compares the escape energy per unit mass at the surface with c^2. Every correction that general relativity makes to the structure of a star is of this size, so it decides whether Newtonian gravity is good enough (P6).

The Sun has $GM/(Rc^2) = 2.1 \times 10^{-6}$, and the white dwarf Sirius B has 2.6×10^{-4}. Newtonian gravity works well for both. For a neutron star of $1.4\,M_\odot$ and 12 km radius, the ratio is 0.17, large enough that relativistic corrections matter. This same ratio also sets the scale of gravitational redshift and of the gravitational energy released by infalling matter.

Object	$GM/(Rc^2)$	surface redshift z	gravity treated by
Sun	2.1×10^{-6}	2.1×10^{-6}	Newton
Sirius B (white dwarf)	2.6×10^{-4}	2.6×10^{-4}	Newton
Neutron star, $1.4\,M_\odot$, 12 km	0.17	0.24	general relativity
Black hole	0.5	–	general relativity

Table 13.1: Compactness and surface redshift, $z = (1 - 2GM/Rc^2)^{-1/2} - 1$.

The redshift of Sirius B, $z = 2.6 \times 10^{-4}$, shifts its spectral lines as a recession at $77\,\mathrm{km\,s^{-1}}$ would,

close to what is measured. Why so dense a star can exist at all is worked out in Section 13.2, and Newtonian gravity is enough for it.

13.2 White Dwarfs

The pressure of degenerate electrons is a power of the density alone, $\rho^{5/3}$ for slow electrons and $\rho^{4/3}$ for fast ones (Eq. (12.16)). A star held up by it is therefore a polytrope (Section 12.5). A white dwarf of the Sun's mass is the size of the Earth. As its mass grows, its electrons are squeezed harder and move faster. The polytrope of slow electrons cannot describe the heaviest ones.

Think of a crowd in a hall, pressed toward the front by a slope. Each person's push on the person behind grows the harder they are squeezed, at first. If everyone is already pushing as hard as they can, adding more people down the slope only crushes the front, and no new push appears to stop it. Electrons near the speed of light behave like that crowd. Their momentum can keep growing, but their speed cannot exceed c, so their pressure grows more slowly with density, $\gamma \to \frac{4}{3}$. By the virial theorem (Section 10.2), a star with $\gamma = \frac{4}{3}$ is barely bound, and by the mass–radius relation (12.29) only one mass fits. Above that mass, degenerate electrons cannot hold a star up. Counting states (P3) and the virial balance of a sphere (P1, P2) give that mass.

For a carbon-oxygen white dwarf, with two atomic mass units per electron ($\mu_e = 2$), the limit is $1.456\,M_\odot$. Figure 13.1 shows the radius shrinking as mass grows toward that limit.

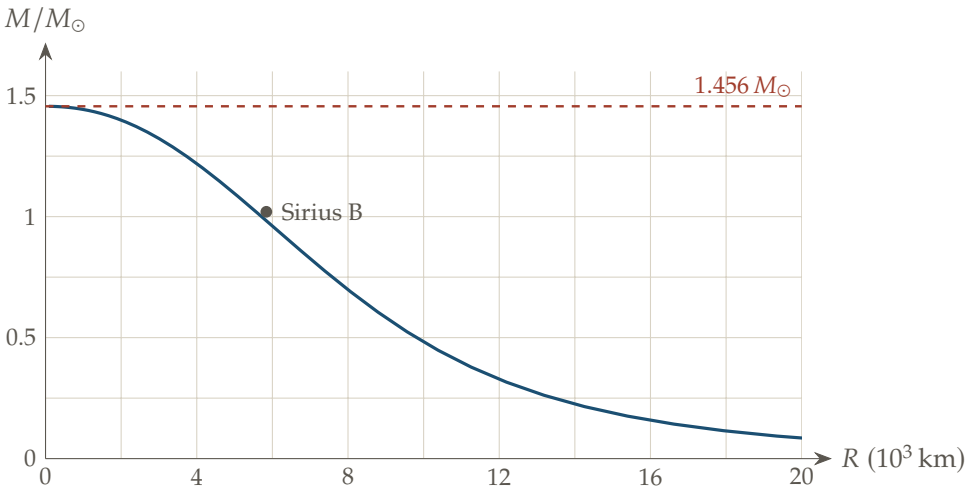


Figure 13.1: The mass of white dwarfs against their radius, computed from the exact pressure of degenerate electrons with $\mu_e = 2$. The dashed line marks $1.456\,M_\odot$. The dot marks the observed mass and radius of Sirius B.

The limit also matters when an iron core collapses in a massive star and when a white dwarf gains matter from a companion before a type Ia supernova (Section 13.2.1). We first derive how white-dwarf radius changes with mass, then find the limiting mass.

Slow electrons. A light white dwarf has slow electrons, so $\gamma = \frac{5}{3}$ and, by Eq. (12.29),

$$R \propto M^{-1/3}. \tag{13.1}$$

Adding mass makes it smaller. The degenerate pressure depends on the density alone, so a heavier star must be denser to hold itself up.

Fast electrons, the Chandrasekhar mass. For fast electrons, $P = \mathcal{K}\rho^{4/3}$ with

$$\mathcal{K} = \frac{hc}{8}\left(\frac{3}{\pi}\right)^{1/3}\frac{1}{(\mu_e m_{\mathrm{u}})^{4/3}}$$

from Eq. (12.16) and $n_e = \rho/(\mu_e m_{\mathrm{u}})$, where m_{u} is the atomic mass unit. With $\gamma = \frac{4}{3}$, Eq. (12.26) gives $r_0^2 = \mathcal{K}\rho_c^{-2/3}/(\pi G)$, so $r_0^3\rho_c = (\mathcal{K}/\pi G)^{3/2}$ does not depend on ρ_c, and the mass formula of Section 12.5 gives one mass only,

$$\begin{aligned} M_{\mathrm{Ch}} &\overset{(1)}{=} 4\pi\Big[-\xi^2\frac{\mathrm{d}\theta}{\mathrm{d}\xi}\Big]_{\xi_1}\left(\frac{\mathcal{K}}{\pi G}\right)^{3/2} \\ &\overset{(2)}{=} 4\pi\times 2.0182\times\left(\frac{hc}{8\pi G}\right)^{3/2}\left(\frac{3}{\pi}\right)^{1/2}\frac{1}{(\mu_e m_{\mathrm{u}})^2}. \end{aligned} \tag{13.2}$$

(1) The mass of a polytrope, $4\pi r_0^3\rho_c[-\xi^2\theta']_{\xi_1}$, with $r_0^3\rho_c = (\mathcal{K}/\pi G)^{3/2}$.

(2) The Lane–Emden value for $\gamma = \frac{4}{3}$ and $\mathcal{K}$ written out.

$$\boxed{M_{\mathrm{Ch}} = \frac{5.83}{\mu_e^2}\,M_\odot = 1.456\,M_\odot \quad (\mu_e = 2).} \tag{13.3}$$

The mass contains h, c, G and the mass per electron, and nothing about the star's history. Quantum mechanics (h), relativity (c) and gravity (G) together fix a mass of the order of the Sun's.

The same mass by an energy argument. Landau found the limit by comparing two energies of one electron in a star of N electrons and radius R. Its share of gravitational energy is about $-GM\mu_e m_{\mathrm{u}}/R$, with $M = N\mu_e m_{\mathrm{u}}$. A fast electron confined to a spacing $R/N^{1/3}$ has, by the uncertainty principle, a momentum of about $\hbar N^{1/3}/R$ and so a kinetic energy of about $\hbar c N^{1/3}/R$. Both scale as $1/R$, so shrinking the star changes their ratio not at all (P6). If gravity wins at one radius, it wins at every radius, and the star collapses. The two are equal when

$$\begin{aligned} &\frac{\hbar c\,N^{1/3}}{R} = \frac{G\,N(\mu_e m_{\mathrm{u}})^2}{R} \\ \Longrightarrow\quad &N = \left(\frac{\hbar c}{G(\mu_e m_{\mathrm{u}})^2}\right)^{3/2}, \qquad M = N\mu_e m_{\mathrm{u}} = \left(\frac{\hbar c}{G}\right)^{3/2}\frac{1}{(\mu_e m_{\mathrm{u}})^2}. \end{aligned} \tag{13.4}$$

This is $0.47\,M_\odot$ for $\mu_e = 2$. The estimate has the right combination of constants and falls short of Eq. (13.3) only by a number, 3.1, that the full polytrope supplies.

The full curve. Real white dwarfs have electrons that are neither slow nor at the speed of light. Figure 13.1 uses the exact pressure of degenerate electrons at any speed, integrated in Newtonian hydrostatic equilibrium. Light white dwarfs follow $R \propto M^{-1/3}$. As the mass rises toward $1.456\,M_\odot$ the radius falls to zero. Sirius B, $1.02\,M_\odot$, lies on the curve at 5566 km, against 5840 km measured, close enough given that the model ignores temperature and composition details. Throughout, $GM/(Rc^2) < 10^{-3}$, so Newtonian gravity is enough. The limit comes from the particles being relativistic, not gravity.

13.2.1 Type Ia Supernovae

A white dwarf of carbon and oxygen has no fuel it can burn at its present density, and left alone it only cools. Two things can push it toward the limit (13.3). It can take gas from a companion star that has swollen enough to spill onto it. Or it can merge with a second white dwarf, after the two have spiralled together by radiating gravitational waves (Section 13.6).

Ignition without a thermostat. As the mass climbs toward the limit, the radius shrinks (Figure 13.1) and the centre is squeezed and heated. Near $1.4\,M_\odot$ the centre reaches a few $10^{12}\,\mathrm{kg\,m^{-3}}$ and a few $10^8\,\mathrm{K}$, and carbon nuclei begin to fuse. Two carbon nuclei, $Z_1 = Z_2 = 6$ with a reduced mass of six atomic mass units, have a Gamow energy $E_G = 7.6\,\mathrm{GeV}$. At $5\times10^8\,\mathrm{K}$, Eq. (12.18) puts the Gamow peak at $E_0 = 1.5\,\mathrm{MeV} = 35\,k_BT$, and the slope $-\frac{2}{3} + E_0/k_BT$ of Section 12.3 makes the rate grow as about T^{35}. A rise of one percent in temperature raises the burning by forty percent.

In an ordinary star such burning is held steady. The extra heat raises the pressure, the core expands and cools, and the burning slows (Section 10.3). In a white dwarf the pressure is that of degenerate electrons, which does not depend on the temperature (Eq. (12.16)). The heat does not expand the core, so nothing cools it, and the thermostat fails (Section 12.3). Hotter gas burns faster, and faster burning makes it hotter. The burning runs away and sweeps through the whole star in about a second, a few times the free-fall time of so dense a star (Section 10.5). By the time the gas is hot enough for its pressure to respond, the fuel is spent.

Nothing is left. Burning carbon and oxygen to iron-group nuclei such as nickel-56 raises the binding energy from about 7.8 to 8.6 MeV per nucleon. About one solar mass burns. A white dwarf near the limit has electrons with γ close to $\frac{4}{3}$, so by Eq. (10.16) its total energy E is close to zero, and its binding energy is only about $5\times10^{43}\,\mathrm{J}$. With $M = 1.4\,M_\odot$ and v the mean speed of the debris, energy conservation (P1) gives

$$\begin{aligned}\tfrac{1}{2}Mv^2 &\overset{(1)}{=} \frac{0.82\,\mathrm{MeV}}{m_\mathrm{u}} \times 1\,M_\odot - |E| \\ &\overset{(2)}{=} 1.6\times10^{44}\,\mathrm{J} - 0.5\times10^{44}\,\mathrm{J} = 1.1\times10^{44}\,\mathrm{J}, \\ v &\overset{(3)}{=} \sqrt{\frac{2\times1.1\times10^{44}\,\mathrm{J}}{2.8\times10^{30}\,\mathrm{kg}}} \approx 9\times10^3\,\mathrm{km\,s^{-1}}. \end{aligned} \tag{13.5}$$

(1) The burning releases 0.82 MeV per nucleon, and a nucleon has a mass of about m_u. Unbinding the star costs $|E|$.
(2) $0.82\,\mathrm{MeV}/m_\mathrm{u} = 7.9\times10^{13}\,\mathrm{J\,kg^{-1}}$.
(3) Solve for v.

The burning releases more than the star's binding energy, so the whole star flies apart and no remnant is left. The speeds in the spectra of these supernovae, about $10^4\,\mathrm{km\,s^{-1}}$, agree.

The light of radioactive nickel. The heat of the explosion itself is spent in the expansion, as the debris grows from the size of the Earth to that of the Solar System. The light seen for months comes from radioactivity. Carbon and oxygen nuclei have equal numbers of protons and neutrons, and so does nickel-56, the most tightly bound nucleus that can be made from them in a second. About $0.6\,M_\odot$ of it is made. It decays to cobalt-56 with a mean life of 8.8 days, releasing 1.75 MeV per decay, and cobalt-56 decays to iron-56 with a mean life of 111 days. With $0.6\,M_\odot/(56\,m_\mathrm{u}) = 1.3\times10^{55}$ nickel nuclei, the nickel heats the debris at the rate

$$\frac{1.3\times10^{55}\times1.75\,\mathrm{MeV}}{8.8\,\mathrm{d}}\,e^{-t/8.8\,\mathrm{d}} = 4.8\times10^{36}\,\mathrm{W}\times e^{-t/8.8\,\mathrm{d}}. \tag{13.6}$$

The light has to diffuse out through the debris (Section 11.4), which takes about three weeks, so the light curve peaks then. The peak luminosity is close to the heating at that moment, from nickel and cobalt together, about $10^{36}\,\mathrm{W}$, or $3\times10^9\,L_\odot$. One star outshines its whole galaxy for weeks. After the peak the light follows the cobalt. If all the decay energy stayed in the debris, the light would fall by $1.0857\times100/111 = 0.98$ magnitudes per 100 days (the factor 1.0857 of

Eq. (4.36)). Real tails fall a little faster, because the debris thins and more of the gamma rays escape without being absorbed.

A measured light curve. Figure 13.2 shows the light of SN 2011fe, a type Ia supernova in the nearby spiral galaxy M101, found in August 2011 within about a day of the explosion and followed in the B and V bands for 635 days after its peak. The B light peaks at $B = 9.92$, and the V light 1.3 days later, as the debris cools and its light shifts to the red. In the first 15 days after the peak the B light falls by $\Delta m_{15}(B) = 1.15$ magnitudes, the measure of how fast a supernova fades that the next paragraph uses. From 60 to 300 days after the peak the light falls on a straight line. A magnitude is a logarithm of the flux, so a straight line in magnitude is an exponential decay of the luminosity, the signature of one radioactive nucleus. A straight-line fit gives 1.87 magnitudes per 100 days in V and 1.46 in B, faster than the 0.98 of cobalt-56 with all its decay energy kept, because more of the gamma rays escape the thinning debris. After about 300 days the fall slows to 1.18 magnitudes per 100 days, closer to the cobalt rate. By then most of the gamma rays escape, and the light comes mainly from the positrons that some cobalt decays emit, which stop inside the debris and keep their share of the energy there. Peak brightness also gives the distance. With the peak $M_B \approx -19.3$ of the next paragraph and the Milky Way's dust in front of M101, $A_B = A_V + E(B-V) = 4.1\,E(B-V) = 0.03$ for $E(B-V) = 0.0077$ (Eq. (4.42)), Eq. (4.44) gives

$$m - M \overset{(1)}{=} B_{\max} - A_B - M_B = 9.92 - 0.03 + 19.3 = 29.19,$$

(1) Eq. (4.44), the measured peak with the dust removed.

a distance of 6.9 Mpc.

A standard candle. Every such explosion starts from a white dwarf of carbon and oxygen close to the same mass, so it makes a similar amount of nickel and reaches a similar peak, $M_B \approx -19.3$ (Section 3.7). The peaks still spread by a few tenths of a magnitude. The spread follows the shape of the light curve. Brighter explosions fade more slowly (the Phillips relation), because more nickel makes hotter, more opaque debris from which light leaks out more slowly. Once each peak is corrected by the width of its light curve, the spread falls to about 0.15 magnitudes, 7% in distance.

No type Ia supernova is close enough for a parallax, so the peak luminosity is calibrated in galaxies that have hosted a type Ia supernova and also contain Cepheids, whose distances come from the parallax of nearby Cepheids (Section 3.7). At its peak it is 10^5 to 10^6 times more luminous than a Cepheid and is seen 500 to 1000 times farther, Eq. (3.52). Cepheids reach about 30 Mpc, and type Ia supernovae reach redshifts of one and more, far enough to show that the expansion of the universe is speeding up (Section 16.7).

13.3 The Relativistic Structure Equations

Compactness tells us when to replace the Newtonian structure equations. For a neutron star, $GM/(Rc^2) \approx 0.2$, so relativistic effects are substantial. We now compare the equations for mass, pressure, energy and heat flow with their Newtonian forms.

Think of weighing a box of hot gas. Heat the gas and the box gets heavier, because energy has mass. Squeeze the gas and its pressure adds to the weight as well. In everyday life the extra weight is far too small to notice. In a neutron star it is not small. The pressure that holds the star up is itself a source of gravity, so pushing harder also pulls harder. General relativity, with conservation of energy and momentum (P1) in curved space-time, gives the equations. They are quoted here from general relativity, not derived again.

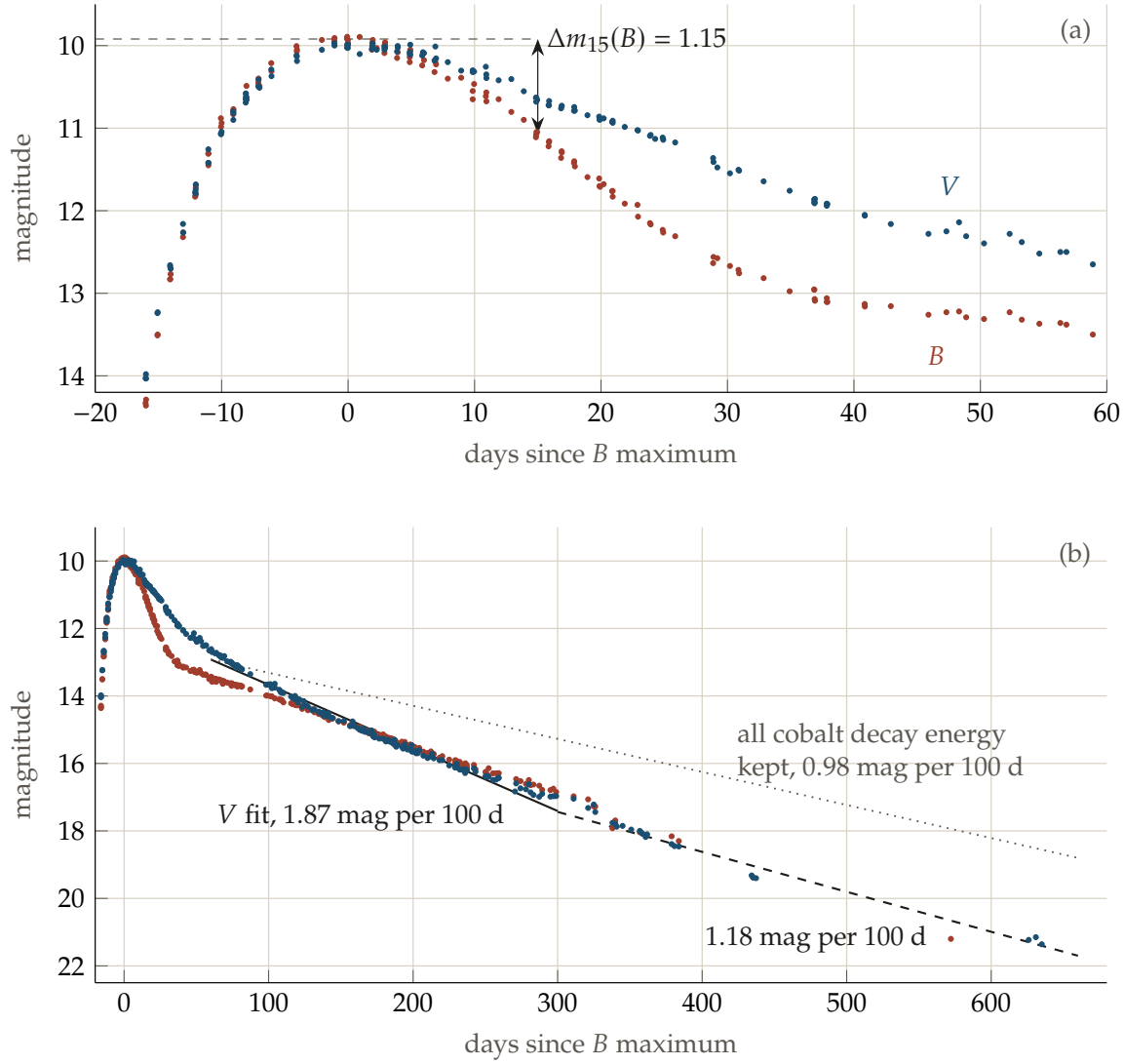


Figure 13.2: B (red) and V (blue) magnitudes of SN 2011fe against days since B maximum, 650 measurements from Munari et al. (2013) and Tsvetkov et al. (2013), taken from the Open Supernova Catalog. (a) The first 60 days, with the drop $\Delta m_{15}(B)$ marked. (b) The whole run, with straight-line fits to V from 60 to 300 days (solid) and from 300 days on (dashed), and the slope of cobalt-56 decay with all its energy kept (dotted).

At a neutron star's centre, pressure can be between about a tenth and a half of its energy density ρc^2. At the Sun's centre, the ratio is only about 2×10^{-6}. The correction factors in the equations below therefore stay close to one for the Sun but change the balance inside a neutron star.

The metric. A static spherical star has the metric

$$ds^2 = -e^{2\Phi/c^2} c^2 dt^2 + \left(1 - \frac{2Gm(r)}{rc^2}\right)^{-1} dr^2 + r^2\left(d\theta^2 + \sin^2\theta\, d\phi^2\right), \tag{13.7}$$

where $\Phi(r)$ becomes the Newtonian potential of Section 9.2 far from the star, and $m(r)$ is the mass inside r.

Reading the corrections. The momentum equation is the Tolman–Oppenheimer–Volkoff (TOV) equation. Each of its three new factors is one physical effect.

	Newtonian	General relativity
Mass	$\frac{dm}{dr} = 4\pi r^2 \rho$	$\frac{dm}{dr} = 4\pi r^2 \rho$, ρc^2 = total energy density
Momentum	$\frac{dP}{dr} = -\frac{Gm\rho}{r^2}$	$\frac{dP}{dr} = -\frac{G\left(\rho + \frac{P}{c^2}\right)\left(m + \frac{4\pi r^3 P}{c^2}\right)}{r^2\left(1 - \frac{2Gm}{rc^2}\right)}$
Potential	$\frac{d\Phi}{dr} = \frac{Gm}{r^2}$	$\frac{d\Phi}{dr} = -\frac{c^2}{\rho c^2 + P}\frac{dP}{dr}$
Energy	$\frac{dL}{dr} = 4\pi r^2 \rho\, \varepsilon$	$\frac{d}{dr}\left(L\, e^{2\Phi/c^2}\right) = \frac{4\pi r^2 \rho\, \varepsilon\, e^{2\Phi/c^2}}{\sqrt{1 - 2Gm/rc^2}}$
Transport	$L = -\frac{16\pi ac\, r^2 T^3}{3\kappa\rho}\frac{dT}{dr}$	$L = -\frac{16\pi ac\, r^2 T^3}{3\kappa\rho}\sqrt{1 - \frac{2Gm}{rc^2}}\, e^{-\Phi/c^2}\frac{d}{dr}\left(T\, e^{\Phi/c^2}\right)$

Table 13.2: The structure equations of a static star in Newtonian gravity and in general relativity (energy equation in a star that does not change).

- $\rho + P/c^2$ in place of ρ. Pressure has inertia. A parcel under pressure is harder to hold up.
- $m + 4\pi r^3 P/c^2$ in place of m. Pressure gravitates. The pressure that holds the star up adds to the pull it must resist.
- $(1 - 2Gm/rc^2)^{-1}$. Space inside the star is curved, and the radial distance is longer than dr.

All three make gravity stronger. Raising the pressure to support more mass also raises gravity, so beyond some mass no equation of state can win. This is why neutron stars, unlike white dwarfs, have a maximum mass set by gravity itself, whatever the particles.

The energy equation conserves the luminosity as seen from far away, $L\, e^{2\Phi/c^2}$, where one factor e^{Φ/c^2} is the redshift of each photon's energy and the other is the slowing of clocks deep in the potential. The square root turns dr into a proper volume.

Equilibrium in a gravitational field. With no heat flow, $L = 0$, the transport equation gives not a uniform T but

$$\boxed{T\, e^{\Phi/c^2} = \text{constant.}} \tag{13.8}$$

This is the Tolman–Ehrenfest relation. In a gravitational field, a body in full thermal equilibrium is hotter at the bottom, by the same factor by which light climbing out is redshifted. Heat flows only when $T\, e^{\Phi/c^2}$ changes from place to place. The statement in Section 10.6 that global equilibrium means one temperature throughout is the Newtonian limit of this. For the Sun the difference is two parts in a million. Across a neutron star it is about twenty percent.

Mass and binding. In general relativity $m(r)$ counts the total energy inside r, including the negative gravitational binding energy. The mass M that governs an orbit around the star is smaller than the rest mass of all its particles, by about $GM/(Rc^2)$ of it. For a neutron star the difference is about a tenth of the rest mass, 3×10^{46} J, released when the core collapses, almost all of it as neutrinos (Problem 13.7.3).

13.4 Computing a Neutron Star

The relativistic equations tell us how pressure supports a neutron star. To obtain its mass and radius, we also need a relation between pressure and density for neutron matter. Once that

relation is chosen, we can build a sequence of models and see where support fails.

Think of stacking sand into a cone. Pour more and the cone grows wider and taller. With dry sand the slope holds, and the cone can grow as large as the sand allows. With wet sand the weight of the top crushes the bottom past some height. Then the pile only spreads, and adding sand does not make it taller. A neutron star is built in the same spirit, one central density at a time. Each central density gives one star, with its own mass and radius. Going to higher central density gives heavier stars, up to a point beyond which the extra gravity of the extra pressure wins and the mass falls. That point is the maximum mass. Finding it is the TOV equation (P1 and P2 in general relativity) solved with the pressure of degenerate neutrons (P3).

For an ideal degenerate neutron gas, the relativistic sequence reaches a maximum of $0.71\,M_\odot$ at radius 9.3 km. With Newtonian gravity, the same gas approaches about $5.69\,M_\odot$, roughly eight times more. Figure 13.3 shows how the relativistic curve turns over as central density rises.

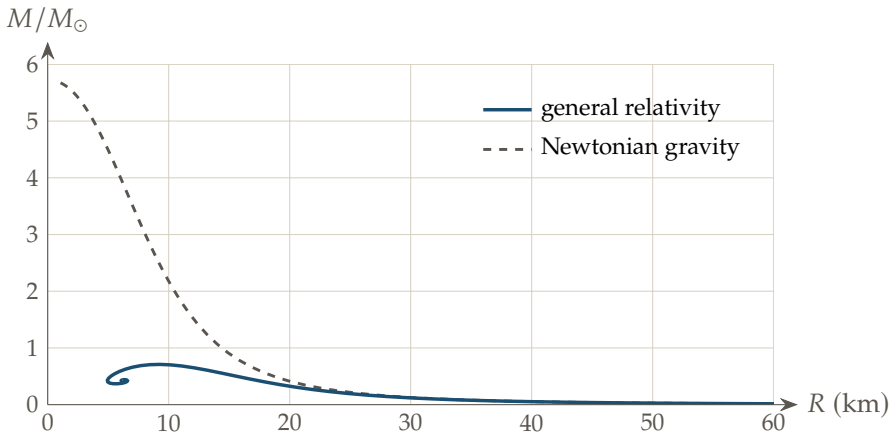


Figure 13.3: Mass against radius for stars of an ideal degenerate neutron gas, computed with the TOV equation (solid) and with Newtonian hydrostatic equilibrium (dashed). Each point is one central density.

Replacing the ideal-gas pressure with an equation of state from nuclear physics gives mass–radius curves that can be compared with measured neutron stars. We begin with the ideal gas to isolate the effect of relativistic gravity.

The procedure. A cold neutron star has no burning and no heat flow that matter for its structure, so only the mass and momentum equations of Table 13.2 are needed, with an equation of state $P(\rho)$.

1. Choose a central density ρ_c, so $P(0) = P(\rho_c)$ and $m(0) = 0$.
2. Integrate $\mathrm{d}m/\mathrm{d}r$ and the TOV equation outward until $P = 0$. That radius is R and $m(R)$ is M.
3. Repeat for a range of ρ_c. Each gives one point (R, M).

Both boundary conditions sit at the centre, unlike the two-ended problem of Section 12.4, so a single outward integration finishes each star. Along the curve, stars with $\mathrm{d}M/\mathrm{d}\rho_c > 0$ are stable. Past the maximum, where $\mathrm{d}M/\mathrm{d}\rho_c < 0$, a small squeeze grows instead of rebounding.

What the curves show. Figure 13.3 uses the exact pressure of an ideal degenerate neutron gas, and its total energy density as the source of gravity. At low mass, where the star is large and $GM/(Rc^2)$ small, the two curves agree. As the mass grows they part. The Newtonian curve climbs toward the Chandrasekhar-like limit for neutrons, about $5.7\,M_\odot$, the $\mu_e = 1$ value of

Eq. (13.3). The relativistic curve turns over at 0.71 $M_\odot$, the Oppenheimer–Volkoff limit, where $GM/(Rc^2) = 0.11$. General relativity has cut the maximum mass by a factor of eight. Pressure that gravitates defeats itself.

Real neutron stars. Pulsars in binaries are weighed with the timing methods of Chapter 6 (Section 6.18), and some are found at about 2 $M_\odot$, far above 0.71 $M_\odot$. Neutrons are not an ideal gas. Nuclear forces repel them strongly at the highest densities and stiffen the pressure, and the best equations of state give maximum masses of about 2 to 2.3 $M_\odot$. The maximum mass itself, and its derivation in general relativity, are taken here from the theory of compact objects, not derived again.

13.5 Black Holes

Above the maximum mass of a neutron star, no pressure can stop a collapse. By the reading of the TOV equation, every increase of pressure adds to the pull it fights.

The collapse continues until the matter has passed inside its own Schwarzschild radius, $r_s = 2GM/c^2$, 3 km for the Sun's mass and 30 km for ten solar masses, where the compactness reaches $\frac{1}{2}$ and nothing, not even light, comes back out. Stars that end their lives with cores well above the neutron-star limit leave black holes. Their properties belong to general relativity and are not repeated here.

13.6 Gravitational Waves from Compact Binaries

In Newtonian gravity, an isolated binary keeps the same orbital energy, $E = -Gm_1m_2/(2a)$ (Eq. (6.34)). General relativity adds a slow energy loss: the orbiting masses emit gravitational waves that carry energy away. The loss is negligible for the Earth and Sun, but it draws close pairs of neutron stars or black holes toward merger.

Think of two children holding hands and whirling round in the middle of a pool. Ripples run out from them to the edge and rock a toy boat floating there. Every ripple carries away a little energy, and that energy comes from the whirling children. A binary of compact stars is the pair of children, space is the water, and the ripples are **gravitational waves**, small changes in the distances between free bodies that travel outward at the speed of light. Einstein predicted them from general relativity in 1916 and 1918. The idea is that the energy they carry is taken from the orbit. The orbit shrinks, and a bound orbit that loses energy speeds up, just as a star that loses energy heats up (Section 10.3). So the waves rise in frequency and in strength until the two bodies merge. The rising tone is called a **chirp**.

The size of the effect depends strongly on the orbit. Earth radiates only 196 W in gravitational waves as it circles the Sun; at that rate its orbit would take 10^{23} years to shrink away. In GW150914, two black holes of 36 and 29 $M_\odot$ radiated about 3 $M_\odot c^2$ in a fraction of a second, reaching a peak power of about 4×10^{49} W. Figure 13.4 shows what a passing wave does to a ring of freely moving particles. We will first ask why an orbit can radiate at all, then follow how that radiation shrinks the orbit and raises the wave frequency.

13.6.1 Why a Quadrupole

The moments of the mass. An accelerated charge radiates through its changing electric dipole moment $\sum_k e_k \vec{x}_k$, the Larmor formula (15.1). Far from a source, its field can be sorted in the same way by the moments of the mass. The particles are labelled k, with masses m_k, positions $\vec{x}_k$ and speeds $\vec{v}_k$.

- The **monopole** is the total mass $\sum_k m_k$, which is conserved (P1). Its field never changes, so it cannot radiate.
- The **mass dipole** is $\sum_k m_k \vec{x}_k$. Its rate of change is $\sum_k m_k \vec{v}_k$, the total momentum, which is conserved (P1). So its second derivative is zero and it cannot radiate. For charges there is no such rule, because an electron and a proton have different ratios of charge to mass, and $\sum_k e_k \vec{x}_k$ can accelerate while the centre of mass does not.
- The other dipole, $\sum_k m_k \vec{x}_k \times \vec{v}_k$, is the total angular momentum, also conserved (P1).

The first moment that can change in an isolated system is the second one, the matrix

$$I_{ij} = \sum_k m_k \, x_{k,i} \, x_{k,j}, \qquad Q_{ij} = I_{ij} - \tfrac{1}{3}\,\delta_{ij} \sum_l I_{ll}, \tag{13.9}$$

where i, j run over the three directions. The trace $\sum_l I_{ll} = \sum_k m_k |\vec{x}_k|^2$ is the second moment I of Route B of the virial theorem (Section 10.2). A change of the trace alone is a sphere breathing in and out, and that does not radiate either. What radiates is Q_{ij}, the **mass quadrupole**, the part of I_{ij} with the trace removed.

The quadrupole formula. General relativity gives the power radiated by a slowly moving source,

$$\boxed{L_{\mathrm{GW}} = \frac{G}{5c^5} \sum_{i,j} \left\langle \ddot{Q}_{ij}^{\,2} \right\rangle,} \tag{13.10}$$

quoted here, with the average taken over several wave periods. A distance d away, along the z axis, the wave appears as the 2×2 matrix

$$\begin{pmatrix} h_+ & h_\times \\ h_\times & -h_+ \end{pmatrix} = \frac{2G}{c^4 d} \begin{pmatrix} \tfrac{1}{2}\left(\ddot{Q}_{xx} - \ddot{Q}_{yy}\right) & \ddot{Q}_{xy} \\ \ddot{Q}_{xy} & -\tfrac{1}{2}\left(\ddot{Q}_{xx} - \ddot{Q}_{yy}\right) \end{pmatrix}, \tag{13.11}$$

also quoted, with Q_{ij} taken at the earlier time $t - d/c$. Only the part across the line of sight appears, and with its trace removed. The two numbers h_+ and $h_\times$ are the **strains** of the two polarisations. They are always written with their subscripts, since a bare h is Planck's constant in Section 13.2.

The factor G/c^5 is tiny. Its inverse, $c^5/G = 3.6 \times 10^{52}\,\mathrm{W}$, is the natural unit of power for gravitational waves, and a source comes near it only when its parts move near the speed of light.

13.6.2 How the Wave Stretches Space

The ring of particles. Free particles sit on a ring in the xy plane, across the wave. A particle at (x, y) from the centre of the ring is moved by the wave to

$$\boxed{\begin{pmatrix} x' \\ y' \end{pmatrix} = \left[\mathbb{1} + \frac{1}{2} \begin{pmatrix} h_+ & h_\times \\ h_\times & -h_+ \end{pmatrix} \right] \begin{pmatrix} x \\ y \end{pmatrix},} \tag{13.12}$$

where the positions are measured by the travel time of light, the way a detector measures them. This too is quoted from general relativity. The wave acts as a tidal field that changes in time, like the pull of the Moon that stretches the Earth along one line and squeezes it across.

With $h_\times = 0$, a particle on the x axis moves from $(x, 0)$ to $\left((1 + \frac{1}{2}h_+)x,\ 0\right)$, and one on the y axis from $(0, y)$ to $\left(0,\ (1 - \frac{1}{2}h_+)y\right)$. One length grows by the fraction $\frac{1}{2}h_+$ and the length across it shrinks by the same fraction. The difference of two perpendicular lengths, divided by the length, is h_+. That is what the word strain means here. As h_+ swings between positive and negative values the ring becomes an ellipse along x, then a circle, then an ellipse along y (Figure 13.4, top). Particles along the line of sight are not moved, so the wave is transverse.

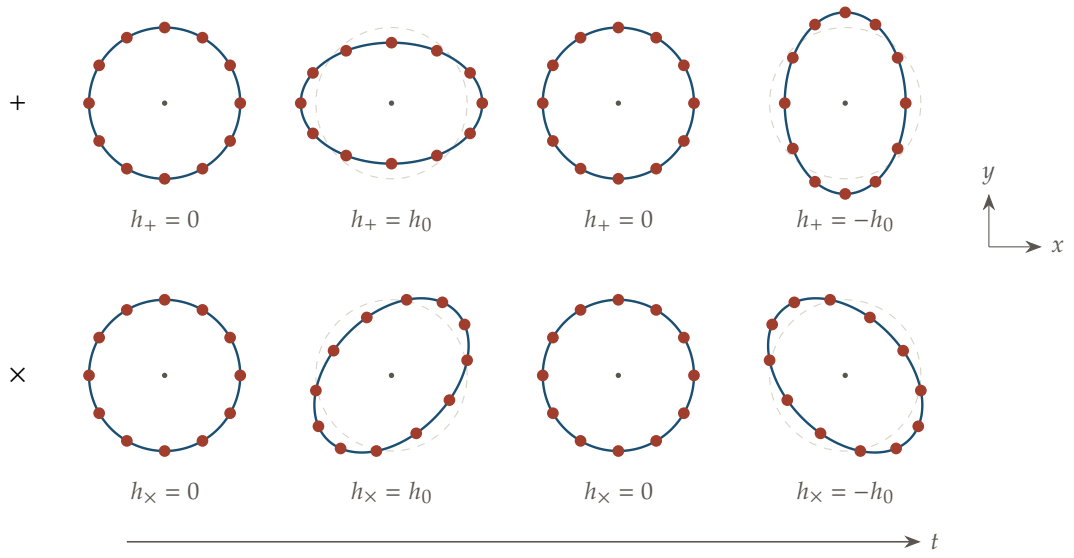


Figure 13.4: A ring of free particles, drawn from Eq. (13.12), at four moments of one wave period, for the + polarisation (top) and the × polarisation (bottom). The wave travels out of the page. The strain is drawn as $h_0 = 0.4$. The dashed circle is the ring with no wave.

Two polarisations, 45° apart. To compare the two patterns, we turn the + pattern through an angle ψ about the line of sight. A matrix that acts on positions turns as $R(\psi)\,(\cdot)\,R(\psi)^{\mathsf{T}}$, where $R(\psi)$ is the rotation matrix written out below, so

$$\begin{aligned} R(\psi)\begin{pmatrix} h_+ & 0 \\ 0 & -h_+ \end{pmatrix} R(\psi)^{\mathsf{T}} &= \begin{pmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{pmatrix}\begin{pmatrix} h_+ & 0 \\ 0 & -h_+ \end{pmatrix}\begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix} \\ &\overset{(1)}{=} h_+\begin{pmatrix} \cos\psi & \sin\psi \\ \sin\psi & -\cos\psi \end{pmatrix}\begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix} \\ &\overset{(2)}{=} h_+\begin{pmatrix} \cos 2\psi & \sin 2\psi \\ \sin 2\psi & -\cos 2\psi \end{pmatrix}. \end{aligned} \tag{13.13}$$

(1) Multiply the first two matrices.
(2) Multiply out, with $\cos^2\psi - \sin^2\psi = \cos 2\psi$ and $2\sin\psi\cos\psi = \sin 2\psi$.

At $\psi = 45°$ the result is $\begin{pmatrix} 0 & h_+ \\ h_+ & 0 \end{pmatrix}$, a pure × wave. At $\psi = 90°$ it is the + wave again with its sign reversed. So the two independent patterns are 45° apart, and the pattern comes back to itself after a turn of 180° (Figure 13.4, bottom). For light, the electric field is a vector, $R(\psi)\,(E, 0)^{\mathsf{T}} = E\,(\cos\psi, \sin\psi)^{\mathsf{T}}$. Its two independent polarisations are 90° apart, and it comes back to itself only after 360°. The angle doubles because the wave of gravity is a matrix, not a vector.

13.6.3 The Power of a Binary

The quadrupole of a circular orbit. Two stars of masses m_1 and m_2, with $M = m_1 + m_2$, move on a circle of radius a (the semi-major axis of Chapter 6) about each other with angular speed ω (the mean motion n of Chapter 6, renamed because n is a number density here). In the orbital plane xy, the separation is $\vec{r} = a(\cos\omega t,\ \sin\omega t,\ 0)$. About the centre of mass the two stars sit

at $\vec{x}_1 = (m_2/M)\vec{r}$ and $\vec{x}_2 = -(m_1/M)\vec{r}$ (Eq. (6.5)). Their second moment, from Eq. (13.9), is

$$I_{ij} = m_1\frac{m_2^2}{M^2}r_ir_j + m_2\frac{m_1^2}{M^2}r_ir_j = \frac{m_1m_2}{M}r_ir_j,$$
$$\begin{pmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{pmatrix} = \frac{m_1m_2a^2}{M}\begin{pmatrix} \cos^2\omega t & \cos\omega t\sin\omega t \\ \cos\omega t\sin\omega t & \sin^2\omega t \end{pmatrix} \overset{(1)}{=} \frac{m_1m_2a^2}{2M}\begin{pmatrix} 1+\cos 2\omega t & \sin 2\omega t \\ \sin 2\omega t & 1-\cos 2\omega t \end{pmatrix}, \tag{13.14}$$

(1) $\cos^2\omega t = \frac{1}{2}(1+\cos 2\omega t)$, $\sin^2\omega t = \frac{1}{2}(1-\cos 2\omega t)$, $\cos\omega t\sin\omega t = \frac{1}{2}\sin 2\omega t$.

and every entry with a z is zero. The trace is $m_1m_2a^2/M$, constant in time, so Q_{ij} and I_{ij} have the same time derivatives. Everything that changes goes as $2\omega t$. The wave has *twice* the orbital frequency, because the binary looks the same after half a turn,

$$f = \frac{2\omega}{2\pi} = \frac{\omega}{\pi}. \tag{13.15}$$

The power. Differentiating three times, with $\frac{d^3}{dt^3}\cos 2\omega t = 8\omega^3\sin 2\omega t$ and $\frac{d^3}{dt^3}\sin 2\omega t = -8\omega^3\cos 2\omega t$,

$$\begin{pmatrix} \dddot{Q}_{xx} & \dddot{Q}_{xy} \\ \dddot{Q}_{xy} & \dddot{Q}_{yy} \end{pmatrix} = \frac{4m_1m_2a^2\omega^3}{M}\begin{pmatrix} \sin 2\omega t & -\cos 2\omega t \\ -\cos 2\omega t & -\sin 2\omega t \end{pmatrix},$$
$$\sum_{i,j}\dddot{Q}_{ij}^2 \overset{(1)}{=} \left(\frac{4m_1m_2a^2\omega^3}{M}\right)^2\left(2\sin^2 2\omega t + 2\cos^2 2\omega t\right) = 32\left(\frac{m_1m_2}{M}\right)^2 a^4\omega^6,$$
$$L_{\text{GW}} \overset{(2)}{=} \frac{32}{5}\frac{G}{c^5}\left(\frac{m_1m_2}{M}\right)^2 a^4\omega^6 \overset{(3)}{=} \frac{32}{5}\frac{G^4m_1^2m_2^2M}{c^5a^5}. \tag{13.16}$$

(1) The four entries squared and added. The sum is the same at every moment, so no average is needed.
(2) The quadrupole formula (13.10).
(3) Kepler's third law, $\omega^2a^3 = GM$ (Eq. (6.50)), so $\omega^6 = G^3M^3/a^9$.

$$\boxed{L_{\text{GW}} = \frac{32}{5}\frac{G^4m_1^2m_2^2M}{c^5a^5} = \frac{32}{5}\frac{c^5}{G}\left(\frac{m_1m_2}{M^2}\right)^2\left(\frac{GM}{ac^2}\right)^5.} \tag{13.17}$$

The second form shows the compactness of Section 13.1, now of the orbit, to the fifth power. For the Earth and the Sun $GM/(ac^2) = 10^{-8}$ and the power is 196 W. When two compact stars are a few of their own radii apart, $GM/(ac^2)$ is a tenth or more and the power approaches c^5/G.

13.6.4 The Shrinking Orbit and the Chirp

The orbit shrinks. The energy carried off by the waves is taken from the orbit (P1), $\mathrm{d}E/\mathrm{d}t = -L_{\text{GW}}$. In one turn the orbit loses only a tiny part of its energy, so it stays close to a circle whose radius slowly falls (P6, as for the slow drifts of Section 6.16). With $E = -Gm_1m_2/(2a)$ from Eq. (6.34),

$$\frac{\mathrm{d}E}{\mathrm{d}t} \overset{(1)}{=} \frac{Gm_1m_2}{2a^2}\frac{\mathrm{d}a}{\mathrm{d}t} \overset{(2)}{=} -\frac{32}{5}\frac{G^4m_1^2m_2^2M}{c^5a^5}$$
$$\Longrightarrow \quad \frac{\mathrm{d}a}{\mathrm{d}t} = -\frac{64}{5}\frac{G^3m_1m_2M}{c^5a^3}. \tag{13.18}$$

(1) Differentiate $E = -Gm_1m_2/(2a)$, as in Eq. (6.158).

(2) $\mathrm{d}E/\mathrm{d}t = -L_{\mathrm{GW}}$ with Eq. (13.17).

$$\boxed{\frac{\mathrm{d}a}{\mathrm{d}t} = -\frac{64}{5}\,\frac{G^3 m_1 m_2 M}{c^5 a^3}.} \tag{13.19}$$

The shrinking speeds up as a^{-3}, so most of the time is spent far apart.

Time to merge. Separating the variables and integrating from the separation a_0 now to zero at the merger, a time t_{merge} later,

$$\begin{aligned} \int_{a_0}^{0} a^3\,da &= -\frac{64}{5}\,\frac{G^3 m_1 m_2 M}{c^5}\int_0^{t_{\mathrm{merge}}} dt \\ \Longrightarrow \qquad -\frac{a_0^4}{4} &= -\frac{64}{5}\,\frac{G^3 m_1 m_2 M}{c^5}\,t_{\mathrm{merge}}, \end{aligned} \tag{13.20}$$

$$\boxed{t_{\mathrm{merge}} = \frac{5}{256}\,\frac{c^5 a_0^4}{G^3 m_1 m_2 M}.} \tag{13.21}$$

The steep a_0^4 decides everything. The Earth and the Sun need 10^{23} years. Two neutron stars of $1.4\,M_\odot$ a solar radius apart need 3×10^7 years, short enough that such pairs, born in the Galaxy, merge within its lifetime.

The chirp. The frequency follows from the separation through Kepler's third law, $\omega = (GM)^{1/2}a^{-3/2}$. As a falls, ω rises,

$$\begin{aligned} \frac{\mathrm{d}\omega}{\mathrm{d}t} &\overset{(1)}{=} -\frac{3}{2}\,\frac{\omega}{a}\,\frac{\mathrm{d}a}{\mathrm{d}t} \overset{(2)}{=} \frac{96}{5}\,\frac{G^3 m_1 m_2 M}{c^5}\,\frac{\omega}{a^4} \\ &\overset{(3)}{=} \frac{96}{5}\,\frac{G^{5/3}}{c^5}\,\frac{m_1 m_2}{M^{1/3}}\,\omega^{11/3}. \end{aligned} \tag{13.22}$$

(1) Differentiate $\omega \propto a^{-3/2}$.
(2) Eq. (13.19).
(3) Kepler's third law (6.50) gives $a^4 = (GM)^{4/3}\omega^{-8/3}$.

The two masses enter only as $m_1 m_2/M^{1/3}$, a mass to the power $\frac{5}{3}$. Its $\frac{3}{5}$ power is the **chirp mass**,

$$\boxed{M_{\mathrm{chirp}} = \frac{(m_1 m_2)^{3/5}}{M^{1/5}}, \qquad \frac{\mathrm{d}\omega}{\mathrm{d}t} = \frac{96}{5}\,\frac{(GM_{\mathrm{chirp}})^{5/3}}{c^5}\,\omega^{11/3}.} \tag{13.23}$$

For two equal masses m, $M_{\mathrm{chirp}} = m/2^{1/5} = 0.87\,m$.

Weighing by listening. A detector measures the wave frequency $f = \omega/\pi$ (Eq. (13.15)) and how fast it rises. With $\omega = \pi f$ in Eq. (13.23),

$$\pi\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{96}{5}\,\frac{(GM_{\mathrm{chirp}})^{5/3}}{c^5}\,(\pi f)^{11/3} \Longrightarrow \boxed{M_{\mathrm{chirp}} = \frac{c^3}{G}\left[\frac{5}{96}\,\pi^{-8/3} f^{-11/3}\,\frac{\mathrm{d}f}{\mathrm{d}t}\right]^{3/5}.} \tag{13.24}$$

Only c and G enter. Neither the distance nor the orientation of the binary is needed. This is how the masses of merging black holes are found. The two masses separately need the smaller terms of higher order in v/c that general relativity adds to Eq. (13.23).

Time left, a second route. Eq. (13.23) gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\omega^{-8/3}\right) = -\frac{8}{3}\,\omega^{-11/3}\,\frac{\mathrm{d}\omega}{\mathrm{d}t} = -\frac{256}{5}\,\frac{(GM_{\text{chirp}})^{5/3}}{c^5},$$

a constant. So $\omega^{-8/3}$ falls in a straight line and reaches zero at the merger, a time

$$t_{\text{merge}} = \frac{5}{256}\,\frac{c^5}{(GM_{\text{chirp}})^{5/3}\,\omega^{8/3}} = \frac{3}{8}\,f\left(\frac{\mathrm{d}f}{\mathrm{d}t}\right)^{-1} \tag{13.25}$$

later. Putting $\omega^{8/3} = (GM)^{4/3}/a_0^4$ and $M_{\text{chirp}}^{5/3}M^{4/3} = m_1m_2M$ gives Eq. (13.21) again.

