Preface

Einstein discovered his theory of gravitation in 1916.

By rights, this theory should not have been discovered until 20 years later, when physicists had acquired a clear understanding of relativistic field theory and of gauge invariance.

Einstein's profound and premature insights into the nature of gravitation had more to do with intuition than with logic.

In contrast to the admirably clear and precise operational foundations on which he based his theory of special relativity, the foundations on which he based general relativity were vague and obscure.

As has been emphasized by Synge and by Fock, even the very name of the theory indicates a misconception: There is no such thing as a relativity more general than special relativity.

But whatever murky roads he may have taken, in the end Einstein's intuition led him to create a theory of dazzling beauty.

If, using Arthur Koestler's image, we regard Copernicus, Kepler, and Newton as sleepwalkers who knew where they wanted to go and managed to get there without quite knowing how, then Einstein was the greatest sleepwalker of them all.

The aim of this book is to develop gravitational theory in the simplest and most straightforward way – in the way it probably would have developed without Einstein's intervention.

This means that we begin with the linear approximation and regard gravitation as the theory of a second-rank tensor field in a flat spacetime background, analogous to electrodynamics.

The geometrical interpretation and the nonlinear Einstein equations gradually emerge from this tensor theory as we attempt to understand and improve the equations of the linear approximation.

This approach is not new: Gupta, Feynman, Thirring, and Weinberg have presented it from somewhat different points of view and with varying amounts of detail.

One advantage of this approach is that it gives a clearer understanding of how and why gravitation is geometry.

Another advantage is that the linear theory permits us to delve immediately into the physics: Gravitational redshift, light deflection, lensing, time delay, Lense-Thirring precession, and gravitational radiation can be treated directly in the context of the linear approximation, without any lengthy preliminary digressions on the mathematics of Riemannian spacetime geometry.

After a full exploitation of the results accessible via the linear approximation (Chapters 1–5), we redevelop the gravitation field equations via the geometrical approach pioneered by Einstein (Chapters 6–9).

This may seem to be a duplication of effort, but it helps students attain a deeper grasp of the principles.

In our exploration of multiple lines of approach, from different perspectives, we are following the example set by Lorentz in his celebrated "Monday lectures" at Leiden, where he would "turn the subject round and round and over and over" to achieve new insights.

As in earlier editions of the book, we enliven the theoretical treatment by presenting relevant experimental and observational results.

In its early years, general relativity acquired the reputation of an abstract, highly mathematical theory, with a limited experimental basis.

But since the 1960s, general relativity has enjoyed a harmonious and invigorating synergy of theory and experiment, with theory motivating experiments and experiments supporting and confirming theory.

The last 30 years have yielded a rich harvest of experimental and observational results, and we try to make the presentation of this information as complete and up to date as possible.

As in earlier editions, we include extensive tables of repetitions of experiments and observations, because testing and retesting are what make experimental results credible (*provando e riprovando*, as says the motto of the Academia dei Lincei, of which Galileo was a founding member).

However, in contrast to the earlier editions, which aimed to include all repetitions of a given experiment or observation, limitations of space compelled us to make some judicious selections, so the entries in our tables are now restricted to the most recent, most precise, and most memorable results.

In this third edition of the book we retain the organization of the second edition, with various shifts of emphasis, additions, and updates – mostly motivated by new experimental measurements and sometimes by improvements in the theoretical treatment.

The following list summarizes the changes relative to the second edition, apart from corrections of various unfortunate misprints.

Chapter 1: New results of measurements of G, new tests of the short-range behavior of the inverse-square law, new data on the quadrupole moment of the Sun from solar oscillations, and new measurements of tidal forces with the Gravity Field and Ocean Circulation Explorer spacecraft.

But we deleted most of the previous material on the fifth force, which can now be regarded as refuted.

Chapter 2: Expanded treatment of special relativity, which now goes beyond the mathematical formalism and provides a brief, self-contained introduction to the theory, with concise derivations of the invariance of the spacetime interval, the energy-momentum of particles and of systems, and the energy-mass relation.

Chapter 3: Improved explanations of the connection between the equation of motion of particles and the field equation, and the connection between the equation of motion and the geometric interpretation of gravity.

Chapter 4: Updated experimental and observational results on the gravitational red-shift, deflection, and time delay of light and radio waves.

Updates on observations of gravitational lensing (especially with the Hubble Space Telescope) and applications to investigations of dark mass and microlensing.

Expanded discussion of the orbital and spin precession according to the Lense-Thirring effect and the Laser Geodynamics Satellite results.

Chapter 5: Explicit discussion of the relationship between the polarization states (or spin states) of gravitational waves, their gauge invariance, and the conservation law for the energy-momentum tensor.

New data on the Hulse-Taylor pulsar and other binary pulsars and the implications for gravitational radiation.

A fuller discussion of sensitivity of LIGO gravitational wave detectors, with omission of most of the previous discussion of the sensitivity of resonant quadrupole detectors, which have now fallen out of favor.

Chapter 6: Geometric interpretation of the Bianchi identities in terms of parallel transport around a parallelepiped and physical interpretation of the Riemann tensor in terms of measurements within small regions; for instance, measurements of small volumes or areas.

Also, a full treatment of Fermi coordinates and Fermi-Walker transport.

Chapter 7: Clearer explanation of the motivation underlying the general-invariance symmetry.

Elimination of the separate treatment of the Birkhoff theorem, which is now incorporated directly into the Schwarzschild solution.

Discussion of the long-awaited final results of the Gravity Probe B experiment on the measurement of the geodetic and Lense-Thirring precession effects.

Chapter 8: Examination of the turning points for motion in the equatorial plane of the Kerr geometry and characterization of the possible circular orbits.

Fuller discussion of the irreducible mass when the black hole includes electric charge, as well as implications of the Cauchy horizon for the maximal Kerr geometry.

Description of recent calculations of the equilibrium configuration of neutron stars and the critical mass limit according to the novel method of Ruffini et al. based on the gravitational Fermi-Thomas model.

Update on the observational evidence for black holes.

Chapter 9: Recent determinations of the Hubble constant, the age of the universe, and the conclusions about the acceleration of the universe extracted from observations of type Ia supernovas by Riess et al. and Perlmutter et al.

In accord with the observational evidence, this chapter now emphasizes the spatially flat Friedmann- Lemaıtre model of the universe with a positive cosmological constant.

Chapter 10: Update of the information on helium abundance and a more detailed treatment – on the basis of the Jeans mass – of the growth of perturbations in the early universe.

Also an improved discussion of inflation, especially in regard to the flatness puzzle and the Grand Unified Theory (GUT) phase transition, and an examination of the implications of small-scale anisotropies in the cosmic background radiation detected by the Wilkinson Microwave Anisotropy Probe satellite, leading to the discovery of baryon acoustic oscillations.

Appendix: Direct derivation of the conservation of the energy-momentum tensor from the general-invariance symmetry, and addition of a new section with the general- relativistic theoretical proof of the equality of inertial and gravitational mass.

Fundamental constants

Speed of light	$c = 3.00 \times 10^{10}$ cm/s
Planck's constant	$\hbar = 1.05 \times 10^{-27}$ erg $\cdot$ s $= 6.58 \times 10^{-22}$ MeV $\cdot$ s
Gravitational constant	$G = 6.67 \times 10^{-8}$ cm^3 g^{-1} s^{-2}
	$\kappa = (16\pi G/c^4)^{1/2} = 2.04 \times 10^{-24}$ (cm $\cdot$ g)$^{-1/2}$ s
Planck length	$l_{Pl} = (\hbar G/c^3)^{1/2} = 1.62 \times 10^{-33}$ cm
Planck time	$t_{Pl} = (\hbar G/c^5)^{1/2} = 5.39 \times 10^{-44}$ s
Planck mass	$m_{Pl} = (\hbar c/G)^{1/2} = 2.18 \times 10^{-5}$ g
	$= 1.22 \times 10^{19}$ GeV/c^2
Electron mass	$m_e = 0.911 \times 10^{-27}$ g $= 0.511$ MeV/c^2
Proton mass	$m_p = 1.67 \times 10^{-24}$ g $= 938$ MeV/c^2
Neutron mass	$m_n = m_p + 2.31 \times 10^{-27}$ g $= m_p + 1.29$ MeV/c^2
Proton charge	$e = 4.80 \times 10^{-10}$ esu
Fine structure constant	$\alpha = e^2/\hbar c = 1/137.0$
Compton wavelength	$\lambda\!\!\!^{-}_C = \hbar/m_e c = 3.86 \times 10^{-11}$ cm
Bohr radius	$a_0 = \hbar^2/m_e e^2 = 0.529 \times 10^{-8}$ cm
Boltzmann constant	$k = 1.38 \times 10^{-16}$ erg/K $= 8.62 \times 10^{-5}$ eV/K
Stefan-Boltzmann constant	$\sigma = \pi^2 k^4/60\hbar^3 c^2 = 5.67 \times 10^{-5}$ g $\cdot$ s^{-3} K^{-4}

Conversion constants

1 year (y) $= 3.16 \times 10^{7}$ s
1 astronomical unit (A.U.) $= 1.50 \times 10^{13}$ cm
1 light year (ly) $= 0.946 \times 10^{18}$ cm
1 parsec (pc) $= 3.26$ ly $= 3.09 \times 10^{18}$ cm
1 second of arc (arcsec) $= 4.85 \times 10^{-6}$ radian
1 electron volt (eV) $= 1.6 \times 10^{-12}$ erg

Astronomical constants

Sun	
mass	$M_\odot = 1.99 \times 10^{33}$ g
radius	$R_\odot = 6.96 \times 10^{10}$ cm
surface gravity	$g_\odot = 2.74 \times 10^4$ cm/s^2
luminosity	$L_\odot = 3.8 \times 10^{33}$ erg/s
Earth	
mass	$M_E = 5.98 \times 10^{27}$ g
equatorial radius	$R_E = 6.38 \times 10^8$ cm
polar radius	$R'_E = R_E - 2.15 \times 10^6$ cm
surface gravity	$g = 9.81 \times 10^2$ cm/s^2
moment of inertia, polar axis	$I^{33} = 0.331\, M_E R_E^2$
moment of inertia, equatorial axis	$I^{22} = I^{11} = 0.329\, M_E R_E^2$
period of rotation 1 sidereal day	$= 8.62 \times 10^4$ s
mean distance to Sun, 1 A.U.	$= 1.50 \times 10^{13}$ cm
orbital period, 1 year	$= 3.16 \times 10^7$ s
orbital speed	$= 29.8$ km/s
Moon	
mass	$M_\leftmoon = 7.35 \times 10^{25}$ g
radius	$R_\leftmoon = 1.74 \times 10^8$ cm
mean distance from Earth	$= 3.84 \times 10^{10}$ cm
orbital period, 1 sidereal month	$= 27.3$ days
Universe	
Hubble constant	$H_0 = 100h$ km/(s $\cdot$ Mpc), with $h \cong 0.70$
critical density	$3{H_0}^2/8\pi G = 1.88 \times 10^{-29}\, h^2$ g/cm^3
	$\cong 0.92 \times 10^{-29}$ g/cm^3
density parameters	$\Omega_{m,0} = 0.27,\ \Omega_{\Lambda,0} = 0.73$

Notation

The components of a 3-D vector **A** with respect to 3-D rectangular coordinates will be indicated by superscripts with the values 1, 2, 3:

$$A^1 = A_x\ , \quad A^2 = A_y, \quad A^3 = A_z$$

For the position vector **x**, the components are

$$x^1 = x, \quad x^2 = y, \quad x^3 = z$$

The symbol A^k , where the Latin superscript *k* takes on the values k = 1, 2, 3, then stands for the *k*th component of the vector.

If no particular value of *k* is specified, the symbol A^k will also stand for the set (A^1,A^2,A^3) of all the components taken together; in the latter case, A^k represents the entire vector **A**.

The Einstein summation convention applies: when a repeated Latin index appears in a term in an equation, a summation is to be carried out over the values 1, 2, 3 of that index, for example

$$A^n B^n \equiv \sum_{n=1}^{3} A^n B^n = A^1 B^1 + A^2 B^2 + A^3 B^3$$

The 3-D Kronecker delta will be written as $\delta_m^n = \begin{cases} 1 & \text{if } m = n \\ 0 & \text{if } m \neq n \end{cases}$

Integration over a 3-D volume will be written as

$$\int f(\mathbf{x}) d^3x \equiv \iiint f(\mathbf{x}) dx dy dz$$

The components of a 4-D vector will be indicated by superscripts with the values 0, 1, 2, 3. In flat 4-D spacetime, with rectangular coordinates ct, x, y, z,

$$A^0 = A_t, \quad A^1 = A_x, \quad A^2 = A_y, \quad A^3 = A_z$$

and

$$x^0 = ct, \quad x^1 = x, \quad x^2 = y, \quad x^3 = z$$

The symbol A^{μ}, where the Greek superscript takes on the values $\mu = 0, 1, 2, 3$, stands for the μth component of the vector; it also stands for the set (A^0, A^1, A^2, A^3), and in the latter case represents the entire 4-D vector.

The definition of the 4-D Kronecker delta is the same as in the 3-D case:

$$\delta_{\mu}^{\nu} = \begin{cases} 1 & \text{if } \mu = \nu \\ 0 & \text{if } \mu \neq \nu \end{cases}$$

When a repeated Greek index appears in a term of an equation, a summation is to be carried out over the values 0, 1, 2, 3 of that index, for example,

$$\eta_{\mu\nu}A^{\nu} \equiv \sum_{\nu=0}^{3} \eta_{\mu\nu}A^{\nu} = \eta_{\mu 0}A^0 + \eta_{\mu 1}A^1 + \eta_{\mu 2}A^2 + \eta_{\mu 3}A^3$$

The Minkowski metric tensor of flat spacetime is taken as

$$\eta_{\mu\nu} = \begin{pmatrix} \eta_{00} & \eta_{01} & \eta_{02} & \eta_{03} \\ \eta_{10} & \eta_{11} & \eta_{12} & \eta_{13} \\ \eta_{20} & \eta_{21} & \eta_{22} & \eta_{23} \\ \eta_{30} & \eta_{31} & \eta_{32} & \eta_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

and the spacetime interval of flat spacetime is

$$ds^2 = \eta_{\mu\nu}dx^{\mu}dx^{\nu} = \sum_{\mu=0}^{3}\sum_{\nu=0}^{3} \eta_{\mu\nu}dx^{\mu}dx^{\nu} = (cdt)^2 - (dx)^2 - (dy)^2 - (dz)^2$$

In the last expression, the parentheses will often be omitted, so the right side becomes $c^2dt^2 - dx^2 - dy^2 - dz^2$ (this somewhat careless notation for a second-order differential imitates what is routinely done in second derivatives, for instance $\frac{d}{dz}\frac{d}{dz}f = \frac{d^2}{dz^2}f$).

The spacetime interval of curved spacetime is

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = \sum_{\mu=0}^{3}\sum_{\nu=0}^{3} g_{\mu\nu}dx^\mu dx^\nu$$
$$= g_{00}c^2dt^2 + g_{01}cdtdx + g_{10}cdxdt + \cdots + g_{33}dz^2$$

If x^0 is the time coordinate, then $g_{00} > 0$ (this is called the timelike sign convention).

In general, indices are raised and lowered with the metric tensor of curved spacetime, for instance,

$$A_\mu = g_{\mu\nu}A^\nu$$

However, in all equations that are written in the linear approximation, the indices are raised and lowered with the Minkowski metric tensor $\eta_{\mu\nu}$.

Partial derivatives are indicated by a comma or by the differential operator ∂,

$$\frac{\partial f}{\partial x^\mu} = f_{,\mu} = \partial_\mu f$$

A dot over a variable indicates a derivative with respect to time (for example, $\dot{z}$ = dz/dt in Chapter 5 and in the Appendix) or a derivative with respect to proper time (for example, $\dot{r}$ = dr/dτ in Chapter 8) or a derivative with respect to a "time parameter" (for example, $\dot{a}$ = da/dη in Chapter 9).

1 - Newton's Gravitational Theory

Few theories can compare in the accuracy of their predictions with Newton's theory of universal gravitation.

The predictions of celestial mechanics for the positions of the major planets agree with observation to within a few arcseconds over time intervals of many years.

The discovery of Neptune and the rediscovery of Ceres are among the spectacular successes that testify to the accuracy of the theory.

But Newton's theory is not perfect: The predicted motions of the perihelia for the inner planets deviate somewhat from the observed values.

In the case of Mercury the excess perihelion precession amounts to 43 arcseconds per century.

This small deviation was discovered through calculations by LeVerrier in 1845, and it was confirmed by Newcomb in 1882.

The explanation of this perihelion precession became one of the early successes of Einstein's relativistic theory of gravitation.

Telescopic observations of planetary angular positions stretching over hundreds of years are needed to detect the excess perihelion precession.

However, with the development of radar astronomy it has become possible to measure the distances to the inner planets directly and very accurately by means of the travel time of a radio signal sent from the Earth to the planet and reflected back.

With such radar observations of distances, the small deviations from Newton's theory can be detected after just a few years of observation.

Although Newton's theory is not perfect, it is in excellent agreement with observation in the limiting case of motion at low velocities in a weak gravitational field.

Any relativistic theory of gravitation ought to agree with Newton's theory in this limiting case.

We therefore begin with a brief exposition of some aspects of Newton's theory.

1.1 The law of universal gravitation

According to Newton, the law governing gravitational interactions is "that there is a power of gravity pertaining to all bodies, proportional to the several quantities of matter which they contain . . .

The force of gravity towards the several equal parts of any body is inversely as the square of the distance of places from the particles" (Newton, 1686).

If one particle (m') is at the origin and the other (m) is at a distance r, then the force is in radial direction, and it has a magnitude

$$F = \frac{Gmm'}{r^2} \tag{1.1}$$

The value of the gravitational constant in Eq. (1.1) is $G = 6.6743 \times 10^{-8}$ dyne·cm^2/g^2.

Strictly speaking, the masses that enter the force law (1.1) are the gravitational masses, which are the sources and the "receptors" of gravitation, in the same way that the electric charge is the source and the receptor of electromagnetic forces.

In principle, the gravitational mass is distinct from the inertial mass, which enters on the left side of the equation of motion, $ma = F$.

Experimentally, these two kinds of masses are found to be equal, and we will examine the experimental evidence for this equality in Section 1.6.

In the following discussion of gravitational fields and potentials (Sections 1.1–1.4), the masses are always gravitational.

If we adopt a naive interpretation of the force law (1.1), gravitation is action-at-distance: A mass at one point acts directly and instantaneously on another mass even though the other mass is not in contact with it.

Newton had serious misgivings about such a ghostly tug-of-war of distant masses and suggested that the interaction should be conveyed by some material medium.

The modern view is that gravitation, like electromagnetism and all other fundamental interactions, acts locally through fields: A mass at one point produces a field, and this field acts on whatever masses with which it comes into contact.

The gravitational field may be regarded as the material medium sought by Newton; the field is material because it possesses an energy density.

The description of interactions by means of local fields has the further advantage of leading to a relativistic theory in which gravitational effects propagate at finite velocity.

Instantaneous action-at-distance makes no sense as a relativistic theory because of the lack of an absolute time; what is instantaneous propagation in one reference frame need not be instantaneous in another.

Of course, in the case of static or quasi-static mass distributions, retardation effects are insignificant, and there is then no practical distinction between local interaction and action-at-distance.

In our Solar System, Newton's theory is an excellent approximation.

The condition for the validity of Newton's theory can be conveniently stated in terms of the potential energy $V(r)$, which for the inverse-square force (1.1) is

$$V(r) = -\frac{Gmm'}{r} \tag{1.2}$$

In general, we can say that relativistic effects will be small, provided that the potential energy of the moving particle is much less than the rest-mass energy and that the speed is much less than the speed of light.

For a mass m orbiting with speed v around a central mass m'. we can express these conditions as

$$|V(r)| \ll mc^2 \text{ and } v \ll c \tag{1.3}$$

where c is the speed of light.

Note that the condition on the potential energy is equivalent to $r \gg Gm'/c^2$.

Hence the deviations from Newton's theory are expected to be very small if the distance from the central mass is sufficiently large and the speed sufficiently low.

For the Sun,with a mass $m' = M_\odot \cong 2.0 \times 10^{33}$g, we have $Gm'/c^2 \cong 1.5$ km, and the condition $r \gg 1.5$ km is obviously very well satisfied, even for comets with a perihelion close to the surface of the Sun.

The gravitational constant G that appears in Eq. (1.1) is not known with the high precision of other fundamental constants.

Whereas the values of e and $\hbar$ are known to eight significant figures, the value of G is known to only five significant figures.

Measurements of G are difficult because of the extreme weakness of the gravitational force between masses of laboratory size.