The strain of a binary. For an observer on the z axis, who sees the orbit face on, differentiating I_{ij} twice and using Eq. (13.11),

$$\begin{pmatrix} \ddot{Q}_{xx} & \ddot{Q}_{xy} \\ \ddot{Q}_{xy} & \ddot{Q}_{yy} \end{pmatrix} = -\frac{2m_1m_2a^2\omega^2}{M}\begin{pmatrix} \cos 2\omega t & \sin 2\omega t \\ \sin 2\omega t & -\cos 2\omega t \end{pmatrix},$$
$$\begin{pmatrix} h_+ & h_\times \\ h_\times & -h_+ \end{pmatrix} \overset{(1)}{=} -h_0\begin{pmatrix} \cos 2\omega t & \sin 2\omega t \\ \sin 2\omega t & -\cos 2\omega t \end{pmatrix}, \qquad h_0 = \frac{4Gm_1m_2a^2\omega^2}{c^4dM}, \tag{13.26}$$

(1) Eq. (13.11). This 2×2 block already has zero trace, so $\frac{1}{2}(\ddot{Q}_{xx} - \ddot{Q}_{yy}) = \ddot{Q}_{xx}$.

with the time t read as $t - d/c$. By Eq. (13.13) this is a + pattern of size h_0 turned through the angle $\psi = \omega t$ (and the sign reversed). The stretching axis follows the two stars round. With $a^2\omega^2 = (GM)^{2/3}\omega^{2/3}$ from Kepler's third law and $\omega = \pi f$,

$$\boxed{h_0 = \frac{4\,(GM_{\text{chirp}})^{5/3}\,(\pi f)^{2/3}}{c^4 d}.} \tag{13.27}$$

An orbit seen at an angle gives a smaller strain. Seen edge on, h_+ is half as large and $h_\times$ is zero. As the frequency rises the strain grows as $f^{2/3}$. Figure 13.5 shows the result, a wave whose frequency and size both grow.

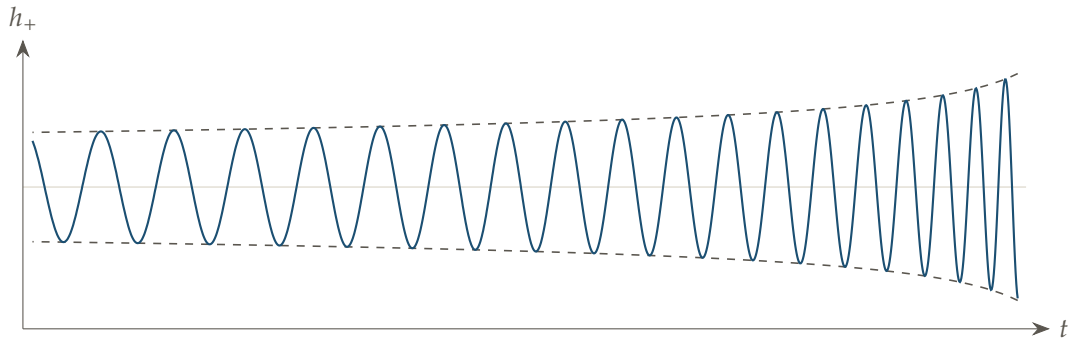


Figure 13.5: The strain h_+ of a circular binary seen face on, from Eqs. (13.23) and (13.27), while the orbital frequency triples. The curve ends at the last stable orbit of an equal-mass pair. The dashed lines are $\pm h_0$.

GW150914 in numbers. The first wave detected, on 14 September 2015, came from black holes of about 36 and 29 $M_\odot$, so $M_{\text{chirp}} = 28\,M_\odot$, at $d = 410$ Mpc. Its frequency rose from 35 to about 150 Hz in 0.2 s. The leading-order formulas hold during the slow spiral, which for this

pair ends near 70 Hz (Eq. (13.30)). The last rise to 150 Hz is the plunge and merger, where they no longer hold. With $GM_\odot/c^2 = 1477\,\mathrm{m}$, Eq. (13.27) at the end of the spiral, $f = 70\,\mathrm{Hz}$, gives

$$h_0 = \frac{4\,(28 \times 1477\,\mathrm{m})^{5/3}\,(\pi \times 70\,\mathrm{Hz}/c)^{2/3}}{410 \times 3.086 \times 10^{22}\,\mathrm{m}} = 1.3 \times 10^{-21}, \tag{13.28}$$

against a measured peak of 1.0×10^{-21}. The orbit was not seen face on, and the detector does not respond fully to every direction. At 35 Hz, Kepler's third law with $\omega = \pi f$ puts the two bodies 890 km apart, under five times their combined Schwarzschild radius, $2GM/c^2 = 192\,\mathrm{km}$, and the spiral goes on down to $6GM/c^2 = 576\,\mathrm{km}$. Nothing but black holes is that compact for its mass.

13.6.5 The First Evidence, a Pulsar Binary

In 1974 Hulse and Taylor found the pulsar PSR B1913+16 in a binary with another neutron star, with an orbital period $P_{\mathrm{orb}} = 7.75$ hours and an eccentricity $e = 0.617$. The arrival times of its pulses carry the Rømer delay of its orbit (Section 6.18), and with the relativistic effects of timing, quoted from general relativity, they give both masses, 1.438 and $1.390\,M_\odot$. With these masses fixed, the decay of the orbit is a prediction with nothing left to adjust.

The predicted decay. For a circular orbit, $P_{\mathrm{orb}} \propto a^{3/2}$ by Kepler's third law, so

$$\frac{1}{P_{\mathrm{orb}}}\frac{\mathrm{d}P_{\mathrm{orb}}}{\mathrm{d}t} \overset{(1)}{=} \frac{3}{2}\frac{1}{a}\frac{\mathrm{d}a}{\mathrm{d}t} \overset{(2)}{=} -\frac{96}{5}\frac{G^3 m_1 m_2 M}{c^5 a^4}, \tag{13.29}$$

(1) Differentiate $P_{\mathrm{orb}} \propto a^{3/2}$.
(2) Eq. (13.19).

With $a = 1.95 \times 10^6\,\mathrm{km}$ from Kepler's third law, this gives $\mathrm{d}P_{\mathrm{orb}}/\mathrm{d}t = -2.03 \times 10^{-13}$. The real orbit is eccentric. The two stars radiate most strongly at their closest approach, and general relativity (Peters, 1964) multiplies the circular rate by

$$\frac{1 + \frac{73}{24}e^2 + \frac{37}{96}e^4}{(1-e^2)^{7/2}} = 11.86 \qquad (e = 0.617),$$

quoted here. The prediction is $\mathrm{d}P_{\mathrm{orb}}/\mathrm{d}t = -2.40 \times 10^{-12}$. The measured value, after a small correction for the acceleration of the pulsar in the Galaxy, agrees within a fraction of a percent. The orbit shrinks by about 3.5 m a year, and the period by 76 μs a year. The effect adds up. Over 30 years the closest approach comes 39 s earlier than a fixed period would predict. This was the first evidence that gravitational waves exist, recognised by the Nobel Prize in 1993, and the two stars will merge in about 300 million years. Many binary pulsars now test Eq. (13.10) in the same way.

13.6.6 Detectors

Laser interferometers. A detector on the ground is a Michelson interferometer with two arms at right angles. A laser beam is split, sent down both arms to mirrors that hang as pendulums and move freely along the arms, and the two returning beams are made to interfere. With the arms along x and y, both of length 4 km, Eq. (13.12) moves the end mirrors to

$$\begin{pmatrix} (1 + \frac{1}{2}h_+) \times 4\,\mathrm{km} \\ \frac{1}{2}h_\times \times 4\,\mathrm{km} \end{pmatrix}, \qquad \begin{pmatrix} \frac{1}{2}h_\times \times 4\,\mathrm{km} \\ (1 - \frac{1}{2}h_+) \times 4\,\mathrm{km} \end{pmatrix}.$$

To first order in the strain, the sideways shifts do not change the lengths. The x arm grows by $\frac{1}{2}h_+ \times 4\,\mathrm{km}$ and the y arm shrinks by the same amount, so the difference of the two arms

is $h_+ \times 4\,\mathrm{km}$. For $h_+ = 10^{-21}$ it is $4 \times 10^{-18}\,\mathrm{m}$, about 1/400 of the diameter of a proton. The interferometer reads the difference, so it is blind to the × wave along its own arms. Detectors with arms turned to other angles see other combinations of h_+ and $h_\times$. The light bounces back and forth hundreds of times in each arm, which multiplies the effect. Below about 10 Hz the shaking of the ground swamps the signal, so these detectors hear from about 10 Hz to a few kilohertz.

The two LIGO detectors, at Hanford in Washington and Livingston in Louisiana, have 4 km arms and lie 3000 km apart, 10 ms for light. A real wave reaches both within that time, and the delay between them helps fix the direction of the source. Virgo in Italy and KAGRA in Japan have 3 km arms. KAGRA is underground and its mirrors are cooled. Planned detectors push further. Cosmic Explorer in the United States would have 40 km arms, the Einstein Telescope in Europe a triangle of 10 km arms underground, and NEMO in Australia aims at the kilohertz waves from the hot neutron-star matter left after a merger.

The mass sets the band. The chirp ends near the last stable circular orbit, at $r = 6GM/c^2$ (quoted from general relativity). With Kepler's third law and $f = \omega/\pi$, the frequency there is

$$f_{\mathrm{end}} = \frac{1}{\pi}\sqrt{\frac{GM}{(6GM/c^2)^3}} = \frac{c^3}{6^{3/2}\pi GM} = 4.4\,\mathrm{kHz} \times \frac{M_\odot}{M}. \tag{13.30}$$

This marks the end of the slow spiral. The plunge and merger that follow carry the frequency up by a factor of two or so. Heavier binaries sing lower (P6). A pair of neutron stars of $2.8\,M_\odot$ in all ends near 1.6 kHz, and GW150914 near 70 Hz (its merger then reached 150 Hz), both in the band of the ground detectors. Black holes of $10^6\,M_\odot$ in the centres of galaxies end at 4 mHz, and pairs of $10^9\,M_\odot$ give nanohertz waves for thousands of years before they merge. Each band needs its own detector.

LISA. The Laser Interferometer Space Antenna is three spacecraft at the corners of a triangle with sides of 2.5 million kilometres, trailing the Earth in its orbit round the Sun. Lasers between them measure changes in the distances between test masses floating freely inside. A strain of 10^{-20} changes a side by $2.5 \times 10^{-11}\,\mathrm{m}$. Far from the shaking of the ground, LISA hears the millihertz band, the mergers of black holes of 10^4 to $10^7\,M_\odot$ and many close pairs of white dwarfs in the Galaxy. It is planned for launch in the 2030s.

Pulsar timing arrays. Millisecond pulsars are clocks steady enough that their pulse arrival times, corrected for the motion of the Earth and for the pulsar's own orbit (Section 6.18), can be predicted to better than a microsecond over years. A wave passing between the Earth and a pulsar changes the travel time of the pulses by a fraction of order h_+, and the change builds up over about one wave period. The timing shift is of order $h_+/(2\pi f)$, 5 ns for $h_+ = 10^{-15}$ at one cycle per year. Many pairs of supermassive black holes add up to a random background of such waves. What marks it out is a definite pattern, the Hellings–Downs curve, in how the timing shifts of two pulsars are correlated, depending on the angle between them on the sky. In 2023 several timing arrays reported evidence for this background, at a strain of about 2×10^{-15} at one cycle per year. The evidence is still being firmed up.

13.6.7 GW170817, Light and Gravity Together

On 17 August 2017 LIGO and Virgo recorded a chirp lasting about a minute and a half in their band, with a chirp mass of $1.19\,M_\odot$. That puts both masses between about 1.2 and $1.6\,M_\odot$, a pair of neutron stars. Then 1.7 s after the merger the Fermi satellite recorded a short burst of gamma rays from the same part of the sky. Eleven hours later telescopes found a new source in

the galaxy NGC 4993, about 40 Mpc away. It was a **kilonova**. Neutron-rich matter thrown out in the merger forms heavy nuclei, whose radioactive decay makes the debris glow for days, like the nickel of a type Ia supernova. One event was seen in gravitational waves, gamma rays, and visible, infrared, ultraviolet, X-ray and radio light. Such events are called multi-messenger.

The speed of gravity. The waves and the gamma rays travelled for $d/c = 40\,\mathrm{Mpc}/c = 4.1{\times}10^{15}\,\mathrm{s}$ and arrived 1.7 s apart. Even if the whole delay came from a difference of speed, with v_{GW} the speed of the waves,

$$\frac{|v_{\mathrm{GW}} - c|}{c} \lesssim \frac{1.7\,\mathrm{s}}{4.1\times10^{15}\,\mathrm{s}} = 4\times10^{-16}.$$

Allowing for a delay of a few seconds in the launch of the burst, gravity travels at the speed of light to about one part in 10^{15}.

A standard siren. Eqs. (13.24) and (13.27) give the distance with no ladder at all. The frequency and its rate give M_{chirp}, and then the measured strain gives d. A merger is a source of known loudness, a **standard siren**. Over cosmological distances the waves are stretched by $1+z$, so f and $\mathrm{d}f/\mathrm{d}t$ give $(1+z)M_{\mathrm{chirp}}$, and the strain then gives the luminosity distance. The redshift itself must come from the host galaxy. For GW170817 the waves alone gave a distance of 44 Mpc, and NGC 4993 recedes at 3000 km s^{-1}, so Hubble's law (Eq. (16.2)) gives $H_0 \approx 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, uncertain by about 15%, mostly because the tilt of the orbit is poorly known. Many such sirens at larger distances will measure the expansion history as well (Section 16.7).

13.6.8 What Has Been Seen

From 2015 to 2020 the ground detectors recorded about ninety mergers, and the run that began in 2023 has added more than two hundred candidates. Most are pairs of black holes, of about 5 to 100 $M_\odot$ each, heavier on average than the black holes found in X-ray binaries. Two are pairs of neutron stars, GW170817 and GW190425. A few, such as GW200105 and GW200115, are a neutron star with a black hole. In each merger of black holes a few percent of the total mass leaves as waves, 3 $M_\odot$ of 65 $M_\odot$ for GW150914, and what is left is one spinning black hole.

13.7 Problem Set: Degenerate and Compact Stars

13.7.1 Problem 1: Gravitational Redshifts

Problem Statement: Light leaving the surface of a static star is redshifted by $1+z = (1-2GM/Rc^2)^{-1/2}$. Find z for (a) the Sun, (b) Sirius B (1.02 $M_\odot$, 5840 km), expressed as a Doppler velocity cz, and (c) a neutron star of 1.4 $M_\odot$ and 12 km.

Essential Ideas:

- For small compactness, $z \approx GM/(Rc^2)$.

Detailed Solutions:

(a) **Sun**.

$$\boxed{z \simeq 2.1\times10^{-6}, \qquad cz \simeq 0.64\,\mathrm{km\,s^{-1}}.}$$

(b) **Sirius B**.

$$\boxed{z \simeq 2.6\times10^{-4}, \qquad cz \simeq 77\,\mathrm{km\,s^{-1}}.}$$

The measured shift of Sirius B's lines, about 80 km s^{-1}, was one of the first confirmations that white dwarfs are as dense as the theory requires.

(c) **Neutron star**. $2GM/Rc^2 = 0.345$, so $1 + z = (0.655)^{-1/2}$.

$$\boxed{z \simeq 0.24.}$$

13.7.2 Problem 2: The Chandrasekhar Mass for Other Compositions

Problem Statement: Using Eq. (13.3), find the limiting mass of a degenerate core made of (a) helium, carbon or oxygen, (b) iron (^{56}Fe, 26 electrons), and (c) pure hydrogen, if such a white dwarf existed.

Essential Ideas:

- $M_{\rm Ch} = 5.83\, M_\odot/\mu_e^2$, where μ_e is the mass per electron in atomic mass units.

Detailed Solutions:

(a) **Helium, carbon, oxygen**. $\mu_e = 2$.

$$\boxed{M_{\rm Ch} \simeq 1.46\, M_\odot.}$$

(b) **Iron**. $\mu_e = 56/26 = 2.15$.

$$\boxed{M_{\rm Ch} \simeq 1.26\, M_\odot.}$$

An iron core grows in a massive star until it nears this mass, and then collapses.

(c) **Hydrogen**. $\mu_e = 1$.

$$\boxed{M_{\rm Ch} \simeq 5.8\, M_\odot.}$$

13.7.3 Problem 3: The Energy of a Supernova

Problem Statement: A neutron star of $1.4\, M_\odot$ and radius 12 km forms from the collapse of a core. (a) Estimate the energy released, using the gravitational energy of a uniform sphere. (b) What fraction of the rest energy is this? (c) Compare with all the light the Sun gives off over 10^{10} years.

Essential Ideas:

- $|U| = \frac{3}{5}GM^2/R$ (Eq. (10.14)). The core started far larger, so almost all of $|U|$ is released.

Detailed Solutions:

(a) **Energy**.

$$\boxed{|U| \simeq \frac{3}{5}\,\frac{6.674\times10^{-11}\times(2.785\times10^{30})^2}{1.2\times10^4}\,\text{J} \simeq 3\times10^{46}\,\text{J}.}$$

(b) **Fraction**. $Mc^2 = 2.5\times10^{47}$ J.

$$\boxed{|U|/Mc^2 \simeq 0.10.}$$

(c) **Against the Sun**. $L_\odot \times 10^{10}\,\text{yr} = 1.2\times10^{44}$ J. A core collapse releases about two hundred times the Sun's entire lifetime output, within seconds. About 99% leaves as neutrinos, which diffuse out of the hot core (P5) in about ten seconds, as was seen from supernova 1987A.

13.7.4 Problem 4: Hotter at the Bottom

Problem Statement: (a) In thermal equilibrium, by how much does the temperature at the bottom of a 10 km column of the Earth's atmosphere exceed that at the top, according to Eq. (13.8)? (b) Across a neutron star, with Φ/c^2 changing by about 0.2, how large is the effect?

Essential Ideas:

- For a small change of potential, $\Delta T/T \approx -\Delta\Phi/c^2$, with $\Delta\Phi = g\,\Delta h$ near a planet's surface.

Detailed Solutions:

(a) **Earth**. $\Delta\Phi/c^2 = 9.81 \times 10^4/(9.0 \times 10^{16})$.

$$\boxed{\Delta T/T \simeq 1.1 \times 10^{-12}.}$$

Utterly negligible. The real temperature difference in the air comes from convection (Problem 12.8.5).

(b) **Neutron star**.

$$\boxed{\Delta T/T \sim 0.2.}$$

Models of cooling neutron stars must use $T\,e^{\Phi/c^2}$, not T, as the quantity that evens out.

13.7.5 Problem 5: Weighing Two Black Holes by Their Chirp

Problem Statement: A chirp like GW150914 passes $f = 50\,\mathrm{Hz}$ while its frequency rises at $\mathrm{d}f/\mathrm{d}t = 290\,\mathrm{Hz\,s^{-1}}$. (a) Find the chirp mass. (b) If the two masses are equal, what is each? (c) How long is left before the merger? (d) How far apart are the two bodies at that moment, compared with $2GM/c^2$? (The chirp mass found this way is $(1+z)M_{\mathrm{chirp}}$, Section 13.6.7. For a source at $z = 0.09$, as GW150914 was, it is about 9% above the true $28\,M_\odot$. Ignore the difference here.)

Essential Ideas:

- The chirp mass from f and $\mathrm{d}f/\mathrm{d}t$, Eq. (13.24), and the time left, Eq. (13.25).
- The orbital angular speed is $\omega = \pi f$ (Eq. (13.15)), and Kepler's third law gives the separation.

Detailed Solutions:

(a) **Chirp mass**. $\frac{5}{96}\pi^{-8/3} f^{-11/3}\,\mathrm{d}f/\mathrm{d}t = 4.20 \times 10^{-7}\,\mathrm{s}^{5/3}$, whose $\frac{3}{5}$ power is $1.49 \times 10^{-4}\,\mathrm{s}$. Dividing by $GM_\odot/c^3 = 4.93 \times 10^{-6}\,\mathrm{s}$,

$$\boxed{M_{\mathrm{chirp}} \simeq 30\,M_\odot.}$$

(b) **Equal masses**. $M_{\mathrm{chirp}} = m/2^{1/5}$, so

$$\boxed{m = 2^{1/5} M_{\mathrm{chirp}} \simeq 35\,M_\odot, \qquad M \simeq 70\,M_\odot.}$$

(c) **Time left**.

$$\boxed{t_{\mathrm{merge}} = \frac{3}{8} \times \frac{50}{290}\,\mathrm{s} \simeq 0.065\,\mathrm{s}.}$$

(d) **Separation**. $\omega = \pi \times 50\,\mathrm{s^{-1}} = 157\,\mathrm{s^{-1}}$ and $a = (GM/\omega^2)^{1/3}$.

$$\boxed{a \simeq 720\,\mathrm{km} \simeq 3.5 \times \frac{2GM}{c^2}.}$$

Two bodies of $35\,M_\odot$ this close, still orbiting, must be smaller than 360 km each. Only black holes fit. The last stable orbit of Section 13.6.6 is reached at 63 Hz, a few orbits later.

Chapter 14

Galaxies and the Motions of Stars

A galaxy is a gravitationally bound collection of stars, gas and dust. Some hold only a thousand stars across about 10^3 pc; others hold 10^{12} stars across about 10^5 pc. Their masses range from roughly 10^5 to $10^{13}\,M_\odot$. Galaxies themselves gather into groups and clusters. At each scale, the motions of the visible members tell us how much gravity holds the system together.

We will use orbital speeds to weigh rotating galaxies and random speeds to weigh systems without orderly rotation. Both methods repeatedly find more mass than the visible stars and gas provide. This unseen mass also shapes the motions of galaxies in clusters. Before weighing those systems, we begin with the Galaxy around us and ask what its stars reveal about its structure.

14.1 Our Galaxy

Galactic coordinates locate objects on the sky (Section 2.8), and the 21 cm line traces gas through the Milky Way's dust (Section 8.4.1). But we observe the Galaxy from inside its disk. To learn its three-dimensional shape and mass, we must use the motions of its stars and gas.

Throw a ball straight up. It climbs until gravity has taken away its speed, and then it falls back. Thrown at the same speed on the Moon, it climbs six times higher, because the Moon pulls six times more weakly. Without any scale, the height reached at a given speed weighs the body underneath. The stars of the Galactic disk do the same. Each one bobs up and down through the plane as it circles the centre, at a few tens of kilometres per second. The disk pulls it back, and how far it rises before turning measures that pull. The pull of a wide flat layer is set by its mass per unit area. So the thickness of the disk weighs the disk.

For a constant pull g, a ball thrown upward at speed v rises to $v^2/2g$. At $10\,\mathrm{m\,s^{-1}}$, that is 5.1 m on Earth and 31 m on the Moon, where $g = 1.62\,\mathrm{m\,s^{-2}}$. Near the Sun, the disk's pull is about $6\times10^{-11}\,\mathrm{m\,s^{-2}}$ (derived below). A star crossing the plane at $20\,\mathrm{km\,s^{-1}}$ therefore reaches roughly 100 pc before returning. Measuring these vertical motions lets us infer the disk's mass per unit area.

The parts. The Galaxy has four parts, listed in Table 14.1. The **disk** holds most of the stars, nearly all the gas and dust, and the spiral arms where new stars form. The **bulge**, with a bar, fills the central few kiloparsecs with old stars. The **stellar halo** is a sparse round cloud of old stars poor in metals, with about 150 globular clusters. The **dark halo** is known only from its pull (Section 14.2), and it holds most of the mass. Seen from outside, the Galaxy would be a barred spiral of type SBbc in the sequence of Section 14.5, a bar across the bulge with fairly open arms. The Sun lies in the disk at $R_0 = 8.2$ kpc from the centre, about 20 pc north of the plane, and it circles at $v_0 \approx 230\,\mathrm{km\,s^{-1}}$, the speed of the local standard of rest of Section 3.6.

Part	Size	Mass ($M_\odot$)	Contents
Thin disk	scale length 2.6 kpc, scale height 0.3 kpc	stars 3.5×10^{10}, gas 1×10^{10}	young and old stars rich in metals, gas, dust, spiral arms
Thick disk	scale height 0.9 kpc	6×10^9	old stars poorer in metals
Bulge and bar	central 2–3 kpc	1.5×10^{10}	old stars
Stellar halo	out to 100 kpc and more	1×10^9	old metal-poor stars, globular clusters
Hot corona	out to 100 kpc and more	a few 10^{10}, uncertain	ionised gas near 2×10^6 K
Dark halo	out to 200–300 kpc	1×10^{12}	unseen
Black hole	inside 120 AU	4×10^6	Sagittarius A*

Table 14.1: The parts of the Milky Way, with rounded sizes and masses.

The picture took a long time to find. Herschel and later Kapteyn counted stars in every direction and placed the Sun near the middle of a small flattened system. They did not know about dust. Extinction (Chapter 4) hides everything beyond a few kiloparsecs in the plane, so in every direction the counts saw the same nearby patch. Shapley used the globular clusters instead. They lie mostly out of the plane, above the dust, and they form a round swarm whose centre is far away in Sagittarius. That centre is the centre of the Galaxy. The 21 cm line passes through dust and maps the gas over the whole disk. Infrared light also passes through, and shows the bulge.

The mass at the centre. The radio source Sagittarius A* sits at the zero point of Galactic coordinates (Section 2.8). Stars close to it move on Kepler orbits that infrared cameras have followed for decades. The star S2 has a period of 16.05 years and a semi-major axis of 0.1255″ on the sky. To weigh what it orbits, we use Kepler's third law $n^2 a^3 = GM$ (Eq. (6.50)) with the mean motion $n = 2\pi/P$ for a period P, and divide by the same law for the Earth round the Sun,

$$\frac{M}{M_\odot} \overset{(1)}{=} \frac{(2\pi/P)^2\, a^3}{(2\pi/1\,\mathrm{yr})^2\,(1\,\mathrm{AU})^3} = \left(\frac{a}{1\,\mathrm{AU}}\right)^3 \left(\frac{P}{1\,\mathrm{yr}}\right)^{-2} \overset{(2)}{=} \frac{(0.1255 \times 8180)^3}{16.05^2} = 4.2 \times 10^6. \tag{14.1}$$

(1) Eq. (6.50) for S2, divided by the same equation for the Earth, $GM_\odot = (2\pi/1\,\mathrm{yr})^2 (1\,\mathrm{AU})^3$.
(2) An angle of 1″ at 1 pc spans 1 AU (Eq. (3.2)), so 0.1255″ at 8180 pc is 1027 AU.

S2 comes within 120 AU of the centre. Four million solar masses inside so small a region, giving out very little light, can only be a black hole. The same orbit fixes R_0 as well. The Doppler shift of S2 gives its speed in kilometres per second and its motion on the sky gives the same speed in seconds of arc per year, and their ratio is the distance, 8.18 kpc.

In 2022 the Event Horizon Telescope, radio dishes across the Earth combined into one interferometer (Chapter 5) working at 1.3 mm, published an image of Sagittarius A*. It shows a bright ring about 52 micro-arcseconds across around a dark centre. General relativity says a black hole of mass M captures every ray aimed within $\sqrt{27}\,GM/c^2$ of it, so it looks like a dark disk of that radius against the glowing gas behind it (quoted here, not derived). For the mass

found from S2,

$$\frac{2\sqrt{27}\,GM/c^2}{R_0} \overset{(1)}{=} \frac{10.39 \times 6.674 \times 10^{-11} \times 4.2 \times 10^6 \times 1.989 \times 10^{30}}{(2.998 \times 10^8)^2 \times 8180 \times 3.086 \times 10^{16}} = 2.55 \times 10^{-10}$$
$$\overset{(2)}{=} 53 \text{ micro-arcseconds.} \tag{14.2}$$

(1) $2\sqrt{27} = 10.39$, $M = 4.2 \times 10^6\, M_\odot$, $R_0 = 8.18\,\text{kpc}$. Here $GM/c^2 = 6.2 \times 10^9\,\text{m}$, or 0.041 AU.
(2) $1\,\text{rad} = 2.063 \times 10^{11}$ micro-arcseconds.

The ring has the size that the mass from S2 predicts. Two methods that share nothing, star orbits a thousand AU out and light bent within a fraction of an AU, find the same black hole.

The pull of a flat layer. Near the Sun the disk is about 0.3 kpc thick and changes over a few kiloparsecs along the plane, so locally it is a wide flat layer. Let Σ be its mass per unit area and let z be the height above its middle. By symmetry the pull points toward the plane and depends only on $|z|$. Take a box of face area A with faces at $+z$ and $-z$, both outside most of the layer, so that it holds the mass ΣA. Gauss's law (9.14) gives

$$-4\pi G\,\Sigma A \overset{(1)}{=} \oint \vec{g} \cdot d\vec{A}$$
$$\overset{(2)}{=} -g(z)\,A - g(z)\,A$$
$$\Longrightarrow \quad g(z) = 2\pi G\Sigma. \tag{14.3}$$

(1) Eq. (9.14) with the mass ΣA inside the box (P2).
(2) On the top face $\vec{g}$ points down and $d\vec{A}$ points up, and on the bottom face the reverse. On the sides $\vec{g}$ lies along the face and gives nothing.

$$\boxed{g = 2\pi G\Sigma} \tag{14.4}$$

Above the layer the pull does not weaken with height, just as the field of a charged plane does not. For $\Sigma = 70\,M_\odot\,\text{pc}^{-2}$ it is $6.1 \times 10^{-11}\,\text{m}\,\text{s}^{-2}$.

A gas of stars above the layer. Now follow one population of stars, say old dwarfs of one type, with mass density $\rho(z)$, and let most of them lie above most of the layer's mass. Let $\sigma_z^2 = \langle v_z^2 \rangle$ be the mean square of their vertical speed at height z. The momentum flux (9.9) counted the momentum that particles carry across a surface as they move through it. No collisions entered that count, so it holds for stars. Its zz entry is $\rho \langle v_z^2 \rangle = \rho\sigma_z^2$ when there is no mean vertical flow. In a steady flat layer only derivatives along z survive, and the z component of Eq. (9.10) reads

$$0 \overset{(1)}{=} \frac{\partial(\rho \langle v_z \rangle)}{\partial t} + \frac{\mathrm{d}}{\mathrm{d}z}\Pi_{zz} - \rho\, g_z$$
$$\overset{(2)}{=} \frac{\mathrm{d}}{\mathrm{d}z}\left(\rho\sigma_z^2\right) + 2\pi G\Sigma\,\rho. \tag{14.5}$$

(1) Eq. (9.10), z component, with no change along x or y (P1).
(2) No mean vertical motion and no change in time, $\Pi_{zz} = \rho\sigma_z^2$, and $g_z = -2\pi G\Sigma$ above the layer by Eq. (14.4).

If σ_z is the same at every height, this is $\mathrm{d}\rho/\mathrm{d}z = -\rho/h_z$, and

$$\boxed{\rho(z) \propto e^{-z/h_z}, \qquad h_z = \frac{\sigma_z^2}{2\pi G\Sigma}.} \tag{14.6}$$

This is the scale height of an atmosphere, Eq. (10.4), with $k_B T/(\mu m_H)$ replaced by σ_z^2 and the surface gravity by $2\pi G\Sigma$. Stars with larger vertical speeds make a thicker layer, and a heavier disk makes a thinner one. For $\sigma_z = 20\,\mathrm{km\,s^{-1}}$ and $\Sigma = 70\,M_\odot\,\mathrm{pc^{-2}}$, $h_z = 210\,\mathrm{pc}$, close to the 300 pc of the old thin disk.

Weighing the disk. Turned around, $\Sigma = \sigma_z^2/(2\pi G h_z)$. Counting stars of one type at increasing heights gives h_z, and their spectra give σ_z. Kuijken and Gilmore did this for about five hundred K dwarfs toward the south Galactic pole and found $\Sigma \approx 70\,M_\odot\,\mathrm{pc^{-2}}$ within 1.1 kpc of the plane. The stars and gas that can be seen near the Sun add up to about $50\,M_\odot\,\mathrm{pc^{-2}}$. The difference is about what a round dark halo puts into a slab of that thickness (Problem 14.8.1). Near the Sun the dark matter is spread through the halo, not packed into the disk.

A thin and a thick disk. Star counts toward the Galactic poles do not follow one exponential. They follow two, with scale heights of about 0.3 kpc and 0.9 kpc (Table 14.1). About 95 per cent of the disk stars near the Sun belong to the **thin disk**. It holds the gas and the dust, it still forms stars, and its stars span a wide range of ages and are rich in metals. The **thick disk** holds stars older than about ten billion years and poorer in metals, and none form there now. By Eq. (14.6) a layer three times thicker in the same pull has vertical speeds $\sqrt{3}$ times larger, and the thick-disk stars do move faster up and down, by about 40 against $20\,\mathrm{km\,s^{-1}}$ (Problem 14.8.2). Old stars have had the longest time to be shaken by clouds and arms (Section 14.3), and the thick disk may also have been puffed up by small galaxies that fell into the young Galaxy.

Stellar populations. Stars are sorted into populations by two properties, their age and their metallicity. The metallicity is the mass fraction Z of all elements heavier than helium (Eq. (12.14)), which astronomers call metals. The two properties go together, for the reason given below, so the populations follow one another in time, as the tree of Figure 14.1 shows. **Population I** stars, like the Sun, are young or of middle age, with Z near 0.02. They live in the thin disk and mark the spiral arms. **Population II** stars are old and poor in metals, with Z from a tenth of the Sun's down to less than a ten-thousandth of it. They fill the stellar halo and the globular clusters, whose turn-off ages are about 1.2×10^{10} years (Eq. (12.42)). Metals track age because each generation of stars adds to them. The Big Bang made hydrogen, helium and a trace of lithium (Section 16.4). Stars burn these into heavier nuclei, and the massive ones return part of what they made to the gas in winds and supernovae within a few million years (Chapter 13). The next generation forms from that enriched gas. A star keeps in its surface the composition of the gas it formed from, so its metals record when it was born. The first stars, formed from gas with $Z = 0$, are called **Population III**. None has been found. A low-mass one would still be shining, since stars below about $0.8\,M_\odot$ live longer than the universe has existed ($10^{10} \times 0.8^{-2.5} = 1.7 \times 10^{10}$ years by the main-sequence lifetime $10^{10}\,\mathrm{yr}\,(M/M_\odot)^{-2.5}$ of Problem 12.8.4), so the first stars were probably all massive and short-lived. Star clusters show this directly. All the stars of a cluster have the same age and the same distance, so its colour–magnitude diagram is a snapshot of one age. Figure 14.2 shows Gaia stars of three clusters, the Pleiades, M67 and NGC 6397, picked out by their common motion on the sky and their common parallax. The parallax gives each cluster's distance, and the stars are drawn at that distance with a small correction for reddening. As a cluster ages, its turn-off slides down the main sequence (Eq. (12.42)). The Pleiades main sequence runs unbroken up to $M_G = -0.4$, and its brightest stars are missing from the data, so its turn-off is at least that bright. M67 turns off at $M_G = 3.0$ and NGC 6397 at $M_G = 3.9$, several magnitudes fainter. The Pleiades are therefore far younger than the other two. Eq. (12.42) gives less than 0.4×10^9 years for the Pleiades, 3.3×10^9 years for M67 and 6.0×10^9 years for NGC 6397. The last is far below the 1.2×10^{10} years of detailed models, because the simple law is too rough for stars near the Sun's mass. NGC 6397 differs in a second way. At the same brightness its main sequence is bluer than

that of the other two. Its stars have few metals, and fewer metals lower the opacity of the gas (Section 12.3), so a star of the same mass is hotter and bluer. Only the old, metal-poor cluster has a horizontal branch, a group of stars burning helium in their cores after the giant branch. The reddening values ($E(B-V) = 0.04$, 0.04 and 0.18), the distances and the ages quoted for these clusters are approximate.

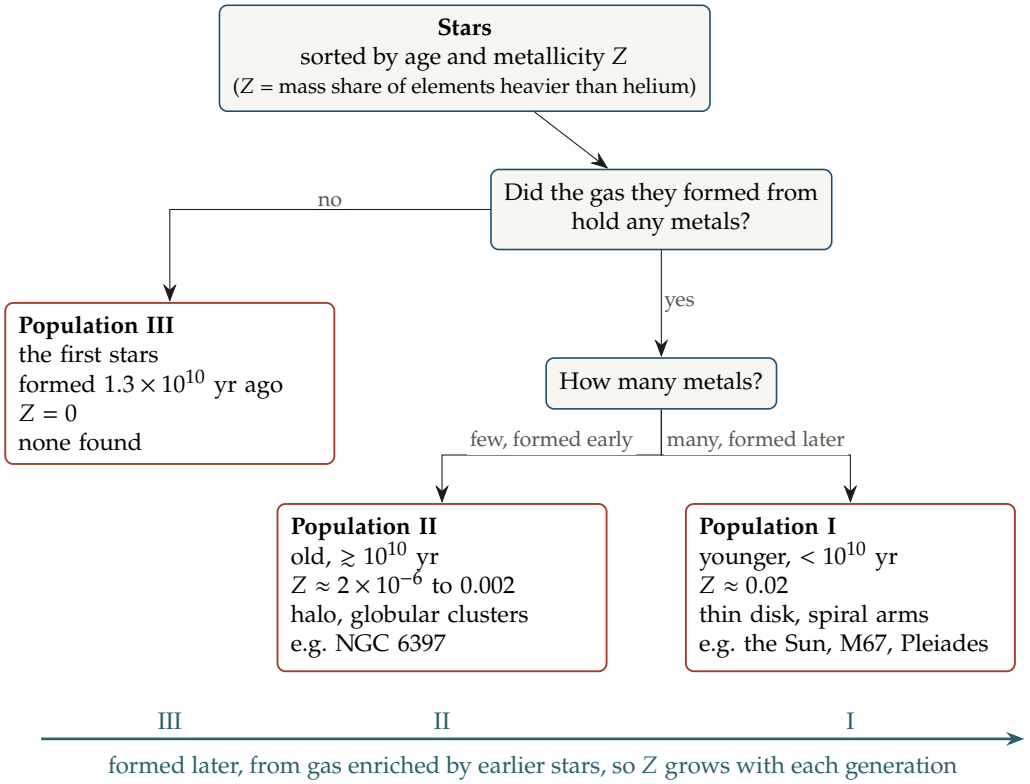


Figure 14.1: Stellar populations as a tree. Stars split first by whether the gas they formed from held any metals and then by how many, with the age, the metal fraction Z, the place in the Galaxy and examples for each population. The arrow at the bottom gives the order in which the populations formed.

Spiral arms. The arms are where stars are being born. The young blue stars, the glowing clouds of ionised gas around them and the dust lanes all lie along the arms, while the old stars are spread smoothly through the disk. The arms cannot be made of the same stars all the time. With a flat rotation curve (Section 14.2), a star at 4 kpc goes round twice as often as one at 8 kpc. In 10^{10} years a star moving at 230 km s^{-1} travels 2352 kpc, and the angle it turns through is this length divided by its radius. So the inner one gains $2352\,\text{kpc} \times (1/4 - 1/8)\,\text{kpc}^{-1} = 294\,\text{rad}$, or 47 turns, on the outer one. An arm made of fixed stars would be wound up into a tight coil, but real arms wind round about once. An arm is a pattern that the stars and gas move through, a **density wave**. A traffic jam on a road behaves the same way. Cars slow down and bunch up as they enter it and speed up as they leave, so the jam stays in place while the cars pass through. In the same way the extra pull of the crowded arm slows the stars and gas that enter it. The gas is squeezed, its dense clouds pass the Jeans limit (Section 10.4), and new stars form. A star of $20\,M_\odot$ lives $10^{10} \times 20^{-2.5} = 6 \times 10^6$ years by the same lifetime, a few per cent of the 220 million years of one orbit. So the brightest stars die close to the arm where they were born, and they make the arm stand out (P6).

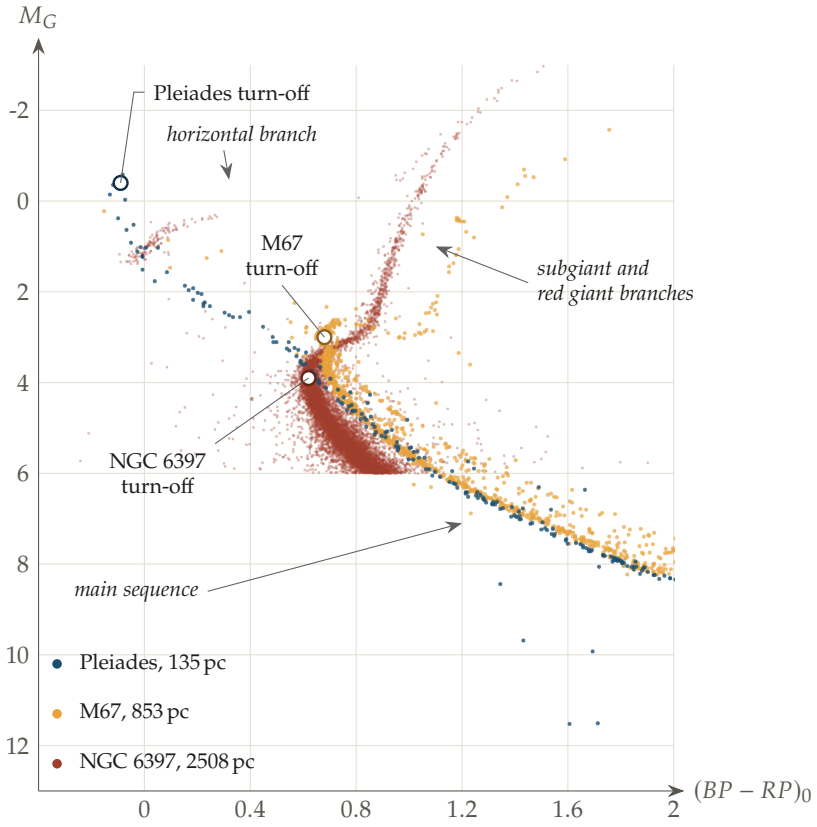


Figure 14.2: Absolute magnitude M_G against colour $(BP - RP)_0$ for Gaia DR3 members of the Pleiades (703 stars, 135 pc), M67 (1040 stars, 853 pc) and NGC 6397 (10574 stars, 2508 pc), corrected for reddening, with brighter stars at the top. Circles mark the turn-offs. The circle on the Pleiades is the brightest populated bin, since the brightest Pleiades stars are not in the data.

The hot corona. The halo also holds thin ionised gas at about 2×10^6 K, the **Galactic corona**, seen by its X-rays and by the absorption lines of highly ionised oxygen in the spectra of distant quasars. Part of it was thrown out of the disk by supernovae, and part has fallen in from outside. The temperature is the one the Galaxy's well gives. Particles held by an isothermal halo, whose spread of speeds along any one direction is σ and whose circular speed is $v_{\rm circ}^2 = 2\sigma^2$ (Section 12.5) have $k_B T = \mu m_{\rm H} \sigma^2$, as for the gas of a cluster (Problem 10.8.2), so for $\mu = 0.6$ and $\sigma^2 = v_0^2/2$

$$T = \frac{\mu m_{\rm H} v_0^2}{2k_B} = \frac{0.6 \times 1.673 \times 10^{-27} \times (2.3 \times 10^5)^2}{2 \times 1.381 \times 10^{-23}}\,{\rm K} = 1.9 \times 10^6\,{\rm K}.$$

The gas is so thin that its total mass is uncertain, but it is a few per cent of the dark halo's at most.

14.2 Rotation Curves and Dark Matter

Planetary orbits reveal the Sun's mass because orbital speed responds to gravity. The same idea can weigh a galaxy, but its mass is spread over many kiloparsecs. We must measure orbital speed at several radii to learn how that mass is distributed. The result need not follow the light we see.

Think of the planets. Mercury moves at 47 km s^{-1}, the Earth at 30, Neptune at only 5.4. Further out, the planets move slower, because all the mass that holds them is in the Sun at the middle.

If Neptune moved as fast as the Earth, the mass inside its orbit would have to be thirty times the mass inside the Earth's orbit, and that extra mass would have to be somewhere between the two. The stars and gas clouds of a spiral galaxy circle its centre in the same way. Their speed at each radius gives the mass inside that radius, whether that mass shines or not. Beyond the edge of the light, the speeds should fall as Neptune's do. They stay flat. The orbital speed weighs everything inside the orbit, seen or unseen.

For a circular orbit around a central mass, $v^2 = GM/r$, so speed falls as $v \propto r^{-1/2}$; Neptune at 30.07 AU moves at $29.78/\sqrt{30.07} = 5.43\,\mathrm{km\,s^{-1}}$. The Sun circles the Galaxy at $v_0 = 230\,\mathrm{km\,s^{-1}}$ and $R_0 = 8.2\,\mathrm{kpc}$. If we approximate the enclosed mass as spherical, its orbit gives $v_0^2 R_0/G = 1.0 \times 10^{11}\,M_\odot$ inside that radius, and one lap takes $2\pi R_0/v_0 = 220$ million years. Farther out, Figure 14.4 shows that the speed stays nearly level even where the visible stars thin out. The mass must keep increasing with radius.

The circular speed. To find the speed of a circular orbit at radius r in any spherical distribution of mass, we equate the acceleration of the orbit with the pull of the mass inside,

$$-\frac{v_{\mathrm{circ}}^2}{r}\,\hat{r} \overset{(1)}{=} \frac{\mathrm{d}\vec{v}}{\mathrm{d}t} \overset{(2)}{=} \vec{g}(r) \overset{(3)}{=} -\frac{G\,m(r)}{r^2}\,\hat{r}. \tag{14.7}$$

(1) Motion on a circle at constant speed v_{circ}.
(2) A star moves under gravity alone (P1 for its momentum).
(3) The gravity of a sphere, Eq. (9.17) (P2).

$$\boxed{v_{\mathrm{circ}}^2 = \frac{G\,m(r)}{r}.} \tag{14.8}$$

The speed at r depends only on the mass inside r, however it is arranged in radius, and the mass outside drops out. That holds exactly for a round halo. For a flat disk the mass outside r also pulls, so for a disk the relation is only a good estimate. The graph of v_{circ} against r is the **rotation curve**.

Reading a rotation curve. Turned around, $m(r) = v_{\mathrm{circ}}^2 r/G$. Each shape of the curve stands for one way of spreading the mass.

Mass	$m(r)$	v_{circ}
all at the centre	constant	$\propto r^{-1/2}$
uniform sphere	$\propto r^3$	$\propto r$
flat curve	$\propto r$	constant

For a flat curve the density follows from $\mathrm{d}m/\mathrm{d}r = 4\pi r^2 \rho$,

$$\rho \overset{(1)}{=} \frac{1}{4\pi r^2}\frac{\mathrm{d}m}{\mathrm{d}r} \overset{(2)}{=} \frac{1}{4\pi r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{v_{\mathrm{circ}}^2\, r}{G}\right) = \frac{v_{\mathrm{circ}}^2}{4\pi G\, r^2}. \tag{14.9}$$

(1) Eq. (9.15).
(2) Eq. (14.8) with v_{circ} the same at every radius.

$$\boxed{\rho = \frac{v_{\mathrm{circ}}^2}{4\pi G\, r^2} \qquad \text{(flat rotation curve).}} \tag{14.10}$$

The isothermal sphere of Section 12.5 gives the same law by another route. Far from its centre it has $\rho = c_s^2/(2\pi G r^2)$ and $v_{\mathrm{circ}}^2 = 2c_s^2$, and putting $c_s^2 = v_{\mathrm{circ}}^2/2$ into the first gives Eq. (14.10). A flat rotation curve is what a halo of particles with one velocity spread at every radius produces.

Measuring the rotation from inside. We see our own Galaxy from a point on one of its orbits. To get v_{circ} at other radii, we measure the Doppler shift of the 21 cm line at Galactic longitude l in the plane, and work out what circular motion predicts for it.

Put the origin at the Galactic centre, with $\hat{x}$ from the Sun toward the centre ($l = 0$), $\hat{y}$ toward $l = 90°$, the direction of rotation at the Sun, and $\hat{z}$ toward the north Galactic pole (Figure 14.3). These are the axes of (U, V, W) of Section 3.6, moved to the centre. The Sun is at $\vec{r}_0 = (-R_0,\, 0)$ in the plane. A cloud at distance d along the direction of sight $\hat{\mathbf{s}} = (\cos l,\, \sin l)$ is at $\vec{r} = \vec{r}_0 + d\,\hat{\mathbf{s}}$, at distance $r = |\vec{r}|$ from the centre. Seen from the north pole the Galaxy turns clockwise.

A body on a circular orbit moves at right angles to the line from the centre, and its speed is $v_{\text{circ}} = \omega r$, where $\omega(r) = v_{\text{circ}}(r)/r$ is its angular speed. So its velocity is the position arrow $\vec{r}$ turned by a right angle, in the direction of rotation, and stretched by ω. Turning an arrow is what the rotation matrix of Eq. (1.3) does. In the plane it is

$$R(\varphi) = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix},$$

which turns an arrow anticlockwise by φ. The Galaxy turns clockwise, so the turn needed is $\varphi = -90°$. With $\cos(-90°) = 0$ and $\sin(-90°) = -1$ this is

$$\mathbf{J} \equiv R(-90°) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \mathbf{J}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ -x \end{pmatrix}.$$

The turned arrow $(y, -x)$ has the same length as (x, y), and it is at right angles to it, since $(x, y) \cdot (y, -x) = xy - yx = 0$. The velocity of a body on a circular orbit at $\vec{r}$ is therefore

$$\vec{v} = \omega(r)\,\mathbf{J}\,\vec{r}. \tag{14.11}$$

At the Sun, $\vec{r}_0 = (-R_0,\, 0)$ turns into $\mathbf{J}\vec{r}_0 = (0,\, R_0)$, so $\vec{v}_0 = \omega_0 R_0\,\hat{y} = v_0\,\hat{y}$, along the direction of rotation as it should be, with $\omega_0 = v_0/R_0$. The matrix gives the velocity of every cloud in the same way, whatever its position, without first finding its angle around the centre. The Doppler shift, Eq. (3.25), measures the velocity of the cloud relative to the Sun along $\hat{\mathbf{s}}$, after the Sun's own small motion relative to the local standard of rest has been removed,

$$\begin{aligned} v_r &= \hat{\mathbf{s}} \cdot \left(\omega\,\mathbf{J}\,\vec{r} - \omega_0\,\mathbf{J}\,\vec{r}_0\right) \\ &\overset{(1)}{=} \omega\,\hat{\mathbf{s}} \cdot \mathbf{J}\vec{r}_0 + \omega\, d\,\hat{\mathbf{s}} \cdot \mathbf{J}\hat{\mathbf{s}} - \omega_0\,\hat{\mathbf{s}} \cdot \mathbf{J}\vec{r}_0 \\ &\overset{(2)}{=} (\omega - \omega_0)\,\hat{\mathbf{s}} \cdot \mathbf{J}\vec{r}_0 \\ &\overset{(3)}{=} (\omega - \omega_0)\begin{pmatrix} \cos l & \sin l \end{pmatrix}\begin{pmatrix} 0 \\ R_0 \end{pmatrix} = (\omega - \omega_0)\,R_0 \sin l. \end{aligned} \tag{14.12}$$

(1) $\vec{r} = \vec{r}_0 + d\,\hat{\mathbf{s}}$.