The gravitational force between masses of planetary size is not weak, but this is of no help in determining G, because only the combination Gm' (where m' is the mass of the attracting body) appears in the equations of motion of bodies with purely gravitational interactions; hence, planetary observations cannot determine the separate values of G and m'.

Table 1.1 gives selected values of laboratory measurements of G.

Experimenter(s)	Year	Method	G (10^{-8} dyne $\cdot$ cm^2/g^2)
Cavendish	1798	Torsion-balance deflection	6.75(±5)**
Poynting	1891	Beam balance	6.70(±4)
Boys	1895	Torsion-balance deflection	6.658(±7)
Eötvös	1896	Torsion-balance period	6.66(±1)
Luther and Towler	1982	Torsion-balance period	6.6726(±5)
Gundlach and Merkowitz	2000	Torsion-balance acceleration	6.6742(±1)
Quinn et al.	2001	Torsion balance deflection	6.6756(±3)
Armstrong and Fitzgerald	2003	Torsion balance, compensated	6.6738(±3)
Schlamminger et al.	2006	Beam balance	6.6743(±1)

Later table…….

Year	G ($10^{-11}\cdot m^3\cdot kg^{-1}\cdot s^{-2}$)	Standard uncertainty
1969	6.6732(31)	460 ppm
1973	6.6720(49)	730 ppm
1986	6.67449(81)	120 ppm
1998	6.673(10)	1,500 ppm
2002	6.6742(10)	150 ppm
2006	6.67428(67)	100 ppm
2010	6.67384(80)	120 ppm
2014	6.67408(31)	46 ppm
2018	6.67430(15)	22 ppm

The values are listed in chronological order; the earlier ones are included for their historical interest, and the more recent values are the best available today.

Figure 1.1 shows the torsion balance used by Cavendish in his pioneering measurements of G late in the 18th century.

A beam with two small masses (B, B) is suspended from a thin fiber.

These small masses are gravitationally attracted by the two large lead spheres (W, W), and this results in a measurable deflection of the beam through some angle around the vertical.

From the known torsional constant of the fiber and the geometry of the balance, the gravitational constant can then be calculated.

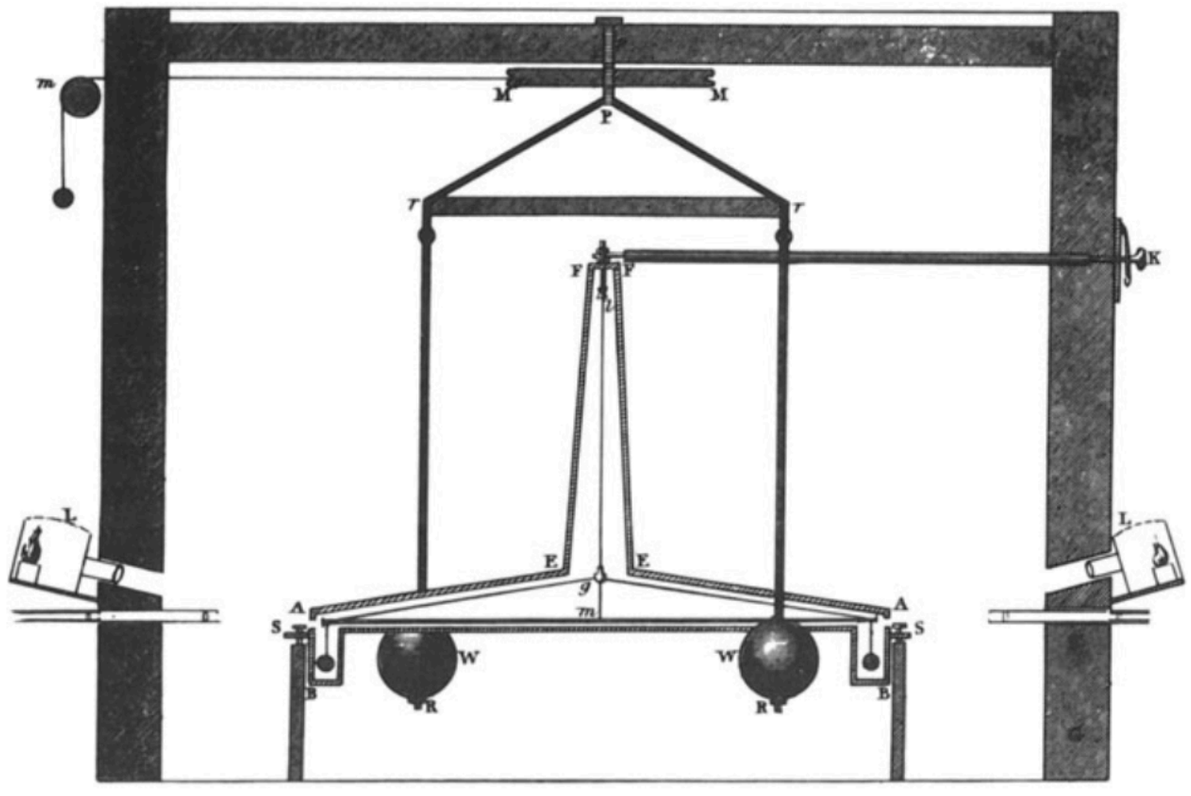


Fig. 1.1 The apparatus used by Cavendish. The large lead spheres (W, W) attract the small spheres (B, B) which are attached to the beam of the torsion balance. (From Cavendish, 1798)

The recent measurements by Gundlach and Merkowitz (2000) and by Schlamminger et al. (2006) have given the most precise value for G.

Surprisingly, these determinations agree almost exactly, although they were performed by entirely different methods.

Gundlach and Merkowitz used a small, delicate torsion balance with four "large" masses of 8 kg each mounted on a rotating turntable.

During each experimental run, the turntable was accelerated at exactly the rate needed to keep the small-mass beam at a fixed angular distance from the large masses, with the torsion fiber in equilibrium (no twist in the fiber).

This procedure eliminates "noise" from the gravitational background and errors arising from irregularities in the torsion contributed by the twisted fiber of the Cavendish arrangement.

In contrast, the apparatus of Schlamminger et al. was colossal, with two large masses of 7.5 metric tons, consisting of pure mercury in two large cylindrical tanks, placed alternatively below or above two test masses of copper of about 1 kg each.

The test masses were hung from the beam of an accurate beam balance, which registered the change in force on the test masses when the large masses were shifted from below the test masses to above them (see Fig. 1.2).

Mercury was selected as the material for the large masses because its uniform density permits accurate calculation of the gravitational force exerted by each large mass on each test mass.

The observed magnitude of the change in force between the two alternative configurations shown in Figure 1.2 then permits the evaluation of the gravitational constant G.

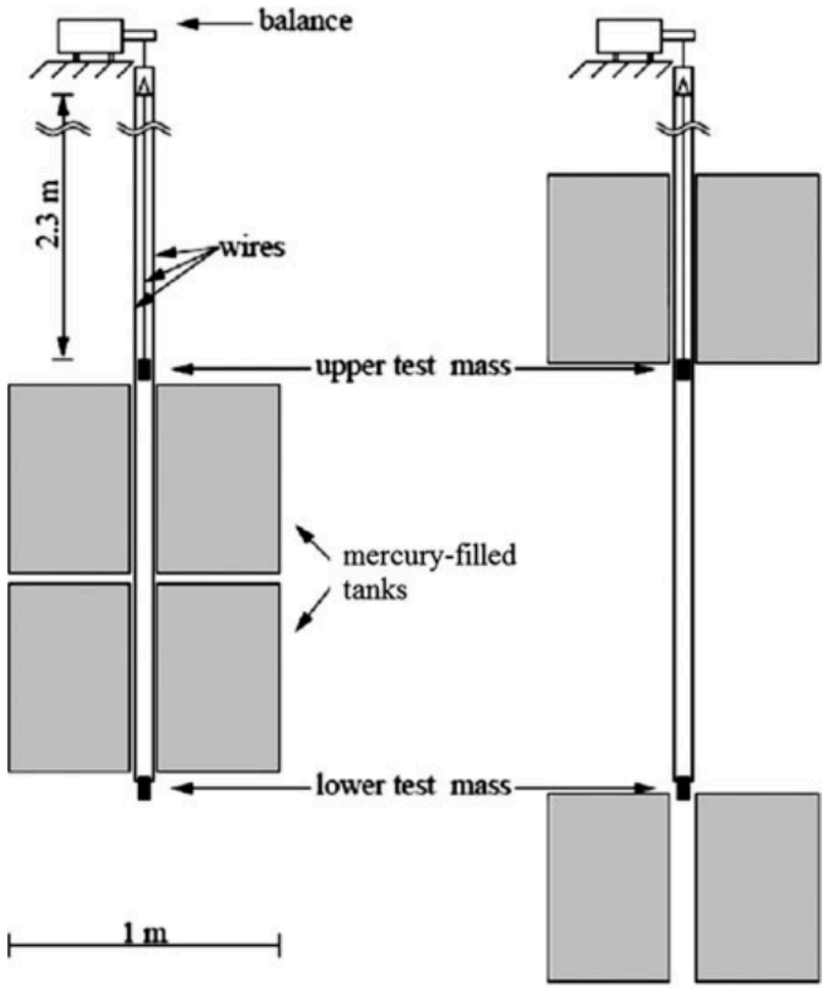


Fig. 1.2 The apparatus used by Schlamminger et al. The two test masses are suspended by wires from the balance arm at the top, and these wires pass through the axial holes of the two large mercury-filled tanks (gray). (a) In the first configuration, the two tanks are adjacent, and the gravitational pull is downward on the upper test mass, upward on the lower test mass. (b) In the second configuration, the two tanks are widely separated, and the pulls on the test masses are reversed. The beam balance detects this change of pull on the test masses. (From Schlamminger et al., 2006)

1.2 Tests of the inverse-square law

Is it possible that there are deviations from the inverse-square law at large distances or at small distances?

By "large distances" we mean distances of up to 10^4 or 10^5 light-years; such distances are large compared with the dimensions of the Solar System, but small compared with the typical dimensions of the universe.

At very large distances (more than 10^7 light-years), there may be cosmological deviations from the $1/r^2$ force (see Section 7.3).

These deviations are not our concern in the present context.

To investigate deviations from the inverse-square law, it is expedient to begin with the general mathematical constraints that relativistic field theory imposes on possible alternatives to the inverse-square law.

It is easiest to express these constraints in terms of the potential.

The inverse-square law has the special potential given by Eq. (1.2), whereas the general potential consistent with field theory turns out to be

$$V(r) = -\alpha \frac{Gmm'}{r} e^{-r/\lambda} \tag{1.4}$$

where α and λ are constants.

This is called a *Yukawa potential*; obviously, the $1/r$ potential (1.2) is a special Yukawa potential with $\alpha = 1$ and $\lambda = \infty$.

The constant λ is called the *range* of the potential – if the distance r appreciably exceeds λ, the potential and the force it produces become negligible.

Besides (1.4), the only other possibility is some superposition of several Yukawa potentials, which would mean that we are dealing with several gravitational fields.

In this case, the long-distance behavior of the net potential is dominated by the Yukawa potential of the largest λ, because this potential will linger to the largest distance.

If we focus on the Yukawa potential of the longest λ, what do the available observational data tell us about the value of this λ?

We know that the range of the gravitational force is very long – we know that our Galaxy as well as clusters of galaxies are held together by gravitation, which implies that the gravitational potential does not deviate much from $1/r$ out to distances of r $\cong$ size of cluster $\cong 10^{24}$ cm.

From this we can conclude that $\lambda > 10^{24}$ cm.

Incidentally, the value of λ is related to the mass of the graviton, a (hypothetical) particle of spin two, which is to gravitation what the photon is to electromagnetism.

According to relativistic quantum theory, the mass of the graviton is inversely proportional to the range of the Yukawa potential,

$$m_\Gamma = \hbar/\lambda c \tag{1.5}$$

If the gravitational force is inverse-square, the range of the force is infinite, and the mass of the graviton is zero.

If we rely on the observational limit $\lambda > 10^{24}$ cm, we obtain $m_\Gamma < 10^{-62}$ g for the graviton mass (Goldhaber and Nieto, 2010).

However, this mass limit rests on the assumption that the observed discrepancies between the observed orbital velocities of stars in the outer reaches of galaxies and the visible mass of these galaxies are accounted for by dark, invisible, mass (sometimes called “missing mass”).

This extra dark mass supposedly makes a large contribution to the total mass *M* of the galaxy, and thereby endows the orbiting stars with a larger orbital velocity, according to the usual relation between centripetal acceleration and gravitational force, $v^2/r = GM/r^2$.

The existence of such dark mass has been challenged, and several schemes have been proposed for modifications of the behavior of gravity at large distances.

For instance, the MOND scheme (*MO*dified *N*ewtonian *D*ynamics) proposed by Milgrom (1983) conjectures that at large distances the strength of the gravitational force is modified from $1/r^2$ to $1/r$, so in the outer reaches of galaxies the force of gravity remains much stronger than expected from Newton’s law.

As a purely ad hoc scheme, MOND has not found much favor among astronomers, who reckon that invisible, dark mass is the lesser of two evils.

In any case, for galaxies, the proposed modification of gravity would come into play only at distances of about 10^{22} cm, so for shorter distances we can still rely on the $1/r^2$ law.

It is of some interest to compare the 10^{24} or 10^{22}-cm limit with the analogous observational limit on the mass of the photon that can be set by examination of galactic magnetic fields.

Such magnetic fields are known to extend over distances of 10^{21} cm, and with this limit on λ we obtain $m\gamma < 10^{-59}$ g for the photon mass.

The mass of the graviton is constrained to a smaller value than the mass of the photon because gravitational fields are observed over larger distances than electromagnetic fields.

Because the value of λ for the long-distance part of the gravitational potential is certainly very large, and because a value $\lambda = \infty$ is consistent with our observational data, we will hereafter assume throughout this chapter that at large distances the gravitational potential reduces to the Newtonian $1/r$ potential, so there are no long-distance deviations from the inverse-square law.

There remains the question of possible deviations from the inverse-square law at short distances, generated by an additional Yukawa potential with a short range λ.

Such deviations arise in speculative theories involving extra dimensions, such as string theories.

These extra dimensions are supposed to be tightly curled up on a short scale of distance, so they remain unobservable.

The extra dimensions probably come into play only on a distance scale of 10^{-33} cm, the Planck distance that characterizes the scale of gravitational quantum fluctuations.

However, according to some radical conjectures, one or more of the extra dimensions might have a distance scale much larger than 10^{-33} cm, maybe as large as a fraction of a millimeter.

An extra dimension with such a length scale would escape observation if all the familiar particles are somehow confined to three dimensions (or, more precisely, four dimensions, if we count the time dimension), and only gravity spreads into the extra millimetric dimension.

In our 3-D space, the only observable effect would then be a modification of the behavior of gravity at millimeter distances.

For distances larger than the size of the extra dimension, the modification of gravity can be approximately represented by a superposition of the Newtonian $1/r$ potential and an extra Yukawa potential with a finite value of λ,

$$V(r) = -\frac{Gmm'}{r} - \alpha \frac{Gmm'}{r} e^{-r/\lambda} \tag{1.6}$$

In theories with an extra dimension, the constant λ is expected to be of the order of magnitude of the size of the extra dimension, and α is expected to be of the order of magnitude of 1.

Independently of the motivation underlying Eq. (1.6), experimenters often use this equation to parametrize deviations from the Newtonian potential, not only on millimetric scales but also on Solar-System scales.

Note that Eq. (1.6) gives an inverse-square force at large distances, but a complicated behavior for distances smaller than λ.

However, for $r \ll \lambda$ the potential reduces to

$$V(r) \cong -\frac{Gmm'}{r}(1 + \alpha) \tag{1.7}$$

so the force reverts to an inverse-square force, with a modified value $(1 + \alpha)G$ for the gravitational constant.

If the range λ of the Yukawa potential in Eq. (1.6) is of the order of, say, a few hundred meters, then the gravitational constant measured in laboratory experiments is $(1 + \alpha)G$, whereas the gravitational constant for interplanetary forces is G.

Limits on λ and on α can be extracted from a variety of orbital, geophysical, and laboratory observations and experiments.

Orbital Observations. High-precision measurements of the distances to Mercury, Venus, Mars, and Jupiter have been obtained by radar ranging, either with radar signals directly reflected by the surface of the planet or with signals returned by a transponder on a spacecraft during a flyby or while in orbit around the planet.

In combination with determinations of planetary orbital periods, obtained by traditional astronomical observations, the distance data permit a rigorous test of Kepler's third law and therefore a test of the inverse-square law.

A recent analysis of all the available data imposes a tight limit on the strength of the extra Yukawa potential, $|\alpha| < 10^{-8}$ for λ between 10^{10} and 10^{14} cm (Fischbach and Talmadge, 1999).

An analogous test can be performed for the orbits of the Moon or of artificial satellites around the Earth.

The distance to the Moon has been measured with high precision by laser ranging, by means of laser pulses reflected by corner reflectors placed on the Moon during the Apollo 11 mission.

Such precise measurements have also been performed on the LAGEOS artificial satellite.

The lunar laser-ranging data show no detectable deviations from the inverse-square law and set a tight limit of $|\alpha| < 10^{-10}$ for $\lambda \approx 10^{10}$ cm (Fischbach and Talmadge, 1999; also reviews by Adelberger, Heckel, and Nelson, 2003; Adelberger et al., 2009; and Newman, Berg, and Boynton, 2009).

Geophysical measurements. Geophysical investigations of the inverse-square law hinge on a method for the determination of G first proposed by Airy in 1856.

This method exploits the variation of the acceleration of gravity with depth below the surface of the Earth (or height above the surface).

If we descend into a deep mine shaft, we find that g varies with depth.

For a uniform-density sphere, g would decrease linearly with depth.

However, the Earth is not of uniform density, and g at first increases with depth and then decreases.

For illustrative purposes, assume that the mass distribution of the Earth is spherical, with a density $\varrho(r)$ and a mass $M(r)$ enclosed within the radius r.

According to the familiar Gauss law (which applies to gravitation as it applies to electrostatics), the acceleration $g(r)$ depends only on the mass enclosed within the radius r,

$$g(r) = G\frac{M(r)}{r^2} \tag{1.8}$$

and

$$\begin{aligned}\frac{dg}{dr} &= -2G\frac{M(r)}{r^3} + \frac{G}{r^2}\frac{dM(r)}{dr} \\ &= -\frac{2g(r)}{r} + \frac{G}{r^2}4\pi r^2\rho(r)\end{aligned} \tag{1.9}$$

With this equation, the value of G can be calculated from the measured values of $g(r)$ and dg/dr, provided we know the density ϱ.

Equation (1.9) is only a rough approximation; for an accurate determination of G via this method, we must also take into account the rotation of the Earth and its ellipsoidal shape.

The Airy method cannot achieve the precision of laboratory determinations of G.

However, it can be exploited to test the inverse-square law, as follows.

Find some region where the density ϱ is known, and measure g as a function of depth in the ground; then calculate G from g and dg/dr, by means of Eq. (1.9) or, rather, by means of the accurate version of this equation.

If the result of this determination of G agrees with the laboratory value $G = 6.6743 \times 10^{-8}$ dyne $\cdot$ cm^2/g^2, then the result verifies the inverse-square law; if not, then it disproves the inverse-square law.

Attempting to apply the Airy method, experimenters have measured gravity as a function of depth in mine shafts (Stacey et al., 1987), in boreholes in the ground (Thomas and Vogel, 1990), in the Greenland icecap (Ander et al., 1989; Zumberge et al., 1990), and underwater in the ocean (Stacey and Tuck, 1981; Zumberge et al., 1991).

In a variant of the Airy method, experimenters have also measured gravity as a function of height on TV transmitter towers several hundred meters high (Eckhardt et al., 1988; Thomas et al., 1989).

Unfortunately, such geophysical tests of the inverse-square law are bedeviled by the presence of underground density variations.

In all these experiments, the investigators seek to detect a deviation from the inverse-square law by comparing their measured values of g with the values calculated from the inverse-square law.

However, the calculations hinge on explicit or implicit assumptions about the homogeneity of the underground material, and it is almost always easy to construct models of slightly inhomogeneous mass distributions that account for the measured data *without* invoking any deviation from the inverse-square law (Parker and Zumberge, 1989).

The measurements of the 1980s were mostly motivated by a proposal by Fischbach et al. (1986), who resuscitated an earlier discarded proposal by Lee and Yang for a "fifth force" proportional to baryon number.

In contrast to Lee and Yang – who had assumed that their baryon force was a $1/r^2$ force, with a $1/r$ potential – Fischbach et al. assumed that their baryon force was based on a Yukawa potential.

Such a baryon force would produce two observable effects: It would alter the behavior of the force with distance [as in Eq. (1.6)], and it would cause inequalities in the free-fall accelerations of bodies toward the ground, because samples of equal masses but different baryon numbers would experience different net forces.

The fifth-force proposal stirred up considerable interest, especially when some measurements of weight as a function of height on towers suggested a deviation from inverse-square.

This deviation was later found to be an illusion arising from problems with the data analysis, and the fifth force was finally laid to rest by comparisons of different mass samples by means of torsion balances, which showed that there was no effect attributable to baryon number (Adelberger et al., 2009; Gundlach, Schlamminger, and Wagner, 2009).

Laboratory Measurements. A simple way to test the inverse-square law is to compare the results of determinations of G by different experimenters.

Most of these determinations were made with torsion balances.

If the force between the masses deviates from the inverse-square law, then the result of a determination of G will depend on the size of the torsion balance.

Cavendish used a large balance, with a beam of about 2 m and a distance of more than 10 cm between the attracting masses; modern versions of the experiment used beams as small as 2 cm and a correspondingly smaller distance between the attracting masses.

The agreement between such determinations of G suggests that there are no substantial deviations from inverse-square.

However, in view of the rather large experimental uncertainties in the determinations of G, the comparison does not yield any stringent limits (de Boer, 1984).

Better limits on deviations from the inverse-square law have been obtained by experiments specifically designed for this purpose.

An elegant experiment by Spero et al. (1980) used a torsion balance to explore the force field inside a long cylindrical shell (see Fig. 1.3).

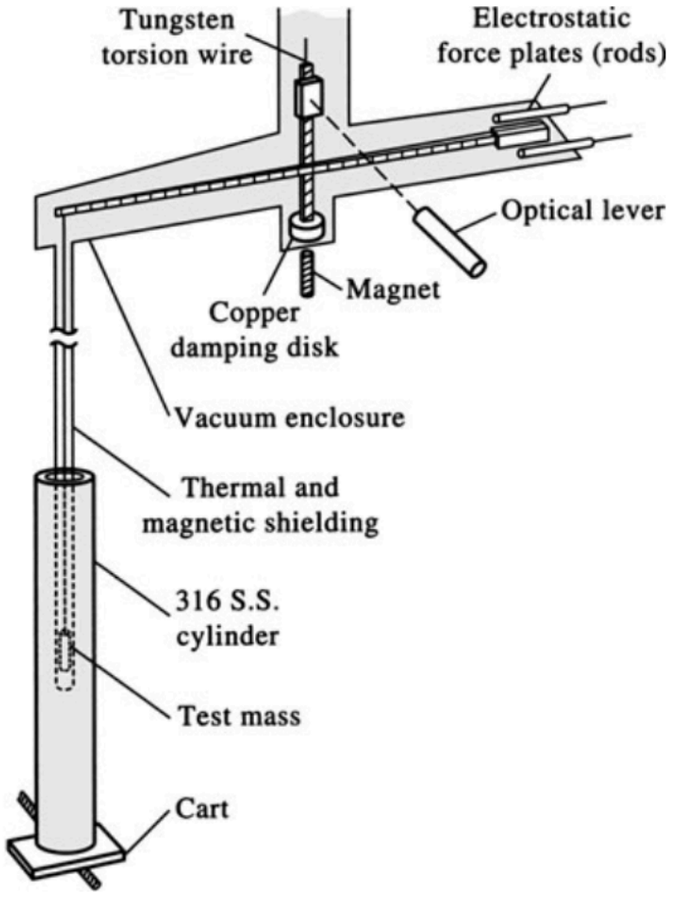


Fig. 1.3 Torsion balance with one mass suspended in the interior of a long cylindrical shell. (From Spero et al., 1980)

If, and only if, the inverse-square law is valid, the force that such a cylindrical shell exerts on a small spherical mass in its interior is exactly zero.

In the experiment, the small mass in Fig. 1.3 was moved back and forth relative to the cylindrical shell, to see whether it experiences any force when near the wall of the cylinder. The absence of any detectable force set a limit of about $|\alpha| < 10^{-4}$ for $\lambda \approx$ a few centimeters.

Similar results were obtained by torsion-balance experiments that compared the force exerted by a small mass placed near a torsion balance with the force exerted by a larger mass placed farther from the torsion balance (Chen, Cook, and Metherell, 1984).

Somewhat larger ranges of λ, reaching somewhere above 10 m, were explored with a "gravity gradiometer" that directly tested that the gradients in the gravitational field of a mass are those appropriate to a $1/r^2$ force (Hoskins et al., 1985).

Less stringent limits on $|\alpha|$, but for larger hypothetical values of λ, were obtained by experiments performed with hydroelectric pumped-storage reservoirs.

The water level of such reservoirs often rises or falls by tens of meters in just a few hours, and the change of gravity that this produces in the region above the water depends on α and λ.

The change of gravity can be measured with a beam balance that has one of its pans above the water level and the other pan below water level, all in a long waterproof tube (Stacey et al., 1987).

Alternatively, the change of gravity can be measured with a high-precision gravimeter, that is, a delicate spring balance (Muller et al., 1990).

Several recent experiments were designed to search for Yukawa potentials with values of λ of a millimeter or less, which are the values of greatest interest for theories with extra dimensions.

An experiment by Kapner et al. (2007) used a torsion balance of a special design (see Fig. 1.4) in which the beams holding the small and the large masses of the Cavendish balance are replaced by plates with circular holes around their circumferences.

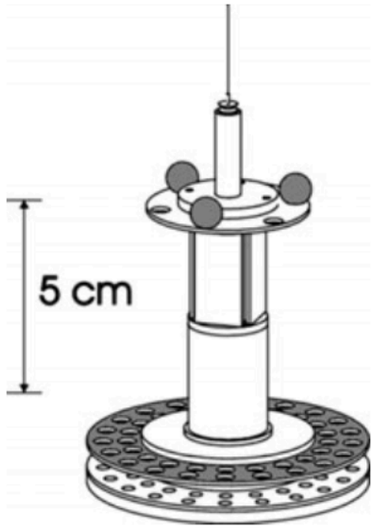


Fig. 1.4 In this torsion balance the upper perforated plate (gray) is attached to a cylinder, which is suspended from the torsional fiber (a short segment of fiber can be seen at the top). The three spherical balls are used for fine adjustments of the instrument. (From Kapner et al., 2007)

The upper plate is suspended from a torsional fiber, but is placed very close to the lower plate (0.05 mm, in some experimental runs).

Whenever the holes in the upper, suspended, plate are not aligned with those in the lower plate, the masses in the interstices between the holes in the upper and lower plates attract each other and exert a detectable torque on each other (this torque can be conveniently described as due to an effective repulsion of the holes).

However, below the bottom plate, there is a second, hidden bottom plate (its edge is barely visible in Fig. 1.4), also with circular holes.

This second bottom plate has its holes aligned with the interstices of the first bottom plate; furthermore, this second bottom plate is more massive, in proportion to the square of the distance from the upper, suspended, plate.

The net result is that the bottom plates in combination exert (almost) no torque on the suspended plate.

But this cancellation of the effects of the two bottom plates fails if the force is not inverse square, and thus the torsion balance is able to detect deviations from the inverse-square law.

This experiment established that $|\alpha| < 10^{-2}$ for $\lambda > 0.2$ mm, from which it can be concluded that the size of the extra dimension, if any, is smaller than that.

The experiment also set limits on Yukawa potentials with shorter ranges, but these limits are not very stringent, because the sensitivity of the torsion balance is poor at shorter ranges.

For instance, the limit is $|\alpha| < 10^{4}$ for $\lambda = 0.01$ mm.

Similar limits were obtained in an experiment with a miniaturized beam balance in the form of a cantilever, 0.25 mm long, micromachined in a crystal of silicon (Smullin et al., 2005).

1.3 Gravitational potential

The Newtonian gravitational force obeys the principle of linear superposition: The gravitational force exerted by a system of particles is the vector sum of the individual forces of the particles.

If N particles are located at positions $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots \mathbf{x}_N$, then the force these exert on a particle of mass m located at $\mathbf{x}$ is given by

$$\mathbf{F}(\mathbf{x}) = -Gm \sum_{i=1}^{N} \frac{m_i}{|\mathbf{x} - \mathbf{x}_i|^3} (\mathbf{x} - \mathbf{x}_i) \tag{1.10}$$

To this force there corresponds a potential energy

$$V(\mathbf{x}) = -Gm \sum_{i=1}^{N} \frac{m_i}{|\mathbf{x} - \mathbf{x}_i|} \tag{1.11}$$

with

$$\mathbf{F}(\mathbf{x}) = -\nabla V(\mathbf{x}) \tag{1.12}$$

In component notation, we can write Eq. (1.12) as

$$F^k = -\frac{\partial V}{\partial x^k}, \text{ with } x^1 \equiv x, x^2 \equiv y, x^3 \equiv z \tag{1.13}$$

The *gravitational field*, which we regard as the carrier of the interaction, is defined as force per unit mass, which equals the acceleration of gravity,

$$\mathbf{g}(\mathbf{x}) \equiv \frac{1}{m}\mathbf{F}(\mathbf{x}) \tag{1.14}$$

The corresponding *gravitational potential* is defined as the potential energy per unit mass,

$$\Phi(\mathbf{x}) \equiv \frac{1}{m}V(\mathbf{x}) = -\sum_{i=1}^{N}\frac{Gm_i}{|\mathbf{x}-\mathbf{x}_i|} \tag{1.15}$$

This definition makes the potential negative, as expected for an attractive force. The gravitational potential is sometimes defined with a sign opposite to that in Eq. (1.15), but we prefer to choose our signs by analogy with electrostatics.

For a continuous mass distribution, the gravitational potential is

$$\Phi(\mathbf{x}) = -\int \frac{G\rho(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}d^3x' \tag{1.16}$$

where $\varrho(\mathbf{x}')$ is the mass density.

Equation (1.16) implies that $\Phi(\mathbf{x})$ obeys the Poisson equation

$$\nabla^2\Phi(\mathbf{x}) = 4\pi G\rho(\mathbf{x}) \tag{1.17}$$

Problems in the theory of the Newtonian gravitational potential involve exactly the same mathematics as in the theory of the electrostatic potential.

Therefore we will not spell out all of the details of the following derivations.

One important difference between gravitation and electrostatics is that the mass density, as opposed to the charge density, can never be negative.

This implies that there can be no shielding of gravitational fields analogous to the shielding of electric fields by conductors.

By ingenious geometric arguments, Newton proved that a spherically symmetric mass distribution behaves in the same way as a point particle located at its center, in that it produces the same field in its exterior and it responds in the same way to fields produced by external sources.

This result, known as *Newton's theorem*, can be obtained very directly by appealing to the uniqueness theorem and the mean-value theorem of potential theory.

The potential energy of a body placed in a gravitational potential $\Phi(\mathbf{x})$ (produced by given external sources) is

$$V = \int \rho(\mathbf{x})\Phi(\mathbf{x})d^3x \tag{1.18}$$

where $\varrho(\mathbf{x})$ is the mass density of the body.

The gravitational self-energy of a continuous mass distribution $\varrho(\mathbf{x})$ is

$$U = \frac{1}{2}\int \rho(\mathbf{x})\Phi(\mathbf{x})d^3x \tag{1.19}$$

where $\Phi(\mathbf{x})$ is the potential produced by the mass distribution itself (the factor of 1/2 is needed to eliminate double counting of the potential energies of pairs of mass elements in the distribution).

This gravitational self-energy can be expressed in the alternative forms

$$U = -\frac{1}{2}\int G\frac{\rho(\mathbf{x})\rho(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}d^3x d^3x' \tag{1.20}$$

and

$$U = \int \frac{(\nabla\Phi)^2}{8\pi G} d^3x + \int \rho\Phi d^3x \tag{1.21}$$

This last equation has an interesting interpretation: The quantity $(\nabla\ \Phi)^2/8\pi G$ may be regarded as the energy density of the gravitational field, and $\varrho\Phi$ may be regarded as an interaction energy density of field and matter.

In the case of electrostatics, all of the electric energy may be regarded as field energy, with an energy density $\mathbf{E}^2/8\pi$ (in cgs units).

In the case of gravity, this is not possible because field energy is positive, and hence something negative must be added to it so as to obtain a negative total energy as given by Eq. (1.20).

1.4 Gravitational multipoles; quadrupole moment of the Sun

The gravitational potential has the simple form $-GM/r$ only in the space surrounding a mass distribution with spherical symmetry.

An arbitrary mass distribution produces a potential

$$\Phi(\mathbf{x}) = -\int \frac{G\rho(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} d^3x' \tag{1.22}$$

If the point $\mathbf{x}$ is outside of the region that contains the mass (see Fig. 1.5) then we can construct a multipole expansion for the potential by using the Taylor-series expansion of $1/|\mathbf{x} - \mathbf{x}'|$ about $\mathbf{x}' = 0$:

$$\frac{1}{|\mathbf{x}-\mathbf{x}'|} = \frac{1}{[(x-x')^2+(y-y')^2+(z-z')^2]^{1/2}}$$
$$= \frac{1}{r} + \sum_k \frac{x'^k x^k}{r^3} + \frac{1}{2}\sum_{k,l}\left(3x'^k x'^l - r'^2\delta_k^l\right)\frac{x^k x^l}{r^5} + \cdots \tag{1.23}$$

where $r = (x^2 + y^2 + z^2)^{1/2}$.

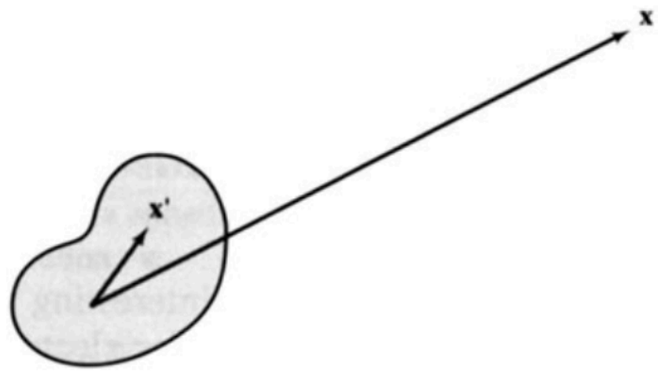


Fig. 1.5 The shaded region contains some mass distribution. The point **x** is outside of this region.

The integral (1.22) can then be written as

$$\Phi(\mathbf{x}) = -\frac{GM}{r} - \frac{G}{r^3}\sum_k x^k D^k - \frac{G}{2r^5}\sum_{k,l} Q^{kl}x^k x^l + \cdots \tag{1.24}$$

where

$$M = \int \rho(\mathbf{x}')d^3x' \tag{1.25}$$

and

$$D^k = \int x'^k \rho(\mathbf{x}')d^3x' \tag{1.26}$$

$$Q^{kl} = \int \left(3x'^k x'^l - r'^2\delta_k^l\right)\rho(\mathbf{x}')d^3x' \tag{1.27}$$

Here $\delta_k^{\ l}$ is the unit tensor (or Kronecker delta), $\delta_k^{\ l} = 1$ if $k = l$, and $\delta_k^{\ l} = 0$ if $k \neq l$.

The vector quantity D^k is the mass dipole moment.

If the origin of coordinates is chosen to coincide with the center of mass, then $D^k = 0$ (no mass dipole); we will usually assume that this is so.

The quantity Q^{kl} is the mass quadrupole tensor.

Equation (1.24) shows that whenever the quadrupole tensor is non-zero, the potential will contain a term $\propto 1/r^3$, and hence the force will deviate from the inverse-square law by an extra term $\propto 1/r^4$.

Most extended mass distributions have a quadrupole tensor; the obvious exception is a mass distribution with spherical symmetry, for which the quadrupole tensor is always zero.

The Earth's polar and equatorial diameters differ by about 3 parts in 10^3.

This deviation from spherical shape produces a quadrupole term in the gravitational potential, which causes perturbations in the elliptical Kepler orbits of satellites.

The main perturbation is a precession of the Kepler ellipse, that is, a slow rotation of the ellipse around the axis of the Earth (see Fig. 1.6).

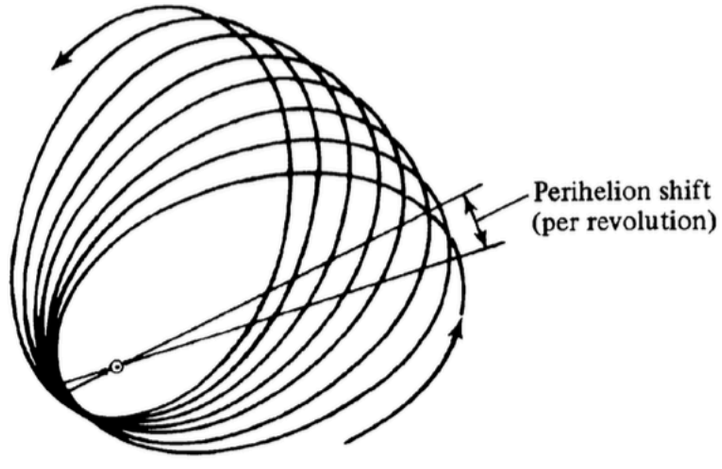


Fig. 1.6 Kepler orbit with precession.

Such observed perturbations of the orbits of satellites have been used for precise determinations of the multipole moments and the mass distribution in the Earth.

The Sun is almost a perfect sphere; its polar and equatorial diameters differ by at most a few parts in 10^5.

However, even such a small difference between the polar and equatorial radii, and the consequent small quadrupole term in the potential, could have important consequences when we attempt to compare the observed perihelion precession of the planets with the theoretical predictions.

According to Newton's theory, the perihelion precession receives a contribution from the quadrupole term and also a contribution from the gravitational perturbations that the planets exert on each other.

This latter contribution is actually the largest by far; for instance, for Mercury, the contribution to the perihelion precession from the quadrupole term is at most a few arcseconds per century, whereas the contribution from interplanetary perturbations is about 500 arcseconds per century.

According to Einstein's theory of general relativity, there is an extra contribution of 43 arcseconds to the perihelion precession of Mercury; this extra precession involves relativistic modifications of the gravitational field and of the equation of motion.

To test Einstein's theory, we want to isolate this extra contribution by subtracting the other two contributions from the observed precession.

The (large) contribution from interplanetary perturbations poses no difficulties in principle – it can be calculated quite precisely.

When this contribution is subtracted from the observed precession, the remainder is 43 arcseconds, in excellent agreement with Einstein's theory.

Thus, if there also were a substantial contribution from a quadrupole term generated by the Sun, the result of the subtraction would be in disagreement with Einstein's theory.

Because the Sun rotates, we expect it to have a small equatorial bulge, or an oblateness.

The observed rotation period at the surface is about 25 days, and under the assumption that the interior rotates at the same rate, we expect that the Sun's equatorial diameter exceeds the polar diameter by about 1 part in 10^5.

Such an oblateness would give the Sun a small quadrupole moment, which can be conveniently characterized by defining a dimensionless parameter J_2,

$$J_2 \equiv -\frac{Q^{33}}{2M_\odot {R_\odot}^2} \tag{1.28}$$

where $M_\odot$ and $R_\odot$ are the mass and the radius of the Sun, respectively, and the z-axis is taken to coincide with the polar axis.

If the Sun were rotating uniformly, the value of J_2 would be about 1×10^{-7}, and the corresponding contribution to the perihelion precession of Mercury would be only a few hundredths of an arcsecond per century; this is smaller than the uncertainty of the observed value of the perihelion precession and therefore would not affect the agreement with Einstein's theory.

Unfortunately, some direct measurements of the shape of the Sun's disk as seen by telescope (Dicke and Goldenberg, 1967) seemed to show an unexpectedly large oblateness and large quadrupole moment, which suggested that the interior of the Sun rotates much faster than the surface.

Other measurements gave conflicting results, and the issue was not resolved until the 1990s when a new method for the determination of the quadrupole moment became available with the discovery of oscillations of the solar surface.

Measurements of the light output, the solar shape, and the Doppler shift of spectral lines all indicate that the body of the Sun oscillates in diverse modes, with periods of several minutes.

A large amount of data on solar oscillations have been collected by telescopes on the ground and by the SOHO (*SO*lar and *H*eliospheric *O*bservatory) satellite, which carries a sophisticated spectrometer that maps the oscillation speeds on the solar surface by precise measurements of the Doppler shifts of spectral lines.

The analysis of these modes of oscillation (solar seismology) has been exploited for an indirect determination of J_2.

If the Sun did not rotate, the modes characterized by spherical harmonics Y_l^m with the same given harmonic index l and different azimuthal indices m would be degenerate – they would have the same frequency of vibration.

But in a rotating Sun, the frequencies of these modes are split by an amount depending on the rate of rotation.