(2) $\hat{\mathbf{s}} \cdot \mathbf{J}\hat{\mathbf{s}} = \cos l \sin l - \sin l \cos l = 0$. A turn by 90° makes any vector perpendicular to the vector it came from.

(3) $\mathbf{J}\vec{r}_0 = (0,\, R_0)$.

$$\boxed{v_r = \left(\frac{v_{\text{circ}}(r)}{r} - \frac{v_0}{R_0}\right) R_0 \sin l.} \tag{14.13}$$

The distance d enters only through r. Two checks. Toward the centre or the anticentre, $\sin l = 0$ and $v_r = 0$, since every orbit there moves across the line of sight. A cloud on the Sun's own circle, $r = R_0$, has $v_r = 0$ too, since it keeps its distance from the Sun.

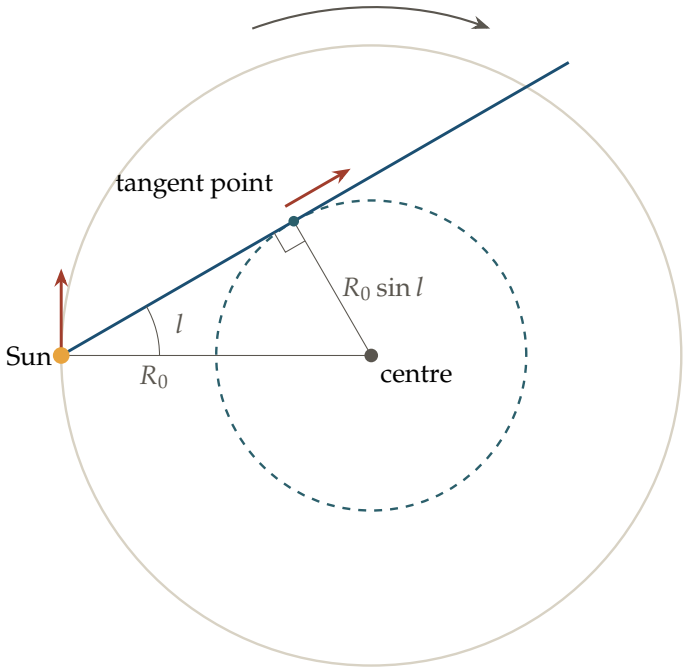


Figure 14.3: The Galactic disk seen from the north Galactic pole, to scale, with $R_0 = 8.2\,\text{kpc}$ and a line of sight at $l = 30°$. The grey circle is the Sun's orbit and the dashed circle has radius $R_0 \sin l$. The red arrows are the velocities of circular orbits at the Sun and at the tangent point. The arc at the top shows the sense of rotation.

The tangent point. Along a line of sight with $0 < l < 90°$, the distance from the centre is smallest where the line touches a circle about the centre, at $r = R_0 \sin l$ (Figure 14.3). If the angular speed $\omega = v_{\text{circ}}/r$ falls outward, as it does for a flat or slowly rising curve, it is largest there, and so is v_r. Setting $r = R_0 \sin l$ in Eq. (14.13),

$$v_{r,\text{max}} = \left(\frac{v_{\text{circ}}(R_0 \sin l)}{R_0 \sin l} - \frac{v_0}{R_0} \right) R_0 \sin l = v_{\text{circ}}(R_0 \sin l) - v_0 \sin l, \tag{14.14}$$

so

$$\boxed{v_{\text{circ}}(R_0 \sin l) = v_{r,\text{max}} + v_0 \sin l.} \tag{14.15}$$

The highest velocity in the 21 cm line at each longitude gives the rotation speed at one radius, with no distance needed except R_0. Sweeping l from 0 to 90° maps v_{circ} from the centre out to R_0. Outside the Sun's circle the method fails, because r grows all along the line. There the distances of the gas or of stars must be found separately, from Cepheids or from *Gaia* parallaxes.

Flat rotation curves. In our Galaxy, v_{circ} stays between about 200 and 240 km s^{-1} from 5 kpc out to 25 kpc. For another spiral seen at an angle, the Doppler shift across its disk gives the curve directly. The light of a disk falls off exponentially with radius, over a few kiloparsecs, so a few such lengths out nearly all the light is inside. If the light traced the mass, $m(r)$ would stop growing there, and v_{circ} would fall as $r^{-1/2}$, like the speeds of the planets. It does not. In the 21 cm line, whose gas reaches much further than the stars, the rotation curves of most spirals stay flat as far out as they can be followed, often several times the radius of the visible disk. By Eq. (14.8) the mass keeps growing as r there, and by Eq. (14.10) the unseen matter has $\rho \propto r^{-2}$.

A model galaxy. Masses add, so the squares of the speeds add,

$$v_{\text{circ}}^2 = \frac{G\big(m_{\text{stars}} + m_{\text{halo}}\big)}{r} = v_{\text{stars}}^2 + v_{\text{halo}}^2. \tag{14.16}$$

For the halo take a density that is level in a core of radius r_c and falls as r^{-2} outside, $\rho = \rho_c/(1 + r^2/r_c^2)$, with ρ_c the central density. Its mass inside r is

$$\begin{aligned} m_{\text{halo}}(r) &\overset{(1)}{=} \int_0^r \frac{4\pi r'^2 \rho_c}{1 + r'^2/r_c^2}\, dr' \\ &\overset{(2)}{=} 4\pi\rho_c r_c^2 \int_0^r \left(1 - \frac{1}{1 + r'^2/r_c^2}\right) dr' \\ &\overset{(3)}{=} 4\pi\rho_c r_c^2 \left(r - r_c \arctan\frac{r}{r_c}\right). \end{aligned} \tag{14.17}$$

(1) Eq. (9.15).
(2) $r'^2/(1 + r'^2/r_c^2) = r_c^2\left[1 - 1/(1 + r'^2/r_c^2)\right]$.
(3) $\int_0^r dr'/(1 + r'^2/r_c^2) = r_c \arctan(r/r_c)$.

By Eq. (14.8),

$$\boxed{v_{\text{halo}}^2 = 4\pi G\rho_c r_c^2 \left[1 - \frac{r_c}{r}\arctan\frac{r}{r_c}\right].} \tag{14.18}$$

Near the centre, $\arctan x \approx x - x^3/3$ gives $v_{\text{halo}}^2 \approx \frac{4}{3}\pi G\rho_c r^2$, the uniform sphere with $v \propto r$. Far out, $\arctan \to \pi/2$ and $v_{\text{halo}}^2 \to 4\pi G\rho_c r_c^2$, the flat curve. Figure 14.4 adds this halo to a round model of the stars. The stars set the curve inside a few kiloparsecs and the halo outside, and between them the total is nearly level. The halo's mass grows as r without end, so a real halo must be cut off somewhere. The motions of satellite galaxies show that the Milky Way's halo reaches 200 to 300 kpc and holds about $10^{12}\,M_\odot$.

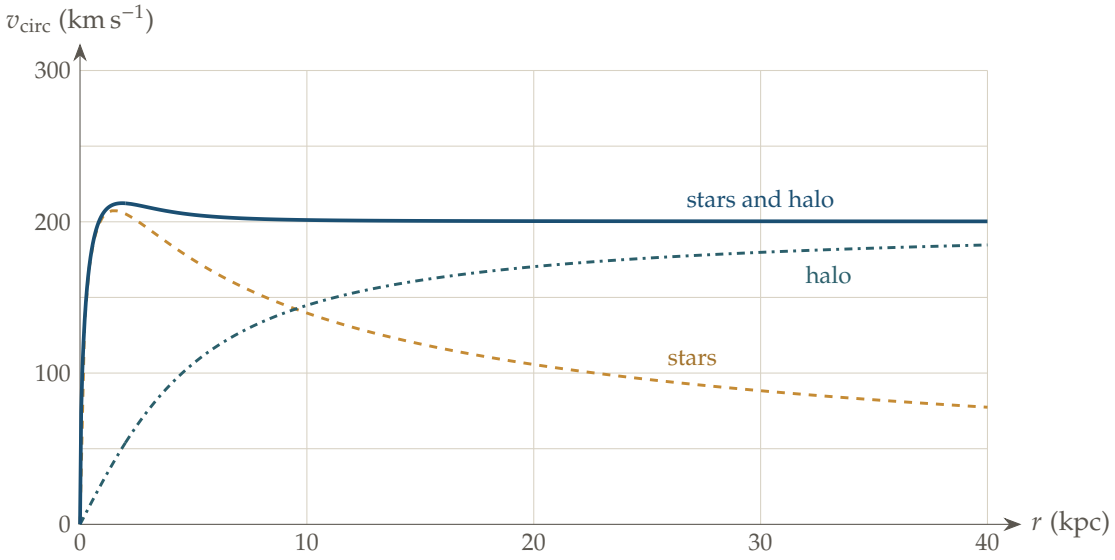


Figure 14.4: The circular speed of a model galaxy against radius. Dashed, stars with enclosed mass $6 \times 10^{10}\,M_\odot\, r^2/(r + 1.5\,\text{kpc})^2$. Dash-dotted, the halo of Eq. (14.18) with $r_c = 4$ kpc and $4\pi G\rho_c r_c^2 = (200\,\text{km s}^{-1})^2$. Solid, both together.

The mass-to-light ratio. A galaxy's mass divided by its luminosity, in units of $M_\odot/L_\odot$, is its **mass-to-light ratio** M/L. The Sun has $M/L = 1$. A population of old stars has a few, because the mass is mostly in faint low-mass stars and the light comes from the few bright ones. With $L \propto M^{3.5}$ near the Sun (Section 12.6), a star of $0.5\,M_\odot$ alone has $M/L = 0.5^{-2.5} = 5.7$. Inside the

visible disk of a spiral, the mass from the rotation curve gives M/L of a few, like its stars. Out to the last point of a flat curve it gives ten to several tens, and for clusters of galaxies about three hundred (Section 14.4). The ratio grows with the radius inside which the mass is measured, because the unseen matter takes over further out. For the Milky Way, a curve flat at 230 km s^{-1} out to 50 kpc would put $6 \times 10^{11}\, M_\odot$ inside that radius, against $6 \times 10^{10}\, M_\odot$ in stars and gas.

What the unseen mass is not. It is not cold gas, since the 21 cm line counts the neutral hydrogen atom by atom (Eq. (8.40)), and hot gas shines in X-rays, which show the corona to be only a small part of the mass. It is not mostly faint stars or dark compact bodies. Such bodies, crossing in front of stars of the Large Magellanic Cloud, would now and then brighten them by bending their light (Eq. (14.41)), and surveys in the 1990s found far too few such events for bodies between about 10^{-7} and $10\, M_\odot$ to make up the halo. The amount of ordinary matter of every kind is fixed by the helium and deuterium made in the first minutes of the universe (Chapter 16), and it falls well short of the mass that galaxies need. The halo is mostly a kind of matter that neither sends out nor absorbs light, and it is felt only through its gravity.

14.3 Stars as a Gas That Rarely Collides

Gas atoms share energy through frequent collisions. We have also described the Galactic disk as a gas of stars, yet its stars almost never touch. Near the Sun they are about a parsec apart and roughly a hundred million times smaller than that distance. If direct collisions cannot mix their motions, can gravity still make the stars forget the orbits on which they formed?

Think of a ship crossing the ocean with its rudder tied. Each wave pushes it a little to one side or the other, in no fixed pattern, and one wave does nothing worth noticing. But the pushes do not cancel exactly, so after many waves the sideways drift has built up, growing as the square root of the number of waves, and in the end the ship is heading somewhere else altogether. A star crossing a galaxy is the ship. Every star it passes tugs it a little, from far away, in a random direction. The tugs add as a random walk in velocity (P5), so the sum of their squares grows steadily, and when that sum reaches the square of the star's own speed the star has forgotten its original orbit. The time this takes is the **relaxation time**, and comparing it with the age of the system decides what kind of system it is (P6).

The time needed to accumulate many small kicks depends strongly on the number of stars. For the ship analogy, sideways kicks of 1 cm s^{-1} take $(500)^2 = 250\,000$ waves to build up to its 5 m s^{-1} speed. A galaxy of 10^{11} stars needs about 5×10^8 crossings to relax (derived below). At roughly 5×10^7 years per crossing, that is 2×10^{16} years, far longer than its age. A globular cluster of 10^5 stars needs only about a thousand shorter crossings and can relax many times over. We will calculate the kick from one passing star (Figure 14.5) and then add the kicks as a random walk. The result explains why a galaxy's stars retain much of their orbital history, while stars in a dense cluster slowly exchange energy and reshape the cluster.

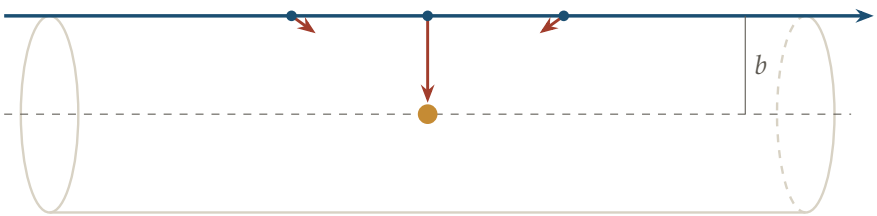


Figure 14.5: A star moving left to right along a straight path (blue) passes a field star (dot on the dashed axis) at impact parameter b. Red arrows are the pull of the field star at three points of the path. The grey cylinder of radius b about the dashed axis contains the path.

One passing star. Let a star move at speed v past a field star of mass $m_\star$, along a straight line that misses it by the **impact parameter** b. We treat the path as straight and the field star as fixed. Both hold while the kick is small compared with v. Put the field star at the origin, the x axis along the path and the y axis from the field star to the point of closest approach, and put the closest approach at $t = 0$. The moving star is then at

$$\vec{r}(t) = \begin{pmatrix} vt \\ b \end{pmatrix}, \qquad d(t) = |\vec{r}| = \sqrt{b^2 + v^2 t^2},$$

where d is its distance from the field star. The field star pulls it with the acceleration of Newton's law, of size $Gm_\star/d^2$ and pointing back toward the origin, along $-\vec{r}/d$. Adding up this pull over the whole passage gives the change of velocity,

$$\begin{aligned} \Delta\vec{v} &\overset{(1)}{=} \int_{-\infty}^{\infty} \vec{a}\, dt \\ &\overset{(2)}{=} \int_{-\infty}^{\infty} \frac{Gm_\star}{d^2}\left(-\frac{\vec{r}}{d}\right) dt = -\int_{-\infty}^{\infty} \frac{Gm_\star}{d^2}\,\frac{1}{d}\begin{pmatrix} vt \\ b \end{pmatrix} dt \\ &\overset{(3)}{=} -\begin{pmatrix} 0 \\ \displaystyle\int_{-\infty}^{\infty} \frac{Gm_\star}{d^2}\,\frac{b}{d}\, dt \end{pmatrix}. \end{aligned} \tag{14.19}$$

(1) The change of velocity is the time integral of the acceleration (P1).
(2) Newton's law of gravity. The size is $Gm_\star/d^2$ and the direction is the unit vector $-\vec{r}/d$, from the moving star to the field star.
(3) The x part, along the path, has vt on top. It is negative before closest approach and positive after, by the same amount at $-t$ and t, so the pull forward on the way in cancels the pull back on the way out (the side arrows of Figure 14.5). The y part has b on top, which does not change sign, so it adds up.

So the kick is across the path, pointing toward the field star (along $-y$). In the y part, $Gm_\star/d^2$ is the full pull and b/d is the fraction of it that points across the path, the cosine of the angle between the pull and the y axis. Its size is

$$\begin{aligned} \Delta v &\overset{(1)}{=} \int_{-\infty}^{\infty} \frac{Gm_\star}{b^2 + v^2t^2}\,\frac{b}{\sqrt{b^2 + v^2t^2}}\, dt \\ &\overset{(2)}{=} \int_{-\pi/2}^{\pi/2} \frac{Gm_\star}{b^2\sec^2\psi}\,\frac{b}{b\sec\psi}\,\frac{b}{v}\sec^2\psi\, d\psi \\ &\overset{(3)}{=} \frac{Gm_\star}{bv}\int_{-\pi/2}^{\pi/2} \cos\psi\, d\psi \\ &\overset{(4)}{=} \frac{2Gm_\star}{bv}. \end{aligned} \tag{14.20}$$

(1) The y part above, with $d = \sqrt{b^2 + v^2t^2}$ written out.
(2) Let ψ be the angle, seen from the field star, between the y axis and the line to the moving star. Then $vt = b\tan\psi$, $d = b\sec\psi$ and $dt = (b/v)\sec^2\psi\, d\psi$. As t runs from $-\infty$ to ∞, ψ runs from $-\pi/2$ to $\pi/2$.
(3) The b's leave $1/b$ and the $\sec\psi$'s leave $1/\sec\psi = \cos\psi$.
(4) $\int_{-\pi/2}^{\pi/2}\cos\psi\, d\psi = \sin\psi\,\big|_{-\pi/2}^{\pi/2} = 2$.

$$\boxed{\Delta v = \frac{2Gm_\star}{b\,v}.} \tag{14.21}$$

The kick is the pull at closest approach, $Gm_\star/b^2$, times about twice the time spent near it, b/v. A closer or slower passage gives a larger kick.

The same kick from Gauss's law. Draw a cylinder of radius b whose axis passes through the field star, parallel to the path (Figure 14.5). The path lies on its surface. Let the cylinder be very long, so its ends carry no flux. With ϕ the angle round the axis, x the length along it and g_r the part of the pull pointing away from the axis,

$$-4\pi G m_\star \overset{(1)}{=} \oint \vec{g}\cdot d\vec{A} \overset{(2)}{=} \int_0^{2\pi} b\,d\phi \int_{-\infty}^{\infty} g_r\,dx \overset{(3)}{=} 2\pi b \int_{-\infty}^{\infty} g_r\,dx. \tag{14.22}$$

(1) Gauss's law (9.14), with $m_\star$ inside (P2).
(2) The area element on the side is $b\,d\phi\,dx$, and only the part of $\vec{g}$ across the side counts.
(3) Every line on the cylinder parallel to the axis is the same as every other, by symmetry about the axis.

So $\int g_r\,dx = -2Gm_\star/b$ along the path, and with $dx = v\,dt$ the kick toward the axis is $2Gm_\star/(bv)$, Eq. (14.21) again. No integral had to be done. The kick depends on the mass of the field star and on b, exactly as the field of a line of charge depends on its charge and the distance. The straight path holds while $\Delta v \ll v$, which is $b \gg 2Gm_\star/v^2$. For a star of one solar mass passed at 200 km s^{-1} that is 0.044 AU. Almost every encounter in a galaxy is far wider than this.

Many passing stars. Take a system of N stars of mass $m_\star$ within a radius R, through which a star crosses at a typical speed v in the **crossing time** $t_{\text{cross}} = R/v$. In one crossing it goes right through the system and passes each of the other N stars once, at some impact parameter. To count how many it passes at each impact parameter, look at the system along the star's path (Figure 14.6). Seen this way the system is a disk of radius R, with its N stars spread over the area πR^2, and the path is a single point in the middle. The impact parameter of each passage is just the distance b of that field star from the point. So the stars passed at impact parameters between b and $b + db$ are the stars in a thin ring of radius b and width db, and their number is the number per unit area times the area of the ring,

$$\frac{N}{\pi R^2} \times 2\pi b\,db = \frac{2Nb\,db}{R^2}.$$

Taking the stars as spread evenly over the disk is rough, but it is enough for an estimate.

Each passage gives a kick of Eq. (14.21), pointing across the path toward the field star that gave it. The field stars lie all around the path, at every angle in the ring, so the kicks point in random directions across it. They add like the steps of a random walk (P5). The kicks themselves mostly cancel, so their vector sum stays small. But the sum of their squares grows with every passage, because in $\left|\sum \Delta\vec{v}_i\right|^2 = \sum |\Delta\vec{v}_i|^2 + \sum_{i\neq j} \Delta\vec{v}_i\cdot\Delta\vec{v}_j$ the cross terms are as often negative as positive and average to zero, as in Eq. (9.41). This sum of squares measures how far the star's velocity has wandered from where it started. In one crossing,

$$\begin{aligned}\sum |\Delta\vec{v}|^2 &\overset{(1)}{=} \int_{b_{\min}}^{R} \left(\frac{2Gm_\star}{bv}\right)^2 \frac{2Nb\,db}{R^2} \\ &\overset{(2)}{=} 8N\left(\frac{Gm_\star}{Rv}\right)^2 \int_{b_{\min}}^{R} \frac{db}{b} = 8N\left(\frac{Gm_\star}{Rv}\right)^2 \ln\frac{R}{b_{\min}}.\end{aligned} \tag{14.23}$$

(1) The square of Eq. (14.21), times the number of passages at each b. No star lies further away than about R. Below $b_{\min} = 2Gm_\star/v^2$ a single kick is as large as v and the straight path fails.

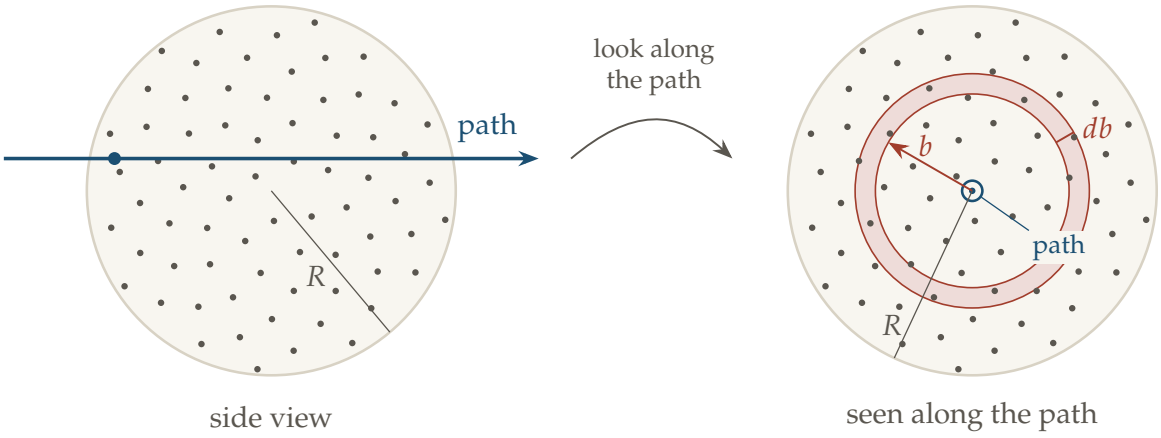


Figure 14.6: A star system of radius R crossed by a star on a straight path (left), and the same system seen along the path (right), where the path is a point at the centre. The shaded ring of radius b and width db holds the field stars passed at impact parameters between b and $b + db$.

(2) Collect the constants.

The integral is a logarithm, so every factor of ten in b adds the same amount. A far passage gives a smaller kick, but there are more stars at large b, and the two effects cancel in each factor of ten. Since $R/b_{\min}$ is a very large number, there are many factors of ten between them. The close passages, near $b_{\min}$, fill only the first few. So the wandering of a star's velocity comes mostly from the many distant, gentle passages, not from the rare close ones. This logarithm is called the **Coulomb logarithm**, because the same sum appears for electrons passing ions in a plasma (Chapter 15).

The relaxation time. To put Eq. (14.23) in terms of N alone, we use the virial theorem (Eq. (10.9) with $\ddot{I} = 0$, $2K = |U|$) in order of magnitude, $v^2 \approx GNm_\star/R$. Then $Gm_\star/R = v^2/N$ and $b_{\min} = 2R/N$, and the relative size of the summed kicks in one crossing is

$$\frac{\sum |\Delta\vec{v}|^2}{v^2} \overset{(1)}{=} \frac{8N}{v^2}\left(\frac{v}{N}\right)^2 \ln\frac{N}{2} \overset{(2)}{=} \frac{8\ln N}{N}. \tag{14.24}$$

(1) Eq. (14.23) with $Gm_\star/(Rv) = v/N$ and $R/b_{\min} = N/2$.
(2) For large N, $\ln(N/2)$ and $\ln N$ differ by less than the cut-offs at R and $b_{\min}$ are known.

Each crossing adds the same share $8\ln N/N$ of v^2 to the sum of squares. The star has forgotten its starting velocity once the sum reaches v^2 itself, that is once the random kicks have changed its velocity by as much as the velocity. With n crossings this happens when $n \times 8\ln N/N = 1$, so after $n = N/(8\ln N)$ crossings. That time is the **relaxation time**,

$$\boxed{t_{\text{relax}} = \frac{N}{8\ln N}\, t_{\text{cross}}, \qquad t_{\text{cross}} = \frac{R}{v}.} \tag{14.25}$$

Only the number of stars and the crossing time enter. For a given crossing time, a system of fewer, heavier stars relaxes faster, because each kick is larger.

What relaxes. Table 14.2 puts in the numbers, with v from $v^2 = GM/R$. The ratio $t_{\text{relax}}/\text{age}$ sorts them (P6). In a galaxy it is a million. Its stars have never traded energy with each other.

System	M ($M_\odot$)	N	R	v (km s^{-1})	t_{cross} (yr)	t_{relax} (yr)	age (yr)
Open cluster	500	10^3	2 pc	1.0	1.9×10^6	3×10^7	10^8
Globular cluster	5×10^4	10^5	5 pc	6.6	7.5×10^5	8×10^8	1.2×10^{10}
Galaxy	10^{11}	10^{11}	10 kpc	207	4.7×10^7	2×10^{16}	10^{10}

Table 14.2: Crossing and relaxation times, Eq. (14.25), for three kinds of star system.

Each moves in the smooth pull of the whole galaxy as if the other stars were spread out into a fluid, and a galaxy is **collisionless**. In a globular cluster the ratio is about a tenth, so the cluster has relaxed many times.

Galaxies remember. A collisionless galaxy keeps the orbits its stars were born on. Near the Sun the spread of stellar velocities is larger toward the Galactic centre than up out of the plane. A relaxed gas would have made it the same in every direction, as collisions do for the pressure of a gas. The vertical balance of Section 14.1 needed only Π_{zz}, so it still holds. The disk does heat slowly, and old stars move faster than young ones, but the heating comes from the much larger kicks of giant molecular clouds and spiral arms, not from single stars. The shapes of galaxies, too, are set when they form and kept since (Section 14.5).

Globular clusters relax, and lose stars. In a cluster that relaxes, kicks drive the speeds toward a Maxwell distribution (P3 and P4). Its fastest stars then have more than the escape speed and leave. To see how many, compare the escape speed with the mean square speed. Let Φ be the gravitational potential, so that a star at rest at infinity has $\frac{1}{2}v_{\text{esc}}^2+\Phi=0$, and $U=\frac{1}{2}\int\Phi\,dm$, since a sum over the potential of every mass element counts each pair twice. Averaging over the mass,

$$\left\langle v_{\text{esc}}^2\right\rangle \overset{(1)}{=} -\frac{2}{M}\int\Phi\,dm \overset{(2)}{=} -\frac{4U}{M} \overset{(3)}{=} \frac{8K}{M} \overset{(4)}{=} 4\left\langle v^2\right\rangle . \tag{14.26}$$

(1) $v_{\text{esc}}^2=-2\Phi$ (P1 for energy), averaged over the mass M.
(2) $\int\Phi\,dm=2U$.
(3) The virial theorem for a bound system, $2\langle K\rangle+\langle U\rangle=0$, Eq. (10.12), with K and U here standing for their long-time averages.
(4) $K=\frac{1}{2}M\left\langle v^2\right\rangle$.

A star escapes once its speed exceeds about twice the root-mean-square speed. In a Maxwell distribution with $\left\langle v^2\right\rangle=3\sigma^2$, where σ is the spread of one velocity component, the fraction of stars faster than $2\sqrt{3}\,\sigma$ is

$$\int_{2\sqrt{3}\,\sigma}^{\infty}\frac{4\pi v^2}{(2\pi\sigma^2)^{3/2}}\,e^{-v^2/2\sigma^2}\,dv=\operatorname{erfc}\sqrt{6}+2\sqrt{\frac{6}{\pi}}\,e^{-6}=0.0074,$$

where the integral was done by parts. Relaxation refills this tail once in every relaxation time, so the cluster loses 0.74 per cent of its stars per t_{relax} and would evaporate in about $135\,t_{\text{relax}}$, some 10^{11} years for the cluster of Table 14.2. The tidal pull of the Galaxy strips the outer stars and shortens this. Relaxation also shares energy between stars of different mass. Heavy stars lose speed, sink toward the centre, and leave the light ones in the outskirts.

Core collapse. The dense core of a globular cluster is a self-gravitating system, with the negative heat capacity of Section 10.3. Its stars move faster than those of the outer parts, so relaxation carries energy outward, as heat flows from a hot body to a cold one. The core loses

energy, and with $C < 0$ it gets hotter and contracts. A hotter core passes energy to the outer parts even faster, and the process runs away. This is the **gravothermal catastrophe**. The core shrinks toward a very small, very dense centre. It is stopped by close pairs of stars that form there. When a third star passes close to a pair, the pair binds tighter and the third star leaves faster, so the pair hands energy to the core and heats it. About a fifth of the globular clusters of the Milky Way have the sharply peaked centres that mark a collapsed core. A galaxy, whose relaxation time is a million times its age, never gets there.

14.4 Weighing Galaxies and Clusters

Rotation curves weigh galaxies whose stars follow roughly circular orbits. In an elliptical galaxy or a galaxy cluster, the members move in many directions instead. Their random speeds still respond to the system's gravity, as the virial theorem predicts (Section 10.2). We observe only their positions on the sky and their speeds along our line of sight, so we need to connect those measured quantities to the full three-dimensional system.

Think of a salad bowl full of marbles, shaken and watched from above. The bowl itself cannot be seen. But marbles that move fast stay inside only if the bowl is deep, and from how fast they move and how far they spread one can judge its depth. The galaxies of a cluster are the marbles, and the cluster's gravity is the bowl. In 1933 Zwicky measured the line-of-sight speeds of galaxies in the Coma cluster, spread by about $1000\,\mathrm{km\,s^{-1}}$ over a few megaparsecs, and found that the bowl must hold hundreds of times the mass of the stars in it. The virial theorem turns random speeds and sizes into a mass.

For a uniform sphere, the virial theorem $2K = |U|$ combines $K = \frac{1}{2}M \cdot 3\sigma^2$ for line-of-sight speed spread σ with $|U| = \frac{3}{5}GM^2/R$ (Eq. (10.14)). It gives $M = 5\sigma^2 R/G$. A cluster with $\sigma = 1000\,\mathrm{km\,s^{-1}}$ and $R = 1.5\,\mathrm{Mpc}$ therefore has mass $1.7 \times 10^{15}\,M_\odot$. Below we refine this estimate using only measured positions and line-of-sight speeds. We then compare it with masses inferred from the hot gas and from gravitational lensing.

From three dimensions to what is seen. Take N galaxies of equal mass $m_\star$, total mass $M = Nm_\star$, in a steady state. Their kinetic energy is $K = \frac{1}{2}M\langle v^2\rangle$. We measure only the line-of-sight component, with mean square σ^2 after the mean motion of the cluster is removed. If the motions favour no direction,

$$\langle v^2\rangle = \langle v_x^2\rangle + \langle v_y^2\rangle + \langle v_z^2\rangle \overset{(1)}{=} 3\sigma^2, \tag{14.27}$$

(1) With no favoured direction, each component has the same mean square as the one along the line of sight.

The gravitational energy is a sum over pairs, $U = -Gm_\star^2 \sum_{\text{pairs}} 1/r$, with $\frac{1}{2}N(N-1)$ pairs. We see each separation r projected on the sky, $r_\perp = r\sin\theta$, where θ is the angle between the separation and the line of sight. The direction of a pair has nothing to do with its separation, so the two can be averaged one at a time over all pairs,

$$\left\langle \frac{1}{r_\perp}\right\rangle \overset{(1)}{=} \left\langle \frac{1}{r}\right\rangle \left\langle \frac{1}{\sin\theta}\right\rangle \overset{(2)}{=} \left\langle \frac{1}{r}\right\rangle \frac{1}{4\pi}\int_0^{2\pi}\!\!\int_0^{\pi} \frac{1}{\sin\theta}\,\sin\theta\,d\theta\,d\phi \overset{(3)}{=} \frac{\pi}{2}\left\langle \frac{1}{r}\right\rangle. \tag{14.28}$$

(1) r and θ are independent.
(2) All directions equally likely, with the solid angle $\sin\theta\,d\theta\,d\phi$ over 4π.
(3) The integrand is 1, so the integral is $2\pi \cdot \pi$.

The virial mass. Setting $2K = |U|$,

$$M \cdot 3\sigma^2 \overset{(1)}{=} G m_\star^2 \frac{N(N-1)}{2} \left\langle \frac{1}{r} \right\rangle \overset{(2)}{=} G m_\star^2 \frac{N(N-1)}{2} \frac{2}{\pi} \left\langle \frac{1}{r_\perp} \right\rangle = \frac{GM^2}{\pi} \left(1 - \frac{1}{N}\right) \left\langle \frac{1}{r_\perp} \right\rangle . \tag{14.29}$$

(1) The virial theorem for a bound system, Eq. (10.12), with K and U from above standing for their long-time averages.
(2) The projection factor $\pi/2$.

For $N \gg 1$,

$$\boxed{M = \frac{3\pi\,\sigma^2}{G\,\langle 1/r_\perp \rangle}.} \tag{14.30}$$

Only line-of-sight speeds and positions on the sky enter, with the distance to the cluster needed to turn angles into $r_\perp$. The result assumes a steady state, motions with no favoured direction, and galaxies that are spread like the mass. It does not depend on what the mass is made of. As a check, a uniform sphere has $\langle 1/r \rangle = |U|/(G m_\star^2 N^2/2) = 6/(5R)$ by Eq. (10.14), so $\langle 1/r_\perp \rangle = 3\pi/(5R)$, and Eq. (14.30) gives $M = 5\sigma^2 R/G$, the value found above.

For a cluster with $\sigma = 1000\,\mathrm{km\,s^{-1}}$ and $1/\langle 1/r_\perp \rangle = 1\,\mathrm{Mpc}$, Eq. (14.30) gives $2.2 \times 10^{15}\,M_\odot$. Coma has a mass of about $10^{15}\,M_\odot$ and a mass-to-light ratio of about 300. Its galaxies shine with a few tenths of a per cent of the light that this mass would give if it were all stars like the Sun.

The hot gas. The same well holds thin hot gas between the galaxies, the **intracluster medium**. Its particles share the energy per unit mass of the galaxies, so $k_B T = \mu m_\mathrm{H} \sigma^2$, about $7 \times 10^7\,\mathrm{K}$ for $\sigma = 1000\,\mathrm{km\,s^{-1}}$ (Problem 10.8.2). This is the virial temperature of Eq. (10.13), $k_B \bar{T} = \mu m_\mathrm{H} |U|/(3M)$. With $|U| \approx GM^2/R$ it reads $k_B \bar{T} \approx GM\mu m_\mathrm{H}/(3R)$, and for the uniform sphere, $|U| = \frac{3}{5} GM^2/R$ and $M = 5\sigma^2 R/G$, it is exactly $\mu m_\mathrm{H} \sigma^2$ again. Gas this hot shines in X-rays by bremsstrahlung (Eq. (15.16) and Problem 15.7.1), and X-ray telescopes measure its temperature and density at each radius. Hydrostatic equilibrium then gives the mass. With $P = \rho k_B T/(\mu m_\mathrm{H})$,

$$-\frac{G\,m(r)\,\rho}{r^2} \overset{(1)}{=} \frac{\mathrm{d}P}{\mathrm{d}r} \overset{(2)}{=} \frac{k_B}{\mu m_\mathrm{H}} \left(T \frac{\mathrm{d}\rho}{\mathrm{d}r} + \rho \frac{\mathrm{d}T}{\mathrm{d}r} \right) \overset{(3)}{=} \frac{\rho\, k_B T}{\mu m_\mathrm{H}\, r} \left(\frac{\mathrm{d}\ln\rho}{\mathrm{d}\ln r} + \frac{\mathrm{d}\ln T}{\mathrm{d}\ln r} \right). \tag{14.31}$$

(1) Hydrostatic equilibrium, Eq. (10.2).
(2) The ideal gas, differentiated.
(3) $\mathrm{d}\rho/\mathrm{d}r = (\rho/r)\ \mathrm{d}\ln\rho/\mathrm{d}\ln r$, and the same for T.

$$\boxed{m(r) = -\frac{k_B T\, r}{G\mu m_\mathrm{H}} \left(\frac{\mathrm{d}\ln\rho}{\mathrm{d}\ln r} + \frac{\mathrm{d}\ln T}{\mathrm{d}\ln r} \right).} \tag{14.32}$$

Only the temperature and the slopes of the profiles enter, not the amount of gas. For gas at one temperature with $\rho \propto r^{-2}$, the bracket is -2 and $m = 2k_B T r/(G\mu m_\mathrm{H})$. With $k_B T = \mu m_\mathrm{H} \sigma^2$ this is $m = 2\sigma^2 r/G$, the isothermal sphere of Section 12.5 with $v_\mathrm{circ}^2 = 2\sigma^2$, which is a second route to the same mass. Inside 1 Mpc it is $4.6 \times 10^{14}\,M_\odot$. The gas itself holds several times the mass of the stars, yet stars and gas together are only about a sixth of the total.

The bending of light. General relativity bends a ray that passes a mass M at impact parameter b through the angle $4GM/(bc^2)$, twice the Newtonian value. For a lens that is not a point, M is the mass inside the circle of radius b on the sky, if the lens is round. Put a source directly behind a lens, at distance D_S from us, with the lens at D_L and D_{LS} between the two (Figure 14.7). A ray we see at a small angle θ from the lens passes it at $b = \theta D_L$. Without bending it would reach the source's distance at the height θD_S. Bent back toward the axis, it reaches that distance at the height $\theta D_S - 4GMD_{LS}/(bc^2)$, and it comes from the source only if this is zero,

$$\theta D_S \overset{(1)}{=} \frac{4GM}{bc^2} D_{LS} \overset{(2)}{=} \frac{4GM}{\theta D_L c^2} D_{LS} \quad \Longrightarrow \quad \theta^2 = \frac{4GM}{c^2} \frac{D_{LS}}{D_L D_S}. \tag{14.33}$$

(1) The height of the bent ray at the source, set to zero, with small angles.
(2) $b = \theta D_L$.

$$\boxed{\theta_E = \sqrt{\frac{4GM}{c^2} \frac{D_{LS}}{D_L D_S}}} \tag{14.34}$$

By symmetry the source appears as a ring of this angular radius, the **Einstein ring**. A source a little off the axis appears as two images or as arcs near the ring. Only the mass inside the ring enters, not how it moves or what state it is in. For $10^{14}\,M_\odot$ with $D_L = 1\,\text{Gpc}$ and $D_S = 2\,\text{Gpc}$, $\theta_E = 20''$, and rich clusters do show arcs of about this size. A galaxy of $10^{11}\,M_\odot$ gives $0.6''$. At such distances the Ds are the distances of Chapter 16 that relate size to angle, for which $D_L + D_{LS} \neq D_S$. The formula keeps its form. Masses from lensing, from the galaxies' speeds and from the hot gas agree within a few tens of per cent. In the colliding pair of clusters called the Bullet Cluster, the lensing mass sits with the galaxies, while most of the ordinary matter, the gas, has been slowed and left between them.

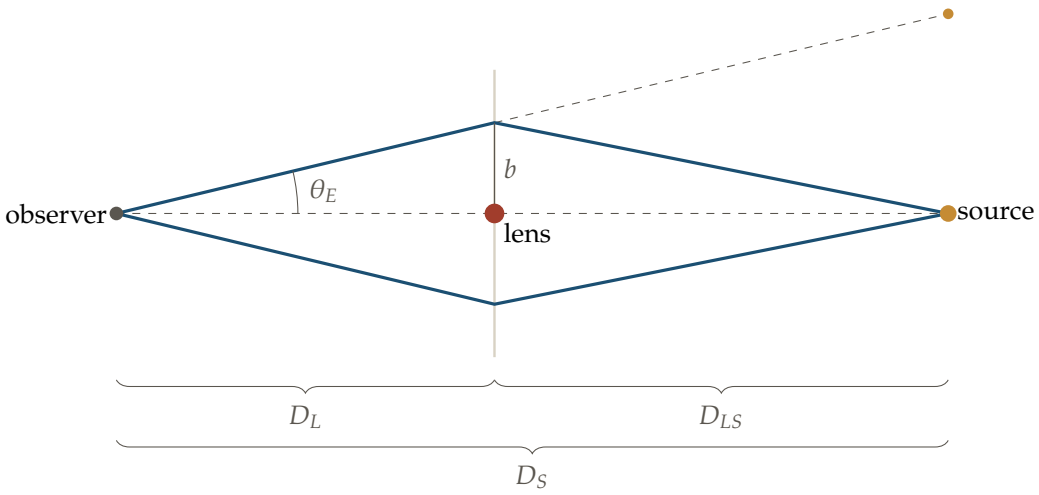


Figure 14.7: A source directly behind a lens. Two rays from the source are bent at the lens plane and reach the observer at the angle θ_E from the lens. The dashed line continues the direction in which the observer sees the upper image. Angles are exaggerated.

A star as a lens. The same bending works for one star. Take a dwarf star of $0.3\,M_\odot$ at $D_L = 4\,\text{kpc}$, in front of a star of the bulge at $D_S = 8\,\text{kpc}$, so $D_{LS} = 4\,\text{kpc}$. Here the distances are small enough that $D_L + D_{LS} = D_S$. Eq. (14.34) gives

$$\theta_E^2 \overset{(1)}{=} \frac{0.3 \times 5.908 \times 10^3\,\text{m}}{8 \times 3.086 \times 10^{19}\,\text{m}} = 7.18 \times 10^{-18}$$
$$\Longrightarrow \quad \theta_E \overset{(2)}{=} 2.68 \times 10^{-9}\,\text{rad} = 0.55\,\text{mas}. \tag{14.35}$$

(1) $4GM_\odot/c^2 = 5.908\,\text{km}$, and $D_{LS}/(D_L D_S) = 1/(8\,\text{kpc})$.
(2) $1\,\text{rad} = 2.063 \times 10^8\,\text{mas}$.

At the lens this angle spans $\theta_E D_L = 2.2\,\text{AU}$. The two images lie about a milliarcsecond apart, far below what a single telescope can separate (Chapter 5). What we see is one star that brightens while the lens passes in front and fades again. This is **microlensing**. As the lens star moves into the line of sight, its gravity bends more and more of the source's light toward the Earth, so the source looks brighter (Figure 14.8). As the lens moves away, the extra light goes. Every lens that passes in front gives one spike of brightness, and a planet around the lens adds a second, shorter spike.

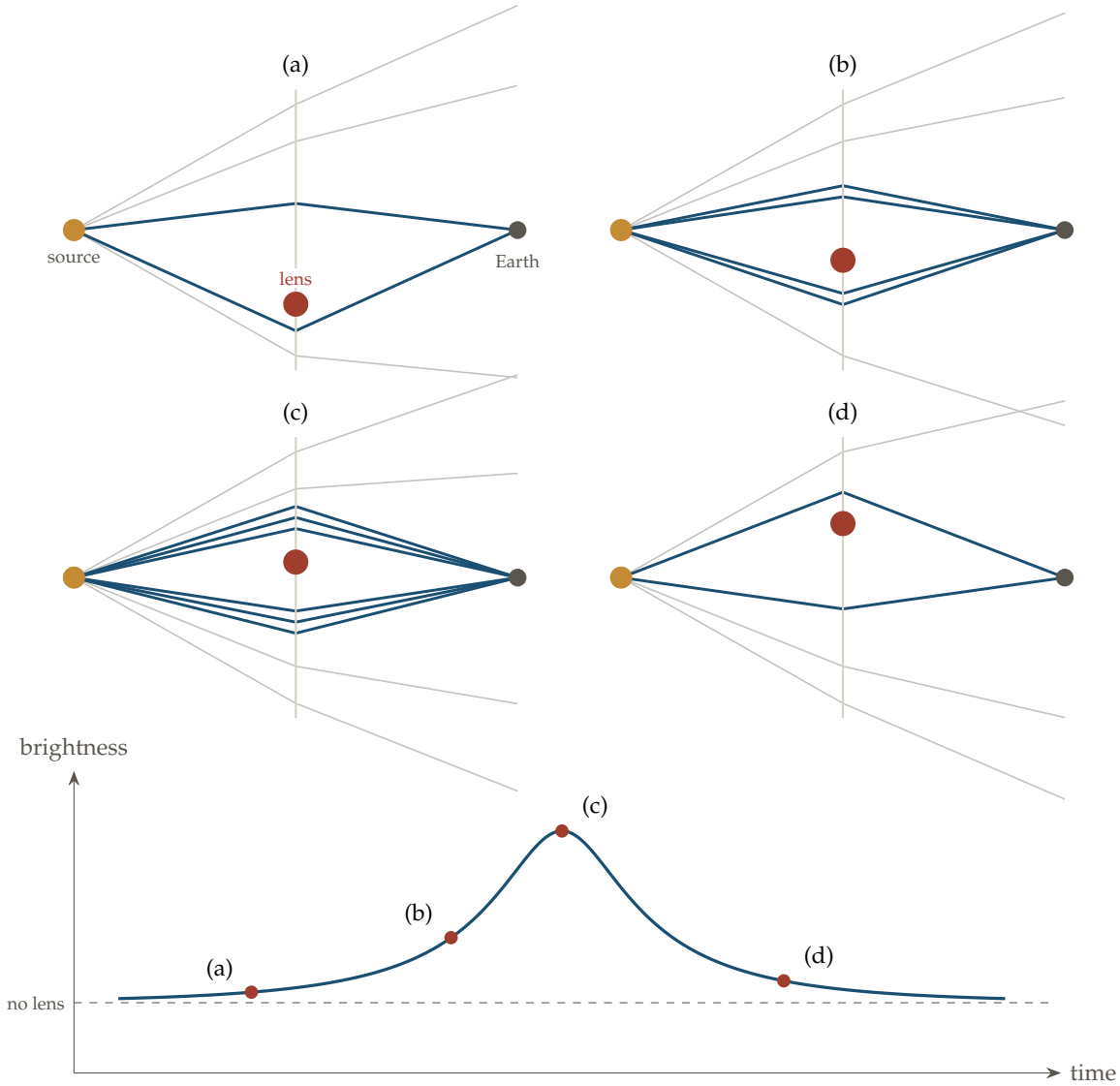


Figure 14.8: Four moments (a) to (d) of a lens passing the line from the source to the Earth, with the brightness curve below. In each panel, rays from the source bend toward the lens. The rays that reach the Earth are dark blue, the others grey. Each dot on the curve marks the moment of one panel. Heights in the panels are exaggerated and not to scale.

The lens equation. Let the source sit a little off the axis, at the angle $u\,\theta_E$ from the lens, so that u is the separation measured in Einstein radii. A ray that we see at the angle θ from the lens, on the same side, passes it at $b = \theta D_L$. As in the chain for θ_E, it reaches the source's distance at

the height $\theta D_S - 4GMD_{LS}/(bc^2)$, and it comes from the source if that height is $u\,\theta_E D_S$. Writing $x = \theta/\theta_E$ for the image position in Einstein radii, with $x < 0$ for an image on the far side of the lens,

$$\begin{aligned} u\,\theta_E D_S &\overset{(1)}{=} \theta D_S - \frac{4GM}{\theta D_L c^2}\,D_{LS} \\ \Longrightarrow \qquad u &\overset{(2)}{=} \frac{\theta}{\theta_E} - \frac{1}{\theta\,\theta_E}\,\frac{4GM}{c^2}\,\frac{D_{LS}}{D_L D_S} \overset{(3)}{=} x - \frac{1}{x}. \end{aligned} \tag{14.36}$$

(1) The height of the bent ray at the source's distance, equal to the source's height, with $b = \theta D_L$.
(2) Divide by $\theta_E D_S$.
(3) The last fraction is θ_E^2 by Eq. (14.34), and $\theta = x\,\theta_E$.

$$\boxed{u = x - \frac{1}{x}, \qquad x_\pm = \frac{u \pm \sqrt{u^2+4}}{2}.} \tag{14.37}$$

The quadratic $x^2 - ux - 1 = 0$ has two roots for every u, so a point lens always makes two images. Their product is $x_+x_- = -1$. One image lies outside the ring on the source's side, the other inside it on the far side. For $u = 0$ both reach $|x| = 1$ and merge into the ring.

The magnification. Bending turns a ray but does not change its intensity. In empty space the intensity is constant along a ray (Eq. (11.3)), and general relativity keeps this for bent rays (quoted, not derived). So each image has the surface brightness of the source, and its flux is that surface brightness times the solid angle it covers. The brightening of an image is the ratio of its solid angle to that of the source. To get it we map a small patch of the image onto the source. Use the direction away from the lens and the direction around it. The lens is round, so an image and its source lie on one line through the lens, a step away from the lens stays away from it, and a turn by a small angle $\delta\phi$ about the lens turns both by the same $\delta\phi$. A patch of the image of size δx by $x\,\delta\phi$ maps onto a patch of the source of size δu by $u\,\delta\phi$,

$$\begin{pmatrix} \delta u \\ u\,\delta\phi \end{pmatrix} = \begin{pmatrix} \mathrm{d}u/\mathrm{d}x & 0 \\ 0 & u/x \end{pmatrix} \begin{pmatrix} \delta x \\ x\,\delta\phi \end{pmatrix}. \tag{14.38}$$

Areas scale by the determinant of the matrix. The image is brighter than the source by the inverse ratio of areas, $1/|\det|$, and

$$\det = \frac{\mathrm{d}u}{\mathrm{d}x}\,\frac{u}{x} \overset{(1)}{=} \left(1 + \frac{1}{x^2}\right)\left(1 - \frac{1}{x^2}\right) = 1 - \frac{1}{x^4}. \tag{14.39}$$

(1) From Eq. (14.37), $\mathrm{d}u/\mathrm{d}x = 1 + 1/x^2$ and $u/x = 1 - 1/x^2$.

We see the two images as one star, so we add their brightenings. With $x_+ > 1$ and $x_- = -1/x_+$,

$$\begin{aligned} A &\overset{(1)}{=} \frac{1}{\left|1 - x_+^{-4}\right|} + \frac{1}{\left|1 - x_-^{-4}\right|} \overset{(2)}{=} \frac{x_+^4}{x_+^4 - 1} + \frac{1}{x_+^4 - 1} \\ &= \frac{x_+^4 + 1}{x_+^4 - 1} \overset{(3)}{=} \frac{x_+^2 + x_-^2}{x_+^2 - x_-^2} \overset{(4)}{=} \frac{u^2 + 2}{u\sqrt{u^2+4}}. \end{aligned} \tag{14.40}$$

(1) The sum over the two images of $1/|\det|$.
(2) $x_-^{-4} = x_+^4$, and $x_+^4 > 1$.
(3) Divide top and bottom by x_+^2, and $1/x_+^2 = x_-^2$.

(4) $x_\pm^2 = \frac{1}{2}\left(u^2 + 2 \pm u\sqrt{u^2+4}\right)$ from Eq. (14.37), so the sum is u^2+2 and the difference is $u\sqrt{u^2+4}$.

$$\boxed{A(u) = \frac{u^2+2}{u\sqrt{u^2+4}}.} \tag{14.41}$$

The brightening depends only on the separation in Einstein radii. The mass and the distances only set the size θ_E. Bending does not depend on wavelength, so the star brightens by the same factor in every colour. Two checks. Far from the lens, $A = 1 + 2/u^4 + \ldots$, which goes to 1. Close to it, $A = 1/u + 3u/8 + \ldots$, which grows without limit for a source exactly behind a point lens. At $u = 1$, $A = 3/\sqrt{5} = 1.34$.

The light curve. The lens and the source drift apart on the sky at a steady angular speed, their relative proper motion μ_{rel} (Section 3.3). Let the closest approach be u_0 Einstein radii, at the time t_0. The path of the source relative to the lens is a straight line, so the separation is the hypotenuse of u_0 and the distance covered along the line,

$$\boxed{u(t) = \sqrt{u_0^2 + \left(\frac{t-t_0}{t_E}\right)^2}, \qquad t_E = \frac{\theta_E}{\mu_{\text{rel}}},} \tag{14.42}$$

where t_E is the time to cross one Einstein radius. A relative speed of $150\,\text{km}\,\text{s}^{-1}$ at 4 kpc is $\mu_{\text{rel}} = 150/(4.74 \times 4000)''/\text{yr} = 7.9\,\text{mas/yr}$ by Eq. (3.21), so $t_E = 0.55/7.9\,\text{yr} = 26$ days. Figure 14.9 puts Eq. (14.42) into Eq. (14.41). The curve rises and falls symmetrically about t_0, it peaks at $A(u_0) \approx 1/u_0$, it is the same in every colour, and it happens once. Variable stars rarely do all four, which is how surveys such as OGLE pick out lensing events among the millions of bulge stars they watch every night. The same events, looked for against the Large Magellanic Cloud, showed that the dark halo is not made of compact bodies (Section 14.2).

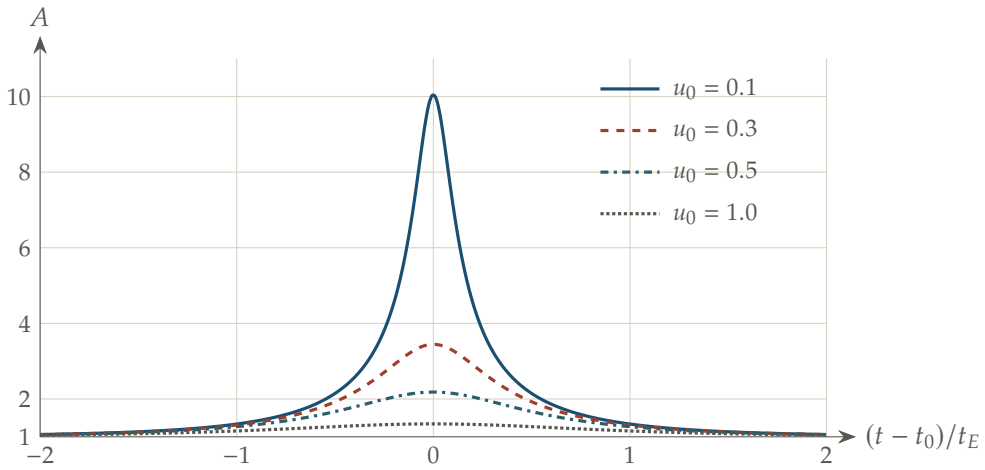


Figure 14.9: The brightening A of a source against time, from Eqs. (14.41) and (14.42), for four closest approaches u_0. Time is in units of the Einstein time t_E.

Planets by microlensing. If the lens star has a planet of mass m_p, the planet is a second, much smaller lens. When one of the two images of the source passes close to the planet, the planet bends its light a little more and adds a short spike or dip to the smooth curve. By Eq. (14.34), the planet's own Einstein angle is $\theta_E\sqrt{m_p/M}$, so the spike lasts about $\sqrt{m_p/M}\,t_E$. Around the star above, an Earth gives $\sqrt{3.0\times10^{-6}/0.3}\times 26\,\text{d} = 2$ hours and a Jupiter $\sqrt{9.5\times10^{-4}/0.3}\times 26\,\text{d} = 1.5$

days. The images sit near the Einstein ring, so the method sees planets most easily at a projected distance of about $\theta_E D_L \approx 2$ AU from their star. For a dwarf star of a few tenths of a solar mass, that is far out and cold. The planet OGLE-2005-BLG-390Lb, found this way in 2005, has about 5.5 Earth masses and lies about 2.6 AU from a dwarf of about 0.2 $M_\odot$, some 6.6 kpc away. The other methods of Section 6.20 would hardly see it. Taking 0.22 $M_\odot$ for the star, the planet moves at $29.8\sqrt{0.22/2.6} = 8.7\,\mathrm{km\,s^{-1}}$, so the star swings by $(5.5\times3.0\times10^{-6}/0.22)\times8.7\,\mathrm{km\,s^{-1}} = 0.65\,\mathrm{m\,s^{-1}}$, over a period of $\sqrt{2.6^3/0.22} = 8.9$ years. A transit needs the orbit to lie within about R_*/a of edge on, with R_* the star's radius and a the orbit's, a few chances in ten thousand. Microlensing needs no light from the planet or from its star, and it reaches across the Galaxy. Its price is that each event happens once, so the planet cannot be looked at again.

14.5 Normal Galaxies

Rotation and random stellar motions give us two ways to weigh a galaxy (Eqs. (14.8) and (14.30)). If we can also connect those motions to its true luminosity, its observed brightness will tell us its distance. That connection depends on how the galaxy is built, so we first distinguish the main kinds of galaxies.

Think of apples from one orchard. They are all about equally dense. Measure the width of an apple and you know its weight without a scale, because the weight grows as the cube of the width. One fixed property turns one measured quantity into another, through a power law. Spiral disks are like the apples. They have nearly the same surface brightness, and their stars nearly the same mass-to-light ratio. With $v_{\mathrm{circ}}^2 = Gm/R$ these two fixed ratios make the luminosity grow as the fourth power of the rotation speed. The width of the 21 cm line then gives the luminosity, and the luminosity with the observed flux gives the distance. A power law follows from the virial theorem and two fixed ratios.

This fourth-power relation has a strong effect: a spiral that rotates twice as fast is $2^4 = 16$ times as luminous, a difference of $2.5\log_{10}16 = 3.0$ magnitudes. For comparison, doubling the width of an apple of fixed density makes it $2^3 = 8$ times as heavy. Elliptical galaxies have a related scaling law based on random speeds. We will use both relations to estimate distances.

The Hubble sequence. Hubble sorted galaxies by their look into a sequence. He drew it as a tuning fork (Figure 14.10). It sorts looks and is not a path of evolution, and the names early and late do not mean age. An elliptical is a smooth oval of stars with no disk and no arms. An ordinary spiral has arms that run in all the way to the central bulge. A barred spiral has an elongated bar of stars across its centre, with the arms attached near the two ends of the bar. The lenticulars sit at the junction where the ellipticals meet the two spiral branches, with a disk like a spiral but no arms. **Ellipticals** are smooth and featureless, labelled E0 to E7 by how flattened they appear. **Lenticulars**, S0, have a disk but no spiral arms. **Spirals** run from Sa, with a large bulge and tightly wound arms, to Sd, with a small bulge and open arms, and those with a bar form a parallel sequence SBa to SBd. **Irregulars** fit nowhere. Along the sequence from ellipticals to late spirals, the fraction of gas and of young blue stars grows, and the support against gravity changes. Ellipticals are held up by random motions of 100 to 300 $\mathrm{km\,s^{-1}}$ and rotate little. Their stars are old and red, and they have little cold gas. Disks are held up by rotation at 100 to 300 $\mathrm{km\,s^{-1}}$, with random motions of only a few tens. A collisionless galaxy keeps the motions it formed with (Section 14.3), so its shape records its history. Mergers of disk galaxies, which mix the orbits violently, can make ellipticals.

Andromeda and the Local Group. The nearest large spiral, the Andromeda galaxy (M31), is about 2.5 million light years, or 770 kpc, away. Nearly every galaxy moves away from us with the expansion of the universe (Chapter 16), but the light of Andromeda is shifted to the

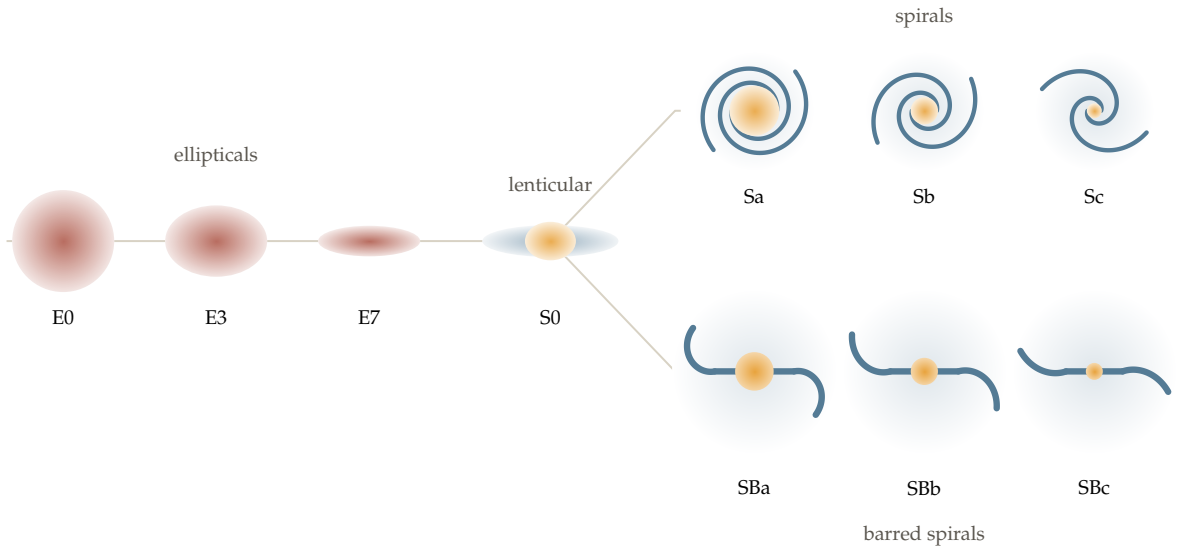


Figure 14.10: The Hubble tuning fork. Ellipticals E0, E3 and E7 on the left, the lenticular S0 at the fork, ordinary spirals Sa, Sb and Sc on the upper branch, and barred spirals SBa, SBb and SBc on the lower branch. Spirals are shown face on. The galaxies are drawn from simple shapes.

blue. Once the Sun's own motion round the Galaxy is taken out, Andromeda and the Milky Way approach each other at about $110\,\mathrm{km\,s^{-1}}$. The expansion alone would carry them apart at $H_0 d = 67.4 \times 0.77 = 52\,\mathrm{km\,s^{-1}}$ (Eq. (16.2)). So the two are bound to each other, the largest members of a small bound group, the **Local Group**. They moved apart after the Big Bang, their gravity stopped them, and now they fall together. A straight-line Kepler orbit (Eqs. (6.42) and (6.52) with $e = 1$) that starts at zero separation and reaches 770 kpc and $110\,\mathrm{km\,s^{-1}}$ after the 13.8 billion years of Eq. (16.13) needs a total mass of about $4 \times 10^{12}\,M_\odot$, many times the mass of the stars of both galaxies, and on that orbit the two meet in about 6 billion years. This is the timing argument mentioned in Problem 6.21.10.

Tully–Fisher from the virial theorem. Let a disk of radius R have mass M inside R and luminosity L. Its rotation speed is $v^2 = GM/R$ by Eq. (14.8). Suppose M/L and the surface brightness, measured by L/R^2, are the same for all such disks. To get L in terms of v alone, we write M and R through these two ratios,

$$L \overset{(1)}{=} \left(\frac{L}{M}\right) M \overset{(2)}{=} \left(\frac{L}{M}\right) \frac{v^2 R}{G} \overset{(3)}{=} \left(\frac{L}{M}\right) \frac{v^2}{G} \left(\frac{R^2}{L}\right)^{1/2} L^{1/2}. \tag{14.43}$$

(1) The mass-to-light ratio.
(2) Eq. (14.8) at the edge of the disk.
(3) $R = (R^2/L)^{1/2} L^{1/2}$.

Dividing by $L^{1/2}$ and squaring,

$$\boxed{L = \left(\frac{L}{M}\right)^2 \left(\frac{R^2}{L}\right) \frac{v^4}{G^2}.} \tag{14.44}$$

This is the **Tully–Fisher relation**, $L \propto v^4$. The power four comes from the virial theorem and the fixed surface brightness. The constant in front holds the two ratios that were assumed fixed. The observed relation is close to $L \propto v^4$ in the near infrared, where the light traces the old stars, and it is tight. Disks do have nearly the same central surface brightness, which explains part of it. Yet v is set mostly by the dark halo and L by the stars, and galaxies of low surface brightness

lie on the same relation. Why the halo and the stars keep the relation so tight is not explained by the virial theorem. Counting the gas with the stars makes the relation tighter still.

Faber–Jackson. For an elliptical galaxy, the uniform-sphere form of the virial theorem from Section 14.4, $M = 5\sigma^2 R/G$, takes the place of $M = v^2 R/G$. The same three steps with v^2 replaced by $5\sigma^2$ give

$$\boxed{L = 25 \left(\frac{L}{M}\right)^2 \left(\frac{R^2}{L}\right) \frac{\sigma^4}{G^2},} \tag{14.45}$$

the **Faber–Jackson relation** $L \propto \sigma^4$. It holds less tightly than Tully–Fisher, because the surface brightness of ellipticals is not fixed. Keeping L/R^2 as a third measured quantity turns the line into a plane in the space of R, σ and surface brightness, the fundamental plane, on which ellipticals lie closely.

Distances. The 21 cm line of a whole spiral is broadened by its rotation. The side moving toward us and the side moving away shift the line by $\pm v \sin \iota$, where ι is the inclination of the disk to the sky ($\iota = 90°$ for a disk seen edge on), so the full width is $W = 2v \sin \iota$. A thin round disk tilted by ι looks like an ellipse with axis ratio $\cos \iota$, which gives ι. Then v gives L by Eq. (14.44), once the constant has been fixed with galaxies of known distance. In magnitudes, $L \propto v^4$ means that the absolute magnitude changes by $-10 \log_{10}$ of the ratio of rotation speeds (Eq. (4.5)), and the distance follows from the distance modulus, Eq. (4.11) (Problem 14.8.5). Tully–Fisher distances reach a few hundred megaparsecs, where the expansion of the universe (Chapter 16) takes over as a measure of distance.

14.6 Telling Stars, Galaxies, Nebulae and Quasars Apart

An image gives an object's position, brightness and a few colours, but a star, a quasar and a distant galaxy can look alike in it. A spectrum adds the evidence needed to tell them apart. Dark lines can mark cooler gas above a stellar surface (Section 11.7); bright lines can mark gas lit by hot stars (Section 8.5); and broad lines can reveal fast gas near a black hole (Section 15.6). The shifts of those lines also give redshifts (Eq. (3.25)). A survey must first choose targets from images and then use their spectra to classify them and measure their distances.

Think of two people seen through a window at night. You see their height and the colour of their coats, and two people of the same height in grey coats cannot be told apart. When they speak, the voice tells more. Its pitch shifts if the speaker walks away, as a siren does. The words say what the speaker is talking about. The roughness of the voice says how much it shakes. In the same way the light of an object is spread into colours, and three things in the spread decide what the object is. They are the overall shape of the colours, the lines on it, and the width and the shift of those lines.

Figure 14.12 compares four objects with redshifts z of -0.000153, 0.000967, 0.07621 and 0.2404. The star and nebula have lines as narrow as the spectrograph can resolve. The galaxy's lines are a few hundred km s^{-1} wide, while the quasar's broad lines span thousands. Line positions give redshift; the kinds of lines and their widths reveal the source. We will first see what the images suggest, then use the spectra to decide what each object is.

What the image says first. The Sloan survey takes its pictures in five filters, u, g, r, i and z, before it takes any spectrum. Two numbers from the pictures sort the objects. The first is the shape. A star is so far away that its image is only the point spread function of the telescope (Chapter 5). The centre of a quasar is also too small to be seen as anything but a point. A galaxy

is larger than the point spread function. The survey measures the brightness of an object twice in the r filter. The PSF magnitude is the light of the best-fitting point, and the model magnitude is the light of the best-fitting smooth galaxy shape. For a point the two agree. For an extended object the point misses the outer light, so $\mathrm{psfMag}_r - \mathrm{modelMag}_r$ is large. A cut at 0.15 works well. Among the objects below, 98 per cent of the stars and 90 per cent of the quasars lie below 0.15, and 100 per cent of the galaxies lie above it. The quasars above the cut are those where the light of the host galaxy adds to the point.

The second number is the colour (Figure 14.11). A star is close to a blackbody (Eq. (4.15)), so its temperature sets all its colours at once. The stars therefore lie on a narrow curve in $u - g$ against $g - r$, the stellar locus. A quasar below z of about 2.2 is bluer in $u - g$ than a star of the same $g - r$. The median $u - g$ is 0.27 for the quasars and 0.95 for the stars with $g - r < 0.4$, and 84 per cent of the quasars have $u - g < 0.6$ against 4 per cent of the stars. This is the ultraviolet excess of the hot disk (Section 15.5). Above z of about 2.2 the Lyman α line moves into the u filter and the forest in front of the quasar dims it (Section 16.6), so those quasars are red in $u - g$ and fall back onto the stellar locus. The galaxies in the figure lie among stars of the same $g - r$. Shape and colour therefore choose what gets a spectrum, and they cannot settle what an object is. The survey also chose its quasar targets partly by these colours, so the figure cannot show how many quasars hide on the stellar locus.

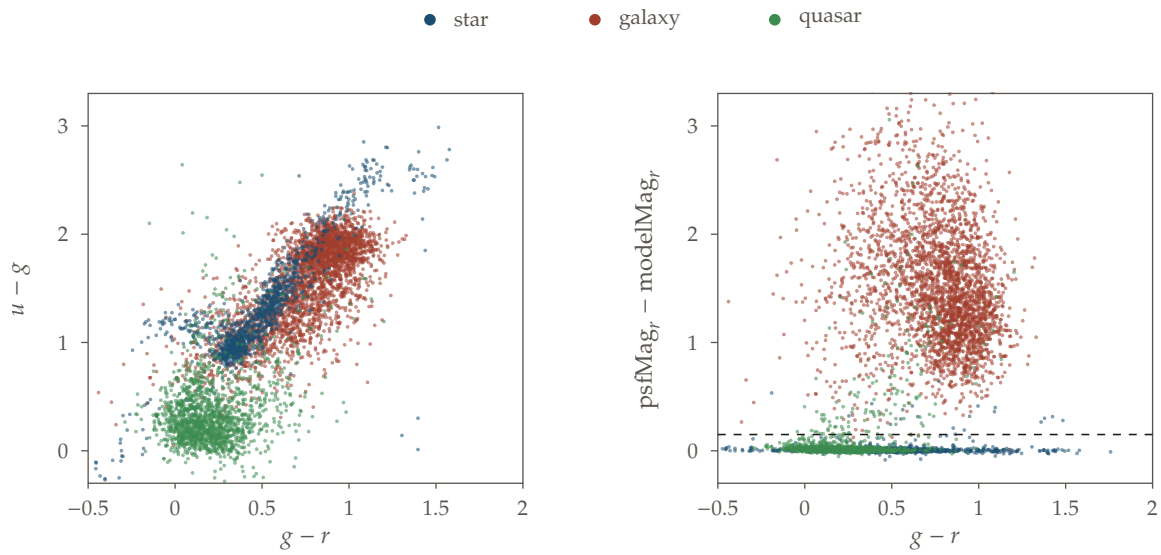


Figure 14.11: Left, $u - g$ against $g - r$ from the PSF magnitudes. Right, $\mathrm{psfMag}_r - \mathrm{modelMag}_r$ against $g - r$, with a dashed line at 0.15. The points are 1151 stars, 2908 galaxies and 1504 quasars of SDSS data release 17, classed by the pipeline from their spectra, with clean photometry, $r < 19.5$ and a random subset of each class. The sample is objects that have spectra, so its mix reflects how SDSS chose its targets.

The spectrum decides. The four spectra of Figure 14.12 are from the SDSS data release 17, all with a pipeline signal to noise of at least 24 per pixel. Their wavelengths are in vacuum, as SDSS gives them, so every rest wavelength used here is a vacuum one. Air wavelengths are shorter by 0.03 per cent, and using them would put about 85 km s^{-1} into every speed.

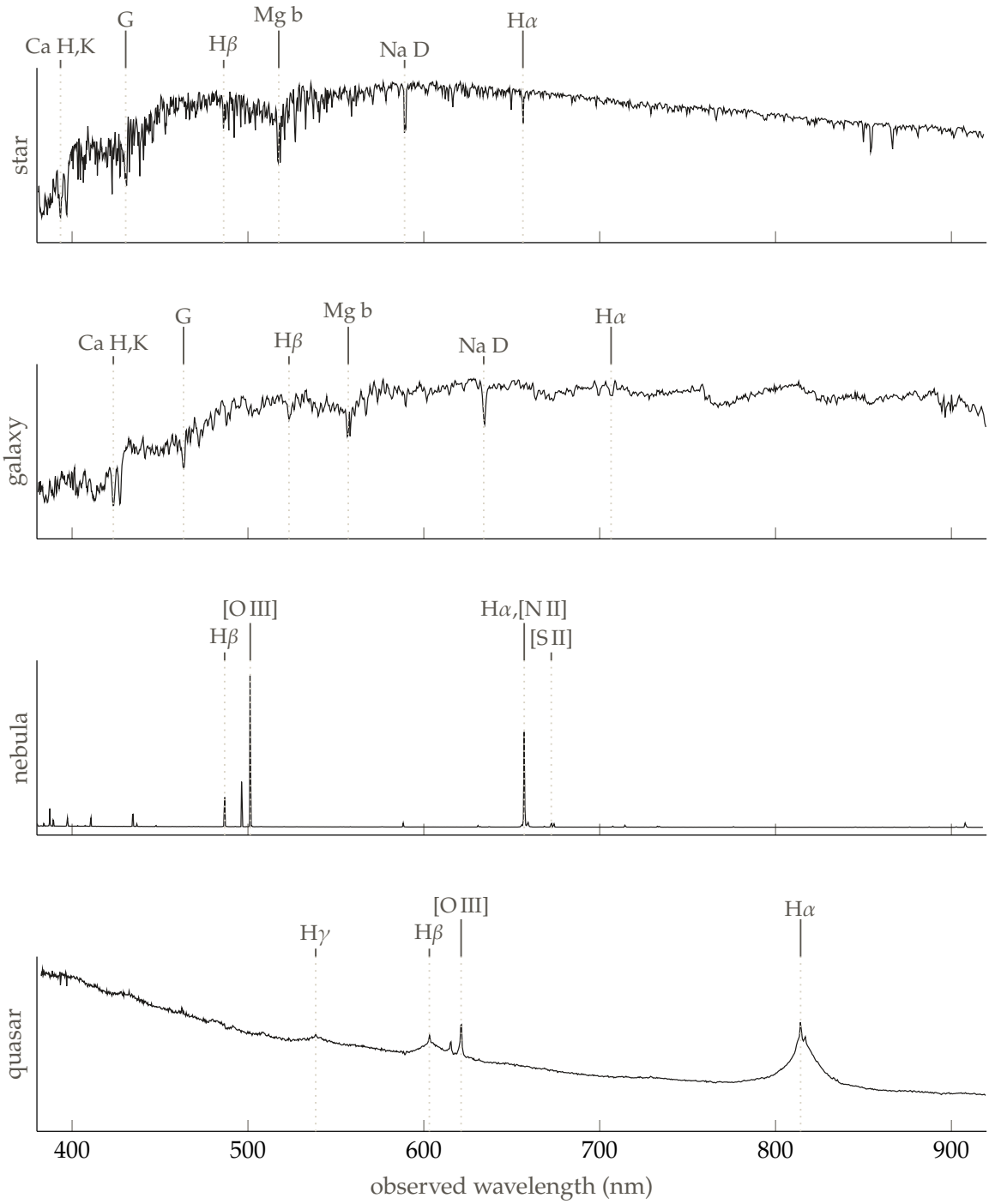


Figure 14.12: Four SDSS spectra on one observed-wavelength axis, each binned by three pixels and divided by its own peak. From the top, a K1 star (plate 2305, MJD 54414, fibre 97), an elliptical-like galaxy (plate 564, MJD 52224, fibre 507), a giant HII region in M101 (plate 1324, MJD 53088, fibre 234) and a quasar (plate 1744, MJD 53055, fibre 630). The marks sit at the observed wavelengths of the named lines, found from the redshifts measured from the lines themselves. The dotted lines run through each panel.

Redshift from a line centre. To get z we fit each line with a Gaussian on a straight continuum and take its centre. Eq. (3.25) then turns the shift into a speed. For the Hα line of the nebula,

$$z \overset{(1)}{=} \frac{\lambda_{\text{obs}} - \lambda_{\text{rest}}}{\lambda_{\text{rest}}} \overset{(2)}{=} \frac{657.104 - 656.461}{656.461} \overset{(3)}{=} 0.000979, \tag{14.46}$$

(1) Eq. (3.25).
(2) The fitted centre and the vacuum rest wavelength of Hα, in nm.
(3) Arithmetic.

and so $v_r = cz = 294\,\text{km s}^{-1}$. The other lines of the nebula give the same z to within a few 10^{-5}. Every spectrum is treated in the same way, and the redshifts of Table 14.3 come from the lines alone, with the pipeline value beside them as a check.

The star. The spectrum of the star (class STAR, subclass K1) is a continuum that rises through the blue and fades in the red, the shape of a surface at a few thousand kelvin (Eq. (4.19)). Dark lines cut it, as expected for cool gas above a hotter layer (Section 11.7). The calcium H and K lines near 393 and 397 nm and the G band near 431 nm are deep in the blue, and Hβ, the magnesium triplet Mg b near 518 nm, the sodium doublet Na D near 589 nm and Hα lie further red. The lines give $z = -0.000153$, which is $-46\,\text{km s}^{-1}$, against $-35\,\text{km s}^{-1}$ from the pipeline. It is a star of our own Galaxy, moving at a few tens of km s^{-1} by its own motion (Chapter 3). The lines are as narrow as the spectrograph makes them. The Hα line has $\sigma = 81\,\text{km s}^{-1}$, only a little above the width of the instrument, which is given below.

How close the star is to a blackbody. The SDSS window, 380 to 920 nm, holds only the top of a blackbody at a few thousand kelvin, so the star's spectrum in Figure 14.12 looks flat rather than like the hump of Eq. (4.19). Figure 14.13 puts the same spectrum on a wide axis with a blackbody fitted to it between 450 and 900 nm, a least-squares fit of the temperature and the height. The best temperature is $T = 4840\,\text{K}$, and the blackbody follows the measured light to within 3% there. Wien's law, $\lambda_{\text{max}} = 2898\,\mu\text{m K}/T$, puts its peak at 0.60 μm, inside the window. Below about 450 nm the star falls well under the curve. A cool star has a great many metal lines in the blue, and together they block much of the light there. Even away from the lines a star is not one blackbody, because the light at each wavelength comes from the depth where the gas turns transparent at that wavelength, and deeper layers are hotter (Section 11.8).

The galaxy. The galaxy is a sum of many stars, mostly old ones, so its spectrum looks like the spectrum of a cool star, with a step at 430 nm that is the 400 nm break of its stars moved to the red. It shows the same dark lines as the star, Ca H and K, the G band, Mg b and Na D, but at other places. The calcium K line has moved from 393.5 to 423.5 nm, and the lines give $z = 0.07621$. A speed $cz = 22847\,\text{km s}^{-1}$ is no longer a small speed compared with c. For so large a shift the expansion of the universe, and not the motion through space, sets most of the redshift (Chapter 16), and the shift is a measure of distance. The lines are also broader than the star's. Single features have σ between 335 and 380 km s^{-1} in our fits, against 81 km s^{-1} for the star. Every star of the galaxy has its own line, shifted by its own speed, and the sum is a line broadened by the random speeds of the stars. Part of the width of a single feature comes from several lines lying together, as Mg b does, and from the wide wings of the calcium lines. The pipeline fits the whole spectrum to a mix of star spectra and finds a velocity dispersion of 230 km s^{-1}. This is the σ of Eq. (14.30), the speed that holds the stars of the galaxy up against its own gravity.

The nebula. The spectrum of the nebula has almost no continuum. It is a set of narrow bright lines, Hβ, [O III] at 496 and 501 nm in rest wavelength, Hα with [N II] on both sides of it, the

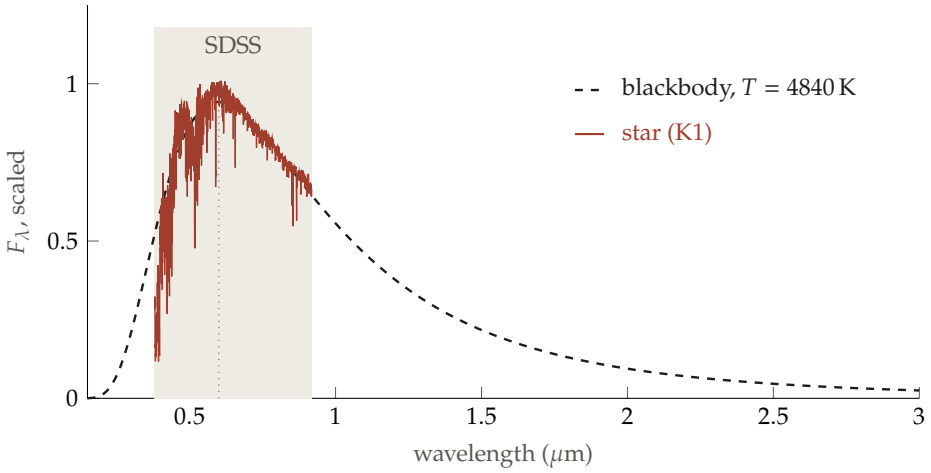


Figure 14.13: The spectrum of the K1 star of Figure 14.12 (red) against wavelength from 0.15 to 3 μm, with the blackbody of $T = 4840$ K fitted to it between 450 and 900 nm (dashed). The shaded band is the range of the SDSS spectrograph and the dotted line marks the peak of the blackbody.

[S II] pair, and in the blue a row of weaker lines. The brackets mark forbidden lines, which a gas so thin that its atoms wait long between collisions can send out (Section 8.4.1). The lines give $z = 0.000967$, or 290 km s^{-1}. A tiny redshift like this is the motion of the gas inside a galaxy next door and is not a measure of distance. M101 as a whole recedes at about 240 km s^{-1}, and its disk rotates. The position of the fibre, RA = 211.1193° and Dec = +54.3980°, is within 1″ of the position that SIMBAD gives for NGC 5471, a giant HII region in the outer disk of M101, about 11′ from the nucleus. The lines come from gas lit by the many hot stars of that region (Section 8.5), which is why starlight adds so little continuum. Its [O II] line at 372.7 nm in rest wavelength falls just short of the blue end of the spectrograph, near 380 nm, and is not in the picture.

The quasar. The spectrum of the quasar has a smooth blue continuum with no absorption lines, falling toward the red. The Hγ, Hβ and Hα lines are broad, and a narrow [O III] stands on the red side of Hβ. The narrow lines give $z = 0.2404$, and the pipeline value is 0.2407. With so large a z the lines of the same atoms have moved far, the Hα line to 814 nm, and the same Balmer lines can be compared in all four panels. The broad lines come from gas moving fast in the pull of the black hole, the narrow ones from gas much further out (Section 15.6).

Line widths and the speed they give. A line that is wider than the thermal motion allows records the spread of speeds of the gas, or of the stars. Each atom or star shifts the line by its own v_r/c, Eq. (3.25), so the width of the line in wavelength is the spread of speeds, in units of c, times the wavelength. For the broad Hβ of the quasar, centred at 603.9 nm with a full width at half maximum of 6096 km s^{-1} in our fit,

$$\Delta\lambda \overset{(1)}{=} \lambda_{\text{obs}} \frac{\Delta v}{c} \overset{(2)}{=} 603.9\,\text{nm} \times \frac{6096}{299\,792} \overset{(3)}{=} 12.3\,\text{nm}. \tag{14.47}$$

(1) Eq. (3.25) applied to a spread of speeds Δv.
(2) The fitted centre of the line and its width.
(3) Arithmetic.

$$\boxed{\frac{\Delta\lambda}{\lambda_{\text{obs}}} = \frac{\Delta v}{c}} \tag{14.48}$$

The pipeline fit of the same line gives 4417 km s^{-1}. The two differ because the line is not a Gaussian, so its fitted width depends on the fit, and only the size of the width, thousands of km s^{-1}, can be trusted. Thermal motion at 10^4 K gives only 9 km s^{-1} (Section 15.6), so the width is bulk motion. The narrow [O III] line has a width of 500 km s^{-1}, a few hundred as Section 15.6 says for the narrow-line region.

What the spectrograph itself does to a line. A spectrograph spreads every sharp line over a small width of its own. For the files used here the resolving power is about 2000, so the narrowest line is no narrower than $c/2000$. Its width in the files is 151 km s^{-1}, and $299\,792/151 \approx 2000$. The lines of the nebula have 181 km s^{-1} for Hα and 164 km s^{-1} for [O III], only a little above the instrument's 153 km s^{-1} at those places. The widths of two independent Gaussians add in square, so taking out the instrument leaves 96 and 53 km s^{-1}. These are rough, since a small error in the instrument width changes them a lot. The measured width of a narrow line is mainly the spectrograph's. It is not the 10 to 30 km s^{-1} of a planetary nebula shell (Section 8.5), which is a small shell of one star, and the gas of a giant HII region, stirred by many stars, may well move faster. The narrow lines of the star are unresolved in the same way.

Kind	Continuum	Lines	Line width	z	z (pipeline)
Star	thermal, a few thousand kelvin	dark, Ca H and K, G, Mg b, Na D, Balmer	instrument's, σ = 81 km s^{-1}	−0.000153	−0.000118
Galaxy	sum of old stars, break at 400 nm	the same dark lines, redshifted	stars' random speeds, 230 km s^{-1}	0.07621	0.07628
Nebula	almost none	bright, narrow, forbidden and Balmer	instrument's, FWHM 181 km s^{-1}	0.000967	0.000970
Quasar	blue, smooth	broad Balmer, narrow [O III]	broad FWHM 4417 to 6096 km s^{-1}, narrow 500	0.2404	0.2407

Table 14.3: The four spectra of Figure 14.12 against the clues of continuum, lines, line width and redshift. The redshift is the mean of the lines fitted here, and the last column is the pipeline's.

How the pipeline decides. The SDSS pipeline does not read a spectrum as a person does. It holds a set of template spectra for each of three classes, STAR, GALAXY and QSO, shifts each template to a trial redshift, scales it to the data with a smooth polynomial added, and compares the two pixel by pixel,

$$\chi^2(z) = \sum_i \frac{[f_i - s\,T(\lambda_i/(1+z))]^2}{\sigma_i^2}. \tag{14.49}$$

Here f_i and σ_i are the flux and its uncertainty in pixel i at wavelength λ_i, T is a template of one class, and s its scale. The class and the z with the least χ^2 over all classes and all trial redshifts are the label of the object. The gap to the second-best fit tells how sure the label is, and a flag is set when the fit is poor. Our nebula carries a flag of 16, which means many pixels lie far from the best fit, as they do when sharp lines meet smooth templates.

SDSS has only these three classes, so a nebula is filed as GALAXY. The spectrum of our HII region in M101 is labelled GALAXY, with the subclass STARBURST. The galaxy templates are the nearest match, because they also contain emission lines from gas. The label therefore says which template fits best and does not say that the object is a galaxy. Astronomers recognise the nebula by three things the pipeline does not use. Its redshift is within a few hundred km s^{-1} of zero, so it lies in our neighbourhood. Its lines are strong and narrow. And it has almost no starlight. It is a part of M101, a galaxy of stars, and so the label is not wrong, but it is not the whole truth. The same care applies to every label, since a class is a statement about the best-matching template and the physics must still be read from the lines. For the maps that

follow, the class and the redshift are what count. The redshift of a galaxy turns into a distance, and every point of the maps of the next section, the walls, the filaments and the voids, is the redshift of one spectrum read in this way.

14.7 Groups, Clusters and the Cosmic Web

We have used the motions of stars to weigh galaxies and the motions of galaxies to weigh clusters (Section 14.4). Those virial estimates assume that each system has had time to settle. On still larger scales, this assumption needs to be checked.

Galaxies are not alone. Many galaxies are held together in a crowd by their own gravity. Up to about fifty galaxies we call the crowd a group, and we live in one, the Local Group. From about fifty to thousands we call it a cluster. Groups and clusters are the same kind of object at two sizes, like a village and a town. Some distance beyond the crowds lie looser regions, the superclusters, where many groups and clusters stand close together, like a district that holds villages and towns. The universe is lumpy, with its mass in these regions and nearly empty voids in between, and the superclusters line up along the filaments and walls of the cosmic web, the largest structures known. It is not clear which of these are real systems that have settled and which are only chance crowds.

Think of dots drawn on a rubber sheet that is slowly stretched. Most dots drift apart as the sheet grows. A few dots are tied together by springs. A tied bunch stays a bunch only if the springs pull it back before the sheet has stretched much. If the bunch swings back and forth many times while the sheet grows a little, it is a bunch for good. If the springs are slow, the sheet wins and the bunch is pulled apart.

The universe is the sheet (Chapter 16) and gravity is the spring. A structure of galaxies has settled if a galaxy crosses it many times within the age of the universe. So the test compares two times (P6), the time a galaxy needs to cross the structure and the age of the universe.

A cluster of radius 1.5 Mpc with galaxies moving at 1000 km s^{-1} is crossed in 1.5 billion years, a tenth of the age of the universe, 13.8 billion years. A supercluster 160 Mpc across with speeds of 600 km s^{-1} needs 130 billion years, about 9.4 times the age. The cluster has had time to settle and the supercluster has not.

The picture is a ladder of five boxes, each showing a region larger than the one before it. It starts with a single galaxy, goes through a group and a cluster, which are one kind of object, and a supercluster, and ends with the filaments and voids (Figure 14.14). A small red square in each box marks the box before it.

The same comparison of two times sorted star systems by their relaxation time (Section 14.3) and decides whether a gas cloud collapses (Section 10.4). Here it sorts the structures of the universe. It gives the ladder of structures from the Milky Way to the cosmic web, a test that says which rungs have settled, and a map of the web itself.

The Milky Way and its satellites. The first rung is one galaxy. The Milky Way is a disk about 30 kpc across in a dark halo of about $10^{12}\,M_\odot$ (Table 14.1). Within about 300 kpc lie dozens of small satellite galaxies. The two largest are the Large and Small Magellanic Clouds, 50 and 60 kpc away, irregular galaxies of about 10^{10} and a few $10^9\,M_\odot$. The rest are dwarf spheroidals, smooth and without gas, with 10^3 to 10^7 stars each and a mass mostly dark.

Groups. A group is a crowd of up to about fifty galaxies bound by their own gravity. We live in one, the Local Group. Its large members are the Milky Way, Andromeda and the spiral M33, with about eighty dwarfs, mostly gathered round the Milky Way and Andromeda, in a region

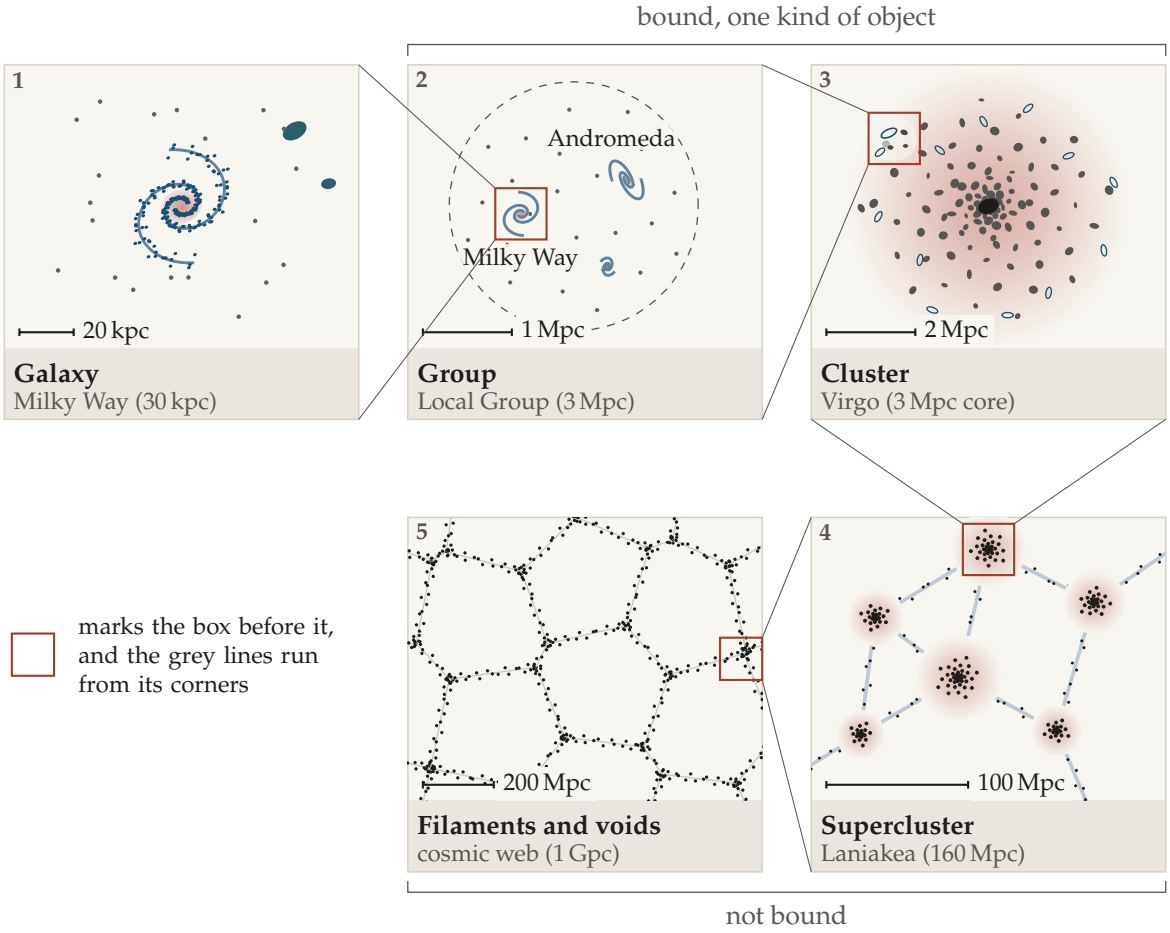


Figure 14.14: A zoom ladder of structure, drawn schematically, with a scale bar in each box. Box 1 shows a two-armed spiral galaxy, the Milky Way, with the Large and Small Magellanic Clouds in teal and a few dwarf galaxies. Box 2 shows a group, the Local Group, with the Milky Way, Andromeda and Triangulum among small dwarfs. Box 3 shows a cluster, the Virgo cluster, as a dense round crowd of mostly smooth oval galaxies that thickens toward the centre, inside a faint glow of hot gas, with a small group falling in at its edge. Box 4 shows a supercluster, Laniakea, as clusters joined by filaments. Box 5 shows the cosmic web as walls and filaments of galaxies that meet at nodes around empty voids. The boxes run along the top row and back along the bottom row. Each red square sits on the object that the box before it shows, enlarged, and the grey lines run from the corners of that box. The squares are not to scale. The brackets group the boxes of bound structures and of unbound ones.

about 3 Mpc across. The timing argument of Section 14.5 gave the mass of the pair as about $4 \times 10^{12}\, M_\odot$. A group holds about 10^{12} to $10^{14}\, M_\odot$.

Clusters. A cluster is the same kind of object as a group but larger, from about fifty to thousands of galaxies. The line between the two is a convention, like the line between a village and a town. A cluster holds 10^{14} to $10^{15}\, M_\odot$ inside a few megaparsecs. Its members move at 500 to 1500 km s^{-1}. Their motions weigh the cluster by Eq. (14.30), and its gas, at 10^7 to 10^8 K, shines in X-rays and weighs it again by Eq. (14.32) (the hot gas in Section 14.4). Most of the cluster is dark matter. The nearest large cluster is Virgo, about 16.5 Mpc away, with galaxies moving at about 700 km s^{-1}. Coma, the cluster that Zwicky weighed, is about 100 Mpc away and has the speed spread of 1000 km s^{-1} used above.

Superclusters. A supercluster is a different kind of object. It is a region where many groups and clusters lie close together. The Local Group is on the edge of the Local or Virgo Supercluster, about 30 Mpc across, with the Virgo cluster near its middle. It is part of a larger region called Laniakea, about 160 Mpc across, with about $10^{17}\,M_\odot$ and about 10^5 galaxies. Its edge is drawn where the galaxies stop moving toward a common centre. A supercluster is not a bound system as a whole. Its galaxies move at a few hundred km s^{-1} and it is still spreading with the expansion of the universe, as the crossing times below show.

A lumpy universe. The mass of the universe is not spread evenly. It sits in groups, clusters and superclusters, and between them lie voids, regions tens of megaparsecs across, up to about 100 Mpc, that are nearly empty of galaxies. Voids fill most of the volume.

Filaments and walls. Superclusters line up along long **filaments** and flat **walls**, which meet at the densest knots and surround the voids. The filaments and walls are the largest structures known, up to a few hundred megaparsecs long. Together with the voids they form the **cosmic web**. Section 16.5 shows how the growth of density ripples makes sheets first, then filaments, then clusters where the filaments meet. Beyond a few hundred megaparsecs the web averages out and the universe looks smooth, as assumed in Chapter 16.

Which structures have settled. Take a structure of radius R whose galaxies move at a typical speed v. The crossing time is R/v (Eq. (14.25)). The age of the universe is close to the Hubble time $1/H_0$, since $H_0 t_0 = 0.95$ for the values of Eq. (16.13). So

$$\frac{t_{\text{cross}}}{t_0} \overset{(1)}{=} \frac{R/v}{t_0} \overset{(2)}{=} \frac{H_0 R}{v}\,\frac{1}{H_0 t_0} \overset{(3)}{=} \frac{H_0 R}{v}\,\frac{1}{0.95}. \tag{14.50}$$

(1) The crossing time, R/v.
(2) Multiply and divide by H_0.
(3) $H_0 t_0 = 0.95$.

$$\boxed{\frac{t_{\text{cross}}}{t_0} \approx \frac{H_0 R}{v}.} \tag{14.51}$$

Here $H_0 R$ is the speed at which the expansion would carry the edge of the structure away from its centre (Eq. (16.2)). The ratio compares that speed with the speed of the galaxies. It uses only the size and the speed, not the mass, and it is good to about 5 per cent. There are three cases.

- **Ratio far below one.** A galaxy crosses the structure many times in the age of the universe. The orbits have mixed and the virial theorem applies. For Coma, $H_0 R = 67.4 \times 1.5 = 101$ km s^{-1} against 1000 km s^{-1}.
- **Ratio near one.** A galaxy has crossed the structure about once. It can be bound, but it has not settled.
- **Ratio above one.** Gravity has not had time to act. The structure is a clump in the expansion that has not collapsed. For Laniakea, $H_0 R = 67.4 \times 80 = 5400$ km s^{-1} against 600 km s^{-1}.

Table 14.4 gives the numbers for the structures above. The Milky Way has crossed itself hundreds of times, and Virgo and Coma about ten times. For them the steady-state assumption behind Eq. (14.30) holds. The Local Group sits near one. This agrees with the timing argument, which has Andromeda and the Milky Way falling together for the first time and meeting in about 6 billion years. A virial mass would be unreliable for such a pair, and that is why its mass was found from the orbit. Superclusters lie above one. They have not settled, and the pull of their own mass is too weak to stop their expansion, so they are not bound. The same holds for

Structure	Size	Members	M ($M_\odot$)	v (km s^{-1})	t_{cross} (Gyr)	t_{cross}/t_0	Bound, settled?
Milky Way	30 kpc	10^{11} stars	10^{12}	220	0.067	0.0048	yes
Local Group	3 Mpc	~ 80 galaxies	4×10^{12}	100	15	1.1	bound, not settled
Virgo cluster	3 Mpc	~ 10^3 galaxies	6×10^{14}	700	2.1	0.15	yes
Coma cluster	3 Mpc	~ 10^3 galaxies	10^{15}	1000	1.5	0.11	yes
Local Supercluster	30 Mpc	~ 10^2 groups	10^{15}	300	49	3.5	no
Laniakea	160 Mpc	~ 10^5 galaxies	10^{17}	600	130	9.4	no

Table 14.4: Rounded sizes, memberships, masses and typical speeds of the structures, with the crossing time R/v for R half the size and its ratio to the age of the universe, $t_0 = 13.8\,\text{Gyr}$.

the filaments and walls, which are longer still. They are clumps of the cosmic web that are still being drawn together (Section 16.5).