Comparisons of these frequencies therefore permit the calculation of the rate of rotation of the solar interior and consequently the calculation of J_2.

Such calculations depend on the details of the model used for the interior of the Sun, and recent results give a value of $J_2 = 2.2 \times 10^{-7}$ (Redouane et al., 2004).

With this value of J_2, the solar quadrupole moment makes only an insignificant contribution to the perihelion precession, leading to excellent agreement between the observed perihelion precession of Mercury and Einstein's theory.

The oscillations of the solar surface do not involve large mass motions, and they there- fore do not produce any detectable time-dependence of the mass multipole moments.

However, in other astrophysical systems, large mass motions involving oscillations or rotations can produce significant time-dependent multipole moments.

Such time- dependent multipole moments are of much interest in the relativistic theory of gravitation because they act as sources of gravitational radiation.

Oscillating or rotating dumbbells (see Fig. 1.7) are simple examples of time-dependent quadrupoles.

The oscillating or rotating masses of such dumbbells can be regarded, respectively, as crude models of an oscillating supernova remnant or a binary star system.

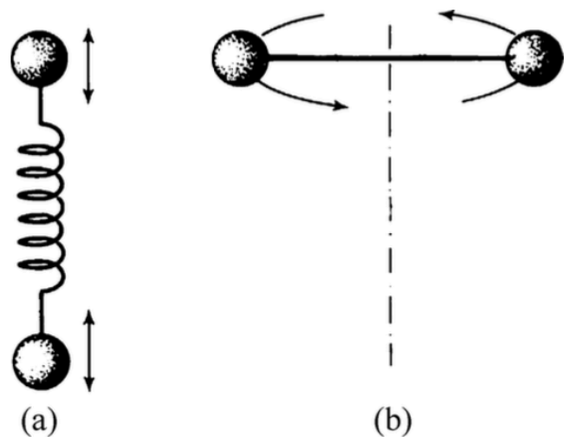


Fig. 1.7 (a) Oscillating masses with a time-dependent quadrupole moment. (b) Rotating masses with a time-dependent quadrupole moment.

Note that the values of the elements of the quadrupole tensor depend on the choice of the origin of coordinates.

Suppose that the coordinates x^k have their origin at the center of mass.

If we shift the origin and introduce new coordinates

$$x'^k = x^k + b^k \tag{1.29}$$

where b^k is a constant, then the quadrupole tensors in the old and new coordinates are related by

$$Q'^{kl} = Q^{kl} + \left(3b^k b^l - b^2 \delta^l_k\right) M \tag{1.30}$$

Equation (1.30) shows that a change of origin modifies Q^{kl} by only an additive constant, and thus the time-dependent part of Q^{kl} is independent of the choice of origin.

This result will be useful later in our study of gravitational radiation (Chapter 5).

1.5 Inertial and gravitational mass

It is implicit in Newton's law of universal gravitation that the mass that acts as the source or the receptor of gravitation is the same as the mass that determines the inertia.

Accordingly, the equation of motion of a pointlike particle in a given gravitational potential is

$$m\frac{d^2x^k}{dt^2} = -m\frac{\partial\Phi}{\partial x^k} \tag{1.31}$$

Here, the mass on the left side of the equation determines the inertia of the particle, and the mass on the right side determines the strength of the gravitational force.

If we cancel the masses in this equation, we obtain

$$\frac{d^2x^k}{dt^2} = -\frac{\partial\Phi}{\partial x^k} \tag{1.32}$$

This says that, in a given gravitational field, all pointlike particles fall with the same acceleration.

For historical reasons, we call this statement *Galileo's principle of equivalence*.

This principle implies that to some extent gravitational forces behave in the same way as the pseudo-forces that result from the use of a non-inertial reference frame.

In an accelerated reference frame, free particles appear to accelerate spontaneously, and at a given point, all particles have exactly the same acceleration.

To be precise, we should also specify that all the particles are placed at the given point with the same initial velocity.

This is necessary because some pseudo-forces – such as the Coriolis force – are velocity dependent.

This restriction on the velocity is also needed for Galileo's principle, because it turns out that the effective "gravitational force" given by general relativity is velocity dependent.

The equivalence between inertial and gravitational effects can make it difficult to distinguish between the two.

Thus an astronaut in a freely falling spaceship will find it difficult to decide whether she is in free fall in a gravitational field, or in unaccelerated motion in a region far away from all fields, if she only performs experiments in the interior of her spaceship.

We must now ask, Is it correct to cancel the masses in Eq. (1.31)?

Are the factors of m on both sides really the same?

The mass that appears on the left side of Eq. (1.31) is the inertial mass.

For any arbitrary body, the *inertial mass* is defined by a procedure first proposed by Mach: We take the body and let it interact, somehow, with the standard kilogram.

Both the body and the standard will accelerate toward or away from one another.

Designating the acceleration of the body and of the standard by a and a_S, respectively, we can then define the inertial mass by

$$m_I = 1\,\text{kg} \times \frac{a_S}{a} \tag{1.33}$$

This gives the inertial mass of the body in kilograms.

The *gravitational mass* can be defined in a similar manner.

We take a standard body and define its gravitational mass to be one unit; for convenience we can use the standard kilogram as a standard for both inertial and gravitational mass.

We now place the arbitrary body at some distance r and let it interact gravitationally with the standard body.

The body will accelerate toward the standard.

We then define the gravitational mass of the body in terms of this acceleration and the distance between the bodies:

$$m_G \equiv \lim_{r\to\infty}\left(\frac{m_I a r^2}{1\,\mathrm{kg}\times G}\right) \tag{1.34}$$

This gives the gravitational mass in kilograms.

Note that since $m_I a$ equals the gravitational force, or the weight, and $1\ \mathrm{kg}\times G/r^2$ is the gravitational field g generated by the standard mass, Eq. (1.34) simply says that the gravitational mass is the weight divided by g.

Alternatively, we can use Eq. (1.33) to write Eq. (1.34) as

$$m_G \equiv \lim_{r\to\infty}\left(\frac{a_S r^2}{G}\right) \tag{1.35}$$

The limiting procedure $r \to \infty$ is needed in Eqs. (1.34) and (1.35) to eliminate the effects of multipole fields, which depend on the mass distribution of the two bodies.

Also, the limiting procedure $r \to \infty$ eliminates the effect of short-range forces (nuclear forces, van der Waals forces, and so on).

At large distances, only the gravitational and electrostatic forces will remain, and the latter can be eliminated by taking the precaution of keeping the standard body neutral.

If we take two identical copies of the standard of mass and let them fall toward each other, the acceleration of each serves to define the constant G:

$$G \equiv \lim_{r\to\infty}\left(\frac{a_{SS} r^2}{1\,\mathrm{kg}}\right) \tag{1.36}$$

With these precise definitions of m_I and m_G, the gravitational force between two particles is

$$F = -\frac{Gm_G m'_G}{r^2} \tag{1.37}$$

and the equation of motion is

$$m_I \frac{d^2x^k}{dt^2} = -m_G \frac{\partial \Phi}{\partial x^k} \tag{1.38}$$

Whether all particles fall with the same acceleration depends on whether all particles have the same value of m_I/m_G.

If this ratio is a universal constant, it must have the value $m_I/m_G = 1$ (the standard body has this value by definition).

The question is then, Is the equation

$$m_I = m_G \tag{1.39}$$

satisfied for all bodies?

We call this equality (1.39) *Newton's principle of equivalence of inertial and gravitational mass*, because Newton first considered the possibility of a difference between mass (inertial) and weight (gravitational) and tested their equality.

Before we turn to the experimental evidence, we remark that for a body of appreciable size the acceleration of the center of mass is given by

$$m_I \frac{d^2x^k}{dt^2} = -\int \rho_G(\mathbf{x}') \frac{\partial \Phi(\mathbf{x}')}{\partial x'^k} d^3x' \tag{1.40}$$

where the integral is over the volume of the body, and ϱ_G is the gravitational mass density

Extended bodies fall at the same rate as point particles only if $\partial \Phi / \partial x'^{k}$ is approximately constant over the volume of the body; in that case we can factor $\partial \Phi / \partial x'^{k}$ out from the integral, and we then obtain (1.38).

Hence Newton's principle of equivalence implies equal accelerations only for bodies of sufficiently small size placed in sufficiently homogeneous gravitational fields.

Bodies of different size and shape placed in an inhomogeneous field generally fall at different rates.

According to the derivation that led us to the equation of motion (1.32), we see that for pointlike particles the Galileo principle implies the Newton principle and vice versa.

This derivation hinges on the validity of Newtonian mechanics; that is, the gravitational fields must be weak ($GM/rc^2 \ll 1$) and the speed slow ($v \ll c$).

If these assumptions are not satisfied, then the Galileo and Newton principles must be regarded as complementary rather than equivalent: In general, the Galileo principle is a statement about a special class of bodies (pointlike particles) moving with arbitrary velocities in arbitrary fields, whereas the Newton principle is a statement about arbitrary bodies moving at low velocity in weak fields.

The experiments to be described in the next section constitute evidence for the Newton principle; but since they were performed with bodies of low velocities in weak fields, they lend only limited support to the Galileo principle

1.6 Tests of equality of gravitational and inertial mass

The earliest recorded experiments on the equality of rates of fall of different bodies were reported by Galileo.

He quickly recognized that dropping two bodies from a height was not the most convenient and accurate method for comparing their rates of fall.

Instead, he compared the oscillations of two pendulums:

> *I took two balls, one of lead and one of cork, the former being more than a hundred times as heavy as the latter, and suspended them from two equal thin strings, each four or five bracchia long. Pulling each ball aside from the vertical, I released them at the same instant, and they, falling along the circumferences of the circles having the strings as radii, passed thru the vertical and returned along the same path. This free oscillation, repeated more than a hundred times,*

We can interpret Galileo's observation to mean that the ratio m_G/m_I is the same for lead and for cork, to within a few parts in a thousand.

The first experiments specifically designed to test the equality of inertial mass ("quantity of matter") and gravitational mass ("weight") are due to Newton.

Like Galileo, he used two pendulums, but he was careful to compensate for friction:

> *I tried the thing in gold, silver, lead, glass, sand, common salt, wood, water, and wheat. I provided two equal wooden boxes. I filled the one with wood, and suspended an equal weight of gold (as exactly as I could) in the centre of oscillation of the other. The boxes, hung by equal threads of 11 feet, made a couple of pendulums perfectly equal in weight and figure, and equally exposed to the resistance of air: and, placing the one by the other, I observed them to play together forwards and backwards for a long while, with equal vibrations. And therefore (by Cor. I and VI, Prop. XXIV, Book II) the quantity of matter in the gold was to the quantity of matter in the wood as the action of the motive force upon all the gold to the action of the same upon all the wood; that is, as the weight of the one to the weight of the other. And by these experiments, in bodies of the same weight, one could have discovered a difference of matter less than the thousandth part of the whole. (Newton, 1686)*

Table 1.2 summarizes the experimental evidence that has accumulated in favor of the equality of inertial and gravitational mass since the days of Galileo and Newton.

These experiments were performed with a large variety of substances.

In setting upper limits on $|m_G - m_I|$, one of the substances used by the experimenter is taken as a standard for which $m_G = m_I$.

Table 1.2 Tests of the equality of m_I and m_G*

Experimenter(s)	Year	Method	$\lvert m_G - m_I \rvert / m_I$
Galileo	1610	pendulum	$<2 \times 10^{-3}$
Newton	1680	pendulum	$<10^{-3}$
Eötvös	1890	torsion balance	$<5 \times 10^{-8}$
Eötvös et al.	1905	torsion balance	$<3 \times 10^{-9}$
Dicke et al.	1964	torsion balance, Sun	$<3 \times 10^{-11}$
Braginsky and Panov	1971	torsion balance, Sun	$<9 \times 10^{-13}$
Koester	1976	free fall of neutron	$<3 \times 10^{-4}$
Keiser et al.	1979	floating mass	$<4 \times 10^{-11}$
Niebauer et al.	1987	free fall	$<5 \times 10^{-10}$
Adelberger et al.	1989	rotating torsion balance	$<1 \times 10^{-11}$
Schlamminger et al.	2008	rotating torsion balance	$<2 \times 10^{-13}$
MICROSCOPE	?	free fall in satellite	$<10^{-15}$?
STEP	?	free fall in satellite	$<10^{-18}$?

* References are given in Will (1993) and Adelberger et al. (2009).

The most precise results have been obtained with torsion balances, by a method first introduced by the Hungarian physicist Eotvos around 1890.

Eotvos began his experimental investigations on the equality of inertial and gravitational mass in response to a prize offered by the University of Gottingen; the high precision of his results so impressed the jury that they awarded him the prize, even though it was originally intended for a theoretical rather than an experimental investigation.

The apparatus of Eotvos is shown in Figs. 1.8 and 1.9.

Two pieces of matter, labeled “weight,” are attached to the arms of a torsion balance.

These weights are made of different substances, for instance, copper and platinum (Eotvos used platinum as the standard and compared other substances with it).



Fig. 1.8 Instrument used by Eötvös and his collaborators for the first gravity survey at Lake Balathon (1902). The instrument used to measure m_G/m_I was of the same type. (Courtesy Hungarian State Geophysical Institute “Eötvös Lorand,” Budapest)

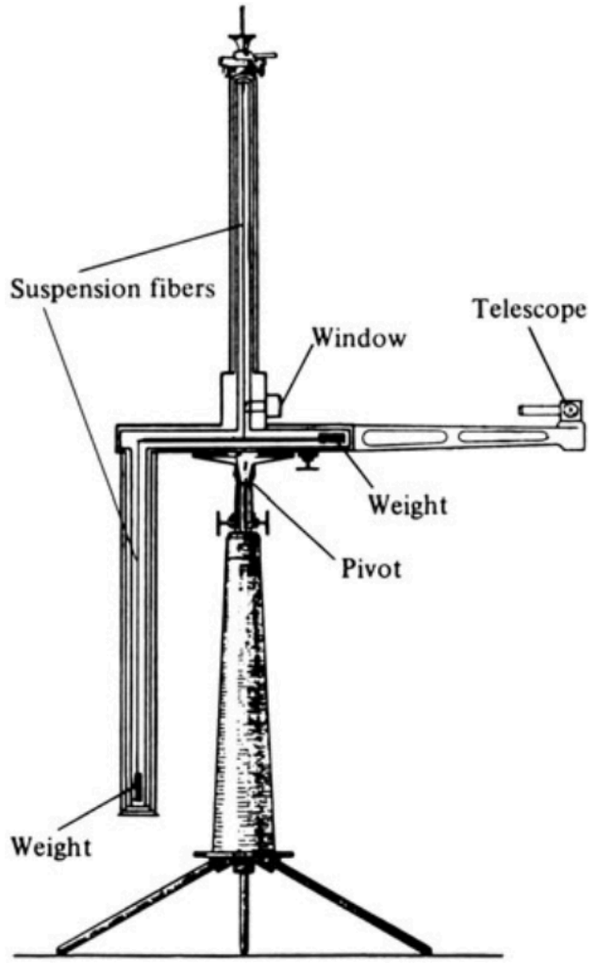


Fig. 1.9 Schematic diagram of the Eötvös apparatus. (After Eötvös, Pekar, and Fekete, 1922)

If m_G/m_I has different values for the two substances, then the balance will experience a torque.

To understand how this comes about, consider the forces that act on the masses.

Seen in the rotating reference frame of the Earth, the forces are of two kinds: There is the gravitational force $m_G\mathbf{g}$ exerted by the Earth and the centrifugal pseudo-force $m_I\mathbf{a}$ produced by the rotation of the Earth.

The quantity **g** is the acceleration of gravity (without centrifugal effects), and **a** is the centrifugal acceleration due to the Earth's rotation at the locality of the experiment. Figure 1.10 shows these forces. In this figure both masses are shown placed directly on the beam of the balance; the asymmetric suspension used in the apparatus of Eotvos is of no relevance to the present experiment (the suspension was originally designed for a different purpose, the detection of gradients in the gravitational field).

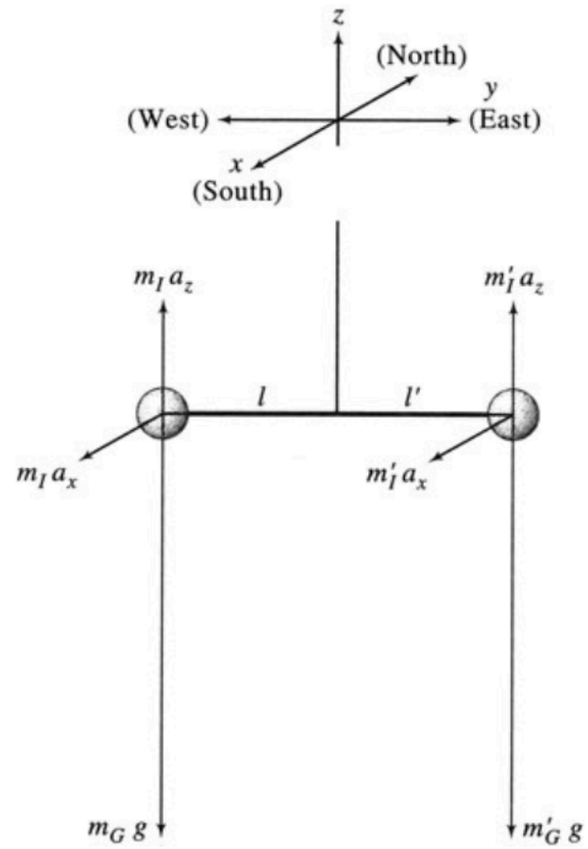


Fig. 1.10 Torsional pendulum suspended near the surface of the Earth at a latitude of about 45° N. Note that the centrifugal force vectors $m_I \mathbf{a}$ are shown exaggerated by a factor of about 100.

In Fig. 1.10, the direction of the z-axis is defined by the direction of **g**, not the direction of a plumb line (which would include centrifugal effects).

The beam of the balance points in the east-west direction.

The centrifugal force then has a vertical (opposite to **g**) component $m_I a_z$ and a horizontal component $m_I a_x$.

The torque about the z-axis is obviously

$$\tau = m_I a_x l - m'_I a_x l' \tag{1.41}$$

We can eliminate l' by using the equilibrium condition for rotation about the x-axis,

$$(m_G g - m_I a_z) l = (m'_G g - m'_I a_z) l' \tag{1.42}$$

and we obtain

$$\tau = m_I a_x l \left(1 - \frac{m'_I}{m_I}\frac{m_G g - m_I a_z}{m'_G g - m'_I a_z}\right) = m_I a_x l \frac{g(m'_G/m'_I - m_G/m_I)}{g m'_G/m'_I - a_z} \tag{1.43}$$

From this it is obvious that a torque exists if and only if $m_G/m_I \neq m'_G/m'_I$.

In equilibrium, the gravitational torque (1.43) is compensated by a torque produced by the suspension fiber.

The presence of the gravitational torque can be detected by rotating the entire apparatus by exactly 180° about the vertical axis (see pivot, Fig. 1.9).

If the equilibrium position of the beam was exactly along the east-west direction before this rotation, then it will be slightly off after the rotation; this change in equilibrium position occurs because turning the apparatus around changes the sign of the torque (1.43) [turning the apparatus around is mathematically equivalent to $l \to -l$ and $l' \to -l'$ in Eqs. (1.41) and (1.42)].

The method of Dicke (1964; Roll, Krotkov, and Dicke, 1964) also uses a torsion balance, but detects the torque, if any, produced by the gravitational force of the Sun and the centrifugal force of the Earth's motion around the Sun.

If, for the sake of simplicity, we imagine a torsion balance of the type shown in Fig. 1.10 suspended at the north pole of the Earth, then this centrifugal force is entirely horizontal.

Assuming equal arms ($l = l'$), the torque about the vertical axis will be

$$\tau = (m_G g - m_I a) l \sin\phi - (m'_G g - m'_I a) l \sin\phi \tag{1.44}$$

where g is the acceleration of gravity produced by the Sun at the place of the experiment, a is the centrifugal acceleration in the reference frame of the Earth orbiting the Sun, and ϕ is the angle between the beam of the balance and the Sun (to a sufficient approximation, $g = a$).

In Dicke's experiment, the beam is held stationary with respect to the Earth and the torque required to do this is measured.

It is obvious from Eq. (1.44) that under these conditions the torque oscillates with a 24-hr period as the Sun angle ϕ increases by 2π.

Any other effect ("noise") disturbing the experiment with a period different from 24 hr can be filtered out by Fourier analysis and discarded.

Note that Dicke's apparatus actually uses a somewhat more complicated torsion balance with a triangular beam holding three masses (see Fig. 1.11).

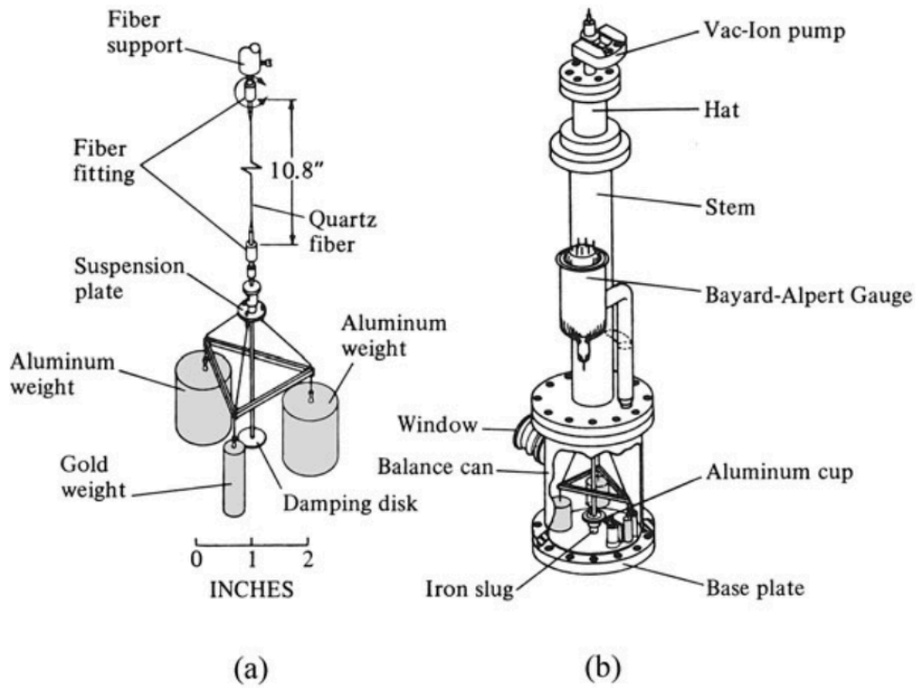


Fig. 1.11 The apparatus used by Dicke et al. (a) The torsion balance has a triangular beam with a mass attached at each vertex. (b) The balance is suspended in an evacuated vessel. (From Dicke, *Gravitation and the Universe*. Reprinted with permission of the American Philosophical Society.)

This arrangement makes the balance less sensitive to gradients in the gravitational field produced by nearby massive bodies, such as the body of an experimenter.

As a further precaution, measurements were carried out by remote control.

The experiment of Braginsky and Panov (1971) used the same technique, but the beam of their balance was star shaped, with eight arms, and the suspension fiber was much longer, which gave the balance a very low restoring torque.

The recent experiment by Schlamminger et al. (2008) of the Eot-Wash group (as well as several earlier experiments by members of the same group) combined some of the best features of the Eotvos and the Dicke experiments.

Like the Eotvos experiment, it seeks to detect the torque produced by the gravitational force produced by the Earth (which is stronger than that produced by the Sun), but it adopts Dicke's method of filtering out unwanted signals by Fourier analysis.

For this purpose, the torsion balance is mounted on a turntable that continuously rotates the entire apparatus about the vertical axis, with a period of about 17 min.

Any noise with a period different from the 17-min period can then be filtered out and discarded.

This experiment verified the equality of the accelerations of free fall for different materials to within ± 4 parts in 10^{13} (for samples of beryllium vs. aluminum).

Concomitantly, it also gave the best evidence against a long-range fifth force coupled to baryon number, proton number, or neutron number, by setting a limit on the strength of such a force of about 10^{-11} times the strength of the gravitational force (Adelberger et al., 2009; Gundlach, Schlamminger, and Wagner, 2009).

The delicate torsion balances used in these experiments are subject to many environmental disturbances.

To avoid such disturbances and attain much higher sensitivity, tests with freely falling bodies in satellites are being developed.

The French MICROSCOPE experiment will use differential accelerometers attached to coaxial cylinders of different metals to test the equality of accelerations to within ±1 part in 10^{15}.

The proposed Italian GG (*G*alileo *G*alilei) experiment is intended to improve on this by a factor of 100, and the proposed American STEP (*S*atellite *T*est of *E*quivalence *P*rinciple) is expected to improve on that by another factor of 10, by using cryogenic SQUID circuits to detect displacements as small as 10^{-13} cm.

All these satellites would be kept effectively drag free by means of thrusters that compensate for atmospheric friction.

The masses would therefore remain in free fall for a long time, and even a very small relative acceleration between the two masses in a pair would ultimately yield a detectable displacement.

The experiments listed in Table1.2 have tested the ratio m_G/m_I for a wide variety of materials.

Thus, Eotvos compared copper, water, copper sulfate, asbestos, snakewood, and so on with platinum; the later, more precise experiments tested gold, platinum, aluminum, copper, titanium, and beryllium.

The results of these tests indicate that $m_G = m_I$ to within the experimental errors.

The nucleus of the atom contains appreciable amounts of energy in its electric, magnetic, “strong,” and “weak” interaction field, and the results can therefore be interpreted to mean that different kinds of energy contribute to the gravitational mass of a system in the same amount as they contribute to the inertial mass; that is, $m_G = m_I$ is valid not only for the samples of materials tested in the experiments but also for the various kinds of energy contained in these materials – all kinds of energy seem to gravitate in the normal way.

Table 1.3 lists different kinds of energy and their contribution to the gravitational mass as calculated from the experimental result of Braginsky and Panov (the slightly higher sensitivity of the experiment of Adelberger et al. does not improve the limits listed in this table, because the slightly higher measurement sensitivity is nullified by a reduced energy-sensitivity of the test samples, aluminum and beryllium for Adelberger, and aluminum and platinum for Braginsky).

Table 1.3 Different forms of energy as sources of gravity*

Form of Energy	$\lvert m_G - m_I \rvert / m_I$
Rest mass, protons + electrons	0 (by definition)
Rest mass, neutrons	$<8 \times 10^{-12}$
Strong fields in nucleus	$<3 \times 10^{-10}$
Electric fields in nucleus	$<4 \times 10^{-10}$
Magnetic fields in nucleus	$<2 \times 10^{-7}$
Weak fields in nucleus	$<5 \times 10^{-3}$
Kinetic energy of nucleons	$<10^{-9}$
Gravitational energy in Earth	$<5 \times 10^{-4}$

* Adapted from Will (1993), with corrections.

The first column in Table 1.3 lists forms of energy that – according to special relativity – contribute to the inertial mass of the test samples.

The second column shows how these different forms of energy gravitate.

For the purposes of this table, a proton paired with an electron has been adopted as the standard body for which $m_G \equiv m_I$; in the samples of neutral atoms used in these experiments, protons and electrons always exist in equal numbers, so the experiments cannot detect a difference between their gravitational accelerations.

The amount of energy in the "weak" interaction in the nucleus (which is the kind of interaction responsible for β decay) is much smaller than the other energies, and the limit we can set on the behavior of this kind of energy is correspondingly worse.

Note that the numbers in Table 1.3 hinge on the assumption that there are no fortuitous cancellations among deviations from $m_G = m_I$.

Free-fall experiments have been performed with individual atoms by an ingenious technique that exploits the quantum-mechanical interference of atoms with themselves ("atom interferometry").

By this method, it proved possible to measure the free-fall acceleration of individual cesium atoms coasting up and down in an "atomic fountain" to a few parts in 10^9 (Peters et al. 2001).

Within this experimental error, the free-fall acceleration of these atoms was found to be in agreement with the free-fall acceleration of a macroscopic test mass measured by conventional interferometry operating with light waves.

Although this experiment was a remarkable tour de force, the precision falls short of what is attainable by an Eotvos balance.

Free-fall experiments have also been performed with individual neutrons and with individual electrons in the gravitational field of the Earth, but the precision of these experiments is relatively low.

For example, measurement of the downward sag of a horizontal beam of low-velocity neutrons emerging from a nuclear reactor has shown that the downward acceleration of these free neutrons agrees with the standard local value of the Earth's gravitational acceleration to within at least 3 parts in 10^4 (Koester, 1976).

Table 1.3 also includes a result for gravitational energy, which was obtained by different means.

The Eotvos experiments do not permit a direct test of the hypothesis that gravitational energy contributes to the gravitational mass, because the ostensible macroscopic amounts of gravitational self-energy in masses of laboratory size are much too small to affect these experiments.

Theoretical considerations suggest that the rest masses of electrons, protons, and neutrons include large amounts of gravitational self- energy, but we do not know how to calculate these self-energies (for the implications of this, see the later discussion).

If we want to discover whether gravity gravitates, we must examine the behavior of large masses, of planetary size, with significant and calculable amounts of gravitational self-energy.

Treating the Earth as a continuous, classical mass distribution (with no gravitational self-energy in the elementary, subatomic particles), we find that its gravitational self-energy is about 4.6×10^{-10} times its rest-mass energy.

The gravitational self-energy of the Moon is smaller, only about 0.2×10^{-10} times its rest-mass energy.

If gravitational self-energy does not contribute in the normal way to the gravitational mass, then the Earth and the Moon would fall at different rates in the gravitational field of the Sun.

The difference in the rates of fall is effectively equivalent to a uniform extra force field pulling the Moon toward the Sun (if gravitational energy gravitates less than normal) or away from the Sun (if gravitational energy gravitates more than normal).

Such an extra force leads to a distortion of the orbit of the Moon relative to the Earth, a distortion called the *Nordvedt effect*.

As Fig. 1.12 shows, the orbit is elongated, or polarized, in the direction of the Sun.

Although the distortion effect is small, very precise measurements of the Earth-Moon distance have been performed by the laser- ranging technique already mentioned in Section 1.2, with a pulse of laser light sent from the Earth to the Moon and reflected back to the Earth by the corner reflectors installed on the Moon during the Apollo mission.

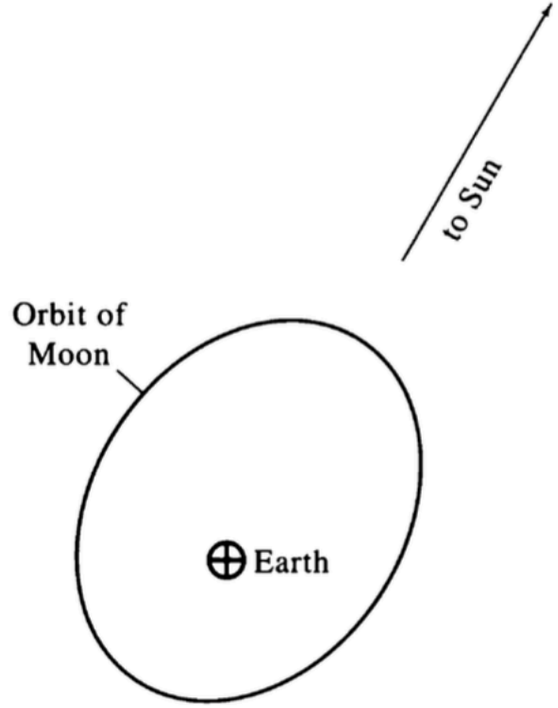


Fig. 1.12 Distortion of the orbit of the Moon produced by a hypothetical uniform extra force field directed toward the Sun. The distortion is shown exaggerated for the sake of clarity (if gravitational energy does not gravitate at all, the lengthening of the orbital diameter would amount to about 5 parts in 10^6).

Measurements of the travel time of the pulse determine the distance to within an uncertainty of a centimeter, and recent improvements are reducing this to a millimeter.

If the uncertainty is taken as 1 cm, the analysis of the orbital data places a direct limit of 5×10^{-4} on the fractional difference between the contributions of gravitational energy to the inertial and the gravitational mass.

Thus, these experiments indicate that gravitational energy gravitates in the normal way.

In setting these limits on how the strong, electromagnetic, weak, and gravitational energies gravitate, we have ignored the self-energies locked up within the rest masses of electrons, protons, and neutrons.

This is a rather questionable attitude, since quantum theory suggests that all these particles contain large amounts of strong, electromagnetic, weak, and gravitational self-energies.

These particles would therefore not be expected to gravitate in the normal way, unless all these forms of energy gravitate in the normal way.

However, from the equal rates of fall of electrons, protons, and neutrons we cannot extract quantitative conclusions for the rates of fall of the self-energies locked up in the rest masses, because we have no way of calculating the magnitudes of these self- energies.

A naive calculation of the, say, gravitational self-energy of the electron gives an infinite value; this, of course, proves only that the calculation is wrong and that our understanding of the quantum dynamics is faulty.

In the absence of any precise knowledge of the amount of self-energy contained in particles, we have to impose a severe restriction on any putative theory of gravitation: The theory must give equal rates of fall for electrons, protons, and neutrons regardless of how much strong, electromagnetic, weak, or gravitational self-energy is contained in the rest masses.

Thus, the theory must make all forms of energy gravitate in the same way.

If the theory does not satisfy this fundamental requirement of universal free fall, the theory is incomplete – it cannot make any definite prediction for the rates of fall of electrons, protons, and neutrons.

So far only one theory of gravitation has been able to fulfill this requirement: Einstein's theory of general relativity.

All the other theories contrived to compete with Einstein's fail to meet the fundamental requirement of universal free fall – they all display a Nordvedt effect, with gravitational energy falling at a different rate [for a catalog of alternative theories of gravitation, see Will (1981)].

A few of these theories contain *several* adjustable coupling constants, and by juggling the values of these constants, the Nordvedt effect can perhaps be made zero for some kinds of bodies; but this does not guarantee that the Nordvedt effect will be zero for every kind of body.

Only Einstein's theory gives us such a guarantee.

Until the inventors of these alternative theories discover some way of calculating the gravitational self-energies of electrons, quarks, gluons, and so on, they cannot predict the rates of fall of these particles.

Because of this serious defect of all the alternative theories gravitation, we will ignore them throughout this book.

As concerns antimatter, we do not yet have any direct experimental evidence on the value of m_G/m_I, although the AEGIS experiment now being developed at CERN is expected to test the rate of fall of a beam of low-energy antiprotons.

In the meantime we have to rely on some circumstantial evidence in favor of a normal gravitational interaction of antimatter.

It was pointed out by Schiff (1959) that if positrons had a negative gravitational mass ("fall upward"), then we would expect nuclei to have anomalous ratios of m_G/m_I.

The reason for this is that a nucleus is surrounded by a fluctuating cloud of virtual electron-positron pairs, a condition known as vacuum polarization.

The number of pairs depends on the charge distribution of the nucleus, and if the gravitational mass of positrons were negative, nuclei surrounded by more pairs would appear to have a lower gravitational mass.

This effect would lead to a difference in m_G/m_I between platinum and aluminum amounting to a few parts in 10^7, in clear contradiction to Eotvos's experimental results (Holzscheiter, 1990).

A much more precise statement about gravitational interactions of antimatter can be made by examining the famous case of the decay of the neutral K^0 meson. The decay of this meson is very sensitive to quantum-mechanical phase factors, and even a small change in the (small) phase factor $\exp(-imG_\Phi t/\hbar)$, arising from the gravitational energy of the meson in the gravitational potential Φ of the Galaxy, would presumably lead to drastic changes in the decay process (Good, 1961). From the absence of such phase effects in the decay, we can conclude that the gravitational masses of the meson K^0 and the antimeson $\bar{K}^0$ differ at most by a few parts in 10^{10} (the inertial masses of the K^0 and the $\bar{K}^0$ are exactly equal; this is a general theorem for every particle and the corresponding antiparticle).

The observed equality of gravitation and inertia for all forms of energy means that any system, containing any combination of the different kinds of energy, will have equal gravitational and inertial masses.

Thus, we will now adopt Newton's principle of equivalence, $m_G = m_I$, as an exact relation that our gravitational theory must satisfy.

1.7 Tidal forces

Consider a spacecraft in orbit around the Earth.

We assume that the spacecraft is not rotating; to be precise, we assume that the spacecraft maintains its orientation relative to the fixed stars.

The spacecraft is in free fall.

The astronauts find themselves in a zero-g environment, they are weightless, they feel no gravitational force.

The reference frame attached to the spacecraft simulates an inertial reference frame: A test particle at rest relative to the spacecraft remains at rest, and a test particle in motion remains in motion with uniform velocity.

However, for an observer at rest relative to the fixed stars, the spacecraft will be in accelerated motion, and such an observer will refute the astronauts' claim that they are in an inertial frame by saying that both the spacecraft and the test particle are falling at the same rate and that the astronauts are being fooled by appearances.

Note that the elimination of the Earth's gravitational field by free fall is impossible unless the equivalence of inertial and gravitational mass holds for all bodies; any body with $m_I \neq m_G$ would appear to accelerate spontaneously with respect to the freely falling spacecraft.

We now ask, Are the gravitational effects of the Earth completely eliminated by free fall?

Is there some local experiment that permits the astronauts to find out that they are falling in a gravitational field rather than staying at rest in some region far away from any attracting mass?

The answer is that the astronauts can detect the gravitational field by the tidal effects it produces.

If the astronauts place a drop of liquid at the center of their spacecraft, they will find that this drop is not exactly spherical, but has two bulges (Fig. 1.13).

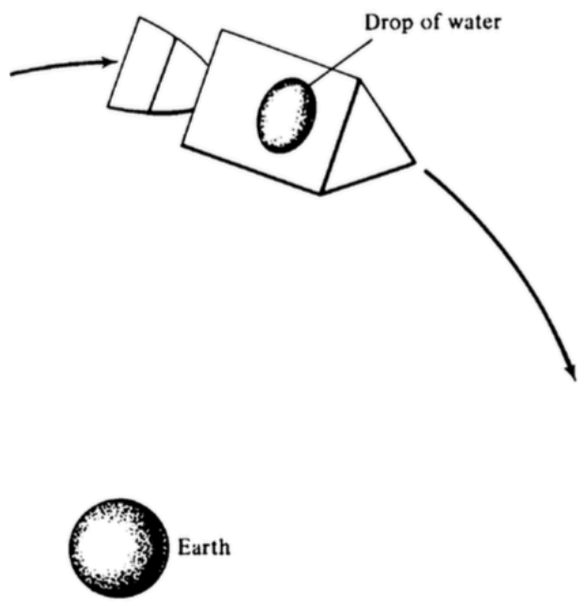


Fig. 1.13 A spacecraft in orbit. The drop of water floating in this spacecraft is distorted by the gravitational pull of the Earth.

One bulge points toward the Earth, one away.

Since, in the absence of external forces, surface tension would make the drop spherical, the deviation from a sphere indicates the existence of a gravitational field.

The bulges result from the inhomogeneity of the gravitational field: The end of the drop nearer the Earth is pulled too much by gravitation, and the other end is not pulled enough.

The force that produces the bulges is called the *tidal force*.

We can calculate it as follows: Consider a reference point moving in free fall; take this point as the origin of a freely falling coordinate system with the z-axis parallel to the radial line (Fig. 1.14).

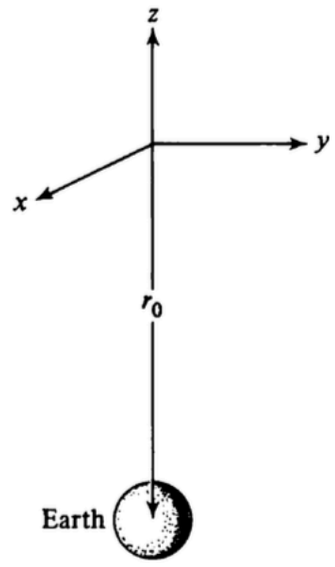


Fig. 1.14 The coordinate system x, y, z belongs to a reference frame in free fall, whose origin is (instantaneously) at a distance r_0 from the Earth.

A particle at position $(0, 0, z)$ in this reference frame experiences a gravitational acceleration $-GM/(r_0 + z)^2$, where r_0 is the distance from the center of the Earth to the origin of our coordinates and M is the mass of the Earth.

Since the origin has acceleration $-GM/r_0^2$, the acceleration of the particle relative to the origin is, in the limit of small z,

$$-\frac{GM}{(r_0+z)^2}+\frac{GM}{{r_0}^2}\cong 2z\frac{GM}{{r_0}^3} \tag{1.45}$$

Hence, relative to our origin the particle moves as though subjected to a force

$$f_z = 2z\frac{GMm}{{r_0}^3} \tag{1.46}$$

This is a tidal force.

Note that it is directly proportional to the distance of the particle from the origin, and it is repulsive.

For a particle at a position $(0, y, 0)$, the tidal force points in the y-direction and is

$$f_y = -y\frac{GMm}{{r_0}^3} \tag{1.47}$$

This is a tidal restoring force, toward the origin.

This force arises from the slight change of angle of the radial line toward the center of the Earth; at (0, 0, 0) the radial line is parallel to the z-axis, but at $(0, y, 0)$ it makes an angle y/r_0 with the z-axis, so the gravitational force acquires a y-component.