The survey in three dimensions. Figure 14.15 shows every galaxy of the survey in one block of the northern sky, 106 405 of them out to 285 Mpc. Earth is at the apex of the cone, and the floor of the cone is the plane of the celestial equator. Seen this way, the walls are sheets and the filaments are threads, and the voids are bubbles between them. Near Earth the galaxies are many and far away they are few. Only the brighter galaxies have a measured spectrum far away, so a distant region is fainter and not emptier (see the thin slice below). The wedge of the next figure is a thin slice along this floor, taken further out, where the walls, filaments and voids are easiest to see.

The web in the thin slice. Figure 14.16 lays flat a thin slice of the sky along the floor of the cone of Figure 14.15, taken out to 846 Mpc, and shows the voids and the filaments.

What the map shows. Each dot is one galaxy. The slice holds the galaxies within 2.5° of the celestial equator and between right ascensions 132° and 234°, out to 846 Mpc. The survey has no spectra of this strip below 130° and covers it only in part above 236°, so the map keeps the full strip between the two limits. We sit at the apex and look outward. The galaxies are not spread evenly. They gather on walls and filaments and leave empty voids between them. The long band near the middle, at about 340 Mpc, is the Sloan Great Wall. The slice is thin, so a wall cut across shows as a line and a filament as a chain of dots, and at 400 Mpc its thickness is $2 \times 400 \times \tan 2.5° = 35\,\text{Mpc}$.

Two features come from how the map is made and not from the universe. The dots thin out with distance because only galaxies brighter than a fixed apparent magnitude have a measured spectrum, so far away only the most luminous are seen. A far region is not emptier, it is fainter. And the dots form thin streaks that point at us. A galaxy moving along the line of sight at 1000 km s^{-1} shifts its redshift as if it were $1000/67.4 = 15$ Mpc nearer or farther (Eq. (16.2)), and the galaxies of a cluster, with speeds like these, are smeared along the line of sight.

How the map is made. The spectrum of each galaxy gives its redshift z from the shift of known lines (Eq. (16.4)). For small z the recession speed is cz and Hubble's law (Eq. (16.2)) gives the distance $d = cz/H_0$, which is $0.1 \times 299\,792/67.4 = 445$ Mpc at $z = 0.1$. The map uses the better distance that allows for the expansion changing while the light travels. Light covers the comoving distance $c\,dt/a$ in a time dt (Section 16.1), so

$$D(z) \overset{(1)}{=} \int_{t_e}^{t_0} \frac{c\,dt}{a} \overset{(2)}{=} \int_{a_e}^{1} \frac{c\,da}{a^2 H} \overset{(3)}{=} \int_0^z \frac{c\,dz'}{H(z')}. \tag{14.52}$$

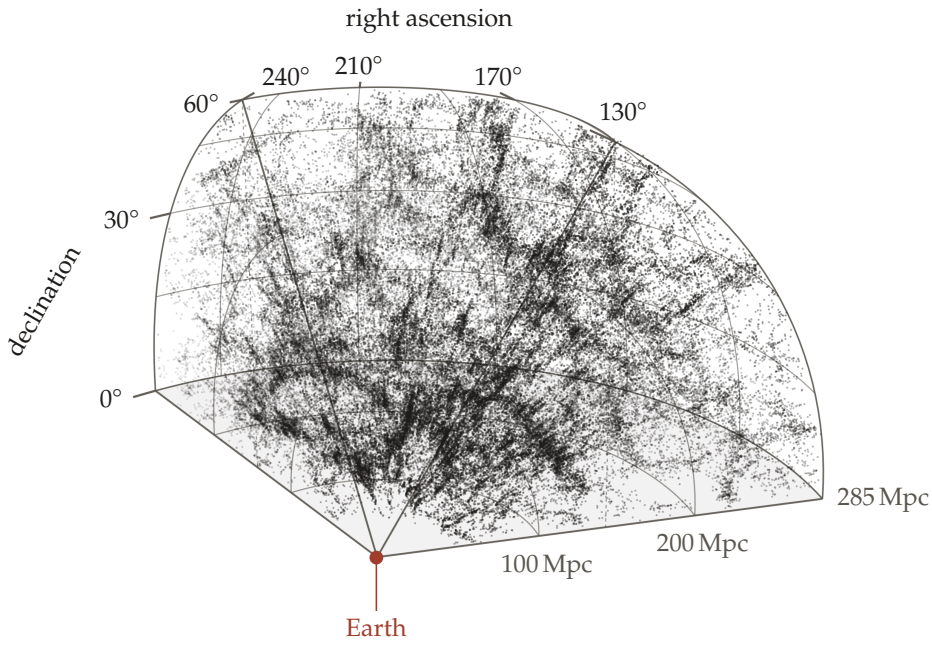


Figure 14.15: 106 405 galaxies of the Sloan Digital Sky Survey (data release 17) with right ascension 130° to 240° and declination 0° to 60°, out to 285 Mpc, placed by their redshifts and seen in perspective. Every galaxy of the block is drawn, and nearer galaxies are darker. The grey lines outline the surveyed volume, a cone with its apex at Earth. The curved far face is a grid of 20° in right ascension and 15° in declination. The shaded floor is the plane of declination 0°, and the arcs on it and on the two side faces are 100 and 200 Mpc from Earth. Tick marks give right ascension along the top edge of the far face and declination up its left edge.

(1) The comoving distance covered by light that left at time t_e.
(2) $dt = da/(aH)$, as in the age integral of Eq. (16.13).
(3) $a = 1/(1+z')$, so $da = -dz'/(1+z')^2$ and $da/a^2 = -dz'$. The minus sign turns the limits around.

Eq. (16.9) gives $H(z) = H_0\sqrt{\Omega_m(1+z)^3+\Omega_\Lambda}$ with $\Omega_m = 0.315$ and $\Omega_\Lambda = 0.685$, and radiation is negligible for $z < 0.2$. The integral is done numerically, and at $z = 0.1$ it gives $D = 434$ Mpc, a little less than cz/H_0 because the expansion rate was higher in the past. Each galaxy is then plotted at the angle given by its right ascension and at the radius $D(z)$.

14.8 Problem Set: Galaxies and the Motions of Stars

14.8.1 Problem 1: The Sun's Orbit and the Local Dark Matter

Problem Statement: The Sun circles the Galactic centre at $v_0 = 230\,\mathrm{km\,s^{-1}}$ at $R_0 = 8.2\,\mathrm{kpc}$. (a) How long does one orbit take? (b) How many orbits has the Sun made in its 4.6×10^9 years? (c) What mass lies inside the Sun's orbit? (d) If the rotation curve is flat and all the mass is in a round halo, what is the density at the Sun, in $M_\odot\,\mathrm{pc^{-3}}$ and in $\mathrm{GeV\,c^{-2}\,cm^{-3}}$? (e) How much of this density lies in a slab within 1.1 kpc of the plane, per unit area? Compare with the $20\,M_\odot\,\mathrm{pc^{-2}}$ by which the weighed disk exceeds the visible stars and gas.

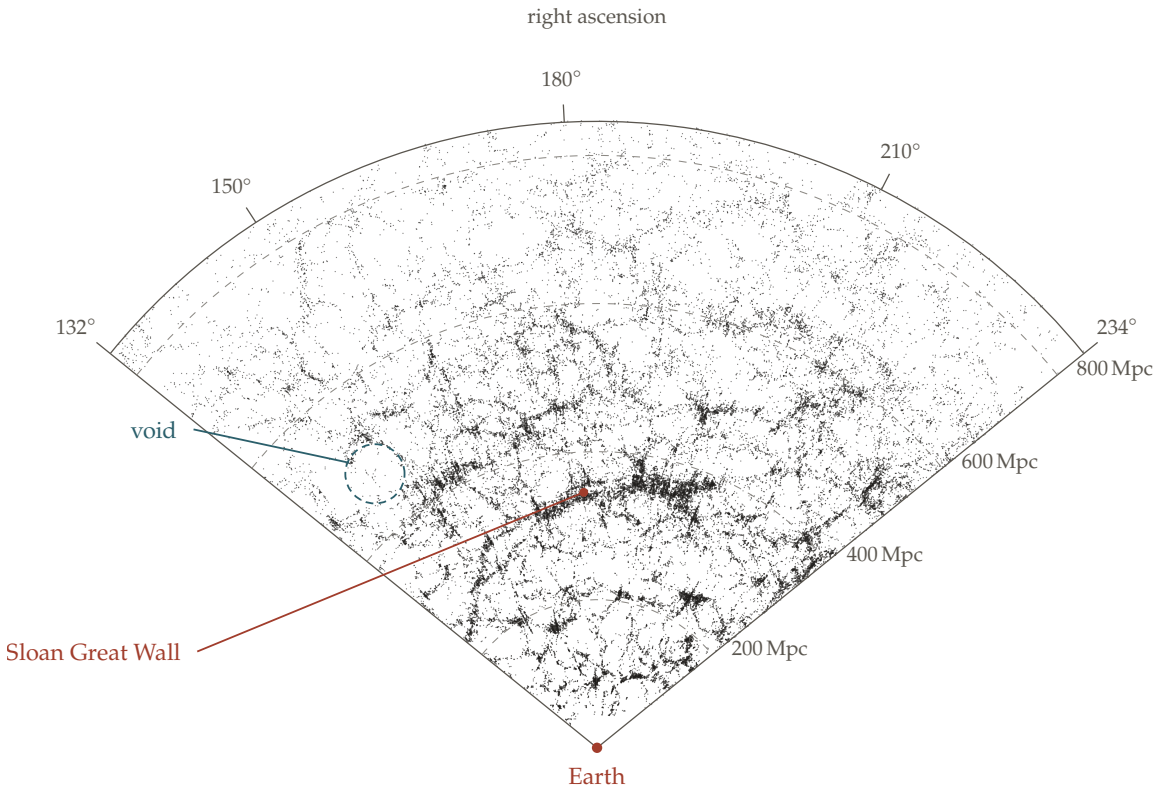


Figure 14.16: 39 263 galaxies of the Sloan Digital Sky Survey (data release 17) in a slice 5° thick in declination, placed by right ascension and by distance from their redshifts. Earth is at the apex and the distance arcs are 200 Mpc apart. A band of galaxies near $z = 0.08$ is marked as the Sloan Great Wall, and a dashed circle marks one void.

Essential Ideas:

- The circular speed weighs the mass inside, $v_{\text{circ}}^2 = Gm(r)/r$, Eq. (14.8).
- A flat curve has $\rho = v_{\text{circ}}^2/(4\pi G r^2)$, Eq. (14.10).

Detailed Solutions:

(a) **Period**. $2\pi R_0/v_0 = 2\pi \times 8.2 \times 3.086 \times 10^{19}\,\text{m}/(2.3 \times 10^5\,\text{m s}^{-1})$.

$$\boxed{2.2 \times 10^8 \text{ years.}}$$

(b) **Number of orbits**. $4.6 \times 10^9/2.19 \times 10^8$.

$$\boxed{\text{about 21.}}$$

(c) **Mass inside**. $m = v_0^2 R_0/G = (2.3 \times 10^5)^2 \times 2.53 \times 10^{20}/6.674 \times 10^{-11}\,\text{kg}$.

$$\boxed{m(R_0) = 1.0 \times 10^{11}\, M_\odot.}$$

(d) **Local density**. $\rho = v_0^2/(4\pi G R_0^2) = 9.9 \times 10^{-22}\,\text{kg m}^{-3}$. Multiplying by c^2 and dividing by $1.602 \times 10^{-10}\,\text{J GeV}^{-1}$ and $10^6\,\text{cm}^3\,\text{m}^{-3}$,

$$\boxed{\rho = 0.015\, M_\odot\,\text{pc}^{-3} = 0.55\,\text{GeV}\,c^{-2}\,\text{cm}^{-3}.}$$

This overestimates the halo, since the disk also supplies part of v_0. Careful models give about two thirds of it.

(e) **In the slab**. $0.015\,M_\odot\,\mathrm{pc}^{-3} \times 2200\,\mathrm{pc}$.

$$\boxed{\approx 30\,M_\odot\,\mathrm{pc}^{-2},}$$

or about 20 with the smaller halo density of (d). This matches the excess found by weighing the disk. The unseen mass near the Sun is what a round halo puts there, with no extra dark matter in the disk itself.

14.8.2 Problem 2: Weighing the Disk

Problem Statement: (a) A population of old stars of the thick disk has a vertical velocity spread $\sigma_z = 40\,\mathrm{km\,s}^{-1}$, the same at every height, and its number falls with height with the scale $h_z = 900\,\mathrm{pc}$. Find the mass per unit area of the disk. (b) Near the plane, take the disk to be a uniform slab of density $\rho_0 = 0.1\,M_\odot\,\mathrm{pc}^{-3}$. Show that a star inside it oscillates up and down harmonically, and find the period. How often does the Sun cross the plane? (c) The derivation of Eq. (14.6) used a pressure, yet stars do not collide. Why is that allowed?

Essential Ideas:

- $h_z = \sigma_z^2/(2\pi G\Sigma)$, Eq. (14.6).
- Gauss's law (9.14) for a box inside the slab.

Detailed Solutions:

(a) **Mass per area**. $\Sigma = \sigma_z^2/(2\pi G h_z) = (4\times10^4)^2/(2\pi\times6.674\times10^{-11}\times2.78\times10^{19})\,\mathrm{kg\,m}^{-2}$.

$$\boxed{\Sigma \simeq 66\,M_\odot\,\mathrm{pc}^{-2}.}$$

(b) **Vertical oscillation**. A box with faces at $\pm z$ inside the slab holds the mass $2z\rho_0 A$, so Gauss's law gives $2g(z)A = 4\pi G\rho_0\,2zA$, and $g = 4\pi G\rho_0 z$ toward the plane. Then $\ddot{z} = -4\pi G\rho_0\,z$, a harmonic oscillation with angular frequency $\sqrt{4\pi G\rho_0}$. With $\rho_0 = 6.77\times10^{-21}\,\mathrm{kg\,m}^{-3}$,

$$\boxed{\text{period} = \frac{2\pi}{\sqrt{4\pi G\rho_0}} \simeq 84 \text{ million years.}}$$

The Sun crosses the plane twice per period, about every 42 million years. The period does not depend on how far the star swings, as long as it stays inside the slab.

(c) **No collisions needed**. The pressure in the momentum equation is the momentum carried across a surface by particles moving through it, Eq. (9.9). Counting that flow needs only the particles' velocities, not collisions. What collisions would add is the same spread of speeds in every direction. The vertical balance uses only $\Pi_{zz} = \rho\sigma_z^2$, so it holds for stars whatever their spread in other directions.

14.8.3 Problem 3: Relaxation and Evaporation of a Globular Cluster

Problem Statement: A globular cluster holds $N = 2\times10^5$ stars with a total mass of $10^5\,M_\odot$ within $R = 5\,\mathrm{pc}$. Estimate (a) the typical speed of its stars, (b) the crossing time, (c) the relaxation time, and (d) the evaporation time. (e) The cluster is 1.2×10^{10} years old. Is it relaxed? Should it still exist?

Essential Ideas:

- $v^2 \approx GM/R$ from the virial theorem in order of magnitude.
- $t_{\mathrm{relax}} = N/(8\ln N)\,t_{\mathrm{cross}}$, Eq. (14.25).

- A relaxed cluster loses 0.74 per cent of its stars per relaxation time, Eq. (14.26) and the Maxwell tail.

Detailed Solutions:

(a) **Speed**. $v = \sqrt{GM/R} = \sqrt{6.674\times10^{-11}\times1.989\times10^{35}/1.543\times10^{17}}\,\mathrm{m\,s^{-1}}$.

$$\boxed{v \simeq 9.3\,\mathrm{km\,s^{-1}}.}$$

(b) **Crossing time**. $R/v = 1.543\times10^{17}/9.28\times10^{3}\,\mathrm{s}$.

$$\boxed{t_{\mathrm{cross}} \simeq 5.3\times10^{5}\ \mathrm{years}.}$$

(c) **Relaxation time**. $N/(8\ln N) = 2\times10^{5}/(8\times12.2) = 2050$.

$$\boxed{t_{\mathrm{relax}} \simeq 1.1\times10^{9}\ \mathrm{years}.}$$

(d) **Evaporation time**. $t_{\mathrm{relax}}/0.0074$.

$$\boxed{t_{\mathrm{evap}} \simeq 1.5\times10^{11}\ \mathrm{years}.}$$

(e) **Age**. The cluster is about 11 relaxation times old, so its stars have shared their energy and its heavy stars have sunk toward the centre. Evaporation alone would take more than ten times its age, so it survives. The Galaxy's tides and the shocks of crossing the disk remove stars faster, and many clusters that were once smaller or looser have already dissolved into the halo.

14.8.4 Problem 4: One Cluster, Three Weighings

Problem Statement: Model a rich cluster as an isothermal sphere, $\rho = \sigma^2/(2\pi G r^2)$, with line-of-sight spread $\sigma = 1000\,\mathrm{km\,s^{-1}}$ for its galaxies and hot gas of $\mu = 0.6$ at $k_BT = \mu m_{\mathrm{H}}\sigma^2$. (a) Find the mass inside 1 Mpc from the X-ray gas. (b) Show that it equals $v_{\mathrm{circ}}^2 r/G$ for this sphere. (c) Find the mass inside a circle of radius b on the sky. (d) Find the Einstein radius of the cluster for a source with $D_{LS}/D_S = \frac{1}{2}$.

Essential Ideas:

- The X-ray mass, Eq. (14.32).
- The Einstein radius, Eq. (14.34), with the mass inside the ring.

Detailed Solutions:

(a) **X-ray mass**. The gas is at one temperature and $\mathrm{d}\ln\rho/\mathrm{d}\ln r = -2$, so

$$m = \frac{2k_BTr}{G\mu m_{\mathrm{H}}} = \frac{2\sigma^2 r}{G} = \frac{2\times10^{12}\times3.086\times10^{22}}{6.674\times10^{-11}}\,\mathrm{kg}.$$

$$\boxed{m(1\,\mathrm{Mpc}) \simeq 4.6\times10^{14}\,M_\odot.}$$

(b) **The same from rotation**. By Eq. (9.15), $m = \int_0^r 4\pi r'^2\sigma^2/(2\pi G r'^2)\,dr' = 2\sigma^2 r/G$, so $v_{\mathrm{circ}}^2 = Gm/r = 2\sigma^2$, the result of Section 12.5. The gas and the galaxies feel the same well.

(c) **Mass on the sky**. The mass per unit area at projected radius $r_\perp$ is $\int_{-\infty}^{\infty} \rho\, dz$ along the line of sight, with $r^2 = r_\perp^2 + z^2$,

$$\int_{-\infty}^{\infty} \frac{\sigma^2\, dz}{2\pi G(r_\perp^2 + z^2)} = \frac{\sigma^2}{2\pi G}\,\frac{\pi}{r_\perp} = \frac{\sigma^2}{2Gr_\perp}.$$

Adding rings of area $2\pi r_\perp\, dr_\perp$ out to b,

$$\boxed{M(<b) = \int_0^b 2\pi r_\perp \frac{\sigma^2}{2Gr_\perp}\, dr_\perp = \frac{\pi\sigma^2 b}{G}.}$$

(d) **Einstein radius**. In Eq. (14.34) put $M = \pi\sigma^2 b/G$ with $b = \theta_E D_L$,

$$\theta_E^2 = \frac{4\pi\sigma^2\theta_E D_L}{c^2}\,\frac{D_{LS}}{D_L D_S} \quad\Longrightarrow\quad \boxed{\theta_E = 4\pi\left(\frac{\sigma}{c}\right)^2 \frac{D_{LS}}{D_S} \simeq 14''.}$$

The ring measures σ directly, without any speed being measured. Three routes, speeds of galaxies, temperature of gas and bending of light, each weigh the same well. When they agree, as they do for most clusters to within a few tens of per cent, the mass is secure.

14.8.5 Problem 5: A Tully–Fisher Distance

Problem Statement: (A constructed example.) A calibrating spiral at 16 Mpc rotates at 200 km s^{-1} and has apparent magnitude 10.0 in the near infrared. A second spiral has a 21 cm line of full width $W = 400$ km s^{-1}, appears as an ellipse with axis ratio 0.5, and has apparent magnitude 12.8 in the same band. (a) Find its inclination and rotation speed. (b) How much brighter is it than the calibrator, in magnitudes? (c) How far away is it?

Essential Ideas:

- $W = 2v\sin\iota$ and $\cos\iota$ equals the axis ratio.
- $L \propto v^4$, Eq. (14.44), so absolute magnitudes differ by $-10\log_{10}(v_1/v_2)$.
- Apparent magnitudes differ by the difference in absolute magnitude plus $5\log_{10}(d_1/d_2)$, Eq. (4.11).

Detailed Solutions:

(a) **Inclination and speed**. $\cos\iota = 0.5$, so $\iota = 60°$, and $v = W/(2\sin\iota) = 400/\sqrt{3}$.

$$\boxed{\iota = 60°, \qquad v \simeq 231\ \text{km s}^{-1}.}$$

(b) **Luminosity**. $-10\log_{10}(231/200)$.

$$\boxed{0.62\ \text{magnitudes brighter.}}$$

(c) **Distance**. The second galaxy appears 2.8 magnitudes fainter although it is 0.62 magnitudes brighter, so $5\log_{10}(d/16\,\text{Mpc}) = 2.8 + 0.62$.

$$\boxed{d = 16\,\text{Mpc} \times 10^{0.685} \simeq 77\,\text{Mpc}.}$$

An error of 10 per cent in W changes v^4 by 46 per cent and the distance by 21 per cent. Accurate inclinations and widths matter as much as the photometry.

14.8.6 Problem 6: A Microlensing Event

Problem Statement: A star of the bulge is lensed by a dwarf star of $0.3\,M_\odot$, with Einstein time $t_E = 26$ days, and the source passes the lens at $u_0 = 0.1$ Einstein radii. (a) Find the peak brightening, as a factor and in magnitudes, and compare it with $1/u_0$. (b) For how long is the source brighter than $A = 1.34$? (c) Find the full width of the light curve at half the peak brightening. (d) The lens has a planet of 10 Earth masses. About how long does its spike last, and how often must the survey look at the star to catch it?

Essential Ideas:

- $A(u) = (u^2+2)/(u\sqrt{u^2+4})$, Eq. (14.41), with $u(t)$ from Eq. (14.42).
- A planet's spike lasts about $\sqrt{m_p/M}\,t_E$.

Detailed Solutions:

(a) **Peak**. At $t = t_0$, $u = u_0 = 0.1$, so $A = 2.01/(0.1\sqrt{4.01}) = 10.04$, and $2.5\log_{10} 10.04 = 2.50$ magnitudes. The small-u form $1/u_0 = 10$ is within half a per cent.

$$\boxed{A_{\max} = 10.04, \quad \text{brighter by } 2.50 \text{ magnitudes.}}$$

(b) **Time above** 1.34. $A(1) = 3/\sqrt{5} = 1.34$ and A falls as u grows, so $A > 1.34$ while $u < 1$, that is while $u_0^2 + (t-t_0)^2/t_E^2 < 1$. This lasts $2t_E\sqrt{1-u_0^2} = 2 \times 26 \times 0.995$ days.

$$\boxed{52 \text{ days.}}$$

(c) **Width at half the peak**. Invert Eq. (14.41). With $w = u^2$, squaring gives $A^2 w(w+4) = (w+2)^2$, which is $(A^2-1)w^2 + 4(A^2-1)w - 4 = 0$. Its positive root is

$$u^2 = w = -2 + \frac{2A}{\sqrt{A^2-1}}.$$

For $A = 5.02$, $A/\sqrt{A^2-1} = 5.02/4.92 = 1.0205$, so $u^2 = 0.0409$. Then $(t-t_0)/t_E = \pm\sqrt{0.0409 - 0.01} = \pm 0.176$, and the full width is $0.352\,t_E$.

$$\boxed{9.1 \text{ days.}}$$

A close passage gives a sharp peak. Most of the brightening happens in a few days around t_0.

(d) **Planet spike**. $m_p/M = 10 \times 3.0 \times 10^{-6}/0.3 = 1.0 \times 10^{-4}$, so the spike lasts about $\sqrt{1.0 \times 10^{-4}} \times 26$ days $= 0.26$ days.

$$\boxed{\approx 6 \text{ hours.}}$$

To sample the spike several times, the star must be observed about once an hour. Surveys watch the bulge every night at a slower rate, and once an event is found, telescopes around the Earth follow it night and day so that a spike at any hour is not missed.

Chapter 15

Radiation from Hot and Fast Particles

Hot, opaque stars shine nearly as blackbodies (Chapter 11). Many other sources do not. Thin gas between the galaxies of a cluster glows in X-rays. Radio galaxies shine across many decades of frequency with no single temperature that could explain their spectra. A quasar can outshine its host galaxy from a region only a few light-days across.

Charged particles make this light when ions, magnetic fields, or other particles accelerate them. We start with the power radiated by an accelerated charge and use the six principles to build the spectra. For particles moving near the speed of light, we also need the Lorentz transformation in matrix form.

15.1 An Accelerated Charge Radiates

In Chapter 8, a passing wave shook a slow electron. Its radiation gave the Thomson cross-section of Eq. (8.7). Ions and magnetic fields can accelerate electrons much more strongly, and electrons in jets can move close to c. Their radiation depends on both the acceleration and the motion of the source toward the observer.

Shake the end of a long rope and a pulse runs away from your hand. A sharper jerk sends a stronger pulse. The charge's electric field behaves in the same way when its motion changes. An electron passing a proton gets one sideways kick. In a magnetic field it turns continuously. In each case, its changing field carries energy away (P1). The radiated power rises with the square of the acceleration. We will compare that loss with the energy of the motion before treating radiation as a small correction (P6).

Consider a classical electron circling a proton at the Bohr radius, 5.29×10^{-11} m. Its acceleration is $9.0 \times 10^{22}\,\mathrm{m\,s^{-2}}$, so it would radiate 4.7×10^{-8} W. At that rate, its 13.6 eV of energy would last 4.7×10^{-11} s. One orbit takes only 1.5×10^{-16} s, so the loss per orbit is 1.6×10^{-6} of its energy. A real hydrogen atom is stable because of quantum mechanics. The classical comparison shows why radiation can often be calculated after finding the particle's motion. The escaping field is strongest at right angles to the acceleration and vanishes along it. The same rule describes radio antennas, X-ray tubes, laboratory synchrotrons, and the scattering in Chapter 8.

The Larmor formula. A charge e moving slowly, $v \ll c$, with acceleration $\dot{\vec{v}}$ radiates the power

$$\mathcal{P} = \frac{e^2\,|\dot{\vec{v}}|^2}{6\pi\varepsilon_0 c^3}, \qquad \frac{d\mathcal{P}}{d\Omega} = \frac{e^2\,|\dot{\vec{v}}|^2}{16\pi^2\varepsilon_0 c^3}\,\sin^2\Theta, \tag{15.1}$$

Here Θ is the angle between the acceleration and the direction of emission. These results come from electromagnetism. Integrating the angular pattern gives the total power, since $\int \sin^2\Theta\, d\Omega = 8\pi/3$. Equal power goes in opposite directions, so the radiation carries energy but no net momentum in this frame.

How fast a charge loses its energy. Take an electron of mass m_e oscillating as $x = x_0 \cos \omega t$. We compare the energy it radiates in one radian of motion with the energy it has (P6). Its acceleration is $\dot{v} = -\omega^2 x$, its energy is $\frac{1}{2} m_e \omega^2 x_0^2$, and a time $1/\omega$ lets it radiate $\langle \mathcal{P} \rangle / \omega$.

$$\frac{\text{energy radiated per radian}}{\text{energy of the motion}} \overset{(1)}{=} \frac{e^2\, \omega^4 x_0^2 / (12\pi\varepsilon_0 c^3)}{\omega \cdot \frac{1}{2} m_e \omega^2 x_0^2} \overset{(2)}{=} \omega \, \frac{e^2}{6\pi\varepsilon_0 m_e c^3} \overset{(3)}{=} \omega\, t_e. \tag{15.2}$$

(1) Eq. (15.1) with $|\dot{v}|^2 = \omega^4 x_0^2 \cos^2 \omega t$, whose average over a cycle is $\omega^4 x_0^2/2$.
(2) The factors x_0^2 and ω^3 cancel.
(3) Define the **radiation time** t_e below.

$$\boxed{t_e = \frac{e^2}{6\pi\varepsilon_0 m_e c^3} = \frac{2}{3}\frac{r_e}{c} = 6.27 \times 10^{-24}\,\text{s}.} \tag{15.3}$$

Here r_e is the classical electron radius of Eq. (8.7). Light takes only a little longer than t_e to cross it. The electron loses a fraction ωt_e of its energy per radian, and $\omega t_e \ll 1$ below about 10^{23} turns per second. We can first find the motion as though the charge did not radiate, then use Larmor's formula to calculate its light.

Fast charges and the boost matrix. Let the charge move along x at speed βc in the observer's frame, which we call the lab. The charge's own frame is primed. To turn a position, time, energy, or momentum from the charge's frame into lab quantities, use the boost

$$\begin{pmatrix} ct \\ x \\ y \end{pmatrix} = \Lambda(\beta) \begin{pmatrix} ct' \\ x' \\ y' \end{pmatrix}, \qquad \Lambda(\beta) = \begin{pmatrix} \gamma_L & \gamma_L\beta & 0 \\ \gamma_L\beta & \gamma_L & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \gamma_L = \frac{1}{\sqrt{1-\beta^2}}, \tag{15.4}$$

The z component is unchanged and omitted. The same matrix transforms four-momentum, with components E/c, momentum along x, and momentum along y. To go from the lab to the charge's frame, use its inverse $\Lambda(-\beta)$. The subscript on the **Lorentz factor** γ_L distinguishes it from the adiabatic exponent γ.

The power is the same in both frames. At one instant, the charge is at rest in its own frame, where Larmor's formula applies. During dt' it radiates energy dE' but no net momentum, by the symmetry of Eq. (15.1). The radiation is emitted at the charge, $x' = y' = 0$. Boost the radiated energy and the time interval to the lab,

$$\Lambda(\beta) \begin{pmatrix} dE'/c \\ 0 \\ 0 \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} \gamma_L\, dE'/c \\ \gamma_L\beta\, dE'/c \\ 0 \end{pmatrix}, \qquad \Lambda(\beta) \begin{pmatrix} c\, dt' \\ 0 \\ 0 \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} \gamma_L\, c\, dt' \\ \gamma_L\beta\, c\, dt' \\ 0 \end{pmatrix},$$
$$\Longrightarrow \quad dE = \gamma_L\, dE', \qquad dt = \gamma_L\, dt'. \tag{15.5}$$

(1) The matrix of Eq. (15.4) times a vector with only its first entry non-zero is that entry times the first column.

The factors γ_L cancel in the ratio,

$$\boxed{\mathcal{P} = \frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\mathrm{d}E'}{\mathrm{d}t'} = \frac{e^2\, |\dot{\vec{v}}'|^2}{6\pi\varepsilon_0 c^3},} \tag{15.6}$$

Here $\dot{\vec{v}}'$ is the acceleration in the charge's frame. We can find the power of a fast charge by calculating that acceleration and applying Larmor's formula there. An observer may receive the light in a shorter burst because the source moves toward them. That arrival-time effect sets the frequencies of synchrotron light in Section 15.3.

Beaming. A photon of frequency ν' emitted in the charge's frame at angle θ' to the direction of motion has four-momentum $(h\nu'/c)(1, \cos\theta', \sin\theta')$. In the lab it has frequency ν and angle θ,

$$\frac{h\nu}{c}\begin{pmatrix}1\\ \cos\theta\\ \sin\theta\end{pmatrix} \overset{(1)}{=} \Lambda(\beta)\,\frac{h\nu'}{c}\begin{pmatrix}1\\ \cos\theta'\\ \sin\theta'\end{pmatrix} \overset{(2)}{=} \frac{h\nu'}{c}\begin{pmatrix}\gamma_L(1+\beta\cos\theta')\\ \gamma_L(\cos\theta'+\beta)\\ \sin\theta'\end{pmatrix}$$

$$\Longrightarrow \qquad \tan\theta \overset{(3)}{=} \frac{\sin\theta'}{\gamma_L(\cos\theta'+\beta)}, \qquad \nu \overset{(4)}{=} \gamma_L(1+\beta\cos\theta')\,\nu'. \tag{15.7}$$

(1) The boost (15.4) applied to the photon's four-momentum.
(2) The matrix product, row by row.
(3) The third row divided by the second.
(4) The first row.

A photon emitted sideways in the charge's frame, $\theta' = \pi/2$, appears at $\tan\theta = 1/(\gamma_L\beta)$ in the lab. The entire forward hemisphere of emission fits inside this angle,

$$\boxed{\theta \lesssim \frac{1}{\gamma_L} \qquad (\gamma_L \gg 1).}$$

For $\gamma_L = 10$, that cone has a half-angle of about 6° (Figure 15.1). The forward photons are also shifted to higher frequencies. Their energy rises by $\gamma_L(1+\beta\cos\theta')$, compared with the all-direction average of γ_L. The forward half therefore carries the fraction $\frac{1}{2} + \frac{\beta}{4}$ of the radiated energy, 0.74 at $\gamma_L = 3$ and nearly $\frac{3}{4}$ at high speed. At $\theta' = 0$, Eq. (15.7) gives $\gamma_L(1+\beta)\nu' = \sqrt{(1+\beta)/(1-\beta)}\,\nu'$, the Doppler shift of an approaching source (Eq. (3.27)).

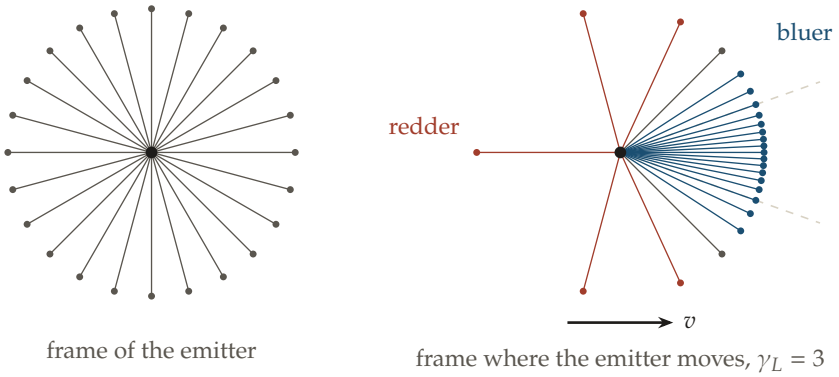


Figure 15.1: Photons emitted at equal steps of 15° in the frame of the emitter (left), and the same photons in a frame where the emitter moves to the right with $\gamma_L = 3$ (right). Each line ends at a dot the same distance from the emitter, since all the photons move at c. On the right, blue photons have a higher frequency than in the emitter's frame and red ones a lower frequency. The dashed lines mark the angle $\tan\theta = 1/(\gamma_L\beta)$.

15.2 Bremsstrahlung

The gas between the galaxies of a cluster is hot, about 7×10^7 K, because it lies deep in the cluster's gravitational well (Problem 10.8.2). It is fully ionised, so it has no bound electrons to make spectral lines. It is also thin enough for light to escape without reaching blackbody equilibrium (Eq. (11.10)). What makes its X-rays?

Think of a car rounding a bend. Even if its speed stays steady, it accelerates sideways because its direction changes. An electron passing an ion also turns. The ion's electric field pulls it inward for a short time, and the accelerating electron sends out a pulse of light (Section 15.1). A closer pass gives a stronger, shorter pulse. The combined light of many such encounters is **bremsstrahlung**, or braking radiation. We find the spectrum of one pulse, count passes at different distances (P6), then average over electron speeds (P3).

For the cluster gas, $k_B T = 6.0\,\text{keV}$, corresponding to 1.5×10^{18} Hz. Its spectrum begins to fall steeply near this X-ray energy. Electrons in the fast tail of the Maxwell distribution can still make higher-energy photons, but each photon must draw its energy from the electron that made it (P1). The same process makes X-rays in an X-ray tube and solar flares, and radio light in ionised gas around hot stars. In reverse, a free electron can absorb a photon while passing an ion. That process contributes to the opacity of stellar interiors (P4).

The ion is much heavier than the passing electron, so we treat it as fixed. The ion's electric field pulls the electron toward it and bends its path (Figure 15.2). The pair's electric dipole moment changes as the electron moves. With the ion fixed, $\boldsymbol{p} = -e\boldsymbol{r}_e + \text{constant}$, so $\ddot{\boldsymbol{p}} = -e\boldsymbol{a}_e$. The acceleration produces the dipole radiation of Eq. (15.1), strongest across the direction of the kick and absent along it.

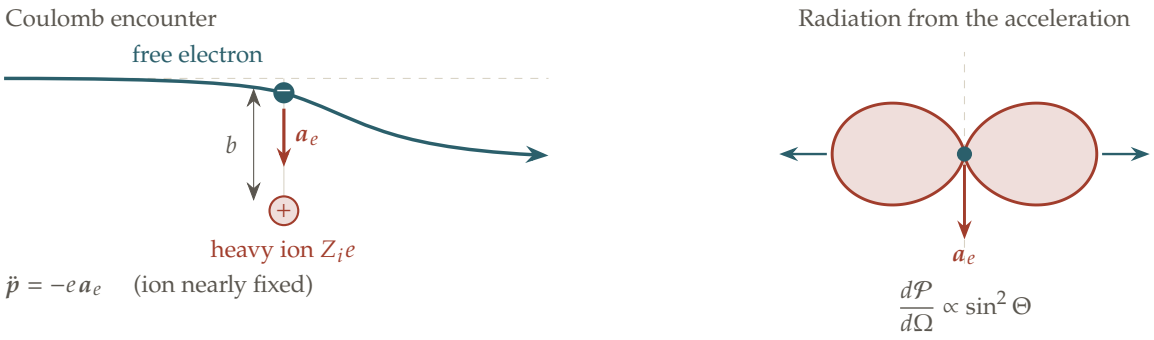


Figure 15.2: A free electron passes a heavy ion and curves toward it (left). The lobes show the angular pattern of its radiation at low speed (right).

A gas of free charges. We can count separate electron–ion encounters only if an electron feels one ion more strongly than the surrounding charges during a close pass. Put an ion of charge $Z_i e$ at the origin, in gas of electron density n_e and temperature T. Its electric potential ϕ draws in electrons. Their density rises by the Boltzmann factor $e^{e\phi/k_B T}$ (P3). Away from the ion, Poisson's equation gives

$$\nabla^2\phi \overset{(1)}{=} -\frac{e}{\varepsilon_0}\left(Z_i n_i - n_e\, e^{e\phi/k_B T}\right)$$
$$\overset{(2)}{=} \frac{n_e e^2}{\varepsilon_0 k_B T}\,\phi \overset{(3)}{=} \frac{\phi}{\lambda_D^2}$$
$$\Longrightarrow \qquad \phi \overset{(4)}{=} \frac{Z_i e}{4\pi\varepsilon_0 r}\, e^{-r/\lambda_D}. \tag{15.8}$$

(1) The charge density is that of the ions, $Z_i e n_i$, minus that of the electrons. The ions are heavy and stay spread evenly on the scale considered.
(2) Far from the ion $e\phi \ll k_B T$, so $e^{e\phi/k_B T} \approx 1 + e\phi/k_B T$. The gas is neutral on average, $Z_i n_i = n_e$.
(3) Define the **Debye length** λ_D below.
(4) The spherically symmetric solution that matches the bare ion's potential at small r. Substituting it back gives $\nabla^2\phi = \phi/\lambda_D^2$.

$$\boxed{\lambda_D = \sqrt{\frac{\varepsilon_0 k_B T}{n_e e^2}} = \frac{\sqrt{k_B T/m_e}}{2\pi\nu_p}.} \tag{15.9}$$

The second form uses the plasma frequency of Eq. (5.2). The **Debye length** is roughly the distance a thermal electron travels during one plasma oscillation. Beyond this distance, the cloud of electrons around an ion screens its field. In cluster gas, $n_e = 10^3\,\mathrm{m}^{-3}$ and $T = 7\times10^7\,\mathrm{K}$ give $\lambda_D = 18\,\mathrm{km}$, with 2.5×10^{16} electrons inside a sphere of that radius. Gas around a hot star, $n_e = 10^8\,\mathrm{m}^{-3}$ and $T = 10^4\,\mathrm{K}$, gives $\lambda_D = 0.7\,\mathrm{m}$ and 1.4×10^8 electrons (P6). Each screening cloud thus contains many electrons. Its average effect is smooth, while a passing electron feels the nearest ion strongly when it gets much closer than λ_D. We can count such encounters one at a time. The gas is a **plasma**, and it blocks radiation below the plasma frequency ν_p (Chapter 5).

One encounter. An electron approaches an ion of charge $Z_i e$ at speed v and impact parameter b. If the deflection is small, its nearly straight path is enough to calculate the force. Set $t = 0$ at closest approach. The sideways force is $Z_i e^2 b/[4\pi\varepsilon_0(b^2 + v^2t^2)^{3/2}]$. Integrating it over the passage gives the sideways change in speed,

$$\Delta v \overset{(1)}{=} \frac{Z_i e^2 b}{4\pi\varepsilon_0 m_e}\int_{-\infty}^{\infty}\frac{dt}{(b^2+v^2t^2)^{3/2}} \overset{(2)}{=} \frac{2Z_i e^2}{4\pi\varepsilon_0 m_e\, bv}. \tag{15.10}$$

(1) Newton's law across the path, integrated over the passage.
(2) $\int_{-\infty}^{\infty} dt/(b^2+v^2t^2)^{3/2} = 2/(b^2 v)$.

The kick tells us the total change in velocity, but the spectrum also depends on how long the kick lasts. We describe the acceleration at each angular frequency ω by its Fourier transform $\tilde{v}(\omega) = \int \dot{v}\, e^{i\omega t} dt$. For a real signal, Parseval's theorem gives $\int \dot{v}^2 dt = \frac{1}{\pi}\int_0^\infty |\tilde{v}|^2 d\omega$. Applying it to Larmor's power, the pulse energy W becomes

$$W \overset{(1)}{=} \frac{e^2}{6\pi\varepsilon_0 c^3}\int_{-\infty}^{\infty}\dot{v}^2\,dt \overset{(2)}{=} \int_0^\infty \frac{e^2\,|\tilde{v}(\omega)|^2}{6\pi^2\varepsilon_0 c^3}\,d\omega$$

$$\Longrightarrow \qquad \frac{\mathrm{d}W}{\mathrm{d}\omega} \overset{(3)}{=} \frac{e^2\,(\Delta v)^2}{6\pi^2\varepsilon_0 c^3}\Big[uK_1(u)\Big]^2, \qquad u = \frac{\omega b}{v}. \tag{15.11}$$

(1) Larmor's power (15.1) integrated over the passage.
(2) Parseval's theorem.
(3) The transform of the force law above is $\tilde{v} = \Delta v\, uK_1(u)$, with K_1 a modified Bessel function. The component of the acceleration along the path changes sign during the passage and adds much less at low frequency, so it is dropped.

The factor $uK_1(u)$ is close to 1 when $u \ll 1$ and falls rapidly, as e^{-u}, when $u \gg 1$. We can therefore keep the low-frequency part and cut the spectrum off near $\omega = v/b$,

$$\boxed{\frac{\mathrm{d}W}{\mathrm{d}\omega} \approx \frac{e^2}{6\pi^2\varepsilon_0 c^3}\left(\frac{2Z_i e^2}{4\pi\varepsilon_0 m_e\, bv}\right)^2 \quad \text{for } \omega \lesssim \frac{v}{b}, \qquad \approx 0 \text{ above.}} \tag{15.12}$$

The electron spends about b/v near the ion. A shorter encounter makes a shorter pulse and reaches higher frequencies (P6). Below the cutoff, each encounter contributes nearly the same energy per unit frequency.

Counting encounters. A single pulse is faint. To get the light from a gas, count how often electrons of speed v pass ions at each impact parameter. In a unit volume, encounters between b and $b + db$ occur at the rate $n_e n_i v\, 2\pi b\, db$. At frequency ω, only encounters with $b < v/\omega$ contribute appreciably, by Eq. (15.12).

$$\begin{aligned}\frac{dW}{dt\, dV\, d\omega} &\overset{(1)}{=} n_e n_i v \int_{b_{\min}}^{b_{\max}} 2\pi b\, \frac{\mathrm{d}W}{\mathrm{d}\omega}\, db \\ &\overset{(2)}{=} n_e n_i v\, 2\pi\, \frac{e^2}{6\pi^2 \varepsilon_0 c^3} \left(\frac{2Z_i e^2}{4\pi\varepsilon_0 m_e v}\right)^2 \int_{b_{\min}}^{b_{\max}} \frac{db}{b} \\ &\overset{(3)}{=} \frac{16}{3}\, \frac{Z_i^2 e^6}{(4\pi\varepsilon_0)^3 c^3 m_e^2}\, \frac{n_e n_i}{v}\, \ln\frac{b_{\max}}{b_{\min}}. \end{aligned} \tag{15.13}$$

(1) Rate of encounters times the energy each gives per unit frequency.
(2) Eq. (15.12). Its b^{-2} against the b of the ring leaves db/b.
(3) Collect constants. $2\pi \times 4/(6\pi^2) = 4/(3\pi)$, and $e^2/\varepsilon_0 = 4\pi e^2/(4\pi\varepsilon_0)$.

The upper limit is $b_{\max} = v/\omega$. Beyond it, the encounter is too slow to radiate much at frequency ω. The lower limit is the larger of two distances. At $Z_i e^2/(4\pi\varepsilon_0 m_e v^2)$ the deflection ceases to be small, and below the quantum wavelength $\hbar/(m_e v)$ the classical path ceases to be a good description. Because the result depends on their ratio only through a logarithm, the exact limits have a modest effect. A full quantum calculation replaces $\ln(b_{\max}/b_{\min})$ by $(\pi/\sqrt{3})\, \bar{g}$. The **Gaunt factor** $\bar{g}$ is of order one. The $1/v$ dependence remains: a slower electron spends longer near the ion and radiates more.

Averaging over electron speeds. Electrons in gas at temperature T have the Maxwell distribution of Eq. (9.28). Their speed intervals carry the weight $4\pi v^2 e^{-m_e v^2/2k_B T} dv$ (P3). A photon of energy $h\nu$ requires an electron with at least that much kinetic energy, so $\frac{1}{2} m_e v^2 \geq h\nu$ (P1). We therefore average the $1/v$ in Eq. (15.13) over electrons above this threshold,

$$\left\langle \frac{1}{v} \right\rangle \overset{(1)}{=} \frac{\int_{v_{\min}}^{\infty} v\, e^{-m_e v^2/2k_B T}\, dv}{\int_0^{\infty} v^2\, e^{-m_e v^2/2k_B T}\, dv} \overset{(2)}{=} \frac{(k_B T/m_e)\, e^{-h\nu/k_B T}}{\sqrt{\pi/2}\, (k_B T/m_e)^{3/2}} \overset{(3)}{=} \sqrt{\frac{2m_e}{\pi k_B T}}\; e^{-h\nu/k_B T}. \tag{15.14}$$

(1) The weight v^2 of the Maxwell distribution times $1/v$, normalised. The lower limit is $\frac{1}{2} m_e v_{\min}^2 = h\nu$.
(2) The two Gaussian integrals, with $m_e v_{\min}^2/2k_B T = h\nu/k_B T$.
(3) Simplify.

Substitute this average into Eq. (15.13) and change from angular frequency to frequency with $d\omega = 2\pi\, d\nu$. We write the power per unit volume and frequency as $4\pi j_\nu$, where j_ν is the emissivity per steradian of Chapter 11.

$$\boxed{4\pi j_\nu = \frac{32\pi^2}{3\sqrt{3}}\, \frac{Z_i^2 e^6}{(4\pi\varepsilon_0)^3 c^3 m_e^2} \sqrt{\frac{2m_e}{\pi k_B T}}\; \bar{g}\, n_e n_i\, e^{-h\nu/k_B T} = 6.8 \times 10^{-51}\, Z_i^2\, \bar{g}\, \frac{n_e n_i}{\sqrt{T}}\, e^{-h\nu/k_B T},} \tag{15.15}$$

The units are $\mathrm{W\,m^{-3}\,Hz^{-1}}$ when n_e, n_i are in $\mathrm{m^{-3}}$ and T is in kelvin. The spectrum is nearly flat below $h\nu \approx k_B T$, apart from the slow change of $\bar{g}$. Above that energy, it falls exponentially

because fast electrons are rare (Figure 15.3). The bend thus measures the gas temperature. At low frequencies, hotter gas is fainter at fixed density, as $T^{-1/2}$, because faster electrons spend less time near each ion.

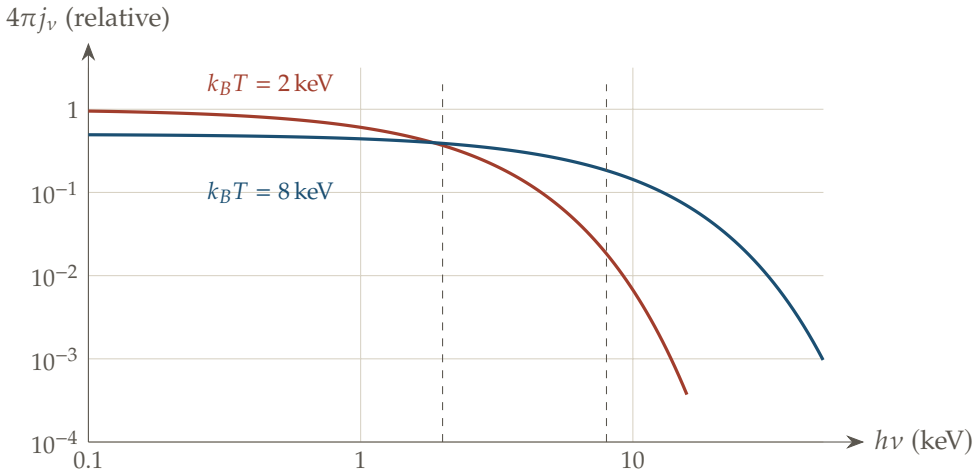


Figure 15.3: Bremsstrahlung emissivity $4\pi j_\nu \propto T^{-1/2} e^{-h\nu/k_BT}$ against photon energy, for the same electron and ion densities at $k_BT = 2$ keV and 8 keV, with $\bar{g}$ held fixed. Dashed lines mark $h\nu = k_BT$.