For a particle at the position $(x, 0, 0)$, there is a similar restoring force,

$$f_x = -x\frac{GMm}{{r_0}^3} \tag{1.48}$$

If the particle has simultaneous x, y, and z displacements, then all three forces – (1.46), (1.47), and (1.48) – are of course present simultaneously.

Note that these expressions for the tidal force only apply if the displacements x, y, z are small compared with r_0.

From these results it is obvious that a drop of liquid – which we can regard as a system consisting of many particles, each pushed or pulled by the tidal force – will be stretched in the radial direction and compressed in the transverse direction.

In general, given an arbitrary gravitational field, the tidal force in a reference frame whose origin is in free fall can be expressed as follows:

$$f^k = \sum_l x^l \frac{\partial F^k(0)}{\partial x^l} = -\sum_l x^l m \frac{\partial^2 \Phi(0)}{\partial x^l \partial x^k} \tag{1.49}$$

where F^k is the gravitational force on the particle and Φ the potential, and where the derivatives are evaluated at the origin ($x^l = 0$).

Equation (1.49) is of course valid only near the origin (small x^l).

We will call the quantity

$$R^k{}_{0l0} = -\frac{1}{mc^2}\frac{\partial F^k(0)}{\partial x^l} = \frac{1}{c^2}\frac{\partial^2 \Phi(0)}{\partial x^l \partial x^k} \tag{1.50}$$

the *tidal force tensor*.

The extra factor mc^2 and the extra indices 00 have been introduced for later convenience.

It turns out that the nine quantities $R^k{}_{0l0}$ are components of the Riemann curvature tensor $R^\mu{}_{\nu\alpha\beta}$.

As we will see in Section 6.6, the latter tensor describes the tidal field in the four-dimensional spacetime of general relativity.

In terms of this tensor, the tidal force that acts on a particle placed at x^l (with $x^l \ll r_0$) is then

$$f^k = -mc^2 \sum_l R^k{}_{0l0} x^l \tag{1.51}$$

and the acceleration produced by this force is

$$\frac{d^2x^k}{dt^2} = -c^2 \sum_l R^k{}_{0l0} x^l \tag{1.52}$$

We will again encounter this kind of tidal acceleration in Section 6.6, where it will emerge as the relative acceleration between two particles in a general-relativistic gravitational field.

In the special case of the gravitational field of a spherical mass, the tensor $R^k{}_{0l0}$ calculated from Eq. (1.50) can be written as follows, in the form of a matrix,

$$R^k{}_{0l0} = \frac{GM}{r_0{}^3 c^2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix} \tag{1.53}$$

Here the first index (k) indicates the row and the second (l) the column of the matrix.

Note that according to Eq. (1.50) $R^k{}_{0l0}$ is always represented by a symmetric matrix, and therefore it can always be diagonalized [as done in Eq. (1.53), by selecting the z- coordinate in the radial direction; see Fig. 1.14].

The tidal force given by Eqs. (1.46), (1.47), and (1.48) satisfies the identity

$$\frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} + \frac{\partial f_z}{\partial z} = -\frac{GMm}{r_0{}^3} - \frac{GMm}{r_0{}^3} + \frac{2GMm}{r_0{}^3} = 0 \tag{1.54}$$

that is, the tidal force field has zero divergence.

This result holds in general for gravitational fields in empty regions of space.

We can prove this by using Eq. (1.49):

$$\sum_k \frac{\partial f^k}{\partial x^k} = -\sum_{k,l} \frac{\partial}{\partial x^k}\left(x^l m \frac{\partial^2 \Phi(0)}{\partial x^k \partial x^l}\right)$$
$$= -m\sum_{k,l} \delta^l_k \frac{\partial^2 \Phi(0)}{\partial x^k \partial x^l} = -m\sum_k \frac{\partial^2 \Phi(0)}{\partial x^k \partial x^k} = 0 \qquad (1.55)$$

The last equality holds true because the gravitational potential satisfies the Laplace equation $\nabla^2 \Phi = 0$ [see Eq. (1.17) with $\varrho = 0$].

Note that the argument shows that the equations $\nabla \cdot \mathbf{f} = 0$ and $\nabla^2 \Phi = 0$ are equivalent.

Hence a gravitational field in vacuo may be characterized by the vanishing of the divergence of the tidal force.

In the presence of a mass density ϱ, the divergence of the tidal force equals

$$\sum_k \frac{\partial f^k}{\partial x^k} = -4\pi m G\rho \qquad (1.56)$$

As we will see later, the tidal forces given by general relativity are somewhat more complicated (these forces are contained in the equation of geodesic deviation of Section 6.6).

They depend on the velocity of the reference point and on the velocity of the particle with respect to that reference point; furthermore, the gravitational field is not that given by Newton.

However, the general-relativistic tidal force agrees with the above Newtonian value in the limit of weak gravitational fields ($Gm/rc^2 \ll 1$) and low speed ($v \ll c$).

The “tidal forces” take their name from the tides they generate on the oceans of the Earth.

These tidal forces are exerted by the Moon and, to a lesser extent, by the Sun; we will ignore the Sun for now.

For a simple estimate of the height of the tide, consider the Earth as a spacecraft in free fall toward the Moon.

According to Eqs. (1.46), (1.47), and (1.48), the tidal force exerted by the Moon on a particle on the surface of the Earth, with coordinates x, y, z (see Fig. 1.15), is

$$f_z = 2z\frac{GMm}{{r_0}^3}, \quad f_y = -y\frac{GMm}{{r_0}^3}, \quad f_x = -x\frac{GMm}{{r_0}^3} \tag{1.57}$$

where M is now the mass of the Moon and r_0 the Earth-Moon distance.

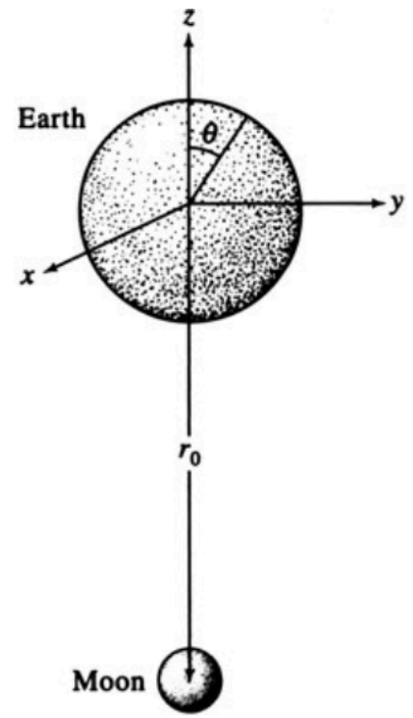


Fig. 1.15 The coordinate system x, y, z is attached to the Earth, which is in free fall towards the Moon.

The tidal potential that gives this force is

$$\begin{aligned} -\frac{GMm}{{r_0}^3}\left(z^2 - \frac{1}{2}y^2 - \frac{1}{2}x^2\right) &= -\frac{GMm}{{r_0}^3}\left(-\frac{1}{2}R^2 + \frac{3}{2}z^2\right) \\ &= -\frac{GMmR^2}{{r_0}^3}\left(\frac{3\cos^2\theta - 1}{2}\right) \end{aligned} \tag{1.58}$$

where $R = (x^2 + y^2 + z^2)^{1/2}$ is the radius of the Earth and θ is the angle shown in Fig. 1.15.

A particle on the surface of the Earth also feels the gravitational attraction of the Earth.

Since we are concerned with particles that remain near the surface, we can write the potential produced by the Earth as mgh where h is the height above the surface (to be precise, we should write $R + h$ in Eq. (1.58), rather than R, but for $h \ll R$, this makes little difference).

The net gravitational potential is then

$$V(h, \theta) = mgh - \frac{GMmR^2}{r_0{}^3}\left(\frac{3\cos^2\theta - 1}{2}\right) \tag{1.59}$$

For a rough estimate of the height of the tide, we assume that the water is in static equilibrium.

Then the surface must be an equipotential, so

$$gh - \frac{GMR^2}{r_0{}^3}\left(\frac{3\cos^2\theta - 1}{2}\right) = \text{constant} \tag{1.60}$$

This gives us the height h of the water as a function of θ.

High tides occur at θ = 0 and at θ = π; low tides at θ = π/2 and θ = 3π/2.

The difference between high and low tide (the "tidal range") is given by

$$\Delta h = \frac{GMR^2}{gr_0{}^3}\left[\frac{3\cos^2\theta - 1}{2}\right]_{\theta=\pi/2}^{\theta=0} = \frac{3GMR^2}{2gr_0{}^3} \tag{1.61}$$

The numerical value of the tidal range given by Eq. (1.61) is 53 cm (the Sun produces an effect about half as large).

This is a reasonable order-of-magnitude estimate for the tide in the open sea.

But our results are not accurate because the water does not reach static equilibrium with the tidal force.

As a consequence of the rotation of the Earth about its axis, the tidal field appears to move once around the Earth every lunar day, and we must seek the response of water to an oscillating force of period 12 hr 25 min.

The time between successive passages of the Moon through a given meridian plane is 12 hr 25 min.

If the axis of rotation of the Earth were perpendicular to the orbit of the Moon, then the two maxima in the tidal potential would simply travel in the equatorial plane of the Earth and produce two equal tides per day, so the period of the tidal driving force would be 12 hr 25 min.

Actually, the axis of the Earth makes an angle with the orbit of the Moon and therefore, at any given latitude, the two daily maxima are not equal.

The force tidal has a strong Fourier component at 12 h 25 min (“semidiurnal tide”) and a weaker component (“diurnal tide”) at 24 hr 50 min.

The Sun contributes other components.

The water in ocean basins has certain natural periods of oscillation, in much the same way as water in a bathtub has a well-defined period for sloshing back and forth; we are therefore dealing with the problem of an oscillator subjected to a periodic driving force.

The response of the oscillator depends critically on how the frequency of the driving forces compares with the natural frequency.

If the natural frequency is much larger than the driving frequency, the oscillations remain in phase with the driving force and quasistatic conditions prevail (the restoring force is almost in static equilibrium with the driving force).

However, for the case of some ocean basins there exist natural modes of oscillation with frequencies smaller than 1 cycle per 12 hr 25 min; this leads to an inverted tide, out of phase with the driving force.

Basins that have a natural frequency very close to the driving frequency can develop enormous tides by resonance.

For example, tides with a range of 15 m are found in the Bay of Fundy.

1.8 Tidal field as a local measure of gravitation

We now look somewhat more closely at the detection of gravitational fields by tides.

If astronauts in orbit wish to detect the gravitational field of the Earth by measuring the tide produced by the Earth on a drop of water, they will find it desirable to use a very large drop of water.

Equation (1.61) shows that the height of tide increases with the size of the drop.

This suggests that if the astronauts have been ordered to confine their experiments to the interior of a sufficiently small spacecraft, then they will not be able to detect the tide, nor the gravitational field.

But suppose that the astronauts use as a measure of the tidal field not the height Δh, but rather the ratio Δh/R, which characterizes the prolateness of the drop.

According to Eq. (1.61) we have tidal

$$\frac{\Delta h}{R} = \frac{3GMR}{2gr_0^3} \tag{1.62}$$

where M is now the mass of the Earth, R the radius of the drop, and g the acceleration of gravity that the drop of water produces at its own surface.

Strictly speaking, Eq. (1.62) is applicable only to tides in a thin layer of liquid covering the surface of a dense rigid sphere.

For a drop made entirely of liquid, the change in the gravitational field of the drop produced by the change of shape must be taken into account; this modifies (1.62) by a numerical factor of no great interest to us.

Note that we are assuming that the drop is held together by gravitation, rather than by surface tension; this assumption is not realistic for small drops of water, but we will pretend that some very powerful detergent has been added to the water so as to make its surface tension very small.

In terms of the density of the fluid, we can write $g = (4\pi R^3/3)\varrho G/R^2$ and

$$\frac{\Delta h}{R} = \frac{9}{8\pi}\frac{M}{\rho r_0{}^3} \tag{1.63}$$

This shows that the shape of the tidal ellipsoid is independent of its size.

Even in the limit $R \to 0$, the tidal deformation remains.

We can therefore regard the prolateness of the tidal ellipsoid as a local measure of the gravitational force.

We can then construct a neat graphical representation of a gravitational tidal field: At each point of space, draw the tidal ellipsoid that gives the shape of a drop of fluid (with zero surface tension) in free fall at that point.

Such a graphical representation of the tidal field of a spherical mass is shown in Fig. 1.16.

We can also construct an alternative graphical representation by appealing to Eq. (1.55).

According to this equation, the tidal force field in vacuo has zero divergence; this implies that the tidal force can be represented by lines of force in exactly the same way as an electric field in vacuo.

Figure 1.17 shows the pattern of lines of force near points in the space surrounding a spherical mass.

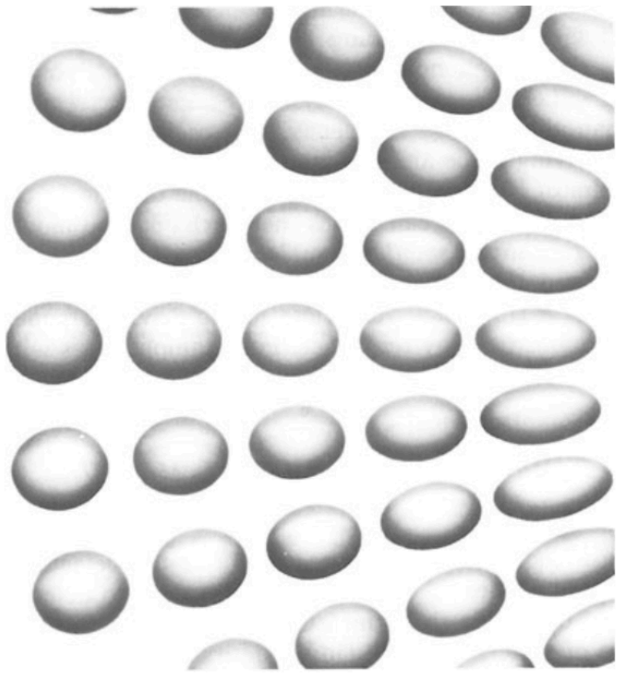
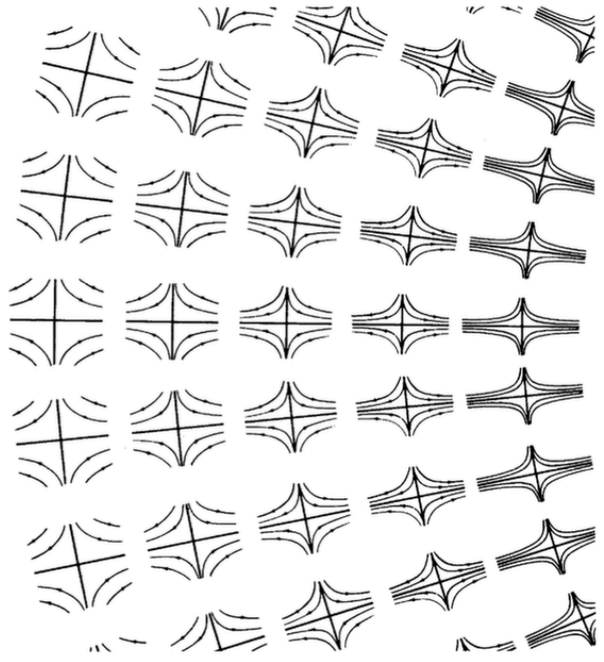


Fig. 1.16 The tidal field of a spherical mass represented by tidal ellipsoids. The mass is beyond the right edge of the drawing.

Fig. 1.17 The tidal field of a spherical mass represented by lines of force. Each of the patterns of lines of force is valid only near its center. Note that this 2-D drawing does not correctly indicate the density of lines of a 3-D field. The density along the radial axis is actually twice as large as the density along the transverse axis.



Note that we have to draw a separate pattern for the vicinity of each point because our expression for the tidal field is valid only for small displacements.

Although the graphical representations shown in Figs. 1.16 and 1.17 are mathematically equivalent, the picture of tidal ellipsoids (Fig. 1.16) has the advantage of emphasizing the local character of the tidal effects.

The picture of lines of force (see Fig. 1.17) suggests that at the center of each pattern the tidal effects are absent.

We must firmly reject this suggestion.

Although some tidal effects do vanish near the origin of a freely falling reference frame, some other effects remain finite, and it is these that ultimately count.

A more familiar graphical representation of gravitational fields uses the lines of force of the ordinary gravitational force, rather than the tidal force (see Fig. 1.18).

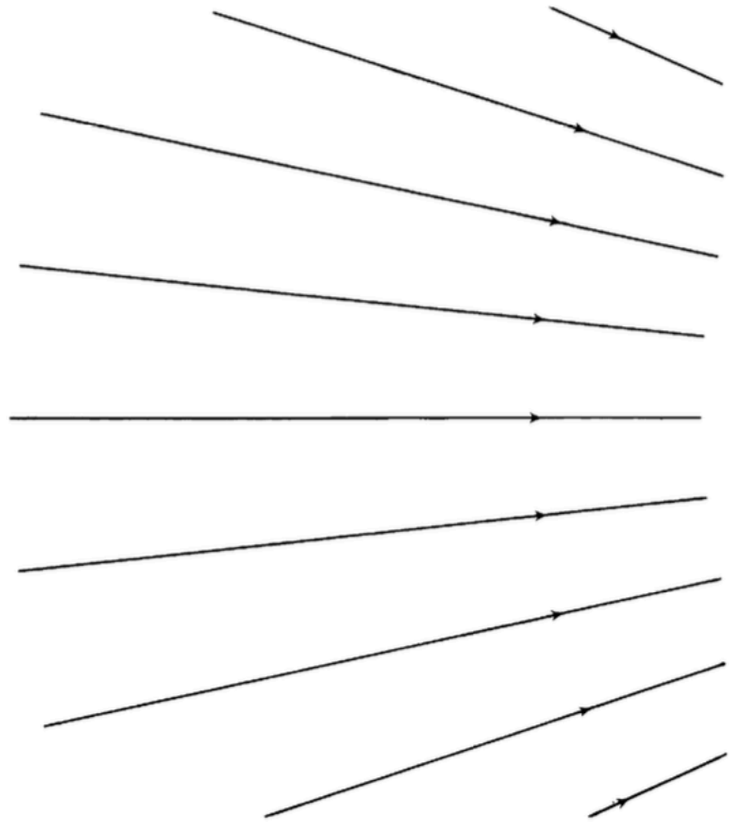


Fig. 1.18 The inverse-square gravitational force represented by lines of force. The diagram is drawn to the same scale as the preceding diagrams.

This is, of course, very useful to an outside observer who wishes to study the motion of a spacecraft or some other body through the field.

But this representation fails to describe what happens locally.

If we want a description of what the astronauts experience in their freely falling spacecraft, then the picture of tidal forces is much more relevant than that of the ordinary force.

The detection of tidal fields by means of the deformation of drops of water is not a practical method, because for a small drop, the surface tension will prevent the formation of tidal bulges.

A somewhat more realistic method is the following: Suppose the astronauts place a freely spinning rigid body in their spacecraft. Figure 1.19 shows a rigid rod at the origin of the coordinates.

The tidal force will exert a torque on the rod and give it an angular acceleration.

For the tidal-force field given by Eqs. (1.46)–(1.48), the torque about the x-axis is

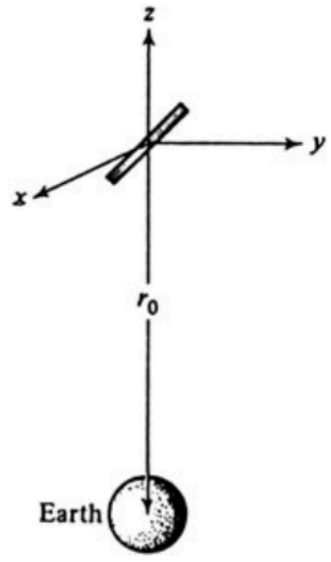


Fig. 1.19 A rigid rod in free fall, instantaneously at rest.

$$\tau_x = \int \left[y\left(2z\frac{GM}{{r_0}^3}\right) - z\left(-y\frac{GM}{{r_0}^3}\right)\right] dm$$
$$= \frac{3GM}{{r_0}^3}\int yz\,dm = \frac{3GM}{{r_0}^3}(-I_{yz}) \tag{1.64}$$

Here $I_{yz} \equiv I^{23}$ is the y-z component of the moment of inertia tensor,

$$I^{kl} = \int \left(r^2\delta^l_k - x^k x^l\right) dm \tag{1.65}$$

the torque exerted by an arbitrary tidal field $R^k{}_{0s0}$ on a body with an inertia tensor I^{ls} is given by

$$\tau^n = c^2 \sum_{k,l,s} \varepsilon^{nkl} R^k{}_{0s0}\left(-I^{ls} + \frac{1}{2}\delta^s_l \sum_r I^{rr}\right) \tag{1.66}$$

where the quantity ε^{nkl} is defined as follows:

$$\varepsilon^{123} = \varepsilon^{231} = \varepsilon^{312} = +1$$
$$\varepsilon^{321} = \varepsilon^{213} = \varepsilon^{132} = -1 \tag{1.67}$$

with all other components zero.