Cooling. The spectrum tells us how much energy the gas loses. Integrating over all frequencies, with $\int_0^\infty e^{-h\nu/k_BT} d\nu = k_BT/h$, gives the radiated power per unit volume,

$$\boxed{4\pi \int_0^\infty j_\nu \, d\nu = 6.8 \times 10^{-51} \frac{k_B}{h} Z_i^2 \, \bar{g} \, n_e n_i \sqrt{T} = 1.4 \times 10^{-40} \, Z_i^2 \, \bar{g} \, n_e n_i \sqrt{T} \ \ \mathrm{W\,m^{-3}}.} \tag{15.16}$$

The power rises with $n_e n_i$ because an electron must pass an ion to radiate. It also rises as $\sqrt{T}$ after we add all frequencies. A hydrogen plasma with $n_e = n_i = n$ holds thermal energy $\frac{3}{2}(n_e + n_i)k_BT = 3nk_BT$ per unit volume. Dividing that energy by the radiated power gives its cooling time,

$$t_{\mathrm{cool}} = \frac{3nk_BT}{1.4 \times 10^{-40} \, \bar{g} \, n^2 \sqrt{T}} = \frac{3k_B\sqrt{T}}{1.4 \times 10^{-40} \, \bar{g} \, n}.$$

For cluster gas with $n = 10^3$ m^{-3} and $T = 7 \times 10^7$ K, this is 2.4×10^{18} s, or 8×10^{10} years when $\bar{g} = 1$. That exceeds the age of the universe (Chapter 16). Most cluster gas therefore remains hot and emits X-rays long after it was heated (P6). A sphere of this gas 1 Mpc in radius radiates 1.5×10^{38} W, comparable to the X-ray luminosity of a rich cluster. Dense cluster centres cool faster, and in some of them the cooling time is shorter than the age of the universe.

Free-free absorption and Kramers opacity. A free electron can also absorb a photon while passing an ion. In thermal equilibrium, Kirchhoff's law relates this reverse process to the emission we have just found (P4). Equation (11.12) gives $\kappa_\nu \rho = j_\nu / B_\nu$.

$$\begin{aligned} \kappa_\nu \rho &\overset{(1)}{=} \frac{6.8 \times 10^{-51}}{4\pi} Z_i^2 \bar{g} \, \frac{n_e n_i}{\sqrt{T}} \, e^{-h\nu/k_BT} \, \frac{c^2 \left(e^{h\nu/k_BT} - 1\right)}{2h\nu^3} \\ &\overset{(2)}{=} \frac{6.8 \times 10^{-51} \, c^2}{8\pi h} Z_i^2 \bar{g} \, \frac{n_e n_i}{\sqrt{T}} \, \frac{1 - e^{-h\nu/k_BT}}{\nu^3}. \end{aligned} \tag{15.17}$$

(1) Eq. (15.15) divided by 4π, over the Planck function (9.31).

(2) $e^{-h\nu/k_BT}(e^{h\nu/k_BT}-1) = 1 - e^{-h\nu/k_BT}$.

At radio frequencies, $h\nu \ll k_BT$, so the last factor approaches $h\nu/k_BT$. The absorption then becomes $1.8\times10^{-12}\,Z_i^2\bar{g}\,n_e n_i T^{-3/2}\nu^{-2}$ m^{-1}. Ionised gas around hot stars absorbs low-frequency radio light strongly but lets higher-frequency radio light through. Where the gas is opaque, its spectrum rises as ν^2, the Rayleigh–Jeans part of $B_\nu(T)$ (Eq. (11.10)). Where it is transparent, $I_\nu \approx \tau_\nu B_\nu \propto \nu^{-2}\nu^2$ is nearly flat.

Inside a star, photons with $h\nu$ of a few k_BT carry most of the radiative flux (Section 11.4). To see how the opacity changes with temperature, write $\nu = x\,k_BT/h$ in Eq. (15.17).

$$\kappa_\nu\rho \propto \frac{n_e n_i}{\sqrt{T}}\,\frac{1-e^{-x}}{x^3}\left(\frac{h}{k_BT}\right)^3 \propto n_e n_i\,T^{-7/2}\,\frac{1-e^{-x}}{x^3}.$$

The Rosseland average over x changes the coefficient but adds no temperature dependence. Since $n_e n_i \propto \rho^2$, the opacity per unit mass obeys

$$\boxed{\kappa_{\mathrm{ff}} \propto \rho\,T^{-7/2}.} \tag{15.18}$$

This is the free–free part of Kramers' law (Section 12.3). Its $T^{-7/2}$ comes from the $T^{-1/2}$ in the emission and the ν^{-3} in the absorption, evaluated at frequencies that rise with T. Bound electrons also absorb light. Their contribution has the same density and temperature powers, but a larger coefficient in gas rich in heavy elements.

15.3 Synchrotron Radiation

We know how to find the power of a fast charge from its acceleration in its own frame (Eq. (15.6)). We also know that its light is beamed into an angle of about $1/\gamma_L$ (Eq. (15.7)). The radio light of the Milky Way and radio galaxies has a different shape from thermal light. It falls roughly as $\nu^{-0.7}$ across many decades of frequency and is strongly polarised. A single temperature cannot produce that spectrum.

Think of a lighthouse. A distant ship sees a flash only while the beam points toward it. A narrower beam makes a shorter flash, even when the lamp turns at the same rate. A fast electron turning in a magnetic field sends most of its light along its velocity, so an observer sees brief flashes as the beam sweeps past. The electron also moves toward the observer during each flash, shortening its arrival time further. Short flashes reach high frequencies. Each electron radiates near γ_L^2 times its slow gyration frequency, with power proportional to $\gamma_L^2 B^2$. Counting electrons of different energies then gives the broad radio spectrum (P3).

The Galactic magnetic field is about 0.5 nT. A slow electron circles in it at 14 Hz. An electron with $\gamma_L = 10^4$ has an energy of 5 GeV and circles 10^4 times more slowly, about once every twelve minutes. Yet it radiates near $10^8 \times 14$ Hz $= 1.4$ GHz, in the radio. Picture a helix around a field line and a narrow beam sweeping with the electron's velocity. Synchrotron light also comes from laboratory electron rings, pulsar nebulae, radio jets and lobes, and gamma-ray burst afterglows. The field turns the electron without doing work (P1). The flash length sets the frequency (P6), while the number of electrons at each energy sets the spectrum (P3).

Gyration. An electron of charge $-e$ and velocity $\vec{v}$ in a field $\vec{B}$ obeys

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\gamma_L m_e \vec{v}\right) = -e\,\vec{v}\times\vec{B}.$$

Taking the dot product with $\gamma_L m_e \vec{v}$,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(\gamma_L^2 m_e^2 v^2\right) \overset{(1)}{=} -e\,\gamma_L m_e\,\vec{v}\cdot(\vec{v}\times\vec{B}) \overset{(2)}{=} 0 \quad\Longrightarrow\quad \gamma_L = \text{const.} \tag{15.19}$$

(1) $\vec{a} \cdot \dot{\vec{a}} = \frac{1}{2}\, d(a^2)/dt$ with $\vec{a} = \gamma_L m_e \vec{v}$.

(2) $\vec{v} \times \vec{B}$ is perpendicular to $\vec{v}$. Since $\gamma_L^2 v^2 = c^2(\gamma_L^2 - 1)$, a constant $\gamma_L^2 v^2$ means a constant γ_L.

The magnetic force is perpendicular to the velocity, so it changes the direction but not the energy (P1). With γ_L constant, the equation becomes $\gamma_L m_e \dot{\vec{v}} = -e\, \vec{v} \times \vec{B}$. Motion along $\vec{B}$ continues unchanged. Across the field, the velocity turns at the angular frequency

$$\boxed{\frac{eB}{\gamma_L m_e} = \frac{2\pi \nu_g}{\gamma_L}, \qquad \nu_g = \frac{eB}{2\pi m_e} = 2.80 \times 10^{10}\,\mathrm{Hz} \times \frac{B}{1\,\mathrm{T}},} \tag{15.20}$$

Here ν_g is the **gyrofrequency** of a slow electron. The combined motion is a helix. A fast electron takes γ_L times longer to circle, though its observed radiation reaches frequencies about γ_L^2 above ν_g (Eq. (15.25)).

Power. Let the electron move along x through a magnetic field B along z. There is no electric field in the lab, but there is one in the electron's frame. Electromagnetism gives the boost of the field pair (E_y, cB_z) by the same matrix as (ct, x). To enter the electron's frame, use $\Lambda(-\beta)$,

$$\begin{pmatrix} E'_y \\ cB'_z \end{pmatrix} = \begin{pmatrix} \gamma_L & -\gamma_L\beta \\ -\gamma_L\beta & \gamma_L \end{pmatrix} \begin{pmatrix} 0 \\ cB \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} -\gamma_L \beta cB \\ \gamma_L cB \end{pmatrix}. \tag{15.21}$$

(1) The second column of the matrix times cB.

The electron is momentarily at rest in this frame, so only the electric field $E' = \gamma_L v B$ pushes it. Its acceleration has magnitude eE'/m_e. Equation (15.6) then gives

$$\begin{aligned} \mathcal{P} &\overset{(1)}{=} \frac{e^2}{6\pi\varepsilon_0 c^3}\left(\frac{e\gamma_L v B}{m_e}\right)^2 \overset{(2)}{=} \frac{e^4}{6\pi\varepsilon_0^2 m_e^2 c^4}\, \varepsilon_0 c\, \gamma_L^2 \beta^2 c^2 B^2 \\ &\overset{(3)}{=} \sigma_T c\, \gamma_L^2 \beta^2 \frac{B^2}{\mu_0} \overset{(4)}{=} 2\sigma_T c\, \gamma_L^2 \beta^2\, u_B. \end{aligned} \tag{15.22}$$

(1) Larmor's formula in the electron's frame, with $|\dot{\vec{v}}'| = eE'/m_e$.

(2) Multiply and divide by $\varepsilon_0 c$, and write $v = \beta c$.

(3) The Thomson cross-section of Eq. (8.7) is $\sigma_T = e^4/(6\pi\varepsilon_0^2 m_e^2 c^4)$, and $\varepsilon_0 c^2 = 1/\mu_0$.

(4) The magnetic energy density $u_B = B^2/(2\mu_0)$.

For an electron moving at an angle to the field, only the velocity across $\vec{B}$ produces the force. The average square of that part over all velocity directions is $\frac{2}{3}\beta^2 c^2$, so

$$\boxed{\mathcal{P} = \frac{4}{3}\, \sigma_T c\, \gamma_L^2 \beta^2\, u_B.} \tag{15.23}$$

The power rises with the square of the electron's energy and with the field energy density u_B. This form will make the comparison with inverse Compton scattering direct.

How long the electrons last. For $\beta \approx 1$, divide the electron's energy $\gamma_L m_e c^2$ by its radiated power to find its cooling time,

$$t_{\mathrm{sync}} = \frac{\gamma_L m_e c^2}{\mathcal{P}} = \frac{3 m_e c}{4\sigma_T \gamma_L u_B},$$

Faster electrons cool sooner. The Crab Nebula is the debris of the supernova seen in 1054. In a field of about 10 nT, its 1 keV X-rays come from electrons with $\gamma_L \approx 2.4 \times 10^7$ (Eq. (15.25)). They last only about 100 years, much less than the nebula's age (P6). The central pulsar must keep supplying them. Electrons making 1 GHz radio light have $\gamma_L \approx 1.5 \times 10^3$ and last over a million years, so they can accumulate throughout the nebula's life.

The frequency from the flash length. Consider an electron moving across the field. We see it only while its velocity lies within about $1/\gamma_L$ of our line of sight (Eq. (15.7)). During that flash, its velocity turns through $2/\gamma_L$. Equation (15.20) gives the lab duration $\Delta t = (2/\gamma_L)/(2\pi\nu_g/\gamma_L) = 2/(2\pi\nu_g)$. The electron also approaches us by $\beta c\,\Delta t$. Light sent at the end of the flash has less distance to travel and catches up with the light sent at the start (P6).

$$\begin{aligned}\Delta t_{\text{obs}} &\overset{(1)}{=} \Delta t - \frac{\beta c\,\Delta t}{c} = \Delta t\,(1-\beta)\\ &\overset{(2)}{=} \frac{\Delta t}{\gamma_L^2(1+\beta)} \overset{(3)}{=} \frac{1}{\gamma_L^2}\,\frac{1}{2\pi\nu_g}.\end{aligned} \tag{15.24}$$

(1) The arrival time of the last light, minus the head start it gains.
(2) $1-\beta = (1-\beta^2)/(1+\beta) = 1/[\gamma_L^2(1+\beta)]$.
(3) $\Delta t = 2/(2\pi\nu_g)$ and $1+\beta \approx 2$.

The received flash is about γ_L^2 times shorter than a turn of a slow electron. Its spectrum therefore reaches frequencies near $\gamma_L^2\nu_g$.

The frequency from the electron's frame. In the electron's frame, the same interval lasts $\Delta t' = \Delta t/\gamma_L$, by Eq. (15.4). Its forward light is shortened again by the Doppler factor $\gamma_L(1+\beta)$ of Eq. (15.7). Thus $\Delta t_{\text{obs}} = \Delta t'/[\gamma_L(1+\beta)] = \Delta t/[\gamma_L^2(1+\beta)]$, exactly as in the lab-frame calculation. Both routes give

$$\boxed{\nu_c \approx \gamma_L^2\,\nu_g.} \tag{15.25}$$

The full field calculation gives the **critical frequency** $\nu_c = \frac{3}{2}\gamma_L^2\nu_g$ for motion across the field. At other angles, multiply ν_g by the sine of the angle between $\vec{v}$ and $\vec{B}$. One electron has the spectrum $F(\nu/\nu_c)$, with $F(x) = x\int_x^\infty K_{5/3}(t)\,dt$. It rises as $\nu^{1/3}$, peaks at $\nu = 0.29\,\nu_c$, then falls as $e^{-\nu/\nu_c}$ (Figure 15.4, top). For the Galactic example, $\gamma_L = 10^4$ in 0.5 nT gives $\nu_c = 2.1$ GHz.

A power law of electrons. Shocks can give electrons a power-law distribution, $n(\gamma_L)\,d\gamma_L \propto \gamma_L^{-p}\,d\gamma_L$, between $\gamma_{\min}$ and $\gamma_{\max}$. Often p lies between 2 and 3. Each electron radiates most strongly near $\nu = \gamma_L^2\nu_g$. At a given ν, the main contributors thus have $\gamma_L = (\nu/\nu_g)^{1/2}$. Count their power (P3),

$$\begin{aligned}4\pi j_\nu &\overset{(1)}{=} \mathcal{P}(\gamma_L)\,n(\gamma_L)\,\frac{\mathrm{d}\gamma_L}{\mathrm{d}\nu}\\ &\overset{(2)}{\propto} \gamma_L^2 B^2\;\gamma_L^{-p}\;\frac{1}{2\sqrt{\nu\nu_g}}\\ &\overset{(3)}{\propto} B^2\left(\frac{\nu}{\nu_g}\right)^{(2-p)/2}\frac{1}{\sqrt{\nu\nu_g}} \overset{(4)}{\propto} B^{(p+1)/2}\,\nu^{-(p-1)/2}.\end{aligned} \tag{15.26}$$

(1) The power at frequencies ν to $\nu + d\nu$ is that of the electrons in the matching $d\gamma_L$.
(2) Eq. (15.23) with $u_B \propto B^2$ and $\beta \approx 1$. From $\gamma_L = (\nu/\nu_g)^{1/2}$, $d\gamma_L/d\nu = 1/(2\sqrt{\nu\nu_g})$.
(3) $\gamma_L^{2-p} = (\nu/\nu_g)^{(2-p)/2}$.
(4) The powers of ν add to $(2-p)/2 - \frac{1}{2} = -(p-1)/2$. Since $\nu_g \propto B$, the powers of B add to $2 - (2-p)/2 - \frac{1}{2} = (p+1)/2$.

$$\boxed{j_\nu \propto \nu^{-(p-1)/2}.} \tag{15.27}$$

The electron distribution sets the slope of the light between its two ends. The one-electron spectrum shapes those ends. Summing the exact $F(x)$ over electrons with $p = 2.5$ and

$\gamma_{\max} = 1000\,\gamma_{\min}$ gives a middle slope of -0.752, close to the -0.75 from Eq. (15.27). Below $\gamma_{\min}^2 \nu_c$ the spectrum rises with the single-electron slope 0.32. Above $\gamma_{\max}^2 \nu_c$ it falls exponentially (Figure 15.4, bottom).

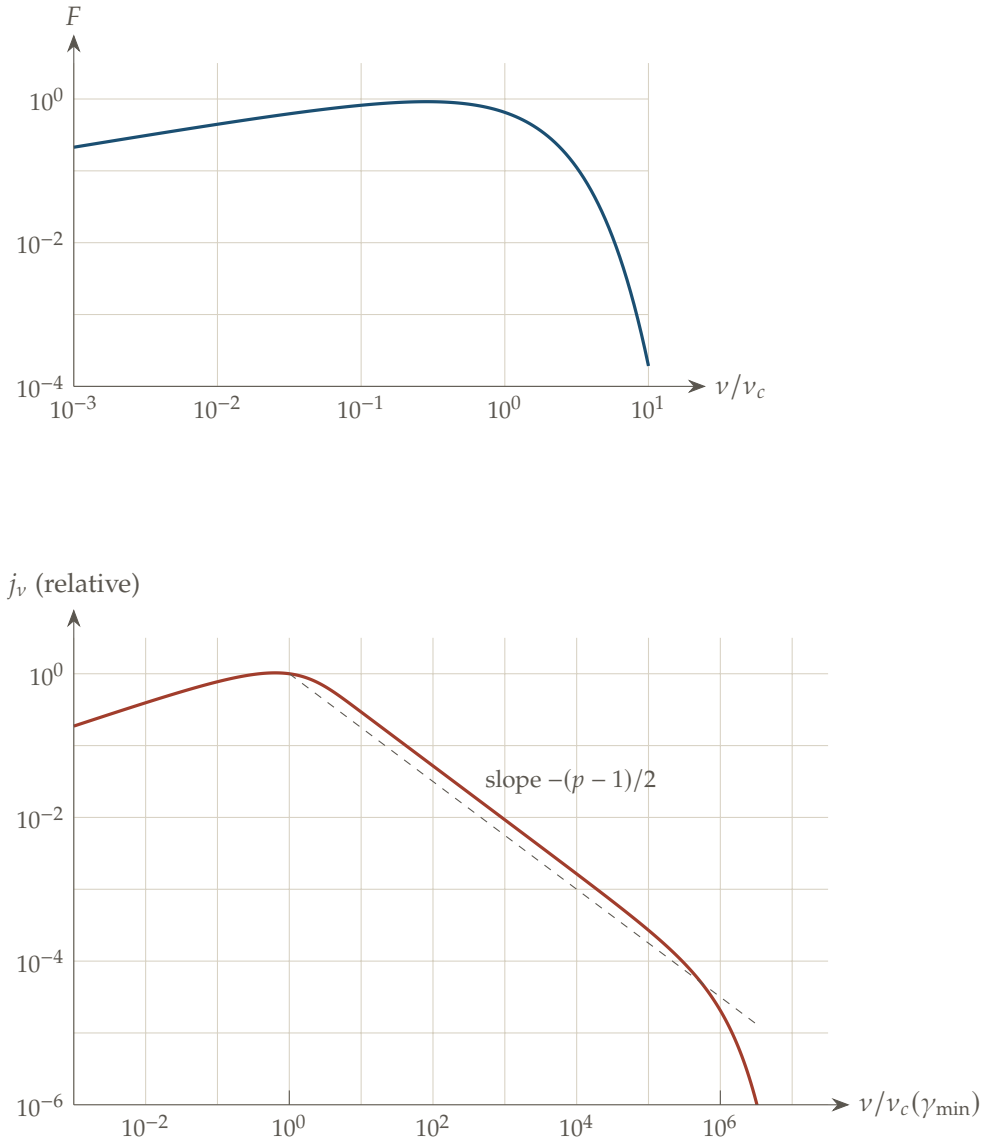


Figure 15.4: Top, the synchrotron spectrum $F(\nu/\nu_c)$ of one electron. Bottom, the spectrum of electrons with $n(\gamma_L) \propto \gamma_L^{-2.5}$ between $\gamma_{\min}$ and $1000\,\gamma_{\min}$, against $\nu/\nu_c(\gamma_{\min})$. Ticks on the frequency axis mark $\nu_c(\gamma_{\min})$ and $\nu_c(\gamma_{\max})$. The dashed line has slope -0.75. Computed with `code/ch15/01_synchrotron.py`.

Most radio galaxies have $\alpha \approx 0.7$ to 0.8 in $j_\nu \propto \nu^{-\alpha}$, corresponding to electron slopes $p \approx 2.4$ to 2.6. The Crab's radio spectrum is flatter, with $\alpha \approx 0.3$ and $p \approx 1.6$. Cooling can steepen either spectrum (P4). Electrons supplied at a rate $\propto \gamma_L^{-p}$ survive for $t_{\text{sync}} \propto 1/\gamma_L$, leaving a population $\propto \gamma_L^{-(p+1)}$ where cooling is fast. Above the resulting break, α increases by $\frac{1}{2}$. Its frequency can reveal the age of a radio source. Synchrotron light is linearly polarised across the magnetic field projected on the sky. Polarisation maps therefore trace cosmic magnetic fields; the strong optical polarisation of the Crab helped identify its light as synchrotron radiation.

15.4 Compton and Inverse Compton Scattering

Equation (8.7) gives the cross-section for light scattered by a free electron. In the Thomson limit, the photon leaves with almost the energy it arrived with. A hard X-ray can instead recoil from the electron and lose energy. A fast electron can give energy to the photon. These are Compton and inverse Compton scattering, the energy-changing processes listed in Table 8.2.

Think of a ball hitting a racket. A still racket sends it back with nearly its arrival speed. A racket moving toward it sends it back faster. The motion is easiest to describe from the racket, where the ball arrives faster, then to translate back to the court. A photon meeting a fast electron behaves similarly. It is blueshifted by about γ_L on the way into the electron's frame and again on the way out. In that frame the scattering is nearly Thomson scattering. The boost matrix of Eq. (15.4) carries the photon there and back.

A microwave-background photon has a mean energy of 6.3×10^{-4} eV. After meeting an electron with $\gamma_L = 10^4$, its average energy can be 85 keV, a hard X-ray. The same exchange of energy between frames appears in the spacecraft gravity assist of Chapter 6. Inverse Compton scattering makes X-rays in radio lobes, gamma rays in pulsar nebulae and blazars, and X-rays above accretion disks. We use conservation of energy and momentum (P1) for one encounter, a thermal distribution for hot gas (P3), and a random walk in photon energy for many encounters (P5). These lead to the Compton shift, inverse Compton power, the Compton y parameter, and the Sunyaev–Zel'dovich effect.

The Compton shift. First let the electron be at rest. A photon of frequency ν leaves the encounter at angle ψ with frequency ν_1. Write a four-momentum as (energy over c, momentum), with dot product $A\cdot B = A^0B^0 - \vec{A}\cdot\vec{B}$. The photon has $k = (h\nu/c)(1,\hat{n})$ before and $k_1 = (h\nu_1/c)(1,\hat{n}_1)$ after. The electron has $q = (m_e c, \vec{0})$ before and q_1 after. Conservation (P1) gives $q + k = q_1 + k_1$.

$$
\begin{aligned}
m_e^2c^2 &\overset{(1)}{=} q_1\cdot q_1 = (q+k-k_1)\cdot(q+k-k_1)\\
&\overset{(2)}{=} m_e^2c^2 + 2\,q\cdot k - 2\,q\cdot k_1 - 2\,k\cdot k_1\\
\Longrightarrow\quad m_e h(\nu-\nu_1) &\overset{(3)}{=} \frac{h^2\nu\nu_1}{c^2}(1-\cos\psi)\\
\Longrightarrow\quad \frac{c}{\nu_1} - \frac{c}{\nu} &\overset{(4)}{=} \frac{h}{m_e c}(1-\cos\psi).
\end{aligned}
\tag{15.28}
$$

(1) The electron's mass is the same before and after. Conservation gives q_1.
(2) Expand. $q\cdot q = m_e^2c^2$, and $k\cdot k = k_1\cdot k_1 = 0$ for photons.
(3) $q\cdot k = m_e h\nu$, $q\cdot k_1 = m_e h\nu_1$, and $k\cdot k_1 = (h^2\nu\nu_1/c^2)(1-\hat{n}\cdot\hat{n}_1)$ with $\hat{n}\cdot\hat{n}_1 = \cos\psi$.
(4) Divide by $m_e h\nu\nu_1/c$.

$$
\boxed{\lambda_1 - \lambda = \lambda_C\,(1-\cos\psi), \qquad h\nu_1 = \frac{h\nu}{1+(h\nu/m_ec^2)(1-\cos\psi)},}
\tag{15.29}
$$

Here the Compton wavelength is $\lambda_C = h/(m_e c) = 2.43\times10^{-12}$ m, and $m_e c^2 = 511$ keV. The wavelength shift depends on the scattering angle, not the incoming wavelength. It becomes important when λ approaches λ_C, or $h\nu$ approaches $m_e c^2$. At $h\nu \ll m_e c^2$, recoil is small and we recover the Thomson limit. A 10 keV X-ray loses about 2% of its energy per scattering from cold electrons.

Inverse Compton, there and back. Now let the electron move along x with Lorentz factor γ_L. A photon of frequency ν approaches at angle θ to its velocity. Boost the photon into the

electron's frame with $\Lambda(-\beta)$,

$$\frac{h\nu'}{c}\begin{pmatrix}1\\ \cos\theta'\\ \sin\theta'\end{pmatrix}\overset{(1)}{=}\begin{pmatrix}\gamma_L & -\gamma_L\beta & 0\\ -\gamma_L\beta & \gamma_L & 0\\ 0 & 0 & 1\end{pmatrix}\frac{h\nu}{c}\begin{pmatrix}1\\ \cos\theta\\ \sin\theta\end{pmatrix}=\frac{h\nu}{c}\begin{pmatrix}\gamma_L(1-\beta\cos\theta)\\ \gamma_L(\cos\theta-\beta)\\ \sin\theta\end{pmatrix}$$

$$\Longrightarrow\qquad \nu'\overset{(2)}{=}\gamma_L(1-\beta\cos\theta)\,\nu. \tag{15.30}$$

(1) The inverse boost of Eq. (15.4) applied to the photon's four-momentum.
(2) The first row.

When $h\nu' \ll m_e c^2$, recoil is small in the electron's frame. The photon keeps frequency ν' and leaves at angle θ_1'. Boosting it back to the lab with $\Lambda(\beta)$ gives, from Eq. (15.7),

$$\boxed{\nu_1=\gamma_L(1+\beta\cos\theta_1')\,\nu'=\gamma_L^2\,(1-\beta\cos\theta)(1+\beta\cos\theta_1')\,\nu.} \tag{15.31}$$

The largest gain occurs when the photon arrives head on, $\theta=\pi$, and leaves forward, $\theta_1'=0$. Then $\nu_1=\gamma_L^2(1+\beta)^2\nu\approx 4\gamma_L^2\nu$. Each boost supplies one factor of γ_L. This Thomson result requires $h\nu'\approx\gamma_L h\nu\ll m_e c^2$ in the electron's frame. At higher energies, recoil matters (Eq. (15.29)), and the photon cannot gain more than about the electron's energy $\gamma_L m_e c^2$.

The power. Let the electron cross a photon field of energy density $u_{\rm rad}$, equal in all directions. Photons arriving at angle θ meet it at the closing speed $c(1-\beta\cos\theta)$. The share arriving within $d\Omega$ carries energy density $u_{\rm rad}\,d\Omega/4\pi$ and scatters at a rate proportional to $\sigma_T c(1-\beta\cos\theta)$. In the electron's frame, Thomson scattering is symmetric forward and backward, so the average $\cos\theta_1'$ in Eq. (15.31) is zero. Subtract the incoming photon power from the outgoing power,

$$\begin{aligned}\mathcal{P}&\overset{(1)}{=}\sigma_T c\int\frac{d\Omega}{4\pi}\,u_{\rm rad}\,(1-\beta\cos\theta)\Big[\gamma_L^2(1-\beta\cos\theta)-1\Big]\\ &\overset{(2)}{=}\sigma_T c\,u_{\rm rad}\Big[\gamma_L^2\Big(1+\frac{\beta^2}{3}\Big)-1\Big]\\ &\overset{(3)}{=}\sigma_T c\,u_{\rm rad}\Big[\gamma_L^2\beta^2+\frac{\gamma_L^2\beta^2}{3}\Big].\end{aligned} \tag{15.32}$$

(1) Rate of scattering times the energy gained, which is $\gamma_L^2(1-\beta\cos\theta)$ times the arriving energy by Eq. (15.31), minus the arriving energy.
(2) $\int(1-\beta\cos\theta)^2\,d\Omega/4\pi=1+\beta^2/3$ and $\int(1-\beta\cos\theta)\,d\Omega/4\pi=1$.
(3) $\gamma_L^2-1=\gamma_L^2\beta^2$.

$$\boxed{\mathcal{P}=\frac{4}{3}\,\sigma_T c\,\gamma_L^2\beta^2\,u_{\rm rad}.} \tag{15.33}$$

This is the synchrotron power of Eq. (15.23) with photon energy density in place of magnetic energy density. The electron's losses to the two processes have the ratio $u_{\rm rad}/u_B$. Dividing the scattered power by the photon scattering rate $\sigma_T c\,u_{\rm rad}/\langle h\nu\rangle$ gives the mean scattered energy $\frac{4}{3}\gamma_L^2\,\langle h\nu\rangle$.

At 2.725 K, the microwave background has $u_{\rm rad}=a_{\rm rad}T^4=4.2\times10^{-14}\,\mathrm{J\,m^{-3}}$ (Eq. (9.30)). A magnetic field of 0.32 nT has the same energy density. In radio lobes, where fields are of order 1 nT, scattering the background can take a tenth or more as much electron energy as synchrotron emission and produce X-rays. In the Crab, electrons scatter their own synchrotron photons up to gamma-ray energies.

Many scatterings in hot gas. Consider slow, thermal electrons with $\beta \ll 1$ and photons with $h\nu \ll k_B T$. We need the mean gain and the spread of frequency changes in one scattering. The mean follows from Eq. (15.33) divided by the scattering rate. For Maxwellian electrons, $\langle \gamma_L^2 \beta^2 \rangle \approx \langle v^2 \rangle / c^2 = 3k_B T/m_e c^2$ (P3). Let $\Delta\nu = \nu_1 - \nu$ mean after minus before one scattering. To first order in β, Eq. (15.31) gives $\Delta\nu/\nu = \beta(\cos\theta_1' - \cos\theta)$. The two directions are independent of the electron velocity and each has $\langle \cos^2 \rangle = \frac{1}{3}$,

$$\left\langle \frac{\Delta\nu}{\nu} \right\rangle = \frac{4}{3} \cdot \frac{3k_B T}{m_e c^2} = \frac{4k_B T}{m_e c^2}, \qquad \left\langle \left(\frac{\Delta\nu}{\nu} \right)^2 \right\rangle = \frac{3k_B T}{m_e c^2} \left(\frac{1}{3} + \frac{1}{3} \right) = \frac{2k_B T}{m_e c^2}. \tag{15.34}$$

Recoil would remove a fraction $h\nu/m_e c^2$ of the photon energy (Eq. (15.29)). It is small here because $h\nu \ll k_B T$. The cloud's effect depends on the number of scatterings multiplied by the small gain per scattering. That product is the **Compton y parameter**,

$$\boxed{y = \int \frac{k_B T}{m_e c^2} \, n_e \sigma_T \, dl,} \tag{15.35}$$

the optical depth weighted by $k_B T/m_e c^2$. A thin cloud gives about $\tau = \int n_e \sigma_T dl$ scatterings per photon. In a thick cloud, the photon wanders and scatters about τ^2 times (Eq. (9.42), P5).

The random walk in frequency. Let $x = h\nu/k_B T_\gamma$, with T_γ fixed. Then x is simply frequency in chosen units. Each scattering changes it by a small, random amount. Equation (15.34) gives the moments of that change per scattering. We first express them per unit y, then track how photons move between frequency intervals.

What Δx means. Follow one photon through the cloud, using y of Eq. (15.35) as a clock that starts at entry. Let $x(y)$ be its frequency variable at that point. Over a short stretch dy, it changes by

$$\Delta x = x(y + dy) - x(y). \tag{15.36}$$

The two Δ symbols cover different intervals. $\Delta\nu = \nu_1 - \nu$ is the change in one scattering, while Δx covers the whole stretch dy. A photon might not scatter during that stretch. If it does, the size of the change depends on the electron it meets. We therefore average over many photons starting at the same x to find $\langle \Delta x \rangle$ and $\langle (\Delta x)^2 \rangle$.

From one scattering to a stretch of y. Let $n_{\rm sc}$ count the expected scatterings. Along dl, it grows by $dn_{\rm sc} = n_e \sigma_T \, dl$. Equation (15.35) then gives $dn_{\rm sc} = (m_e c^2/k_B T)\, dy$. In a sufficiently short stretch, the photon scatters once with probability $dn_{\rm sc}$ and twice only with probability of order $dn_{\rm sc}^2$. With no scattering, $\Delta x = 0$. With one, $\Delta x = x\, \Delta\nu/\nu$, since x and ν change by the same fraction. Averaging these possibilities gives

$$\begin{aligned} \langle \Delta x \rangle &\overset{(1)}{=} dn_{\rm sc} \; x \left\langle \frac{\Delta\nu}{\nu} \right\rangle \\ &\overset{(2)}{=} \frac{m_e c^2}{k_B T} \, dy \; x \, \frac{4k_B T}{m_e c^2} \\ &\overset{(3)}{=} 4x \, dy, \end{aligned} \tag{15.37}$$

$$\left\langle (\Delta x)^2 \right\rangle \overset{(4)}{=} dn_{\rm sc} \; x^2 \left\langle \left(\frac{\Delta\nu}{\nu} \right)^2 \right\rangle \overset{(5)}{=} \frac{m_e c^2}{k_B T} \, dy \; x^2 \, \frac{2k_B T}{m_e c^2} = 2x^2 \, dy.$$

(1) The step is $x\, \Delta\nu/\nu$ with chance $dn_{\rm sc}$ and zero otherwise.

(2) $dn_{\rm sc} = (m_e c^2/k_B T)\,dy$ from the definition of y, and the mean of $\Delta\nu/\nu$ from Eq. (15.34).
(3) The factors $k_B T/m_e c^2$ cancel. This is why y is the right clock.
(4) The same for the square of the step.
(5) The mean square of $\Delta\nu/\nu$ from Eq. (15.34).

$$\boxed{\frac{\langle \Delta x\rangle}{dy} = 4x, \qquad \frac{\langle (\Delta x)^2\rangle}{dy} = 2x^2.} \tag{15.38}$$

Per unit y, a photon starting at x moves upward by $4x$ on average, with a mean square step of $2x^2$. These are local rates, measured over a short stretch. The second is the mean square of one step; it is not the width of the full photon spectrum.

How the photon numbers change. Let $N(x, y)\,dx$ count photons between x and $x + dx$. To connect single-photon steps to this distribution, choose a smooth test function $w(x)$ that vanishes outside a finite range. Sum its value over all photons, $W = \int N\,w\,dx$. When each photon moves from x to $x + \Delta x$ during dy, the sum changes by

$$\begin{aligned} dW &\overset{(1)}{=} \sum_{\text{photons}} \big[w(x+\Delta x) - w(x)\big] \\ &\overset{(2)}{=} \sum_{\text{photons}} \Big[\Delta x\, w'(x) + \frac{1}{2}(\Delta x)^2\, w''(x)\Big] \\ &\overset{(3)}{=} \int N\Big[4x\, w' + x^2\, w''\Big]dx\; dy, \\ \Longrightarrow \quad \int w\,\frac{\partial N}{\partial y}\,dx &\overset{(4)}{=} \int N\Big[4x\, w' + x^2\, w''\Big]dx \\ &\overset{(5)}{=} \int w\Big[-\frac{\partial}{\partial x}\big(4xN\big) + \frac{\partial^2}{\partial x^2}\big(x^2 N\big)\Big]dx. \end{aligned} \tag{15.39}$$

(1) W counts w once for each photon, so it changes by the change of w at each photon.
(2) Taylor expansion of w about the starting x. Each step $\Delta\nu/\nu$ is of order β, so the third and higher moments per unit y are smaller by further factors of $\sqrt{k_B T/m_e c^2} \ll 1$ and are dropped. This small-step approximation is the whole of the physics assumed.
(3) The $N\,dx$ photons near x all start at x, so their steps average to Eq. (15.38), $\langle \Delta x\rangle = 4x\,dy$ and $\frac{1}{2}\langle (\Delta x)^2\rangle = x^2\,dy$.
(4) Divide by dy. The function w does not depend on y, so $dW/dy = \int w\,\partial N/\partial y\,dx$.
(5) Integrate by parts, once in the first term and twice in the second. The end terms vanish because w and w' are zero at the ends of its range.

Because the last line holds for every smooth w, the expressions it weighs must agree at each x,

$$\frac{\partial N}{\partial y} = -\frac{\partial}{\partial x}\Big[4xN - \frac{\partial}{\partial x}\big(x^2 N\big)\Big]. \tag{15.40}$$

This is a continuity equation in frequency space (P1). Scattering moves photons between frequencies without creating or destroying them. The bracket is the flux toward higher x. Its first term, $4xN$, describes the average upward drift. The second describes random spreading (P5), like diffusion in Eq. (9.44), with larger steps at higher x. Write $N = x^2 f$, where f is the

photon occupation number of Section 9.3,

$$\frac{\partial N}{\partial y} \overset{(1)}{=} -\frac{\partial}{\partial x}\left[4x^3 f - \frac{\partial}{\partial x}\left(x^4 f\right)\right] \overset{(2)}{=} \frac{\partial}{\partial x}\left(x^4 \frac{\partial f}{\partial x}\right). \tag{15.41}$$

(1) $N = x^2 f$ in Eq. (15.40).
(2) $\partial(x^4 f)/\partial x = 4x^3 f + x^4\, \partial f/\partial x$, and the $4x^3 f$ terms cancel.

$$\boxed{\frac{\partial\left(x^2 f\right)}{\partial y} = \frac{\partial}{\partial x}\left(x^4 \frac{\partial f}{\partial x}\right).} \tag{15.42}$$

This is the Kompaneets equation in the limit of negligible recoil. Multiply by x and integrate over frequency to get $d\langle x\rangle/dy = 4\langle x\rangle$. The mean photon energy then grows as e^{4y} while the approximation holds. In a hot corona above an accretion disk, repeated scattering can lift cooler seed photons into an X-ray spectrum.

The Sunyaev–Zel'dovich effect. Microwave-background photons crossing a hot galaxy cluster have $y \ll 1$. Choose T_γ as the background temperature. Before the cluster, their occupation number is $f_0 = 1/(e^x - 1)$. To first order in y, Eq. (15.42) gives

$$\frac{\Delta f}{f_0} \overset{(1)}{=} \frac{y}{x^2 f_0}\frac{\mathrm{d}}{\mathrm{d}x}\left(x^4 \frac{\mathrm{d}f_0}{\mathrm{d}x}\right) \overset{(2)}{=} y\,\frac{x e^x}{e^x - 1}\left(x\,\frac{e^x + 1}{e^x - 1} - 4\right) \xrightarrow{x \ll 1} -2y. \tag{15.43}$$

(1) Integrate Eq. (15.42) across the cluster with the right-hand side evaluated on f_0.
(2) $df_0/dx = -e^x/(e^x - 1)^2$, differentiated once more. For $x \ll 1$ the first factor tends to 1 and the bracket to $2 - 4$.

At low frequencies, intensity is proportional to temperature. The background seen through the cluster therefore appears cooler,

$$\boxed{\frac{\Delta T}{T_{\mathrm{CMB}}} = -2y \qquad (h\nu \ll k_B T_{\mathrm{CMB}}).}$$

Scattering shifts photons upward in frequency without removing them. A deficit at low frequencies therefore comes with an excess at high frequencies. The change crosses zero at $x = 3.83$, or 217 GHz (Figure 15.5). For a cluster with $n_e = 10^3\,\mathrm{m}^{-3}$, $k_B T = 6\,\mathrm{keV}$, and a path length of 1 Mpc, we have $\tau = 2.1 \times 10^{-3}$, $k_B T/m_e c^2 = 0.012$, and $y = 2.4 \times 10^{-5}$. The low-frequency background appears dimmer by 130 μK. This fractional distortion does not fade with the cluster's distance (P6), which lets surveys find distant clusters.

15.5 Accretion and Active Galaxies

A quasar can shine at 10^{40} W, about a thousand times a large galaxy of 10^{10} to 10^{11} Suns, while its observed brightness rises or falls noticeably over a few days. The region producing that variable light can be only a few light-days across. Hydrogen burning releases nuclear binding energy because the fusion products have slightly less mass than the fuel. That mass difference is about 0.7% of the fuel's rest mass (Section 12.3), too little to power a quasar from so small a region. Accretion draws on a different source, gravitational energy. As gas spirals inward, its

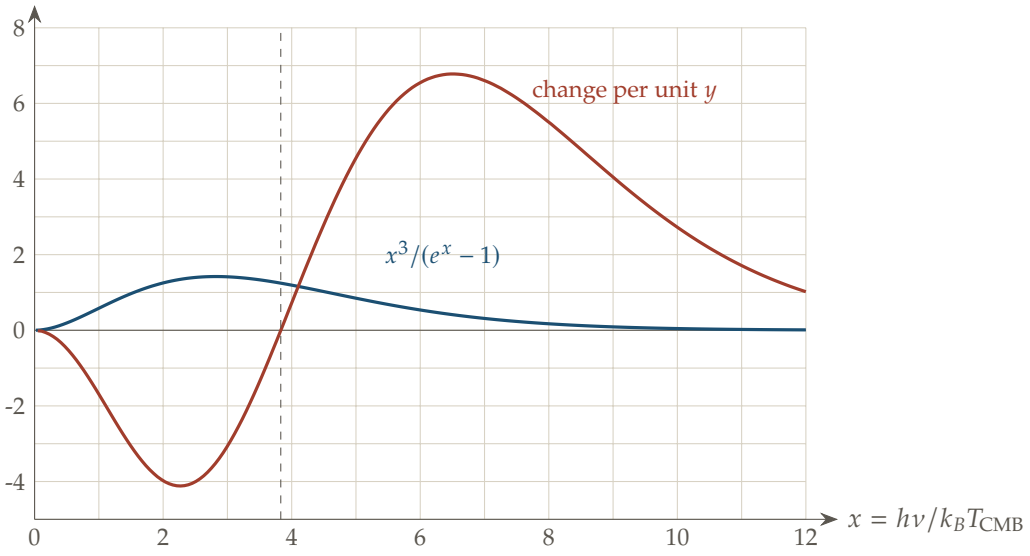


Figure 15.5: The microwave background intensity shape $x^3/(e^x-1)$, and its change per unit y after crossing hot gas, $x^4e^x(e^x-1)^{-2}[x(e^x+1)/(e^x-1)-4]$, both in units of $2(k_BT_{\rm CMB})^3/(hc)^2$. The dashed line marks $x = 3.83$.

gravitational potential energy falls even as its orbital speed rises. Friction between disk rings turns the decrease in total orbital energy into heat, which the gas radiates.

Water falling over a dam turns a turbine. The power depends on both the height of the fall and how much water flows each second. Accreting gas follows the same energy accounting. The depth of the gravitational well sets the energy available per kilogram, while the inflow rate sets how many kilograms arrive. In a disk, friction transfers orbital energy between rings and heats the gas, which radiates that heat (P1, P4).

Falling onto a neutron star can release gravitational energy equal to 0.17 of the gas's rest-mass energy, 25 times the yield per kilogram from burning hydrogen. A quasar shining at 10^{40} W with efficiency 0.1 needs $1.1\times10^{24}\,\mathrm{kg\,s^{-1}}$, or 18 solar masses a year. Picture a thin disk growing hotter toward the centre, sometimes with narrow jets along its axis. Accretion also powers X-ray binaries, white-dwarf outbursts, and disks around young stars. Energy and angular momentum conservation (P1), gravity (P2), local heating and cooling balance (P4), and light-travel times (P6) give the luminosity, disk spectrum, and apparent jet speeds.

Accretion luminosity. Let gas fall from far away onto a body of mass M and radius R at a rate $\dot{M}$. Outside the body, the gravitational potential energy per kilogram is $\Phi(r) = -GM/r$ (P2). The gas starts nearly at rest, where $\Phi \approx 0$, and gains speed as it falls. When it reaches the surface, that motion turns to heat. In steady state the heat leaves as radiation at the rate it is made (P4). Multiply the energy released per kilogram by $\dot{M}$,

$$\begin{aligned} L &\overset{(1)}{=} \left[\Phi(\infty) - \Phi(R)\right]\dot{M} \\ &\overset{(2)}{=} \left[0 - \left(-\frac{GM}{R}\right)\right]\dot{M} \\ &\overset{(3)}{=} \frac{GM\dot{M}}{R}. \end{aligned} \tag{15.44}$$

(1) Energy conservation (P1). Each kilogram releases the drop of its potential energy on the way down, and in a steady state all of it is radiated (P4).

(2) $\Phi(r) = -GM/r$ outside the body, so $\Phi(\infty) = 0$ and $\Phi(R) = -GM/R$.

(3) The two minus signs cancel.

Divide the radiated power by $\dot{M}c^2$, the rest-mass energy carried in each second. This defines the **efficiency** η. The denominator sets a scale for comparison. The light comes from gravitational energy lost during infall, and the ratio is

$$\boxed{L = \frac{GM\dot{M}}{R}, \qquad \eta \equiv \frac{L}{\dot{M}c^2} = \frac{GM}{Rc^2}.} \tag{15.45}$$

This η equals the compactness in Table 13.1. It is 2.6×10^{-4} for a white dwarf like Sirius B and 0.17 for a neutron star. A black hole has no surface, so gas radiates as it spirals toward the last stable orbit. For a black hole without spin, general relativity puts that orbit at $6GM/c^2$ and gives efficiency $1 - \sqrt{8/9} = 0.057$. For the fastest allowed spin, it can reach 0.42. These efficiencies come from general relativity. The Newtonian disk estimate below releases half of GM/R by $R = 6GM/c^2$, giving $1/12 = 0.083$.

The Eddington limit and black-hole masses. Radiation pushes back on the infalling gas. Equation (11.27) limits the steady luminosity of mass M to $1.26 \times 10^{31}\,\mathrm{W}\,(M/M_\odot)$. A quasar at 10^{40} W therefore needs at least $8 \times 10^8\,M_\odot$ (Problem 11.9.4). If it brightens within one day, the emitting region cannot be much larger than one light-day, 2.6×10^{13} m (P6). The Schwarzschild radius of $10^9\,M_\odot$ is 3×10^{12} m. The source thus holds roughly a billion solar masses within only a few tens of Schwarzschild radii, pointing to a black hole.

The thin disk, angular momentum. Gas approaching the hole has angular momentum, so it settles into a disk rather than falling straight in. At radius r, a ring orbits at angular speed $\sqrt{GM/r^3}$ and has angular momentum $\sqrt{GMr}$ per unit mass. It can move inward only by passing angular momentum to outer rings. Let $\mathcal{T}(r)$ be the torque exerted by gas inside r on gas outside. In steady flow, the inward angular momentum carried by gas minus the outward amount carried by the torque is constant with radius (P1). At the inner edge R_{in}, the gas plunges toward the hole and we take the torque to vanish.

$$\begin{aligned} \dot{M}\sqrt{GMr} - \mathcal{T}(r) &\overset{(1)}{=} \dot{M}\sqrt{GMR_{\mathrm{in}}} - 0 \\ \Longrightarrow \qquad \mathcal{T}(r) &= \dot{M}\sqrt{GM}\left(\sqrt{r} - \sqrt{R_{\mathrm{in}}}\right). \end{aligned} \tag{15.46}$$

(1) The constant evaluated at the inner edge, where $\mathcal{T} = 0$.

The thin disk, energy. Energy also flows between rings. Gas moving inward carries orbital energy $-GM/2r$ per kilogram, equivalent to an outward energy flow $\dot{M}GM/2r$. The torque does work on outer gas at a rate $\mathcal{T}\sqrt{GM/r^3}$. A ring between r and $r + dr$ radiates the energy it takes from these flows (P1). Both faces emit locally as blackbodies at temperature $T(r)$ (P4, Eq. (4.24)).

$$\begin{aligned} 2 \cdot 2\pi r\,dr\;\sigma_{\mathrm{SB}}T^4 &\overset{(1)}{=} -\frac{\mathrm{d}}{\mathrm{d}r}\left[\frac{GM\dot{M}}{2r} + \mathcal{T}\sqrt{\frac{GM}{r^3}}\right]dr \\ &\overset{(2)}{=} -GM\dot{M}\,\frac{\mathrm{d}}{\mathrm{d}r}\left[\frac{1}{2r} + \frac{1}{r} - \frac{\sqrt{R_{\mathrm{in}}}}{r^{3/2}}\right]dr \\ &\overset{(3)}{=} GM\dot{M}\left[\frac{3}{2r^2} - \frac{3\sqrt{R_{\mathrm{in}}}}{2r^{5/2}}\right]dr = \frac{3GM\dot{M}}{2r^2}\left(1 - \sqrt{\frac{R_{\mathrm{in}}}{r}}\right)dr. \end{aligned} \tag{15.47}$$

(1) Radiation from the two faces of the ring equals the energy flow into it from inside minus the flow out of it.