If no other forces act, Eq. (1.64) implies that the x-component of the spin of the rigid body changes at a rate

$$\frac{dS_x}{dt} = -\frac{3GM}{{r_0}^3} I_{yz} \tag{1.68}$$

Similar formulas can be obtained for the rates of change of the other components of the spin.

This rate of change of the spin can serve as measure of the tidal field.

For a local measurement, we must proceed to the limit of small size of the rigid body.

The inertia is of the order $I \approx ml^2$ (where l is a typical dimension of the body), and the spin is $S \approx ml^2\omega$.

In the limit $l \to 0$, both torque and spin vanish, but $d\omega/dt$ remains finite.

It is easy to see that $d\omega/dt$ depends only on the shape of the rigid body, not on the size.

We can therefore use the angular acceleration to detect the tidal force.

If the body is originally spinning, then the tidal force will cause a precession of the spin.

A well-known example is the equinoctial precession of the spin of the Earth, which is caused by the lunar and solar tidal forces acting on the equatorial bulge of the Earth.

The Stanford gyroscope experiment (Gravity Probe B) provides a good illustration of the importance of tidal torques on a rotating body.

This experiment, first proposed by Schiff, was developed by Fairbanks, Everitt, and colleagues over many years, and it was finally sent into orbit with the space shuttle in 2004.

The experiment measured an extremely small relativistic precession of a freely spinning gyroscope predicted by general relativity.

Since the experimenters did not want their gyroscopes to pre- cess from tidal torques, they had to manufacture spherically symmetric gyroscopes.

The shape of these gyroscopes had to be spherical to within a few parts in 10^6, otherwise the precession caused by tidal torques would have washed out the relativistic effects the experimenters sought to measure (for more about this experiment, see Chapter 7).

There exist other practical methods for local measurement of the tidal field.

The Eotvos balance shown in Fig. 1.8 was originally designed as a gravity gradiometer to measure gradients in the gravitational field of the Earth by means of the torque they produce on the balance.

One of the weights of this balance is placed lower than the other to make the balance sensitive to vertical gradients in the horizontal component of gravitational force.

With this arrangement, Eotvos found that tidal fields as small as $R^k{}_{0l0} = (1/c^2)\partial^2\Phi/\partial x^k \partial x^l \approx 10^{-32}/\text{cm}^2$ were detectable.

The beam of the Eotvos balance was 40 cm long.

To see how good the balance is at measuring $R^k{}_{0l0}$ locally, consider the limit $l \to 0$ (where l is a typical dimension of the balance).

The tidal torque on the balance will then tend to zero.

The equilibrium position of the beam is determined by the condition that the torque exerted by the tidal forces be equal to the restoring torque exerted by the fiber.

If we scale down the size of the balance, we must also decrease the torque constant of the fiber to retain the sensitivity of the balance to a given tidal field.

This means that the thickness of the fiber must be reduced (it can be shown that a reduction of the size of the apparatus by a factor of, say, two requires a reduction in fiber thickness by a factor of $2^{3/2}$).

There are, of course, practical limitations that prevent us from reducing the size of our apparatus ad infinitum and obtaining the value of R^k_{0l0} at exactly one point; these limitations must, however, not be blamed on the gravitational field, but rather on the properties (ultimately quantum properties) of materials.

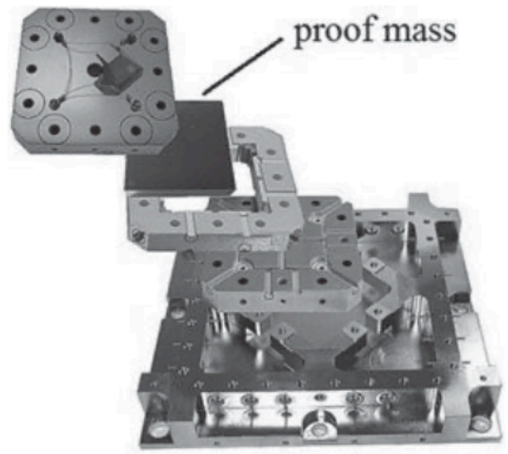
Compact and sturdy gravity gradiometers have been developed for geophysical mea-surements of tidal force fields on the ground and in spacecraft.

A very sensitive electrostatic gravity gradiometer (EGG) was built for the European Space Agency's GOCE (*G*ravity Field and *O*cean *C*irculation *E*xplorer) spacecraft launched in 2009. GOCE's mission was the compilation of a worldwide map of accurate values of surface gravity, especially on the ocean surface; these values can be calculated from the values of gravity gradients measured by the spacecraft.

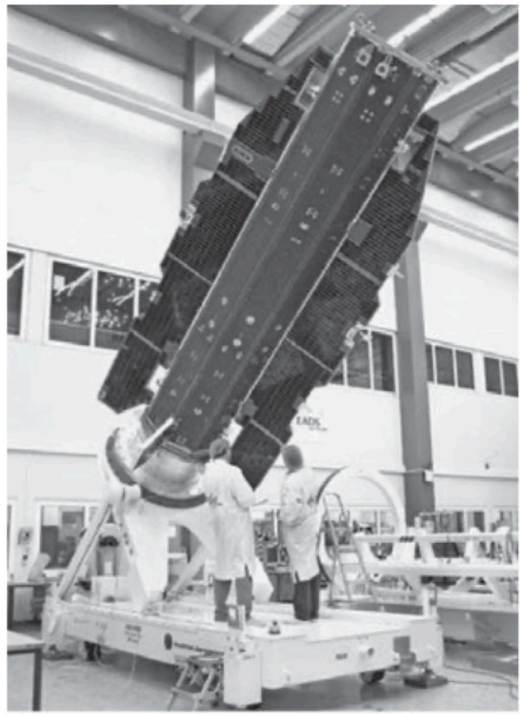
The EGG used for this purpose consists of six electrostatic accelerometers arranged in pairs and separated by 50 cm along the three perpendicular axes defined by the spacecraft (see Fig. 1.20b).

Each accelerometer is itself triaxial; that is, it measures the components of the acceleration along the three perpendicular directions.

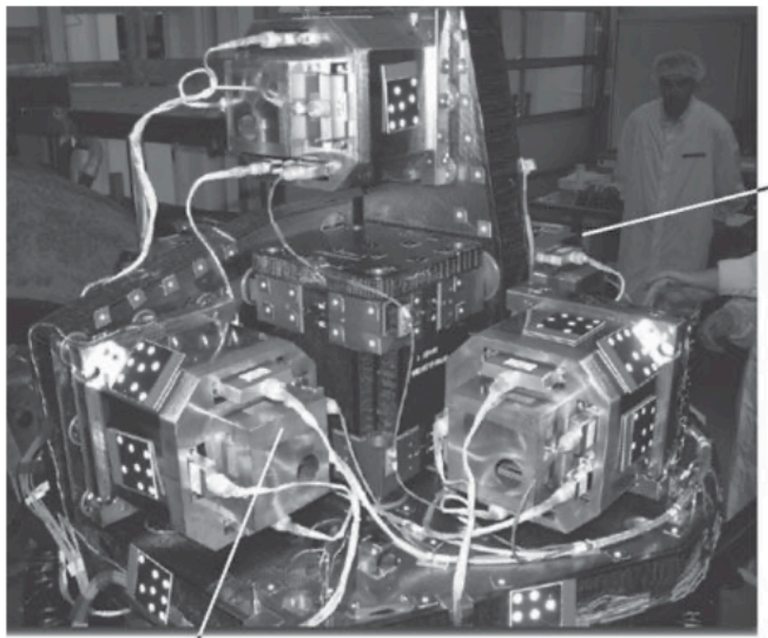
Each accelerometer consists of a proof mass (a block of metal, 1 cm × 1 cm × 4 cm; see Fig. 1.20a) within a tight cage, so the six faces of the block and the adjacent walls of the cage effectively act as capacitor plates.



(a)



(c)



(b)

Fig. 1.20 (a) One of the accelerometers (disassembled) used in the EGG of the GOCE satellite. The proof mass is a block of metal, of about 320 g. (b) Accelerometers mounted in the EGG apparatus. Four of the six accelerometers are visible here. The left accelerometer and the right accelerometer (barely visible) form one pair, separated by 50 cm. The other pairs are top-bottom and front-back (only the top and front accelerometers are visible). (c) The GOCE spacecraft. The EGG apparatus is placed in the middle of the satellite, at the center of mass. (Credits: ESA/ONERA and GOCE/ESA.)

The block floats in the cage, held in equilibrium by the electrostatic forces acting between these capacitor plates.

When the cage is subjected to an acceleration or to a gravitational field, feedback circuits increase the charges in the capacitors so as to maintain the block in equilibrium, and the electric currents required to achieve this indicate the magnitudes of the gravitational field components along the directions perpendicular to the faces of the block.

The differences between the electric signals from the two accelerometers in a pair directly indicate the gradients of the gravitational field components over the 50-cm distance.

The EGG is capable of detecting differences in acceleration as small as 10^{-10} cm/s^2 over a distance of 50 cm, that is, a gradient of 2×10^{-12} /s^2 .

This means that the sensitivity of the EGG is slightly better than the Eotvos balance; however, the latter balance is cumbersome and stationary, and it requires a long time to come to equilibrium, whereas the EGG performs measurement in flight, in a fraction of a second, while speeding through the inhomogeneities of the Earth's gravitational fields.

Since the center-to-center distance between the pairs of accelerometers making up the gradiometer is only about 50 cm, measurements with this instrument are quite local.

And, in principle, a reduction of the size of the instrument is possible.

At present the sensitivity is limited by thermal fluctuations ("noise") in the electric circuitry.

Superconducting circuitry would permit a reduction of size, while maintaining the same sensitivity for the measurement of field gradients.

From this survey of actual experiments and Gedankenexperiments we see that there exist several methods for measuring the tidal field locally, in a small neighborhood of a given point.

The limitations on the minimum size of the neighborhood needed to perform measurements of a given precision do not arise from any intrinsic properties of the gravitational field; rather, these limitations arise from the quantum nature of matter, which prevents us from constructing an apparatus of arbitrarily small size.

The tidal field is no less a local quantity than, say, the electric field. Mathematically, this local nature of the tidal field is perfectly obvious: The tidal field is the derivative of the force field, which means that knowledge of the latter in an arbitrarily small neighborhood of a point is sufficient to determine the former at that point.

Local experiments can distinguish between a reference frame in free fall in a gravitational field and a truly inertial reference frame placed far away from all gravitational fields.

Local experiments can distinguish between a reference frame at rest in a gravitational field and an accelerated reference frame far away from all gravitational fields.

Gravitational effects are not equivalent to the effects arising from an observer's acceleration. It is worth emphasizing this, because in his original paper on the theory of general relativity Einstein wrote,

> Let K' be a system of reference such that relative to K' a mass sufficiently distant from other masses has an accelerated motion such that its acceleration and direction of acceleration are independent of its material composition and physical state.
>
> Does this permit an observer at rest relative to K' to draw the conclusion that he is on a "really" accelerated system of reference? The answer is in the negative; for the above mentioned behavior of freely moving masses relative to K' may be interpreted equally well in the following way. The system of reference K' is unaccelerated, but the spacetime region being considered is under the sway of a gravitational field, which generates the accelerated motion of the bodies relative to K'. (Einstein,)

This statement, which is one formulation of the principle of equivalence of gravitation and acceleration, is true only in a limited sense.

Gravitation and acceleration are equivalent only as far as the translational motion of point particles is concerned (this amounts to what we called the Galileo principle of equivalence, sometimes also called the "weak" principle of equivalence, or WEP).

If the rotational degrees of freedom of the motion of masses are taken into consideration, then the equivalence fails.

Unfortunately, Einstein's statement has often been generalized to sweeping assertions about all laws of physics being the same in a laboratory freely falling in a gravitational field, and in another laboratory far away from any field (this is called the "strong" principle of equivalence, or SEP).

Such generalizations are unwarranted because, as we have seen, even quite simple devices signal the presence of a true gravitational field by their sensitivity to tidal forces and therefore permit us to discriminate between a gravitational field and the pseudo-force field of acceleration.

The confusion surrounding the principle of equivalence led Synge to remark:

> I have never been able to understand this principle. . . . Does it mean that the effects of a gravitational field are indistinguishable from the effects of an observer's acceleration? If so, it is false. In Einstein's theory, either there is a gravitational field or there is none, according as the Riemann tensor [tidal-force tensor] does or does not vanish. This is an absolute property; it has nothing to do with any observer's worldline. . . . The Principle of Equivalence performed the essential office of midwife at the birth of general relativity. . . . I suggest that the midwife be now buried with appropriate honours and the facts of absolute spacetime be faced. (Synge)

To avoid such confusion, we will base our further development of gravitational theory on the precise and unambiguous equality $m_I = m_G$. This equality is necessary and, to a large extent, sufficient for the construction of the relativistic theory.

2 - The Formalism of Special Relativity

> Raffiniert ist der Herrgott,
> Aber boshaft ist Er nicht.
> [God is cunning, but not malicious.] Albert Einstein

Throughout this book we are concerned with fields – that is, functions of space and time, such as gravitational fields, electromagnetic fields, and velocity fields – and density distributions that describe masses and charges.

Spacetime is the arena in which these fields perform their joint evolutions.

It is therefore clear that we must first get to know the structure and geometry of spacetime.

Unfortunately, because the velocity of light is so large, everyday experience leads us to acquire various misconceptions about the geometry of spacetime.

This set of misconceptions goes under the name of Newtonian, or Galilean, spacetime.

The true (or more true) geometry of spacetime was discovered through the development of Einstein's theory of special relativity, starting in 1905.

The keystone of this theory is the principle of relativity, according to which the laws of physics are the same in all inertial reference frames.

Einstein was led to this principle by his investigation of Maxwell's equations.

As he wrote in his autobiographical notes,

> After ten years of reflection such a principle (the principle of special relativity) resulted from a paradox upon which I had already hit at the age of sixteen: If I pursue a beam of light with the velocity c (velocity of light in a vacuum), I should observe such a beam as an electromagnetic field constant in time, periodic in space. However, there seems to exist no such thing, neither on the basis of experience, nor according to Maxwell's equations. (Einstein, 1951)

This left Einstein with two alternatives: Either Maxwell's equations were wrong in all reference frames except one (the ether frame), or else something was wrong with the geometry of spacetime suggested by everyday experience.

Einstein had a strong conviction that Maxwell's equations should not permit any intrinsic distinction between two inertial reference frames and that therefore the equations should have the same form in all such reference frames.

Accordingly, he chose the second alternative: Newton's spacetime geometry is wrong.

He was then faced with the task of finding a spacetime geometry compatible with Maxwell's equations.

In fact, it is not hard to find what spacetime geometry will make Maxwell's equations true in all inertial reference frames.

The desired geometry follows quite directly from the requirement that the velocity of light be the same in all inertial reference frames.

The drastic modification of the geometry of space and time demands a corresponding modification of mechanics.

Newton's laws of mechanics are true in all inertial reference frames if, and only if, spacetime is Newtonian.

Einstein and his followers had to develop a new set of mechanical equations, designed to have the same form in all inertial reference frames in relativistic spacetime.

In this chapter we review the special theory of relativity and develop the mathematical formalism of four-vectors and tensors.

Detailed discussions of the physical basis of special relativity and of the experiments that compel us to accept relativistic spacetime can be found in many textbooks that are devoted to special relativity.

2.1 The spacetime of special relativity

There are two aspects to the structure of spacetime: the topological structure and the geometric structure.

Crudely speaking, the topological structure tells us how many dimensions spacetime has, how the different parts of spacetime are connected (that is, which points are in the neighborhood of which), and how smooth the spacetime is.

Topology might be called "rubber-sheet geometry," because it is concerned with those properties of a space that are unchanged by arbitrary "smooth" deformations of the space.

The geometric structure tells us how to construct straight and parallel lines, and circles and other geometric figures, and it tells us the distances between points in spacetime.

Experience teaches us that spacetime is a 4-D continuum.

All the points in spacetime – such as, say, the point where and when a firecracker exploded, or the points where and when the fragments landed on the ground – can be smoothly parametrized by four real numbers, which we call the coordinates of the points.

In Newtonian physics, as well as in special relativistic physics, this parametrization of all points of spacetime by four coordinates can be performed globally; if we use rectangular coordinates x, y, z, t, each of which ranges from $-\infty$ to $+\infty$, then these coordinates span the entire spacetime.

In general-relativistic physics, where spacetime is curved, such a global parametrization is not possible; attempts at extending the coordinates in all directions usually result in singularities in the coordinates.

Such coordinate singularities are similar to what mapmakers find when they attempt to use the longitude and latitude angles as coordinates for the curved 2-D surface of the Earth – these coordinates develop singularities at the poles of the Earth, where the longitude coordinate fails to be unique.

Nevertheless, even for the curved spacetime of general relativity, it is still possible to construct well-behaved 4-D coordinate patches locally, for a finite neighborhood of any given point.

In the same way, mapmakers can draw a map with well-behaved coordinates for any part of the Earth; they merely must make sure to avoid latitude and longitude angles when drawing the map of the polar regions and to use instead, say, a rectangular grid in those regions.

Mathematically, a space that can be covered by coordinate patches with real numbers as coordinates is called a *manifold*.

Thus, spacetime is a 4-D manifold.

The topology of the spacetime of Newtonian physics and also the topology of the spacetime of special-relativistic physics is that of the Euclidean 4-D space R^4, that is, the set of all quadruplets of real numbers.

This simple topology permits us to parametrize these spacetimes globally, with a single coordinate patch for the entire spacetime.

The neighborhood of a given point is the set of all points such that their coordinates differ only little from those of the given point. In view of the direct correspondence between points of spacetime and points of R^4, we can afford to use the careless terminology "the point t, x, y, z" rather than the more precise "the point whose coordinates are t, x, y, z."

All of this seems trivial, but is not because it is quite possible that the topology of spacetime is not Euclidean.

One possible violation could occur at large distances: Our universe might be closed and have the topology of a sphere.

A violation of Euclidean topology is also possible at very small distances.

When dealing with distances of 10^{-13} cm, 10^{-23} cm, or even less, it is not possible to set up a rigid coordinate system of clocks and meter sticks.

We tend to believe that it cannot be done because of the atomic structure of matter, but maybe there is a more fundamental reason why it cannot be done.

Perhaps spacetime at a subnuclear level has a very pathological structure and is multiply connected, full of wormholes and bubbles; perhaps it is not even a continuum.

We can test for such pathologies only indirectly: We tentatively assume spacetime is an Euclidean continuum and use coordinates t, x, y, z.

If our theories based on the use of such coordinates are successful, then we can feel reasonably confident that we have made the right assumption.

At really small distances, of the order of 10^{-33} cm, a Euclidean topological structure is quite unlikely.

At such distances the fluctuations of quantum gravitation are extremely violent and probably produce an ever-changing dynamic topology (Wheeler, 1968).

The length $\sqrt{\hbar G/c^3} = 1.6 \times 10^{-33}$ cm is called the *Planck length*; it gives the characteristic size of the fluctuations in quantum gravitation.

The topological structures of the spacetimes of Newton and of special relativity are the same.

However, these spacetimes differ in the definitions of distances, that is, in the metric structures.

Before we write down an expression for the distance between two points, we must impose a restriction: choose the t-, x-, y-, z-axes of spacetime in such a way that we obtain an inertial reference frame, so the trajectories of free particles are straight lines in space and also in a plot of position vs. time.

This means that to obtain an inertial reference frame we must choose the axes, and the units along these axes, so that the trajectories of free particles have the equations of straight lines in four dimensions, that is, $x = a_1t + b_1$, $y = a_2t + b_2$, $z = a_3t + b_3$.

Furthermore, to obtain an inertial reference frame suitable for the study of dynamics, we must impose an additional restriction: choose the synchronization of clocks in such a way that Newton's second law is valid at low speeds (a "bad" synchronization can spoil Newton's second law while leaving Newton's first law unaffected; see Ohanian, 2004).

The trajectory of a particle in 4-D spacetime is called the *worldline* of the particle.

Figure 2.1 shows some examples of worldlines of free particles in an inertial reference frame.