(2) The torque (15.46), so $\mathcal{T}\sqrt{GM/r^3} = GM\dot{M}\left(r^{-1} - \sqrt{R_{\rm in}}\, r^{-3/2}\right)$.
(3) Differentiate term by term.

$$\boxed{\sigma_{\rm SB}T^4 = \frac{3GM\dot{M}}{8\pi r^3}\left(1 - \sqrt{\frac{R_{\rm in}}{r}}\right), \qquad T \propto r^{-3/4} \text{ for } r \gg R_{\rm in}.} \tag{15.48}$$

The product $M\dot{M}$ sets the temperature at a given radius. Far from the inner edge, a ring radiates three times the energy released locally by gas moving through it. The torque transports the other two thirds from smaller radii.

Integrating over the disk checks the energy budget,

$$\int_{R_{\rm in}}^{\infty} \frac{3GM\dot{M}}{2r^2}\left(1 - \sqrt{\frac{R_{\rm in}}{r}}\right) dr = \frac{GM\dot{M}}{2R_{\rm in}},$$

This is half the energy $GM\dot{M}/R_{\rm in}$ available from the fall. The gas still carries the other half as orbital kinetic energy at $R_{\rm in}$, as required for a circular orbit (Section 10.2). A star releases that energy where the gas reaches its surface. A black hole swallows it.

Define $T_* = (3GM\dot{M}/8\pi\sigma_{\rm SB}R_{\rm in}^3)^{1/4}$. Then $T = T_*(R_{\rm in}/r)^{3/4}(1 - \sqrt{R_{\rm in}/r})^{1/4}$, with a peak at $r = \frac{49}{36}R_{\rm in}$ and $T = 0.488\,T_*$ (Figure 15.6). A $10^8\,M_\odot$ black hole fed $1\,M_\odot$ a year, with $R_{\rm in} = 6GM/c^2$, has a hottest ring at 1.1×10^5 K. Its disk shines mainly in the ultraviolet with 4.7×10^{38} W. For a $10\,M_\odot$ black hole fed $10^{-8}\,M_\odot$ a year, the peak is 3.5×10^6 K, or $k_BT = 0.30$ keV, in X-rays. Because $R_{\rm in} \propto M$, we have $T_* \propto (\dot{M}/M^2)^{1/4}$. At a fixed fraction of the Eddington limit, $\dot{M} \propto M$ and $T_* \propto M^{-1/4}$. Smaller black holes therefore have hotter disks.

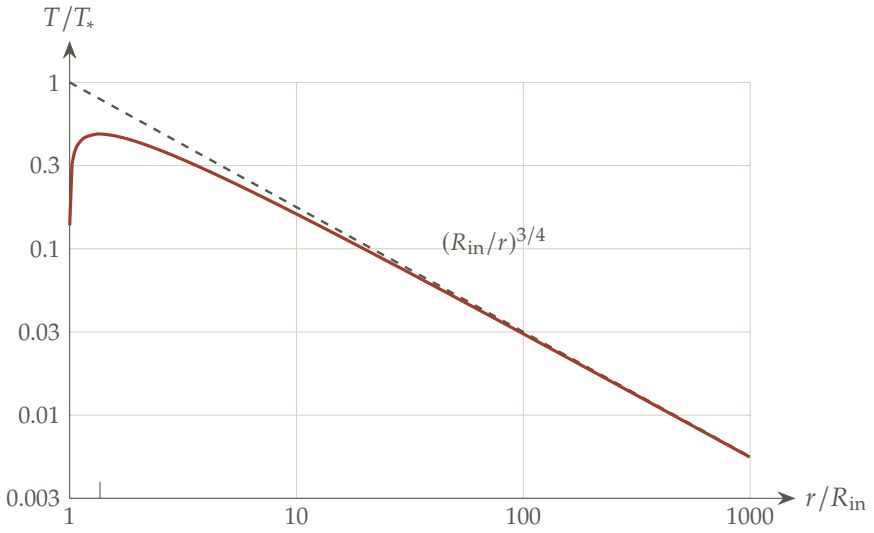


Figure 15.6: Temperature of a thin accretion disk, $T/T_* = (R_{\rm in}/r)^{3/4}(1 - \sqrt{R_{\rm in}/r})^{1/4}$, against radius, with the power law $(R_{\rm in}/r)^{3/4}$ dashed. The tick marks $r = \frac{49}{36}R_{\rm in}$.

The spectrum of the disk. Each face of a ring emits flux $\pi B_\nu(T(r))$ (Eq. (4.20)). Far from the inner edge, $T = T_*(R_{\rm in}/r)^{3/4}$. To add the rings at one frequency ν, use $u = h\nu/k_BT(r) =$

$(h\nu/k_BT_*)(r/R_{\text{in}})^{3/4}$ in place of radius. Then $r = R_{\text{in}}(uk_BT_*/h\nu)^{4/3}$.

$$\begin{aligned} L_\nu &\overset{(1)}{=} \int 2 \cdot 2\pi r \; \pi B_\nu\big(T(r)\big)\, dr \\ &\overset{(2)}{=} 4\pi^2 \frac{2h\nu^3}{c^2} \int \frac{r\, dr}{e^u - 1} \\ &\overset{(3)}{=} 4\pi^2 \frac{2h\nu^3}{c^2} \frac{4}{3} R_{\text{in}}^2 \left(\frac{k_BT_*}{h\nu}\right)^{8/3} \int_0^\infty \frac{u^{5/3}\, du}{e^u - 1} \;\propto\; \nu^{1/3}. \end{aligned} \tag{15.49}$$

(1) Two faces of each ring of area $2\pi r\, dr$.
(2) The Planck function (9.31) with $h\nu/k_BT(r) = u$.
(3) $r\, dr = \frac{4}{3}R_{\text{in}}^2(k_BT_*/h\nu)^{8/3}u^{5/3}du$. For $k_BT(\text{outer edge}) \ll h\nu \ll k_BT_*$ the limits of u are 0 and ∞, and the integral is the number $\Gamma(\frac{8}{3})\zeta(\frac{8}{3}) = 1.93$.

Each ring is nearly a blackbody, but the whole disk contains many temperatures. Between the outer and inner disk frequencies, their sum follows $L_\nu \propto \nu^{1/3}$. It cuts off near the frequency of the hottest ring. In quasars, this broad hump appears in the ultraviolet.

Faster than light. Radio interferometers track bright knots moving along jets from active nuclei (Chapter 5). Some appear to cross the sky faster than light. Let a knot leave the core at $t = 0$ with speed βc at angle θ to our line of sight. After time t, it has moved $\beta ct\cos\theta$ toward us and $\beta ct\sin\theta$ across the sky. Its later light starts closer to us, so the observed interval between the two positions is shorter than t (P6).

$$\begin{aligned} \Delta t_{\text{obs}} &\overset{(1)}{=} t - \frac{\beta ct\cos\theta}{c} = t\,(1 - \beta\cos\theta) \\ \Longrightarrow \quad \beta_{\text{app}} &\overset{(2)}{=} \frac{\beta ct\sin\theta}{c\,\Delta t_{\text{obs}}} = \frac{\beta\sin\theta}{1 - \beta\cos\theta}. \end{aligned} \tag{15.50}$$

(1) The time between emissions minus the head start of the later light.
(2) The apparent speed across the sky, in units of c, is the sideways distance over the observed time.

$$\boxed{\beta_{\text{app}} = \frac{\beta\sin\theta}{1 - \beta\cos\theta}.} \tag{15.51}$$

The apparent speed peaks where the derivative $\beta(\cos\theta - \beta)/(1 - \beta\cos\theta)^2$ vanishes, at $\cos\theta = \beta$. There $\sin\theta = 1/\gamma_L$ and $\beta_{\text{app}} = \beta\gamma_L$, which exceeds 1 for $\beta > 1/\sqrt{2}$ (Figure 15.7). The knot itself still moves below c; its later light has less distance to travel. This is the same arrival-time compression that shortened the synchrotron flash. Since $\beta_{\text{app}} \le \beta\gamma_L < \gamma_L$, an apparent speed gives a lower limit on the jet's Lorentz factor. Features in M87's jet move at about $6c$, implying $\gamma_L \ge 6$ (Problem 15.7.5). Beaming also brightens the approaching jet and dims the receding one (Eq. (15.7)), making many jets look one-sided.

15.6 Active Nuclei Seen From Different Angles

Accretion onto a black hole explains the power of a quasar (Section 15.5). Yet active galaxies look different from one another. Some show both broad and narrow emission lines; others show only narrow lines. Some have vast radio lobes, while others flicker within hours from radio to gamma rays. The difference depends on what surrounds the hole, how large those parts are, and the angle from which we see them.

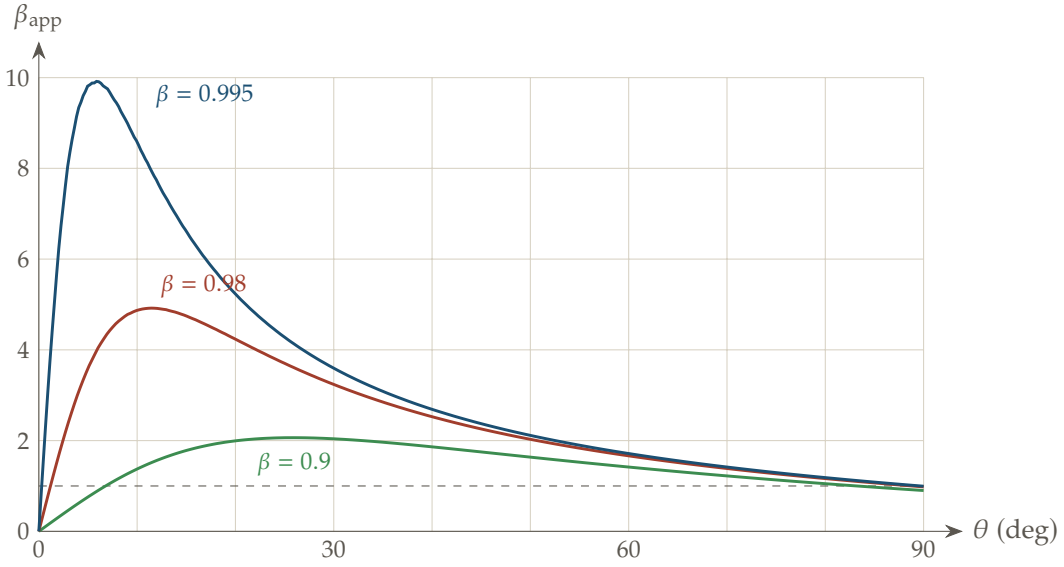


Figure 15.7: Apparent speed across the sky, $\beta_{\text{app}} = \beta \sin\theta/(1-\beta\cos\theta)$, against the angle θ between the motion and the line of sight, for $\beta = 0.9$, 0.98 and 0.995. The dashed line marks $\beta_{\text{app}} = 1$.

Imagine a lamp inside a thick, ring-shaped shade, open at the top and bottom. From above, the bulb is visible. From the side, the shade hides it, though light still reaches a distant wall. If the bulb flashes, the wall brightens later, and the delay tells us how far away it is. In an active nucleus, the hot disk is the bulb, a ring of dust is the shade, and more distant gas catches the light. Delays locate the gas. Its speed weighs the black hole, and the temperature at which dust survives locates the ring. The view changes with the sight line and with the brightness of the engine.

Take a nucleus shining at 10^{37} W. Its broad emission line follows changes in the continuum 20 days later, placing the line-making gas about 20 light-days from the centre. The line width gives a line-of-sight speed spread of 1500 km s^{-1}. Holding gas that fast at that distance requires a black hole of $2.6 \times 10^7\, M_\odot$. Dust inside 32 light-days would rise above 1500 K and evaporate, so the dusty ring begins beyond the fast gas. Figure 15.8 shows the disk and hole at the centre, fast clouds nearby, slower clouds farther out, and jets along the axis. The delay uses the light-travel argument of the radar echo (Eq. (3.40), P6). The line width uses the virial theorem (Eq. (14.30), P2). Dust temperature follows radiative balance (Eq. (6.245), P4). Together, these measurements reveal the structure and mass of the nucleus.

The light of an active nucleus. Starlight fills a normal galaxy (Chapter 14). An active nucleus adds radiation from a hot disk, free electrons, and jets. The disk makes the ultraviolet hump of Eq. (15.49). Hot electrons above it can scatter disk photons into X-rays when $y \gtrsim 1$. Jets and lobes make radio synchrotron light with spectra like Eq. (15.27); inverse Compton scattering can add X-rays and gamma rays. Gas falling toward a black hole of 10^6 to $10^{10}\, M_\odot$ supplies the energy with the efficiency of Eq. (15.45). For the size estimates below, use the 10^{37} W, $2.6 \times 10^7\, M_\odot$ example.

The hole and the disk. For $2.6 \times 10^7\, M_\odot$, the Schwarzschild radius is $2GM/c^2 = 7.8 \times 10^{10}$ m, or four light-minutes. At efficiency 0.1, a luminosity of 10^{37} W requires $\dot{M} = L/(0.1c^2) = 1.1 \times 10^{21}$ kg s^{-1}, or $0.018\, M_\odot$ a year (Eq. (15.45)). Far from the disk's inner edge, Eq. (15.48)

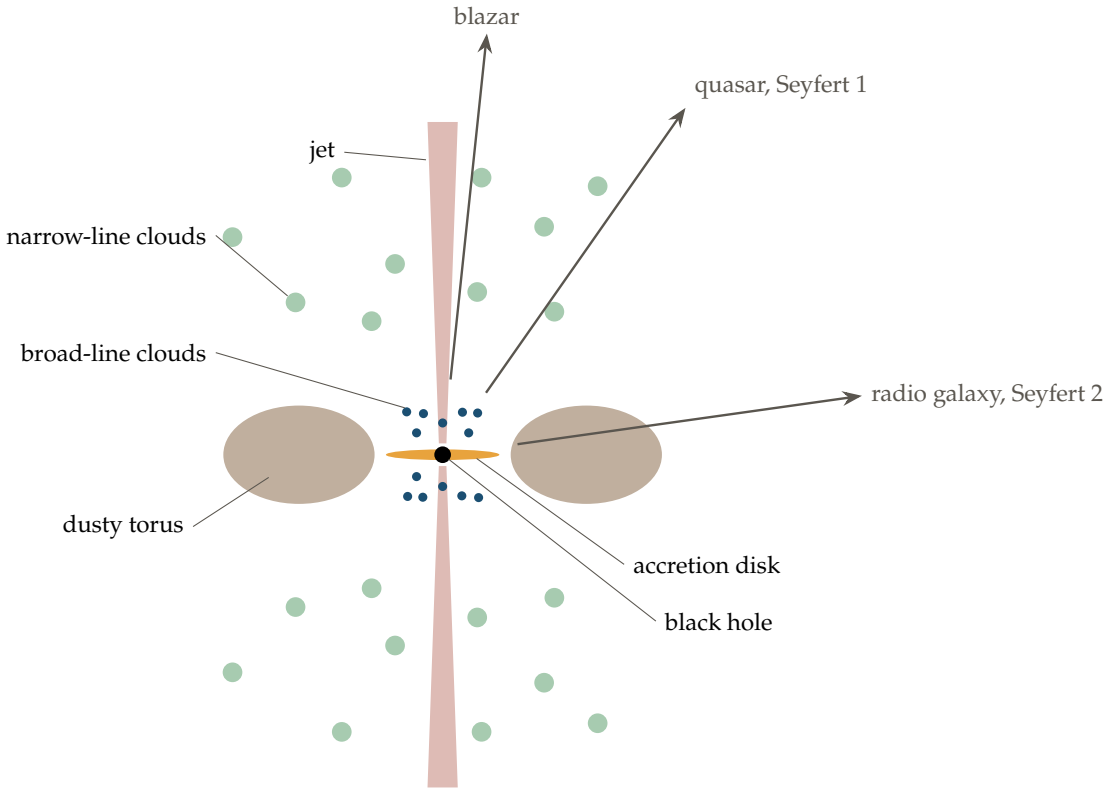


Figure 15.8: Cut through an active nucleus along the axis of its disk, not to scale. Black hole and accretion disk at the centre, broad-line clouds above and below the disk, dusty torus, narrow-line clouds in two cones, and two jets. Arrows are sight lines to observers who see a blazar, a quasar or Seyfert 1, and a radio galaxy or Seyfert 2.

gives the radius at temperature T,

$$r = \left(\frac{3GM\dot{M}}{8\pi\sigma_{\mathrm{SB}} T^4} \right)^{1/3} .$$

The disk reaches 10^4 K at 9.4×10^{12} m, or 0.36 light-days and 120 Schwarzschild radii, where it shines in visible light. Its hottest ring, for $R_{\mathrm{in}} = 6GM/c^2$, reaches 7.8×10^4 K and shines in the ultraviolet. The visible and ultraviolet emitting disk fits within about a light-day.

The broad-line region, weighed. Many active nuclei have emission lines thousands of km s^{-1} wide. The Doppler shift turns that width into a spread of line-of-sight speeds σ (Eq. (3.25)). Thermal motion at 10^4 K gives only $\sqrt{k_B T/m_{\mathrm{H}}} = 9$ km s^{-1} (Eq. (8.30)). The much larger observed width comes from clouds orbiting the hole. Let their masses, speeds, and distances from a hole of mass M be m_i, v_i, and r_i (P2).

$$\begin{aligned} \sum_i m_i \left\langle v_i^2 \right\rangle &\overset{(1)}{=} \sum_i GMm_i \left\langle \frac{1}{r_i} \right\rangle \\ \Longrightarrow \qquad 3\sigma^2 &\overset{(2)}{=} GM \left\langle \frac{1}{r} \right\rangle \overset{(3)}{=} \frac{GM}{r}. \end{aligned} \tag{15.52}$$

(1) The virial theorem for a bound system, $2\langle K \rangle + \langle U \rangle = 0$, Eq. (10.12), with $K = \sum_i \frac{1}{2} m_i v_i^2$ and $U = -\sum_i GMm_i/r_i$. The hole is so heavy that it hardly moves, and the pull of the clouds on one another is small next to the pull of the hole.

(2) Divide by $\sum_i m_i$. For motions with no favoured direction the mean square speed is three times its part along the line of sight, $\langle v^2 \rangle = 3\sigma^2$, as for galaxies in a cluster (Section 14.4).
(3) The clouds lie at about one distance r.

$$\boxed{M = \frac{3\sigma^2 r}{G}}. \tag{15.53}$$

Only cloud speed and distance enter. The clouds weigh the hole as long as they are gravitationally bound to it.

The size from the echo. These clouds lie only a few hundredths of a parsec from the hole, too close to resolve in a distant galaxy. A time delay measures their distance instead. When the disk brightens, the line-making clouds brighten after its light reaches them. Let the disk flash at $t = 0$. A cloud at distance r and angle θ from our line of sight sends us the delayed light.

$$\begin{aligned} \Delta t(\theta) &\overset{(1)}{=} \frac{r}{c} - \frac{r\cos\theta}{c} = \frac{r}{c}(1 - \cos\theta), \\ \langle \Delta t \rangle &\overset{(2)}{=} \frac{r}{c} \int_0^\pi (1 - \cos\theta) \frac{2\pi \sin\theta \, d\theta}{4\pi} \overset{(3)}{=} \frac{r}{c}. \end{aligned} \tag{15.54}$$

(1) The flash first travels r out to the cloud. The cloud is $r\cos\theta$ nearer to us than the hole, so its light then gains $r\cos\theta/c$ on the direct light.
(2) Clouds spread evenly over a shell. A fraction $2\pi \sin\theta \, d\theta/4\pi$ of them lies between θ and $\theta + d\theta$.
(3) $\int_0^\pi \sin\theta \, d\theta = 2$ and $\int_0^\pi \cos\theta \sin\theta \, d\theta = 0$.

The delay ranges from 0 for a cloud in front of the hole to $2r/c$ for one behind it. For clouds spread around the hole, the mean is r/c. Measuring the lag Δt between continuum and line light is **reverberation mapping**. Put $r = c\,\Delta t$ into Eq. (15.53),

$$\boxed{M = \frac{3\sigma^2 c\,\Delta t}{G}}. \tag{15.55}$$

For the example nucleus, a lag of $\Delta t = 20$ days gives $r = c\,\Delta t = 5.2 \times 10^{14}\,\mathrm{m} = 0.017\,\mathrm{pc}$. With $\sigma = 1500\,\mathrm{km\,s^{-1}}$, the mass is $M = 3 \times (1.5 \times 10^6)^2 \times 5.2 \times 10^{14}/(6.67 \times 10^{-11})\,\mathrm{kg} = 5.2 \times 10^{37}\,\mathrm{kg} = 2.6 \times 10^7\,M_\odot$. A circular orbit at that radius has speed $\sqrt{GM/r} = 2600\,\mathrm{km\,s^{-1}} = \sqrt{3}\,\sigma$, as expected. Equation (11.27) puts the luminosity at 3% of the Eddington limit. The main uncertainty lies in the factor 3. The lag measures a mean radius, while the virial relation uses a mean inverse radius. Flattened cloud orbits would also change the factor by order unity.

The dusty torus. Farther from the hole, dust absorbs light from the nucleus. A grain intercepts flux $L/(4\pi r^2)$ across its cross-section and radiates from its whole surface, as a planet does in the balance before Eq. (6.245). Dust survives only where its temperature stays below about 1500 K. The inner edge therefore lies where absorbed and radiated power balance at that temperature (P4).

$$\begin{aligned} T &\overset{(1)}{=} \left(\frac{L}{16\pi\sigma_{\mathrm{SB}} r^2} \right)^{1/4} \\ \Longrightarrow \quad r &\overset{(2)}{=} \sqrt{\frac{L}{16\pi\sigma_{\mathrm{SB}} T^4}}. \end{aligned} \tag{15.56}$$

(1) The planet's balance with the grain in place of the planet and the albedo set to zero, since dust absorbs nearly all the ultraviolet and visible light that falls on it. The size of the grain drops out, as the planet's radius did.

(2) Raise to the fourth power and solve for r.

$$\boxed{r_{\mathrm{dust}} = \sqrt{\frac{L}{16\pi\sigma_{\mathrm{SB}}\,(1500\,\mathrm{K})^4}} \;\propto\; L^{1/2}.} \tag{15.57}$$

For $L = 10^{37}\,\mathrm{W}$, the denominator is $16\pi\sigma_{\mathrm{SB}}(1500\,\mathrm{K})^4 = 1.44 \times 10^7\,\mathrm{W\,m^{-2}}$. It gives $r_{\mathrm{dust}} = 8.3 \times 10^{14}\,\mathrm{m} = 0.027\,\mathrm{pc}$, or 32 light-days, outside the broad-line clouds at 20 light-days. For a $10^{40}\,\mathrm{W}$ quasar, the blackbody estimate is 0.85 pc. Small grains radiate infrared light less efficiently than blackbodies and run hotter, placing the real inner edge a few times farther out. Dust absorbs ultraviolet light and reradiates it in the infrared, near 2 μm at the hot inner edge (Eq. (4.19)). A thick ring of this dust, the **torus**, hides the centre when seen from the side. The central flux also sets the ionisation of the broad-line clouds, so their distance should grow roughly as $L^{1/2}$. Measured lags show about that trend.

The narrow-line region. Beyond the torus, gas in two illuminated cones extends hundreds of parsecs and makes lines only a few hundred $\mathrm{km\,s^{-1}}$ wide. The hole's pull is weak at those distances. At 100 pc, a $2.6 \times 10^7\,M_\odot$ hole gives an orbital speed of only $\sqrt{GM/r} = 34\,\mathrm{km\,s^{-1}}$. Stars in the central galaxy and outflowing gas set the larger speeds. The gas is thin, with about 10^9 to 10^{10} particles per cubic metre. Equation (8.39) compares radiative decay with collisions. Below the critical density of Section 8.4.1, an excited ion can emit even through a slow, forbidden transition before another collision reaches it. Thus the narrow-line region shows forbidden lines, especially the line of doubly ionised oxygen at 500.7 nm. Broad-line clouds are much denser, near 10^{15} to $10^{16}\,\mathrm{m^{-3}}$. Collisions empty these levels before they radiate, leaving allowed lines. The narrow-line region sits outside the torus and remains visible from every angle.

Jets and lobes. Some nuclei launch jets along the disk axis at nearly the speed of light. Their knots can show apparent speeds of several c (Eq. (15.51)). Synchrotron light makes the jets bright in radio (Eq. (15.27)); the lobes they feed can reach hundreds of kiloparsecs beyond the galaxy. The two **Fanaroff–Riley** classes describe where a radio source is brightest. In class I, jets brighten near the core and spread into fading plumes. M84 (3C 272.1) in Virgo is one example, about 14 kpc across on the sky. In class II, the jets stay narrow and end in bright hot spots where they meet the surrounding gas. Class II sources are, on average, more powerful radio emitters.

One engine, many names. Table 15.1 shows how the view changes with the sight line. Through the torus opening, we see the hot disk and both broad and narrow lines. From the side, the torus hides the disk and broad-line clouds; narrow lines and dust infrared remain. When a jet points nearly toward us, beaming brightens and compresses its light in time (Eq. (15.7)). Such a source is a blazar. Its synchrotron continuum can overwhelm the lines, while inverse Compton scattering supplies gamma rays (Eq. (15.31)).

Luminosity and viewing angle account for many of the names in the table. Jet strength matters too. Most nuclei lack strong jets. A nucleus with one can appear as a radio galaxy from the side, a radio-loud quasar from closer to its axis, or a blazar when the jet points toward us. Figure 15.9 sorts the names by asking three questions in turn. The first is whether the nucleus has a strong jet. Without one, the luminosity separates a Seyfert from a quasar, and for a Seyfert the sight line separates type 1 from type 2. With a jet, the angle between the jet and our sight line decides the name.

A Seyfert 2 should hide the broad-line clouds behind its torus. Some of their light can still escape through the opening and scatter toward us from free electrons above the torus. Scattering selects the component of the electron's motion across our line of sight (Eq. (15.1)), making the

Kind	Sight line	What is seen
Seyfert 1	through the opening of the torus	broad and narrow lines, a bright nucleus in a visible galaxy
Seyfert 2	through the torus	narrow lines only, infrared from warm dust, broad lines in polarised light
Quasar	through the opening, nucleus far brighter	broad and narrow lines, the nucleus outshines its galaxy
Radio galaxy	from the side, strong jets	narrow lines, jets and lobes in the radio, Fanaroff–Riley I or II
Blazar	down a strong jet	featureless continuum from radio to gamma rays, changes within hours to days

Table 15.1: Kinds of active galaxies by sight line and by what is seen.

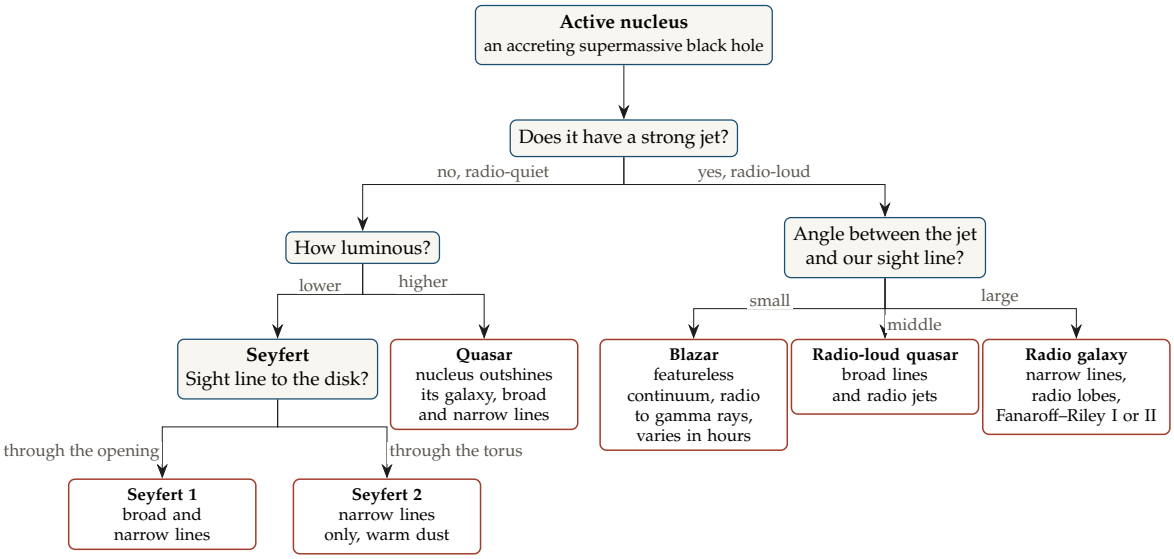


Figure 15.9: The kinds of active nucleus as a tree, sorted by jet strength, then by luminosity or by the angle between the jet and our sight line, and for a Seyfert by whether the sight line passes through the opening of the torus. Each leaf gives what is seen.

scattered light linearly polarised. Broad lines appear in this polarised light in some Seyfert 2s, including NGC 1068, even though they are hidden in direct light.

Size from variability. A source that changes within a day cannot be much larger than a light-day across (Section 15.5). For a distant jet, cosmic expansion lengthens the observed interval, while motion toward us shortens it. Let a blob of radius R move at speed βc at angle θ in a galaxy at redshift z. A change takes at least $\Delta t' \approx R/c$ to cross the blob in its own frame (P6). Call the interval in the host galaxy Δt_{lab} and the interval at our telescope Δt_{obs}.

$$\frac{\Delta t_{\text{obs}}}{1+z} \overset{(1)}{=} (1 - \beta\cos\theta)\,\Delta t_{\text{lab}} \overset{(2)}{=} \gamma_L(1 - \beta\cos\theta)\,\Delta t' \overset{(3)}{=} \frac{\Delta t'}{\mathcal{D}} \overset{(4)}{\gtrsim} \frac{R}{c\,\mathcal{D}}. \tag{15.58}$$

(1) The expansion stretches every interval by $1 + z$ (the chain before Eq. (16.4)). In the host's frame the later light starts $\beta c\,\Delta t_{\text{lab}}\cos\theta$ nearer to us, as in the chain before Eq. (15.51).
(2) The time row of the boost (15.4) for two events at the blob, $x' = 0$.
(3) Define the **Doppler factor** $\mathcal{D}$ below.

(4) $\Delta t' \gtrsim R/c$.

$$\boxed{\mathcal{D} = \frac{1}{\gamma_L(1-\beta\cos\theta)}, \qquad R \lesssim \frac{c\,\Delta t_{\rm obs}\,\mathcal{D}}{1+z}.} \tag{15.59}$$

The frequency transformation offers a check. Equation (15.30) gives $\nu' = \gamma_L(1-\beta\cos\theta)\,\nu$, or $\nu = \mathcal{D}\,\nu'$. The same factor raises frequencies and shortens time intervals. For a source at rest, $\mathcal{D} = 1$ and we recover the ordinary light-crossing bound. At $\cos\theta = \beta$, $\mathcal{D} = \gamma_L$; directly along the motion, $\mathcal{D} = \gamma_L(1+\beta) \approx 2\gamma_L$. A blazar at $z = 0.5$ varying within an hour, with $\mathcal{D} = 10$, has $R \lesssim 3.0\times10^{8}\times3600\times10/1.5\,\mathrm{m} = 7.2\times10^{12}\,\mathrm{m} = 48\,\mathrm{AU}$. Without beaming, the bound would be 4.8 AU, smaller than the 20 AU Schwarzschild radius of a $10^9\,M_\odot$ hole. The rapid changes therefore support a beamed jet.

Fresh gas can change the accretion rate, and an unstable disk can flicker. In a jet, even a small change of speed or direction changes $\mathcal{D}$. Because the observed brightness depends on a high power of $\mathcal{D}$, this can make the jet vary strongly.

Pairs of black holes. When galaxies merge, their central black holes can form a bound pair. Interactions with surrounding stars and gas drain orbital energy and draw the holes inward. At smaller separations, gravitational waves carry off energy (Section 13.6). Consider a circular pair of total mass $M = 10^9\,M_\odot$ separated by 0.01 pc.

$$f_{\rm GW} \overset{(1)}{=} 2\times\frac{1}{2\pi}\sqrt{\frac{GM}{(0.01\,\mathrm{pc})^3}} \overset{(2)}{=} \frac{1}{\pi}\sqrt{\frac{1.33\times10^{29}}{2.94\times10^{43}}}\ \mathrm{s^{-1}}$$
$$\overset{(3)}{=} \frac{6.72\times10^{-8}}{\pi}\ \mathrm{s^{-1}} = 2.1\times10^{-8}\,\mathrm{Hz}. \tag{15.60}$$

(1) The wave frequency $f_{\rm GW}$ is twice the orbital frequency, Eq. (13.15) (the f there, with a subscript here because f is the photon occupation number of Eq. (15.42)), and the orbital angular speed is $\sqrt{GM/(\text{separation})^3}$ by Kepler's third law, Eq. (6.50).
(2) $GM = 6.674\times10^{-11}\times1.989\times10^{39} = 1.33\times10^{29}\,\mathrm{m^3\,s^{-2}}$, and $(0.01\,\mathrm{pc})^3 = (3.086\times10^{14}\,\mathrm{m})^3 = 2.94\times10^{43}\,\mathrm{m^3}$.
(3) $\sqrt{4.52\times10^{-15}} = 6.72\times10^{-8}$.

The orbit takes 3.0 years, so the wave period is 1.5 years and its frequency is 21 nanohertz. For two equal holes, Eq. (13.21) gives a remaining time of $(5/256)\,c^5(0.01\,\mathrm{pc})^4/(G^3 m_1 m_2 M)$. Using $c^5 = 2.42\times10^{42}$, $(0.01\,\mathrm{pc})^4 = 9.07\times10^{57}$, $G^3 = 2.97\times10^{-31}$, and $m_1 m_2 M = 1.97\times10^{117}$ in SI units gives 7.3×10^{11} s, about 2×10^4 years. Many pairs can radiate at such periods at once, adding to a background of low-frequency gravitational waves. Pulsar timing arrays search for waves with periods of years (Section 13.6.6). Several arrays reported evidence for a background in 2023; its sources and their numbers remain under study.

15.7 Problem Set: Radiation from Hot and Fast Particles

15.7.1 Problem 1: X-rays From a Cluster of Galaxies

Problem Statement: Ionised hydrogen fills a galaxy cluster out to radius $1\,\mathrm{Mpc} = 3.09\times10^{22}$ m. Its density is $n_e = n_i = 10^3\,\mathrm{m^{-3}}$, its temperature is $T = 7\times10^7$ K, and $\bar{g} = 1$. Find (a) the radiated power per unit volume, (b) the total X-ray luminosity, (c) the cooling time compared with the universe's age of 1.4×10^{10} years, and (d) the photon energy where the spectrum falls to $1/e$ of its low-frequency value.

Essential Ideas:

- The cooling rate $1.4 \times 10^{-40}\, n_e n_i \sqrt{T}\ \mathrm{W\,m^{-3}}$, Eq. (15.16).
- The thermal energy $3nk_BT$ per unit volume, and the spectrum $\propto e^{-h\nu/k_BT}$ of Eq. (15.15).

Detailed Solutions:

(a) **Power per volume**. $1.4 \times 10^{-40} \times 10^6 \times \sqrt{7 \times 10^7}$.

$$\boxed{1.2 \times 10^{-30}\,\mathrm{W\,m^{-3}}.}$$

(b) **Luminosity**. The volume is $\frac{4}{3}\pi(3.09 \times 10^{22})^3 = 1.2 \times 10^{68}\,\mathrm{m^3}$.

$$\boxed{L_X \simeq 1.5 \times 10^{38}\,\mathrm{W}.}$$

(c) **Cooling time**. $3nk_BT = 2.9 \times 10^{-12}\,\mathrm{J\,m^{-3}}$, divided by the power per volume.

$$\boxed{t_{\mathrm{cool}} \simeq 2.4 \times 10^{18}\,\mathrm{s} \simeq 8 \times 10^{10}\ \mathrm{years}.}$$

The cooling time exceeds the age of the universe by more than a factor of five. Most of the gas can still hold the heat it gained when the cluster formed.

(d) **Turn-down energy**. The spectrum falls by e at $h\nu = k_BT$.

$$\boxed{h\nu = 6.0\,\mathrm{keV}, \qquad \nu = 1.5 \times 10^{18}\,\mathrm{Hz}.}$$

The energy where the X-ray spectrum turns down measures the gas temperature. Together with Eq. (10.2), that temperature helps determine the cluster mass.

15.7.2 Problem 2: Electrons in a Radio Lobe

Problem Statement: A radio lobe has magnetic field $B = 1\,\mathrm{nT}$. At 1.4 GHz its spectrum follows $j_\nu \propto \nu^{-0.75}$. Find (a) the Lorentz factor of electrons radiating at that frequency, using $\nu = \frac{3}{2}\gamma_L^2 \nu_g$, (b) their synchrotron lifetime, (c) the electron index p, and (d) the effect of scattering the microwave background, whose energy density is $u_{\mathrm{rad}} = 4.2 \times 10^{-14}\,\mathrm{J\,m^{-3}}$.

Essential Ideas:

- $\nu_g = 2.80 \times 10^{10}\,\mathrm{Hz}\,(B/1\,\mathrm{T})$, Eq. (15.20).
- $t_{\mathrm{sync}} = 3m_e c/(4\sigma_T \gamma_L u_B)$ with $u_B = B^2/2\mu_0$, and the inverse Compton power has u_{rad} in place of u_B (Eq. (15.33)).
- $\alpha = (p-1)/2$, Eq. (15.27).

Detailed Solutions:

(a) **Lorentz factor**. $\nu_g = 28.0\,\mathrm{Hz}$, so $\gamma_L = \sqrt{1.4 \times 10^9/(1.5 \times 28.0)}$.

$$\boxed{\gamma_L \simeq 5.8 \times 10^3, \qquad \gamma_L m_e c^2 \simeq 3\,\mathrm{GeV}.}$$

(b) **Lifetime**. $u_B = (10^{-9})^2/(2 \times 1.26 \times 10^{-6}) = 4.0 \times 10^{-13}\,\mathrm{J\,m^{-3}}$.

$$\boxed{t_{\mathrm{sync}} = \frac{3 \times 9.11 \times 10^{-31} \times 3.00 \times 10^8}{4 \times 6.65 \times 10^{-29} \times 5.8 \times 10^3 \times 4.0 \times 10^{-13}}\,\mathrm{s} \simeq 1.3 \times 10^{15}\,\mathrm{s} \simeq 4 \times 10^7\ \mathrm{years}.}$$

(c) **Electron index**. $0.75 = (p-1)/2$.

$$\boxed{p = 2.5.}$$

(d) **Scattering the background**. The two losses add, so the lifetime is $t_{\rm sync}$ divided by $1 + u_{\rm rad}/u_B = 1 + 0.105$.

$$\boxed{t \simeq 3.8 \times 10^7 \text{ years.}}$$

Synchrotron cooling dominates at 1 nT. Below 0.32 nT, scattering the background would dominate and produce more X-rays relative to the lobe's radio light.

15.7.3 Problem 3: Photons Kicked and Thrown

Problem Statement: (a) A 511 keV gamma ray scatters through 90° from a stationary electron. Find its outgoing energy. (b) Electrons with $\gamma_L = 10^4$ scatter microwave-background photons of mean energy 6.3×10^{-4} eV. Find the mean outgoing energy and the maximum for a photon that began at the mean energy. (c) Check the Thomson limit for part (b). (d) For the cluster of Problem 15.7.1, find y across 1 Mpc and the low-frequency temperature dip of the background, taking $T_{\rm CMB} = 2.725$ K.

Essential Ideas:

- $h\nu_1 = h\nu/[1 + (h\nu/m_e c^2)(1 - \cos\psi)]$, Eq. (15.29).
- Mean gain $\frac{4}{3}\gamma_L^2$ and largest gain $4\gamma_L^2$ (Eqs. (15.31) and (15.33)), valid while $\gamma_L h\nu \ll m_e c^2$.
- $y = (k_B T/m_e c^2)\, n_e \sigma_T L$ and $\Delta T/T = -2y$, Eqs. (15.35) and (15.43).

Detailed Solutions:

(a) **Compton shift**. $h\nu/m_e c^2 = 1$ and $1 - \cos 90° = 1$.

$$\boxed{h\nu_1 = \frac{511}{2}\,\text{keV} \simeq 256\,\text{keV}.}$$

(b) **Inverse Compton**. $\frac{4}{3} \times 10^8 \times 6.3 \times 10^{-4}$ eV, and $4 \times 10^8 \times 6.3 \times 10^{-4}$ eV.

$$\boxed{\langle h\nu_1 \rangle \simeq 85\,\text{keV}, \qquad h\nu_{1,\rm max} \simeq 250\,\text{keV}.}$$

(c) **Thomson limit**. In the electron's frame the photon has about $\gamma_L h\nu = 6.3$ eV, far below 511 keV.

$$\boxed{\text{Yes, by a factor of about } 10^5.}$$

(d) **Sunyaev–Zel'dovich dip**. $\tau = 10^3 \times 6.65 \times 10^{-29} \times 3.09 \times 10^{22} = 2.1 \times 10^{-3}$, and $k_B T/m_e c^2 = 6.0/511 = 0.0118$.

$$\boxed{y \simeq 2.4 \times 10^{-5}, \qquad \Delta T = -2y T_{\rm CMB} \simeq -130\,\mu\text{K}.}$$

The gas produces both the X-rays of Problem 15.7.1, proportional to n_e^2, and this microwave dip, proportional to n_e. Together the signals constrain gas density and depth. The angular size then helps determine the cluster distance.

15.7.4 Problem 4: Feeding Compact Stars

Problem Statement: A neutron star has mass $1.4\,M_\odot$, radius 12 km, and accretes gas. Find (a) its efficiency compared with hydrogen burning, (b) its luminosity at $\dot{M} = 10^{-9}\,M_\odot\,\text{yr}^{-1}$ as a fraction of the Eddington limit, and (c) its temperature if the whole surface radiates as a blackbody. (d) A $10\,M_\odot$ black hole accretes at $10^{-8}\,M_\odot\,\text{yr}^{-1}$ through a thin disk with $R_{\rm in} = 6GM/c^2$. Find the temperature of its hottest ring.

Essential Ideas:

- $\eta = GM/(Rc^2)$ and $L = GM\dot{M}/R$, Eq. (15.45). Hydrogen burning gives 0.007.
- $L_{\rm Edd} = 1.26 \times 10^{31}\,{\rm W}\,(M/M_\odot)$, Eq. (11.27).
- $L = 4\pi R^2 \sigma_{\rm SB} T^4$, and the hottest ring of a disk at $0.488\,T_*$ (Eq. (15.48)).

Detailed Solutions:

(a) **Efficiency**. $6.67 \times 10^{-11} \times 2.79 \times 10^{30}/(1.2 \times 10^4 \times 9.0 \times 10^{16})$.

$$\boxed{\eta \simeq 0.17, \quad \text{about 25 times hydrogen burning.}}$$

(b) **Luminosity**. $\dot{M} = 6.3 \times 10^{13}\,{\rm kg\,s^{-1}}$, so $L = 0.17\,\dot{M}c^2$.

$$\boxed{L \simeq 9.8 \times 10^{29}\,{\rm W} \simeq 2.5 \times 10^3\,L_\odot \simeq 0.06\,L_{\rm Edd}.}$$

(c) **Surface temperature**. $T = [L/(4\pi R^2 \sigma_{\rm SB})]^{1/4}$.

$$\boxed{T \simeq 1.0 \times 10^7\,{\rm K}, \qquad k_B T \simeq 0.85\,{\rm keV}.}$$

The light comes out in X-rays, which is how accreting neutron stars in binaries were found.

(d) **Disk of a black hole**. $R_{\rm in} = 89\,{\rm km}$ and $T_* = (3GM\dot{M}/8\pi\sigma_{\rm SB}R_{\rm in}^3)^{1/4} = 7.1 \times 10^6\,{\rm K}$.

$$\boxed{T_{\rm max} = 0.488\,T_* \simeq 3.5 \times 10^6\,{\rm K}, \qquad k_B T \simeq 0.3\,{\rm keV}.}$$

Both the stellar black hole and the neutron star glow in soft X-rays at these inflow rates. The much larger disks around quasar black holes are cooler and glow in the ultraviolet.

15.7.5 Problem 5: The Jet of M87

Problem Statement: Knots in M87's jet appear to move across the sky at about $6c$. Find (a) the smallest Lorentz factor and true speed consistent with this apparent speed, (b) the largest angle the jet can make with our line of sight, and (c) the two viewing angles that give $\beta_{\rm app} = 6$ if $\beta = 0.99$.

Essential Ideas:

- $\beta_{\rm app} = \beta \sin\theta/(1 - \beta\cos\theta)$, at most $\beta\gamma_L$, Eq. (15.51).
- As $\beta \to 1$, $\beta_{\rm app} \to \sin\theta/(1 - \cos\theta) = \cot(\theta/2)$.

Detailed Solutions:

(a) **Smallest Lorentz factor**. $\beta\gamma_L \geq 6$, and $\beta^2\gamma_L^2 = \gamma_L^2 - 1$, so $\gamma_L \geq \sqrt{37}$.

$$\boxed{\gamma_L \geq 6.1, \qquad \beta \geq 6/\sqrt{37} = 0.986.}$$

(b) **Largest angle**. Even at $\beta \to 1$, $\cot(\theta/2) \geq 6$ requires $\theta \leq 2\,{\rm arccot}\,6$.

$$\boxed{\theta \leq 19^\circ.}$$

(c) **Angles for** $\beta = 0.99$. Solving $0.99\sin\theta/(1 - 0.99\cos\theta) = 6$,

$$\boxed{\theta \simeq 4.6^\circ \text{ or } \theta \simeq 14.4^\circ.}$$

Below the peak $\beta\gamma_L = 7.0$, the same apparent speed occurs at two angles, one on either side of $\cos\theta = \beta$. The motion alone limits γ_L from below and θ from above. The brightness ratio of the approaching and receding jets helps narrow them further.

15.7.6 Problem 6: Weighing a Black Hole by its Echo

Problem Statement: In a Seyfert 1 nucleus, broad hydrogen lines follow continuum changes 12 days later. Their line-of-sight speed spread is $\sigma = 2000\,\mathrm{km\,s^{-1}}$, and the nucleus shines at 4×10^{37} W. Find (a) the cloud distance, (b) the black-hole mass, (c) the Eddington fraction, and (d) the inner edge of dust that survives below 1500 K. Compare the dust edge with the cloud distance.

Essential Ideas:

- The mean echo delay of a shell of clouds is r/c, and the virial theorem gives $M = 3\sigma^2 c\,\Delta t/G$, Eq. (15.55).
- $L_{\rm Edd} = 1.26\times10^{31}\,\mathrm{W}\,(M/M_\odot)$, Eq. (11.27).
- $r_{\rm dust} = \sqrt{L/(16\pi\sigma_{\rm SB}T^4)}$, Eq. (15.57).

Detailed Solutions:

(a) **Distance**. $r = c\,\Delta t = 3.00\times10^8\times12\times86400\,\mathrm{m}$.

$$\boxed{r \simeq 3.1\times10^{14}\,\mathrm{m} \simeq 0.010\,\mathrm{pc}.}$$

(b) **Mass**. $3\sigma^2 = 1.2\times10^{13}\,\mathrm{m^2\,s^{-2}}$, so $M = 1.2\times10^{13}\times3.1\times10^{14}/(6.67\times10^{-11})\,\mathrm{kg} = 5.6\times10^{37}\,\mathrm{kg}$.

$$\boxed{M \simeq 2.8\times10^7\,M_\odot.}$$

(c) **Eddington fraction**. $L_{\rm Edd} = 1.26\times10^{31}\times2.8\times10^7\,\mathrm{W} = 3.5\times10^{38}\,\mathrm{W}$.

$$\boxed{L/L_{\rm Edd} \simeq 0.11.}$$

(d) **Inner edge of the dust**. With $16\pi\sigma_{\rm SB}(1500\,\mathrm{K})^4 = 1.44\times10^7\,\mathrm{W\,m^{-2}}$,

$$\boxed{r_{\rm dust} = \sqrt{\frac{4\times10^{37}}{1.44\times10^7}}\,\mathrm{m} \simeq 1.7\times10^{15}\,\mathrm{m} \simeq 64\ \text{light-days}.}$$

The dust edge is about 5 times farther out than the broad-line clouds. From the side, the torus can therefore hide those clouds while the more distant narrow-line gas remains visible.

Chapter 16

The Expanding Universe

Distant galaxies are moving away from us, and the farther they are, the faster they go. Run the film backwards and everything was once closer, denser and hotter. The whole history of the universe on the largest scales follows from the same principles that built a star. Energy conservation for a shell of matter and the gravity of a sphere (P1, P2) give the expansion. Counting states (P3) gives the light left over from the hot beginning. Balance of rates against the rate of expansion (P4, P6) decides when atoms formed and how much helium was made. Gravity working against expansion decides how galaxies grew from tiny ripples.

General relativity, which the reader knows, gives the same equations for a uniform universe. The Newtonian route is used here because it shows where each term comes from with the least machinery.

16.1 Hubble's Law and the Scale Factor

Radial velocities from the Doppler shift were derived in Chapter 3 (Eq. (3.25)). Measured for distant galaxies, these velocities are almost all away from us. A single velocity for each galaxy says nothing about what the whole universe is doing.

Think of raisins in a rising loaf of bread. As the dough swells, every raisin moves away from every other raisin. A raisin twice as far away moves away twice as fast, and every raisin sees the same pattern, so none of them is the centre. The universe expands like the dough. All distances between galaxies grow by the same factor at the same time. A uniform stretching makes velocity proportional to distance, which is Hubble's law, and it stretches light too, which is the cosmological redshift.

In numbers, today's expansion rate is $H_0 = 67.4\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. A galaxy 100 Mpc away recedes at $6740\,\mathrm{km\,s^{-1}}$. The time $1/H_0$ is 14.51 billion years, the age the universe would have if it had always expanded at today's speed. The picture is a grid drawn on a balloon. Inflating the balloon carries every grid point away from every other, while each point keeps its place on the grid. The same uniform stretching appears in the dough, in an expanding gas and in any self-similar flow. It gives Hubble's law and the redshift in terms of one function of time, the scale factor.

The scale factor. Give every galaxy fixed coordinates $\vec{x}$ on the stretching grid, its **comoving** position. The true distance between two galaxies is then $a(t)$ times their comoving separation, where $a(t)$ is the **scale factor**, set to 1 today. A uniform universe has only this one function of time.

Hubble's law. For a galaxy at true distance $d = a\,x$,

$$\frac{\mathrm{d}d}{\mathrm{d}t} \overset{(1)}{=} \frac{\mathrm{d}a}{\mathrm{d}t}\,x \overset{(2)}{=} \frac{\dot{a}}{a}\,d. \tag{16.1}$$

(1) The comoving separation x does not change.
(2) $x = d/a$.

$$\boxed{v = H\,d, \qquad H(t) \equiv \frac{\dot{a}}{a}.} \tag{16.2}$$

The velocity is proportional to the distance, with the same H for every galaxy. Every observer finds the same law, so it picks out no centre.

Redshift as stretching. Light travels on the grid at c in true distance, so in comoving distance it covers $c\,dt/a(t)$ in a time dt. Two crests leave a galaxy a time Δt_e apart, at scale factor a_e, and arrive a time Δt_0 apart today. Both crests cover the same comoving distance, so

$$\int_{t_\mathrm{e}}^{t_0} \frac{c\,dt}{a} = \int_{t_\mathrm{e}+\Delta t_\mathrm{e}}^{t_0+\Delta t_0} \frac{c\,dt}{a} \quad\Longrightarrow\quad \frac{\Delta t_\mathrm{e}}{a_\mathrm{e}} \overset{(1)}{=} \frac{\Delta t_0}{a_0}. \tag{16.3}$$

(1) Subtract the common part of the two integrals. What is left are two short pieces, of length Δt_e at the start and Δt_0 at the end, during which a hardly changes.

The wavelength is c times the time between crests, so it grows in proportion to the scale factor,

$$\boxed{1 + z \equiv \frac{\lambda_\mathrm{obs}}{\lambda_\mathrm{emit}} = \frac{a_0}{a_\mathrm{e}} = \frac{1}{a_\mathrm{e}}.} \tag{16.4}$$

The redshift measures how much the universe has grown since the light left, not a velocity. For small distances it reduces to the Doppler form $z \approx v/c$ with $v = Hd$.

16.2 The Friedmann Equations

Hubble's law describes the expansion today. Chapter 6 and Section 10.5 followed a single body or a single shell moving under gravity. The expansion is not free. Every galaxy pulls on every other, so gravity should slow the expansion, and how $a(t)$ changes depends on what the universe contains.

Throw a ball straight up. If it goes fast enough it escapes the Earth, and if not it falls back. The answer depends only on its energy, kinetic against the depth of the gravitational well. Take any sphere in a uniform universe. By P2, a shell on its surface feels only the mass inside, so it moves exactly like the thrown ball, with energy conserved (P1). Writing that energy in terms of the scale factor gives the Friedmann equation, the equation of motion of the whole universe.

The ball in numbers. The escape speed from the Earth is $11.2\,\mathrm{km\,s^{-1}}$. For the universe the same balance defines a critical density, $3H_0^2/(8\pi G) = 8.5 \times 10^{-27}\,\mathrm{kg\,m^{-3}}$, about five hydrogen atoms per cubic metre. The picture is a sphere carved out of the universe, its surface shell flying outward and slowing under the pull of what it contains. The same equation comes out of general relativity for any uniform universe, with the energy constant reinterpreted as the curvature of space. It gives the expansion rate, how each kind of content dilutes, whether the expansion speeds up, and the age.

The energy of a shell. Take a sphere of comoving radius x, true radius $r = a\,x$ and density ρ. The mass inside, $M = \frac{4}{3}\pi\rho a^3 x^3$, stays constant as the shell moves with the expansion. Energy conservation per unit mass of the shell (P1, with P2 for the gravity) reads

$$\begin{aligned} \frac{1}{2}\dot{r}^2 - \frac{GM}{r} &= \text{constant} \\ \implies \quad \frac{1}{2}\dot{a}^2x^2 - \frac{4}{3}\pi G\rho\, a^2x^2 &\overset{(1)}{=} \text{constant} \\ \implies \quad \left(\frac{\dot{a}}{a}\right)^2 &\overset{(2)}{=} \frac{8\pi G}{3}\,\rho - \frac{kc^2}{a^2}. \end{aligned} \tag{16.5}$$

(1) $r = ax$ and $M = \frac{4}{3}\pi\rho a^3 x^3$.

(2) Divide by $\frac{1}{2}a^2x^2$, and write the constant as $-kc^2/(2x^2)$, which defines k. It is the same for every shell because the universe is uniform.

$$\boxed{H^2 = \frac{8\pi G}{3}\,\rho - \frac{kc^2}{a^2}.} \tag{16.6}$$

This is the **Friedmann equation**. General relativity gives exactly this form, with ρ the total energy density divided by c^2 and k the curvature of space. The universe is observed to be flat to good accuracy, so from here on $k = 0$, and then the density is fixed by the expansion rate,

$$\rho_{\text{crit}} = \frac{3H^2}{8\pi G}, \qquad \Omega_i \equiv \frac{\rho_{i,0}}{\rho_{\text{crit},0}}. \tag{16.7}$$

Each content i (matter, radiation, the cosmological constant) is described by its share Ω_i of the critical density today.

How each content dilutes. The Friedmann equation needs $\rho(a)$. Take a comoving volume $V \propto a^3$. Its energy $\rho c^2 V$ changes only by the work done by its pressure as it expands (P1, the first law with no heat flowing in a uniform universe),

$$\begin{aligned} d(\rho c^2 a^3) &= -P\, d(a^3) \\ \implies \quad \dot{\rho} + 3\frac{\dot{a}}{a}\left(\rho + \frac{P}{c^2}\right) &\overset{(1)}{=} 0. \end{aligned} \tag{16.8}$$

(1) Divide by $c^2a^3\,dt$ and use $d(a^3) = 3a^2\,da$.

With $P = w\rho c^2$ for a fixed w, this integrates to $\rho \propto a^{-3(1+w)}$. Three cases matter.

- Matter, $w = 0$ (pressure negligible), $\rho \propto a^{-3}$. The particles are only spread over a larger volume.
- Radiation, $w = \frac{1}{3}$ (Eq. (9.26)), $\rho \propto a^{-4}$. Each photon is also redshifted, losing one more power of a.
- A cosmological constant, $w = -1$, $\rho =$ constant. Its energy density does not dilute at all.

So for a flat universe

$$\boxed{H^2 = H_0^2\left(\frac{\Omega_m}{a^3} + \frac{\Omega_r}{a^4} + \Omega_\Lambda\right), \qquad \Omega_m + \Omega_r + \Omega_\Lambda = 1.} \tag{16.9}$$

Radiation rules when a is small, then matter, then the cosmological constant. Matter and radiation were equal at $1 + z = \Omega_m/\Omega_r$, $z \approx 3461$, when the universe was about 4.96×10^4 years old.

Does the expansion slow down? Differentiate Eq. (16.6) (with $k = 0$) in time and use the fluid equation,

$$2\frac{\dot a}{a}\left(\frac{\ddot a}{a} - \frac{\dot a^2}{a^2}\right) \overset{(1)}{=} \frac{8\pi G}{3}\dot\rho$$
$$\overset{(2)}{=} -8\pi G\,\frac{\dot a}{a}\left(\rho + \frac{P}{c^2}\right)$$
$$\Longrightarrow \qquad \frac{\ddot a}{a} \overset{(3)}{=} -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right). \tag{16.10}$$

(1) $\mathrm{d}/\mathrm{d}t = \ddot a/a - \dot a^2/a^2$.
(2) The fluid equation (16.8).
(3) Divide by $2\dot a/a$ and replace $\dot a^2/a^2$ by $(8\pi G/3)\rho$.

$$\boxed{\frac{\ddot a}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right).} \tag{16.11}$$

Ordinary matter and radiation slow the expansion. Pressure adds to the pull, as it did in the TOV equation (Section 13.3). A content with $w < -\frac{1}{3}$ makes $\rho + 3P/c^2 < 0$ and speeds the expansion up. A cosmological constant, $w = -1$, does.

Single-content universes. With one content at a time, Eq. (16.9) integrates directly. For matter only, $\dot a = H_0 a^{-1/2}$, so $a^{1/2}da = H_0\,dt$ and

$$a = \left(\tfrac{3}{2}H_0 t\right)^{2/3}, \qquad t_0 = \frac{2}{3H_0} = 9.67\ \text{Gyr}. \tag{16.12}$$

For radiation only, $a = (2H_0 t)^{1/2}$ and $t_0 = 1/(2H_0)$. For a cosmological constant only, $a = e^{H_0(t-t_0)}$, with no beginning at all. Expansion that is always slowing has an age shorter than $1/H_0$.

Our universe. For the Planck values $H_0 = 67.4\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, $\Omega_m = 0.315$, $\Omega_r = 9.1\times10^{-5}$ and $\Omega_\Lambda = 0.685$, the age is found by integrating $dt = da/(aH)$ from $a = 0$ to 1 on the cluster,

$$\boxed{t_0 = \int_0^1 \frac{da}{a\,H(a)} = 13.79\ \text{Gyr}.} \tag{16.13}$$

This is longer than the matter-only age, because the cosmological constant has sped up the expansion lately, and close to $1/H_0$ by coincidence. Figure 16.1 compares the four histories. The oldest stars, about 13 billion years old, fit inside this age, as they must.

16.3 The Microwave Background and Recombination

Light in thermal equilibrium has the Planck spectrum (Eq. (9.31)), and the universe was once dense and hot. Today the universe is cold, thin and transparent. If its early light is still around, it has been stretched by expansion for billions of years, and it is not obvious that it should still look thermal, or when it last met matter.

Think of a guitar string in a box that slowly grows. Each note the string can play gets lower as the box grows, but the set of notes keeps the same pattern, only shifted. Light in an expanding universe is like that. Every mode is stretched by the same factor. The early universe was full of light in equilibrium with matter. As it expanded, every photon's wavelength grew with a, the number of photons in each mode stayed the same (P1), and the Planck spectrum stayed a

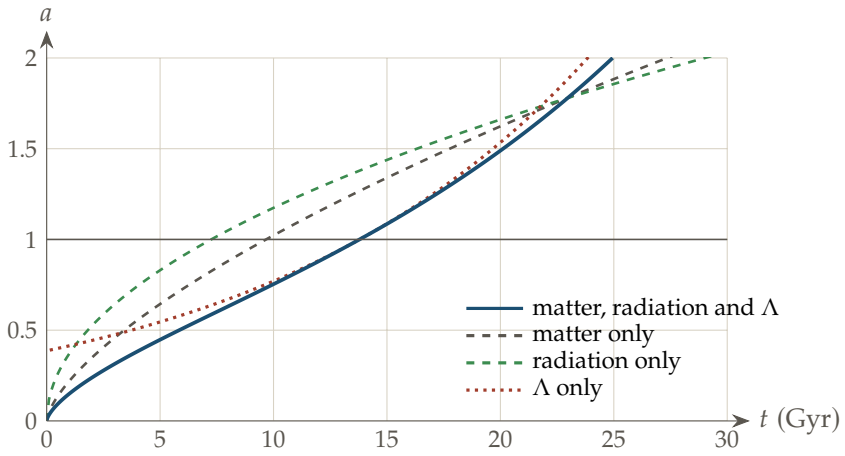


Figure 16.1: The scale factor against time for a flat universe with matter, radiation and a cosmological constant (Planck 2018 values), and for universes with only matter, only radiation or only a cosmological constant, each with the same H_0 when $a = 1$.

Planck spectrum at a temperature falling as $1/a$ (P3). When the universe cooled enough for electrons to join protons, the light was set free. The temperature at which that happened is set by counting states (P3), detailed balance (P4) and one ratio, the enormous number of photons per atom (P6).

In numbers, the light is seen today at 2.7255 K, 411 photons per cubic centimetre. It was released at about 3000 K, far below the 1.58×10^5 K at which a typical photon carries hydrogen's ionisation energy. The picture is a fog that suddenly clears. Before, light scatters off free electrons at every step. After, the electrons are locked in atoms and the light flies straight, carrying a snapshot of the moment the fog lifted. The same counting gives the ionisation of gas in stellar atmospheres and nebulae, and the same photon excess sets when the first nuclei formed (Section 16.4). It gives the temperature law, the photon density and the Saha equation.

A blackbody stays a blackbody. Between interactions, each photon's momentum falls as $1/a$, by Eq. (16.4). The number of photons in a given set of modes does not change, and the modes themselves are carried along with $p \propto 1/a$. The occupation of a mode, $f = 1/(e^{pc/k_BT} - 1)$ in equilibrium (Section 9.3), is therefore unchanged if pc/k_BT stays the same for each mode,

$$\boxed{T \propto \frac{1}{a}, \qquad T = T_0(1+z).} \tag{16.14}$$

The spectrum keeps the Planck form exactly, at a temperature falling as the universe grows. It is measured today at $T_0 = 2.7255$ K, the most perfect blackbody known. Hot gas in galaxy clusters scatters a small part of it to higher energies, so at low frequencies clusters appear as faint cool spots on the microwave sky, the Sunyaev–Zel'dovich effect of Eq. (15.43).

How many photons. With Eq. (9.23), two polarisations and $x = pc/k_BT$,

$$\begin{aligned} n_\gamma &\overset{(1)}{=} \int_0^\infty \frac{2}{h^3} \frac{4\pi p^2\, dp}{e^{pc/k_BT} - 1} \\ &\overset{(2)}{=} 8\pi \left(\frac{k_BT}{hc}\right)^3 \int_0^\infty \frac{x^2\, dx}{e^x - 1} \\ &\overset{(3)}{=} 16\pi\zeta(3) \left(\frac{k_BT}{hc}\right)^3 . \end{aligned} \tag{16.15}$$

(1) Counting states, Eq. (9.23), with the Planck occupation.
(2) The substitution $x = pc/k_BT$.
(3) $\int_0^\infty x^2 dx/(e^x - 1) = 2\zeta(3) = 2.404$, where $\zeta(3) = 1.202$.

At $T_0 = 2.7255\,\mathrm{K}$ this gives $n_\gamma = 411\,\mathrm{cm}^{-3}$. The energy density is $a_{\mathrm{rad}} T_0^4$ (Eq. (9.30), with the radiation constant written a_{rad}, since a is now the scale factor), which is a share $\Omega_\gamma = 5.4 \times 10^{-5}$ of the critical density. Neutrinos add 68% more, giving the $\Omega_r = 9.1 \times 10^{-5}$ used above. Ordinary matter, $\Omega_b = 0.049$, has $\eta = n_b/n_\gamma = 6.1 \times 10^{-10}$ baryons per photon. This tiny ratio decides the next result.

The Saha equation. Hydrogen is ionised and re-formed by $\mathrm{p} + \mathrm{e} \rightleftharpoons \mathrm{H} + \gamma$. In equilibrium, each process is balanced by its reverse (P4), so the populations take their equilibrium ratio from counting states (P3). A bound atom in its ground state has energy $-B$, with $B = 13.6\,\mathrm{eV}$. Ionising it frees an electron, which can then take any of the translational states available to it. An electron sharing space with the other free electrons has a volume $1/n_e$ to itself, and by Eq. (9.22) the number of its momentum states up to a thermal momentum is about $(2\pi m_e k_B T/h^2)^{3/2}/n_e$ (the Maxwell distribution integrated over d^3p/h^3). The Boltzmann factor for the energy B completes the ratio (the spin states of electron, proton and atom cancel, $2 \times 2/4 = 1$),

$$\boxed{\frac{n_p n_e}{n_{\mathrm{H}}} = \left(\frac{2\pi m_e k_B T}{h^2}\right)^{3/2} e^{-B/k_B T}.} \tag{16.16}$$

This is the **Saha equation**. With the ionised fraction $x = n_e/n_{\mathrm{H,tot}}$ and $n_p = n_e$, it becomes

$$\frac{x^2}{1-x} = \frac{1}{n_{\mathrm{H,tot}}} \left(\frac{2\pi m_e k_B T}{h^2}\right)^{3/2} e^{-B/k_B T}, \qquad n_{\mathrm{H,tot}} = (1 - Y)\,\eta\, n_\gamma,$$

with $Y = 0.245$ the helium mass fraction (Section 16.4).

Why so cold? The bracket grows as $T^{3/2}$, while $n_{\mathrm{H,tot}} = (1 - Y)\eta n_\gamma$ grows as T^3, so the right side is roughly

$$\frac{1}{\eta} \left(\frac{m_e c^2}{k_B T}\right)^{3/2} e^{-B/k_B T} \times (\text{a number of order one}).$$

The factor $1/\eta \approx 1.6 \times 10^9$ is huge. Each atom that forms is surrounded by over a billion photons, and even far out on their Planck tail enough of them carry 13.6 eV to break it again. Hydrogen stays ionised until $e^{-B/k_B T}$ has fallen far enough to beat 10^9 (P6). The script solves for x against z (Figure 16.2). The fraction falls through one half at $z = 1370$, $T = 3736\,\mathrm{K}$, where $B/k_B T = 42.3$, against the $1.58 \times 10^5\,\mathrm{K}$ at which $k_B T = B$.

Real recombination lags a little behind Saha's equilibrium. Each atom that forms in its ground state emits a photon that can ionise its neighbour, so the last stages are slowed and the light decouples at $z \approx 1090$, $T \approx 3000\,\mathrm{K}$, where the photons' mean free path grows past the size of the universe (P6, as in Section 11.2). That moment, about 380 000 years after the start, is what the microwave background shows.

16.4 Freeze-Out and the First Three Minutes

Recombination happened when a reaction balance shifted as the universe cooled (Section 16.3). Earlier, at temperatures of billions of kelvin, the same kind of balance governed the nuclei. One more effect then matters. A reaction can only keep its balance if it runs faster than the universe changes.

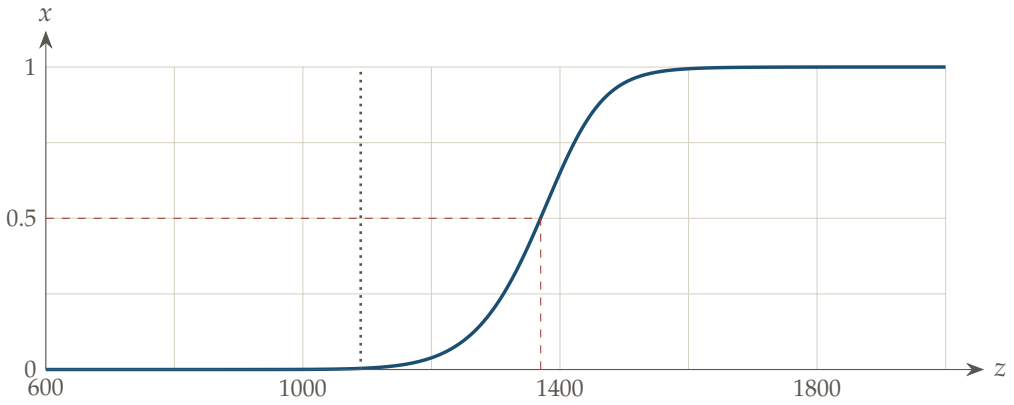


Figure 16.2: The ionised fraction of hydrogen against redshift from the Saha equation. The dashed lines mark $x = \frac{1}{2}$. The dotted line marks $z = 1090$.

Think of a party where guests keep swapping partners. While the room is crowded, swaps are frequent and partners stay mixed. As guests leave and the room empties, encounters become rare, and at some point everyone is stuck with whoever they last danced with. The mix freezes in at the moment encounters become rarer than the room changes. In the early universe, neutrons and protons turned into each other through weak reactions with electrons and neutrinos. While those reactions ran faster than the expansion (P4), the neutron-to-proton ratio followed its equilibrium value from the Boltzmann factor (P3). When the expansion overtook them (P6), the ratio froze, and almost all the neutrons that remained ended up in helium.

In numbers, the ratio froze at about one neutron for five protons. Some neutrons decayed before nuclei could form, leaving one for seven, which makes a quarter of the mass into helium. The picture is two curves crossing, the reaction rate falling steeply as the universe cools and the expansion rate falling gently, with the freeze-out where they cross. The same freeze-out decides how much dark matter is left over if it is a particle, when neutrinos stopped interacting, and the amounts of deuterium and lithium. It gives the time–temperature relation, the freeze-out ratio, the helium fraction and the deuterium delay.

Time and temperature. Early on, radiation dominates and $\rho \propto a^{-4}$, so $\dot{a}/a \propto a^{-2}$ gives $a \propto t^{1/2}$ and $H = 1/(2t)$. The Friedmann equation then ties the time to the temperature. With $\rho c^2 = g\,\frac{1}{2} a_{\mathrm{rad}} T^4$, where $g = 10.75$ counts photons, electrons, positrons and three kinds of neutrino at a few MeV (fermions counting $\frac{7}{8}$ of a boson),

$$\frac{1}{4t^2} \overset{(1)}{=} \frac{8\pi G}{3}\,\frac{g\,a_{\mathrm{rad}} T^4}{2c^2}$$

$$\Longrightarrow \qquad t = \left(\frac{3c^2}{16\pi G\, g\, a_{\mathrm{rad}}}\right)^{1/2} \frac{1}{T^2} \overset{(2)}{=} 0.74\,\mathrm{s}\left(\frac{1\,\mathrm{MeV}}{k_B T}\right)^2 . \tag{16.17}$$

(1) $H = 1/(2t)$ in the Friedmann equation, with each species' share of the radiation energy density.

(2) Numbers for $g = 10.75$.

A second after the beginning the temperature was about 10^{10} K, and three minutes after it about 10^9 K.

The freeze-out condition. The weak reactions $\mathrm{n} + \nu \rightleftharpoons \mathrm{p} + \mathrm{e}$ and $\mathrm{n} + \mathrm{e}^+ \rightleftharpoons \mathrm{p} + \bar{\nu}$ have a cross-section that grows as the square of the energy, so their rate per nucleon grows as T^5

(density $\propto T^3$, speed c, cross-section $\propto T^2$). The expansion rate grows only as T^2. The ratio of the two falls as T^3, and the reactions stop keeping up when

$$\boxed{\Gamma \approx H.} \tag{16.18}$$

With the measured strength of the weak interaction this happens at $k_B T_f \approx 0.8\,\mathrm{MeV}$.

Neutrons and helium. Before freeze-out the reactions in each direction balance (P4), and the Boltzmann factor of Eq. (9.39) gives

$$\frac{n}{p} = e^{-Q/k_B T}, \qquad Q = (m_n - m_p)c^2 = 1.293\,\mathrm{MeV}.$$

At freeze-out this is $e^{-1.293/0.8} = 0.20$. Free neutrons then decay, with a mean life of 879 s. Almost every neutron left when nuclei finally form, at about $t \approx 200\,\mathrm{s}$, ends up in helium-4, which holds two neutrons and two protons. With n/p the ratio at that moment, the mass fraction in helium is

$$Y \overset{(1)}{=} \frac{4\,(n/2)}{n+p} = \frac{2\,(n/p)}{1+n/p} \overset{(2)}{=} \frac{2\times\frac{1}{7}}{1+\frac{1}{7}} = 0.25. \tag{16.19}$$

(1) Each helium nucleus uses two neutrons and has four times the nucleon mass, so the helium mass is $4\times n/2$ in units of the nucleon mass, out of a total $n+p$.
(2) After freeze-out and decay, $n/p \approx \frac{1}{7}$.

The measured primordial helium fraction is 0.245. A crude model with freeze-out at 0.8 MeV and nuclei forming at 180 s gives $n/p = 0.16$ and $Y = 0.28$. The full calculation, which follows the reaction rates through the freeze-out instead of cutting them off, gives $n/p \approx \frac{1}{7}$ and agrees with the measurement.

The deuterium bottleneck. Helium forms through deuterium, bound by only 2.22 MeV. As with hydrogen in Section 16.3, a billion photons per nucleon can break every deuteron as long as enough photons on the Planck tail carry 2.22 MeV. Deuterium survives only once $e^{-B_D/k_B T}$ has fallen below about η,

$$k_B T \approx \frac{B_D}{\ln(1/\eta)} = \frac{2.22\,\mathrm{MeV}}{21.2} \approx 0.1\,\mathrm{MeV}, \tag{16.20}$$

and the full Saha-type balance, which adds the slowly varying factors, lowers this to 0.07 MeV, about 8×10^8 K. By the time-temperature relation that is about three minutes, the time the neutrons had to decay. The same photon excess that delayed recombination delayed the first nuclei. The tunnelling that makes fusion possible at these energies is that of Section 12.3.

16.5 The Growth of Structure

The Jeans criterion (Section 10.4) showed that in a static gas, large enough overdensities grow exponentially. The universe is not static. Expansion pulls every overdense region apart even as its own gravity pulls it together, and the microwave background shows that at $z \approx 1090$ the density varied by only one part in 10^5. Something must turn such tiny ripples into galaxies.

Think of trying to roll a snowball while someone keeps stretching the field it rolls on. It still grows, but much more slowly than on a fixed field. An overdense sphere in an expanding universe is a small closed universe inside a larger one. It expands a little more slowly than its

surroundings, so its density contrast grows, by P1 and P2 for its surface shell. In the matter era it grows only in proportion to a, not exponentially.

In numbers, from $z = 1090$ to today a grows by 1091, so a contrast of 10^{-5} would grow only to about 10^{-2}, far short of the value near one at which galaxies form. The picture is two spheres expanding side by side, the denser one falling slowly behind. The same equation describes the growth of every structure in the universe, from galaxy clusters to the cosmic web, and its numbers are part of the case for dark matter. It gives the growth equation, its solution in the matter era, and why ordinary matter alone cannot explain galaxies.

A slightly overdense sphere. Take a sphere whose density is $\bar\rho(1+\delta)$, with $\delta \ll 1$ and $\bar\rho$ the mean density. Its mass is the mean-universe mass of its comoving region, so its surface sits at $r = a\,x(1-\delta/3)$, slightly inside where it would be in the mean universe, since $r^3(1+\delta)$ must equal a^3x^3. By P2 the surface obeys $\ddot r = -GM/r^2$, with $M = \frac{4}{3}\pi\bar\rho a^3x^3$. Expanding to first order in δ,

$$\begin{aligned} \ddot r &\overset{(1)}{=} \ddot a\, x\left(1-\tfrac{\delta}{3}\right) - \tfrac{2}{3}\dot a\, x\,\dot\delta - \tfrac{1}{3}a\,x\,\ddot\delta, \\ -\frac{GM}{r^2} &\overset{(2)}{=} -\frac{4}{3}\pi G\bar\rho\, a\, x\left(1+\tfrac{2\delta}{3}\right). \end{aligned} \tag{16.21}$$

(1) Differentiate $r = ax(1-\delta/3)$ twice.
(2) $(1-\delta/3)^{-2} \approx 1 + 2\delta/3$.

At zeroth order the two sides give $\ddot a = -\frac{4}{3}\pi G\bar\rho\, a$, the mean universe (Eq. (16.11) for matter). At first order, after multiplying by $-3/(ax)$ and using that result,

$$\boxed{\ddot\delta + 2H\dot\delta = 4\pi G\bar\rho\,\delta.} \tag{16.22}$$

Without the term $2H\dot\delta$ this is the Jeans growth of Section 10.4 with no pressure, $\delta \propto e^{\sqrt{4\pi G\bar\rho}\,t}$. The expansion adds a drag.

Growth in the matter era. With matter only, $a \propto t^{2/3}$, $H = 2/(3t)$ and $4\pi G\bar\rho = \frac{3}{2}H^2 = 2/(3t^2)$. Trying $\delta \propto t^s$ in Eq. (16.22),

$$\begin{aligned} s(s-1) + \tfrac{4}{3}s &\overset{(1)}{=} \tfrac{2}{3} \\ \Longrightarrow \quad s &= \tfrac{2}{3} \quad \text{or} \quad s = -1. \end{aligned} \tag{16.23}$$

(1) Each term of Eq. (16.22) is a number times t^{s-2}.

The growing solution is $\delta \propto t^{2/3} \propto a$. Density contrasts grow only as fast as the universe expands.

Why ordinary matter is not enough. Before recombination, ordinary matter is tied to the light by Thomson scattering. Its pressure is that of the radiation, its Jeans length is larger than any region in causal contact, and by Section 10.4 it oscillates instead of growing. Only after $z \approx 1090$ can it grow, and then by at most a factor 1091 by today, from the measured 10^{-5} to 10^{-2}. Matter that does not feel the light, dark matter, can grow from the equality of matter and radiation at $z \approx 3461$, and by recombination its contrast is already large. Ordinary matter then falls into the dark matter's wells. Galaxies exist today because most of the matter does not interact with light, the same conclusion that rotation curves reach in Chapter 14.

The cosmic web. Once a region's contrast nears one, it stops expanding, turns around and collapses (the free fall of Section 10.5). What forms depends on the scale, and the result is the hierarchy of Section 14.7 (Figure 14.14). Galaxies gather into groups and clusters, held together by their own gravity. These lie in superclusters, which are not, and the superclusters lie along filaments and walls around nearly empty voids. Maps of galaxy redshifts show this cosmic web (Figure 14.16), and computer simulations of dark matter growing by Eq. (16.22) reproduce it. Two facts explain its shape. The small scales grow first, because the dark-matter ripples are largest on small scales, so small halos form first and merge into larger ones over time. And a slightly uneven region collapses fastest along its shortest axis, so it becomes a sheet first, then a filament where sheets meet, then a knot, a cluster, where filaments meet. On scales above a few hundred megaparsecs the web averages out and the universe is uniform, as the Friedmann equations assume.

16.6 The Lyman-α Forest

Most of the ordinary matter in the universe lies between the galaxies, in thin gas that gives off almost no light of its own. Gas that does not shine can still be seen if something bright lies behind it. A distant quasar is such a lamp, and the hydrogen in front of it absorbs the Lyman α line, the jump of its electron from $n = 1$ to $n = 2$ at $\lambda_\alpha = 121.567\,\text{nm}$ (Section 8.3). But the gas lies at many distances, each with its own redshift, and it is not obvious where in the quasar's spectrum each part of it will show.

Think of a strip of elastic with a ruler printed on it, stretched steadily while you watch. A friend walks along it and cuts a notch at the mark 121.6 at several places along the way. Every notch is cut at the same mark, but the stretching goes on after each cut, so the notches cut first have been stretched the most by the time you look. Each cloud absorbs only at 121.6 nm in its own frame. The light reaching it left the quasar bluer than that and has been stretched on the way, so the cloud removes the photons that happen to be passing through 121.6 nm as they cross it. After the cloud, those missing photons keep being stretched along with the rest. Today the gap sits at $\lambda_\alpha(1 + z_c)$, set by the redshift z_c of the cloud, not of the quasar. Every cloud cuts its own dip, at its own place.

In numbers, a quasar at $z = 3$ shows its own Lyman α emission at 486.3 nm. Clouds at $z = 2.55$, 2.75 and 2.90 in front of it absorb at 431.6, 455.9 and 474.1 nm. A real line of sight crosses hundreds of clouds, 296 in the computed example, and their dips crowd together into a "forest". The picture is Figure 16.3, three clouds that each absorb at the same wavelength in their own frame and leave their dips at three different places in the spectrum we see.

The same absorption maps the gas between the galaxies, measures how the dark matter clumps on small scales, tests whether dark matter is cold and how heavy neutrinos are, and shows when the universe was reionised. It gives where each dip falls, how much a neutral universe would absorb, and what the dips trace.

Where each dip falls. A photon leaves the quasar at scale factor a_q with wavelength λ_e. By Eq. (16.4) its wavelength grows in proportion to the scale factor, so when it crosses a cloud at scale factor a_c its wavelength is $\lambda_e a_c/a_q$. The cloud absorbs it only if that equals λ_α. A photon that survives would arrive today, $a = 1$, with λ_e/a_q. So the photons removed by the cloud are those that would have arrived at

$$\lambda_{\text{obs}} \overset{(1)}{=} \frac{\lambda_e}{a_q} \overset{(2)}{=} \frac{\lambda_\alpha}{a_c} \overset{(3)}{=} \lambda_\alpha(1 + z_c). \tag{16.24}$$

(1) Wavelength grows as a from a_q to 1.
(2) Absorption at the cloud, $\lambda_e = \lambda_\alpha a_q/a_c$.

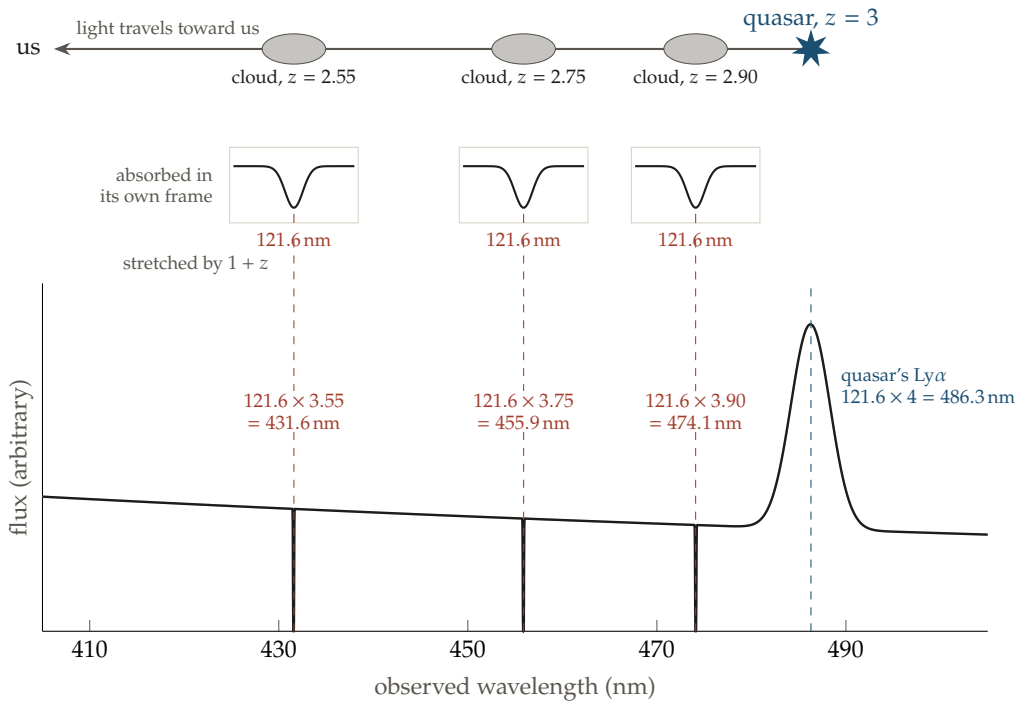


Figure 16.3: Top, a quasar at $z = 3$ and three hydrogen clouds between it and us, each with the spectrum it sees in its own frame, absorbed at 121.6 nm. Bottom, the spectrum of the quasar that reaches us. Each cloud's dip lies at $121.6\,(1 + z)$ nm for that cloud's redshift z, and the quasar's own Lyman α emission at $121.6\,(1 + z_q)$ nm.

(3) $1 + z_c = 1/a_c$, Eq. (16.4).

$$\boxed{\lambda_{obs} = \lambda_\alpha\,(1 + z_c).} \tag{16.25}$$

The quasar's redshift has dropped out. Three consequences follow.

- Every cloud lies between us and the quasar, $z_c < z_q$, so every dip lies blueward of the quasar's own Lyman α emission at $\lambda_\alpha(1 + z_q)$. Light redward of it was redder than λ_α when it left and only got redder, so no cloud can absorb it, and the spectrum there is clean.
- Each wavelength in the forest belongs to one redshift, and so to one distance along the line of sight. The spectrum is a map of the hydrogen along a line through the universe, read as a function of wavelength.
- The other Lyman lines (Lyman β at 102.6 nm and so on) make dips of their own further blue. Between the quasar's Lyman β and Lyman α emission, only Lyman α dips appear, which is the region usually used.

Figure 16.4 shows the forest of a quasar at $z = 3$ computed for a realistic population of clouds. On average a fraction 0.88 of the light gets through between its Lyman β and Lyman α emission, and the spectrum redward of its Lyman α peak is untouched.

How much a neutral universe would absorb. Suppose the hydrogen between the galaxies were spread evenly, with n_{HI} neutral atoms per unit volume. A photon seen today at frequency ν_0 had frequency $\nu = \nu_0(1 + z)$ at redshift z. By Eq. (11.6) the fraction it loses over a path dl is $n_{HI}\sigma(\nu)\,dl$, where $\sigma(\nu)$ is the cross-section of the line. The path covered while the redshift

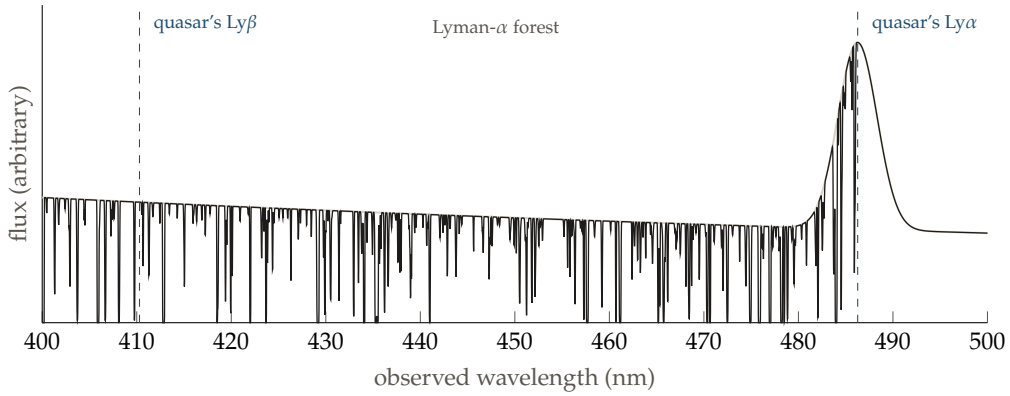


Figure 16.4: The spectrum of a quasar at $z = 3$ seen through 296 hydrogen clouds between $z = 2$ and $z = 3$, with column densities and numbers per unit redshift as observed. The grey curve is the quasar's light without the clouds.

drops by dz is $dl = c\,dt = c\,dz/[(1+z)H]$, since $1 + z = 1/a$ gives $dz/dt = -(1+z)H$. So

$$\tau(\nu_0) \overset{(1)}{=} \int n_{\mathrm{HI}}\,\sigma\big(\nu_0(1+z)\big)\,\frac{c\,dz}{(1+z)H} \overset{(2)}{=} \frac{n_{\mathrm{HI}}\,c}{H\,\nu_\alpha}\int \sigma(\nu)\,d\nu \overset{(3)}{=} \frac{e^2 f}{4\varepsilon_0 m_e c}\,\frac{\lambda_\alpha\,n_{\mathrm{HI}}}{H}. \tag{16.26}$$

(1) Add up the fractions lost along the whole path.

(2) The line is so narrow that σ is non-zero only over a tiny range of z around the redshift where $\nu_0(1+z) = \nu_\alpha$. Over that range n_{HI}, H and $1+z$ are constant. Change variable to $\nu = \nu_0(1+z)$, so $dz = (1+z)\,d\nu/\nu$ with $\nu \approx \nu_\alpha$.

(3) $c/\nu_\alpha = \lambda_\alpha$, and the area under the line's cross-section, from atomic physics, is $\int \sigma\,d\nu = (e^2/4\varepsilon_0 m_e c)\,f$ with oscillator strength $f = 0.416$ for Lyman α.

$$\boxed{\tau_{\mathrm{GP}} = \frac{e^2 f}{4\varepsilon_0 m_e c}\,\frac{\lambda_\alpha\,n_{\mathrm{HI}}(z)}{H(z)}.} \tag{16.27}$$

This is the **Gunn–Peterson** optical depth. The shape of the line does not matter, only its area. Each wavelength samples the gas in a slab whose depth is set by how fast the expansion sweeps the line across it, which is why H appears. If all the hydrogen between the galaxies were neutral, with the measured density of ordinary matter, τ_{GP} would be 1.6×10^5 at $z = 3$ and 3.8×10^5 at $z = 6$. Not one photon in a million million would get through blueward of the quasar's Lyman α line. Yet quasars at $z = 3$ let most of that light through. So the gas between the galaxies is almost completely ionised. A neutral fraction of only 6×10^{-6} already gives $\tau = 1$ at $z = 3$.

Reionisation. After recombination (Section 16.3) the gas was neutral. Something ionised it again, the ultraviolet light of the first stars, galaxies and quasars. In quasars beyond $z \approx 6$ the forest merges into a band of total darkness, a Gunn–Peterson trough. The neutral fraction there has risen above about 10^{-4}. The trough marks the end of reionisation, about a billion years after the Big Bang.

What the dips trace. Today's ionised gas is held in balance by two rates (P4). Ultraviolet light ionises each neutral atom at a rate Γ_{ion}, and protons and electrons recombine at a rate

$\alpha_{\rm rec}(T)\, n_e$ per proton. Balancing ionisations against recombinations in nearly fully ionised gas, $n_e \approx n_p \approx n_{\rm H}$,

$$\Gamma_{\rm ion}\, n_{\rm HI} = \alpha_{\rm rec}(T)\, n_{\rm H}^2 \quad \Longrightarrow \quad n_{\rm HI} = \frac{\alpha_{\rm rec}(T)}{\Gamma_{\rm ion}}\, n_{\rm H}^2 \propto (1+\delta)^2,$$

where δ is the density contrast of Section 16.5. The neutral hydrogen, and so the local optical depth, grows as the square of the density. Each dip is an overdense place on the line of sight, and the depth of the absorption at each wavelength measures the density of gas at that distance. The forest is a map of the density of the universe along a line, at redshifts 2 to 5, where no galaxy survey reaches so finely.

Dark matter and neutrinos. On scales larger than its Jeans length (Section 10.4), a few hundred kiloparsecs for gas at 10^4 K, the gas has fallen into the wells of the dark matter and traces them (Section 16.5). So the dips trace the clumping of the dark matter itself, on scales from about one to a hundred megaparsecs. Three things follow from measuring how strongly the forest fluctuates on each scale.

- **Dark matter is cold.** Particles born with random speeds fly straight out of any overdensity smaller than the distance they travel, and erase it before it can grow. Fast, light particles erase small structures. The forest shows structure down to about a megaparsec, so the dark-matter particles were slow. This rules out warm dark matter lighter than a few keV.
- **Neutrinos are light.** Neutrinos with mass behave as hot dark matter and smooth out small-scale growth in the same way. The forest, combined with the microwave background, limits the sum of the three neutrino masses to about 0.1 eV.
- **The expansion history.** With many quasars close together on the sky, the forests can be compared across lines of sight, mapping the gas in three dimensions. The clustering of the gas has a preferred scale, 150 Mpc, set by sound waves before recombination. Seen at $z \approx 2.3$, its apparent size measures $H(z)$ and the distance to that redshift, a standard ruler for the expansion at an epoch no supernova reaches.

The widths of the narrowest dips give one more number. Their Doppler width is set by the thermal speed of the hydrogen atoms (Section 8.3), and it gives the temperature of the gas, about 10^4 K, set by the heating of photoionisation.

16.7 The Accelerating Expansion

The acceleration equation, Eq. (16.11), said that ordinary matter slows the expansion and a cosmological constant speeds it up. Which wins today cannot be read off from the expansion rate alone. It needs distances to objects far enough away that the expansion history has changed along the way.

Think of judging whether a car is speeding up by looking at two milestones far apart. If the second is reached sooner than the first speed predicts, the car was slowing. If it takes longer, the car was speeding up. A distant galaxy at a given redshift appears fainter if the universe has been speeding up, because it is farther away. Type Ia supernovae are such milestones. They are white dwarfs pushed to the Chandrasekhar mass (Section 13.2), so they explode with nearly the same peak luminosity (Section 13.2.1), and their brightness gives their distance. At a given redshift they are fainter than a slowing universe predicts.

In numbers, at $z = 1$ the luminosity distance in the Planck universe is 6802 Mpc, against 5211 Mpc in a matter-only universe with the same H_0. A supernova there is 0.58 magnitudes fainter, an effect first seen in 1998. The picture is a scale factor curve that bends upward in its

recent part (Figure 16.1). The same deceleration parameter is measured with the sizes of galaxy clustering patterns and with gravitational-wave sources of known loudness (Section 13.6.7). It gives the deceleration parameter and when the acceleration began.

The deceleration parameter. The dimensionless measure of whether the expansion slows is $q \equiv -\ddot{a}a/\dot{a}^2$. Dividing Eq. (16.11) by H^2, with $\rho_i = \Omega_i \rho_{\text{crit}}$ today and $P = w\rho c^2$,

$$q_0 \overset{(1)}{=} \frac{4\pi G}{3H_0^2} \sum_i \rho_i (1 + 3w_i) \overset{(2)}{=} \frac{1}{2} \sum_i \Omega_i (1 + 3w_i)$$
$$= \frac{\Omega_m}{2} + \Omega_r - \Omega_\Lambda. \tag{16.28}$$

(1) Eq. (16.11) with $P_i = w_i \rho_i c^2$, divided by H_0^2.
(2) $\rho_{\text{crit}} = 3H_0^2/(8\pi G)$, Eq. (16.7).

For the Planck values $q_0 = -0.53$, so the expansion is speeding up today. It began to speed up when $\Omega_m a^{-3}/2 = \Omega_\Lambda$, at $z = 0.63$, a little before the cosmological constant came to dominate the density at $z = 0.30$.

16.8 Problem Set: The Expanding Universe

16.8.1 Problem 1: Hubble Time and Age

Problem Statement: (a) Convert $H_0 = 67.4\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1}$ into a Hubble time $1/H_0$ in years. (b) What is the age of a flat matter-only universe with this H_0? (c) Why is it a problem that the oldest stars are about 13 billion years old, and how does the cosmological constant solve it?

Essential Ideas:

- $1\,\text{Mpc} = 3.086 \times 10^{19}\,\text{km}$. Matter-only age $2/(3H_0)$.

Detailed Solutions:

(a) **Hubble time**. $1/H_0 = 3.086 \times 10^{19}/67.4\,\text{s} = 4.58 \times 10^{17}\,\text{s}$.

$$\boxed{1/H_0 \simeq 14.51\ \text{Gyr}.}$$

(b) **Matter only**.

$$\boxed{t_0 = \tfrac{2}{3}\,14.51\ \text{Gyr} \simeq 9.67\ \text{Gyr}.}$$

(c) **Old stars**. A universe younger than its oldest stars is impossible. The cosmological constant has sped up the recent expansion, so for the same H_0 today the past expansion was slower and took longer, giving 13.79 Gyr (Eq. (16.13)).

16.8.2 Problem 2: How Thin Is the Universe?

Problem Statement: Find the critical density today in $\text{kg}\,\text{m}^{-3}$ and in protons per cubic metre. How many protons per cubic metre are there in ordinary matter ($\Omega_b = 0.049$)?

Essential Ideas:

- $\rho_{\text{crit}} = 3H_0^2/(8\pi G)$, Eq. (16.7).

Detailed Solutions:

$$\boxed{\rho_{\text{crit}} = 8.5 \times 10^{-27}\,\text{kg}\,\text{m}^{-3} = 5.1\ \text{protons per m}^3, \qquad \text{ordinary matter } 0.25\ \text{per m}^3.}$$

Ordinary matter averages one proton in four cubic metres, while each cubic metre holds 4.1×10^8 photons of the microwave background, giving $\eta = 6.1 \times 10^{-10}$.

16.8.3 Problem 3: The Light at Recombination

Problem Statement: (a) What was the temperature of the microwave background at $z = 1090$? (b) What was the mean energy of its photons, $2.70\,k_BT$? (c) Compare with the ionisation energy of hydrogen, and explain why hydrogen could recombine.

Essential Ideas:

- $T = T_0(1 + z)$, Eq. (16.14).
- The photon excess $1/\eta$ in the Saha equation.

Detailed Solutions:

(a) **Temperature**.

$$\boxed{T = 2.7255 \times 1091\,\mathrm{K} \simeq 2970\,\mathrm{K}.}$$

(b) **Mean photon energy**. $2.70 \times (8.617 \times 10^{-5}\,\mathrm{eV\,K^{-1}}) \times 2970\,\mathrm{K}$.

$$\boxed{\bar{E} \simeq 0.69\,\mathrm{eV}.}$$

(c) **Why recombination happens**. The mean photon has only $\frac{1}{20}$ of the 13.6 eV needed. The fraction of photons above 13.6 eV falls as e^{-B/k_BT}, and once it is below $\eta \approx 6 \times 10^{-10}$ there are fewer ionising photons than atoms. At higher temperatures, a fraction as small as 10^{-9} of a billion photons per atom still sufficed to keep hydrogen ionised.

16.8.4 Problem 4: Helium From the Neutron Ratio

Problem Statement: (a) If the neutron-to-proton ratio is $\frac{1}{6}$ when nuclei form, what is the helium mass fraction? (b) If neutrons did not decay, with $n/p = \frac{1}{5}$, what would it be? (c) Why does a faster expansion rate (more kinds of light particles) raise Y?

Essential Ideas:

- $Y = 2(n/p)/(1 + n/p)$.
- Freeze-out at $\Gamma \approx H$, Eq. (16.18).

Detailed Solutions:

(a) **With decay**.

$$\boxed{Y = \frac{2/6}{7/6} \simeq 0.29.}$$

(b) **Without decay**.

$$\boxed{Y = \frac{2/5}{6/5} \simeq 0.33.}$$

(c) **Faster expansion**. A larger H at a given temperature makes $\Gamma \approx H$ happen at a higher temperature, when e^{-Q/k_BT} is larger, and leaves less time for decay. Both raise n/p and Y. The measured Y was used this way to count the light neutrino species before particle accelerators did.

16.8.5 Problem 5: Growth Since Recombination

Problem Statement: In a matter-dominated universe, by what factor does a density contrast grow from $z = 1090$ to $z = 0$? Starting from 10^{-5}, does ordinary matter alone reach $\delta \sim 1$ today? From when could dark matter grow?

Essential Ideas:

- $\delta \propto a$ in the matter era, Eq. (16.22).

Detailed Solutions:

$$\boxed{\frac{\delta_0}{\delta_{1090}} = \frac{a_0}{a_{1090}} = 1091, \qquad \delta_0 \simeq 10^{-2}.}$$

This is far below one, and the cosmological constant, which slows growth in the last few billion years, makes the gap larger. Dark matter is not held up by the light, so it can grow from matter–radiation equality, $z \approx 3461$. By recombination it has grown by a further factor of a few and its wells are ready for ordinary matter to fall into.