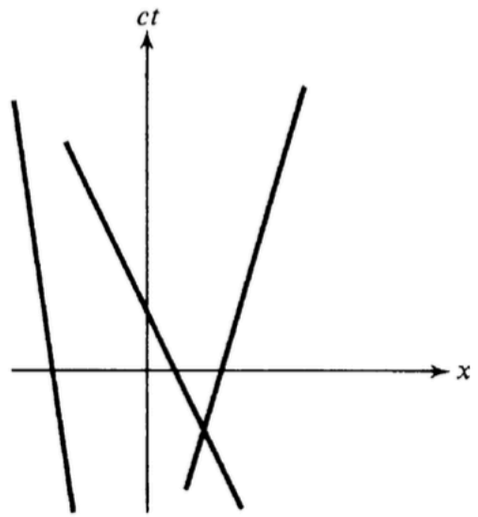


Fig. 2.1 Worldlines of free particles. For such a particle, the worldline, or the plot of position vs. time, is represented by an equation linear in $x, y, z,$ and t.

Consider now two spacetime points separated by a small interval dt, dx, dy, dz (the use of infinitesimal intervals will be helpful for our later work).

To define the geometry, we write down a quadratic form in dt, dx, dy, dz that specifies the distance.

Newtonian spacetime has two distances and relativistic spacetime only one:

NEWTONIAN SPACETIME:

$$\text{time interval, } dt$$
$$\text{space interval, at fixed time, } dl^2 = dx^2 + dy^2 + dz^2 \tag{2.1}$$

RELATIVISTIC SPACETIME:

$$\text{spacetime interval, } ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2 \tag{2.2}$$

In the last equation, c is the speed of light in vacuum, approximately $c = 3.00 \times 10^{10}$ cm/s.

Thus, Newtonian spacetime contains two separate geometries: a three-dimensional Euclidean geometry for space and a 1-D geometry for time.

Relativistic spacetime contains only one geometry, which combines both space and time.

As Minkowski expressed it,

> "Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality" (Minkowski, 1908).

Note that the spacetime interval of the relativistic geometry is not always positive.

Intervals with a positive, negative, or zero value of ds^2 are called, respectively, timelike, spacelike, and lightlike. Figure 2.2 shows examples of displacements that have time-, space-, and lightlike intervals; in this figure lightlike displacements make an angle of 45° with the spacetime axes.

Along the worldline of a massive particle, the interval must be timelike.

This follows from Eq. (2.2) if we take into account that the speed $(dx^2 + dy^2 + dz^2)^{1/2}/dt$ for such a particle is less than c.

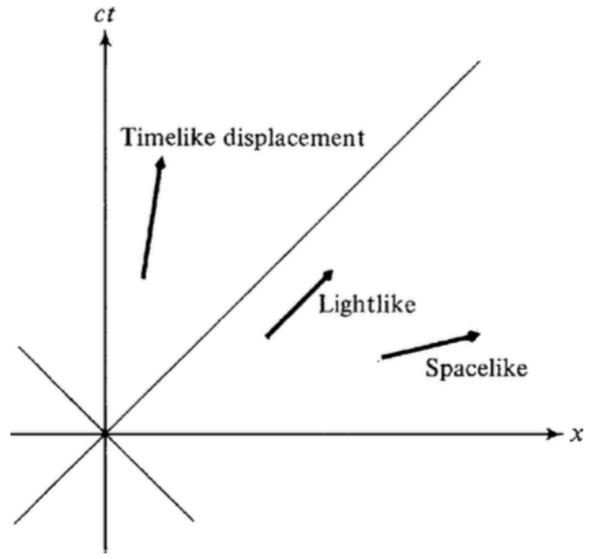


Fig. 2.2 Timelike, lightlike, and spacelike displacements. The lightlike displacements make an angle of 45° with the t-axis, because for these displacements $dx = \pm cdt$.

Along the worldline of a light signal, the interval is of course lightlike.

The set of all worldlines of light signals leaving or arriving at a given point forms a 3-D "surface" in the 4-D spacetime.

If the given point is the origin, this surface has the equation

$$c^2t^2 - x^2 - y^2 - z^2 = 0$$

and it is called the *light cone* corresponding to the given point.

Since we cannot make a drawing of a 4-D space, let us omit the z-coordinate; then the light cone is simply the ordinary cone $c^2t^2 - x^2 - y^2 = 0$ shown in Fig.2.3.

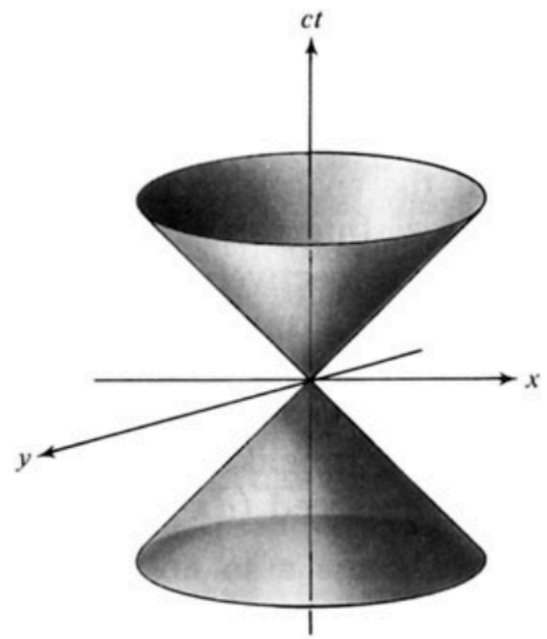


Fig. 2.3 The light cone.

The upper ($t > 0$) and lower ($t < 0$) parts of this cone are called the forward and backward light cones (or the future and past light cones), respectively.

Particles that leave the origin must have worldlines inside the forward light cone; particles that arrive at the origin must have worldlines inside the backward light cone.

An essential property of distance is that it is independent of the choice of coordinates.

In Newtonian spacetime, the absolute character of time intervals and of lengths (at a fixed time) is taken as a fundamental assumption, and accordingly, time intervals and lengths are independent of the choice of coordinates – that is, they are *invariant* under transformations of the coordinates.

In relativistic spacetime, neither time intervals nor lengths are invariant. Instead, the speed of light is invariant; that is, the speed of light remains invariant under transformations of the coordinates from one inertial reference frame to another (more generally, all laws of physics remain unchanged under such transformations, and there is no intrinsic distinction between different inertial reference frames; this is Einstein's principle of relativity; see the later discussion).

As a corollary of the invariance of the speed of light, the spacetime interval is invariant, so

$$ds^2 = ds'^2 \tag{2.3}$$

or, more explicitly,

$$c^2dt^2 - dx^2 - dy^2 - dz^2 = c^2dt'^2 - dx'^2 - dy'^2 - dz'^2 \tag{2.4}$$

where t, x, y, z are the spacetime coordinates in one inertial reference frame and t', x', y', z' those in another.

To deduce Eq. (2.4) from the invariance of the speed of light, let us assume that the second reference frame moves with speed V along the x-axis of the first (and, conversely, the first moves with speed $-V$ along the $-x'$ -axis of the second).

In this case, the lengths dl = $\sqrt{dy^2 + dz^2}$ and dl′ = $\sqrt{dy'^2 + dz'^2}$ are perpendicular to the relative velocity of the reference frames.

If this perpendicular length dl′ is contracted as measured in the first reference frame, then the principle of relativity demands that the perpendicular length dl is likewise contracted as measured in the second reference frame, because the relative behavior of these lengths must be the same.

These mutual contraction effects are logically contradictory, and therefore no such contraction or elongation is possible, that is, the transverse coordinates y and z are invariant.

Equation (2.4) therefore simplifies to $c^2dt^2 - dx^2 = c^2dt'^2 - dx'^2$.

To prove this simplified equation, we note that when we express $c^2dt^2 - dx^2$ in terms of the differentials dt′ and dx′ of the transformed coordinates, we necessarily obtain a quadratic polynomial expression,

$$c^2dt^2 - dx^2 = Ac^2dt'^2 + Bcdt'dx' + Cdx'^2 \tag{2.5}$$

For the special case of a light signal, cdt = ±dx, and then the left side of this equation is zero.

This implies that the value of dt′, for some assumed fixed value of dx′, is a root of the quadratic equation $0 = Ac^2dt'^2 + Bcdt'dx' + Cdx'^2$.

But from the invariance of the speed of light, we know that the roots must be cdt′ = ±dx′, and therefore the polynomial must be of the form $A(cdt' - dx')(cdt' + dx') = A(c^2\,dt'^2 - dx'^2)$.

Accordingly, $B = 0$, and Eq. (2.5) reduces to

$$c^2dt^2 - dx^2 = A(c^2dt'^2 - dx'^2) \tag{2.6}$$

By the principle of relativity, when we perform the inverse transformation and express $c^2dt'^2 - dx'^2$ in terms of the differentials *dt* and *dx*, we must obtain an equation similar to (2.6), but with the same factor *A* on the left side (the coefficient *A* depends on the magnitude of the relative velocity of the reference frames, but not on the direction, because reversal of the direction of the velocity is equivalent to a reversal of the directions of the *x* and x′-axes, which leaves all the terms involving *x* and x′ in Eq. (2.6) unchanged and therefore leaves the coefficient *A* unchanged).

To avoid a contradiction, we must then set A = 1, which completes the proof of Eq. (2.4).

The invariance of the infinitesimal spacetime interval ds^2 and, more generally, the invariance of the finite spacetime interval $s^2 = c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$, permit us to measure such a spacetime interval by means of the following general "radar-ranging" procedure: Suppose that *P* and P′ are two arbitrary spacetime points separated by some finite interval.

Take a clock with a straight worldline (a freely moving clock) that passes through the point *P* (see Fig. 2.4a), and send a light signal from this clock to the point P′, where the light signal is reflected and returns to the clock.

If the time registered by the clock is $T = 0$ when it passes through the point *P*, $T = T_1$ when the light signal is emitted, and $T = T_2$ when the reflected signal is received, then the spacetime interval between *P* and P′ equals (Synge, 1956; Marzke and Wheeler, 1964)

$$s^2 = c^2 T_1 T_2 \tag{2.7}$$

To establish this formula, it is convenient to go to a reference frame in which the clock is at rest, so the triangle $T_1P'T_2$ becomes an equilateral right triangle (see Fig. 2.4b).

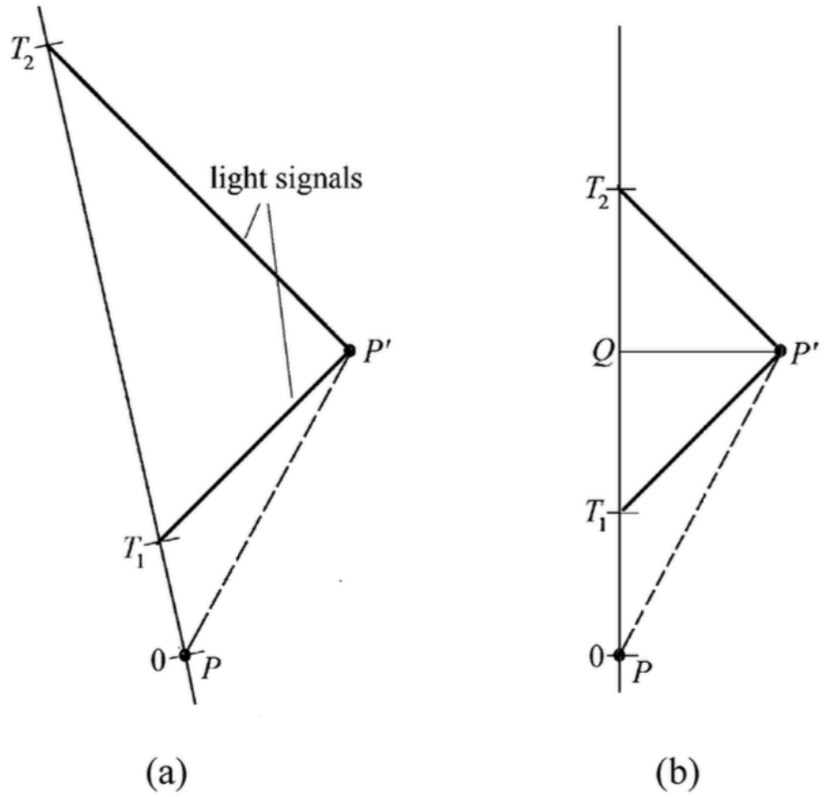


Fig. 2.4 (a) Procedure for the direct measurement of the spacetime interval $P\,P'$. The clock passes through the point P at a clock time $T = 0$, it emits a light signal at the clock time T_1, and it receives the echo at the clock time T_2. (b). In a reference frame in which the clock is at rest, the point P' has a time coordinate $0Q$ relative to the point P and a space coordinate QP'. From the geometry of this diagram, we see that these time and space coordinates can be expressed as $\frac{1}{2}(T_2 + T_1)$ and $\frac{1}{2}c\,(T_1 - T_2)$, respectively.

The time and space coordinates of the point P′ relative to the point P are then $^1/_2\,(T_2 + T_1)$ and $^1/_2\,c(T_1 - T_2)$, respectively, and therefore the spacetime interval PP′ equals $^1/_4\,c^2(T_1 + T_2)^2 - {}^1/_4\,c^2(T_1 - T_2)^2 = c^2T_1T_2$, as stated in Eq. (2.7).

Because s^2 is invariant, this formula is valid in all reference frames.

It is instructive to consider two special cases of Eq. (2.7).

If the worldline of the clock passes through both of the points P and P′ (which means that the spatial distance between P and P′ is zero in the reference frame of this clock), then $T_1 = T_2$ and Eq. (2.7) reduces to the self-evident result $s^2 = c^2T_1{}^2$.

And if the worldline of the clock is at right angles to the displacement from P to P' (which means that P and P' are simultaneous in the reference frame of the clock), then $T_1 = -T_2$, and Eq. (2.7) reduces to $s^2 = -c^2T_1^2$.

This result implies that the spatial distance is $l = cT_1$, as was to be expected, since in this case our generalized radar-ranging procedure reduces to ordinary radar ranging.

Note that this radar-ranging procedure is a direct implementation of the modern definition of the unit of length, which states that the meter is the distance traveled by light in 1/299,792,458 s.

As an alternative to radar ranging, we could calculate the spacetime interval by means of Eq. (2.2), from separate measurements of the space and time intervals x and t.

However, such a calculation is less direct than the radar-ranging procedure, and furthermore, it requires us to set up a coordinate system, so we can define x and t.

The radar-ranging procedure permits us to bypass the coordinate system, so we can measure the spacetime interval without the use of coordinates.

This gives the spacetime interval a direct physical meaning.

The coordinates measured in any two inertial coordinate systems must be related by linear transformations, because the validity of Newton's first law in both of these coordinate systems demands that straight worldlines represented by a linear relationship between the t, x, y, z coordinates transform into straight worldlines represented by a linear relationship between the t', x', y', z' coordinates.

The requirement that coordinate transformations between two inertial reference frames be such as to leave invariant the distances specified in Eq. (2.1) or (2.2) then uniquely determines the form that these transformations may take.

If the origin of the x′, y′, z′ coordinates moves with velocity V along the x-axis of the x, y, z coordinates, then the Newtonian and the relativistic transformations between these coordinates are, respectively,

NEWTONIAN SPACETIME – GALILEAN TRANSFORMATION

$$\text{Invariant distances } dl = dl' \text{ and } dt = dt' \Rightarrow \begin{cases} t' = t \\ x' = x - Vt \\ y' = y \\ z' = z \end{cases} \tag{2.8}$$

RELATIVISTIC SPACETIME – LORENTZ TRANSFORMATION

$$\text{Invariant distance } ds^2 = ds'^2 \Rightarrow \quad \begin{cases} t' = \dfrac{t - Vx/c^2}{\sqrt{1 - V^2/c^2}} \\ x' = \dfrac{x - Vt}{\sqrt{1 - V^2/c^2}} \\ y' = y \\ z' = z \end{cases} \tag{2.9}$$

It can be shown that these are the only linear transformations that leave the distances (2.1) and (2.2) invariant, except for trivial shifts of origin or rotations of the axes.

The combination of two Lorentz transformations in succession is itself a Lorentz transformation.

If the individual speeds of the two transformations are V and U , then it is easy to check that the speed of the combined Lorentz transformation is $(V + U)/(1 + VU/c^2)$.

Not surprisingly, this net speed is as expected from the relativistic combination law for speeds.

The fact that two Lorentz transformations in succession lead to a combined Lorentz transformation means that the totality of all these transformations, including transformations with different directions of velocity, form a group, called the *Lorentz group*.

We can enlarge this group by including ordinary space rotations and space translations, (that is, shifts of origin).

This comprehensive group of all the transformations that leave ds^2 invariant is called the *Poincaré group*.

We will now concentrate on relativistic spacetime and the Lorentz transformation.

This transformation imposes severe restrictions on acceptable laws of physics.

The reason is this: the spacetime structure has been designed so that, as far as the behavior of free particles and light is concerned, there is no distinction between different inertial reference frames – a free particle obeys the same laws in all inertial frames.

For the behavior of light and other electromagnetic waves, the absence of any distinction between different inertial reference frames is experimentally well established by the Michelson-Morley experiment and other similar experiments. (For other physical phenomena, the experimental evidence is less firm; only a few experiments have been specifically designed to test relativity with phenomena other than light, and the evidence that emerges from such other phenomena is mostly indirect, that is, the theories of interactions we construct on the basis of relativity are in excellent agreement with observation.)

We therefore take our cue from the experiments with light, and we adopt the postulate that no experiment can make any intrinsic distinction between different inertial reference frames, or, equivalently, that all physical systems obey the same laws in all inertial frames.

This is a symmetry principle imposed on the laws of physics, called the *principle of relativity*.

The laws of physics in different reference frames are related by Lorentz transformations, so these transformations restrict the acceptable forms of the law of physics:

Principle of relativity: *All laws of physics must be invariant under Lorentz transformations*.

In this context, "invariant" means that each law retains the same mathematical form and that all numerical constants retain the same values.

The restrictions placed on the laws of physics by this symmetry principle are so stringent that, for example, in electrodynamics we can derive all the Maxwell equations from Coulomb's law alone (Schwartz, 1972; Ohanian, 1988).

The importance of these restrictions is even greater in quantum theory: the principle of relativity and the structure of spacetime tell us that only particles of integer or half-integer spin can exist (for instance, spin $\hbar/3$ is forbidden), that there must be a connection between spin and statistics, and so on.

What concerns us most is that the principle of relativity also places very tight restrictions on possible theories of gravitation.

We will exploit these restrictions in the next chapter.

2.2 Tensors in spacetime

In this section we develop the definitions of tensors in spacetime.

Mathematicians often define tensors as linear operators that act on vectors (for more on this, see Section 2.6).

But here we prefer to adopt a simpler definition based on the transformation properties under Lorentz transformations.

These transformation properties are essential for the formulation of the principle of relativity.

We begin with some convenient notations.

First, we replace t, x, y, z by

$$x^0 \equiv ct, \quad x^1 \equiv x, \quad x^2 \equiv y, \quad x^3 \equiv z \tag{2.10}$$

We can then write Eq. (2.9) as

$$ds^2 = \sum_{\mu=0}^{3} \sum_{\nu=0}^{3} \eta_{\mu\nu} dx^\mu dx^\nu \tag{2.11}$$

where $\eta_{\mu\nu}$ is a matrix with components

$$\eta_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \tag{2.12}$$

This is called the *metric tensor*, or the Minkowski tensor, of spacetime.

We will use Einstein's summation convention, according to which any repeated Greek super- or subscript appearing in a term of an equation is to be summed from 0 to 3.

Accordingly, Eq. (2.11) can also be written as

$$ds^2 = \eta_{\mu\nu} dx^\mu dx^\nu. \tag{2.13}$$

It is also convenient to define

$$x_\mu \equiv \eta_{\mu\nu} x^\nu \tag{2.14}$$

so

$$x_0 = ct, \quad x_1 = -x, \quad x_2 = -y, \quad x_3 = -z \tag{2.15}$$

Note that in Eq. (2.14) the summation convention affects only the repeated index ν, not the index μ; the latter simply tells us that the equation has four components ($\mu = 0, 1, 2, 3$).

An index such as ν, repeated in a given term in an equation, is called a *dummy index* because it disappears when the summation is written out in full,

$$\eta_{\mu\nu}x^{\nu} = \eta_{\mu 0}x^{0} + \eta_{\mu 1}x^{1} + \eta_{\mu 2}x^{2} + \eta_{\mu 3}x^{3}$$

Hence $\eta_{\mu\nu}x^{\nu}$, $\eta_{\mu\alpha}x^{\alpha}$, $\eta_{\mu\beta}x^{\beta}$, etc., are exactly the same thing.

An index such as μ that appears only once in each term of an equation is called a *free index*.

The quantities x^{μ} and x_{μ} are called, respectively, the *contravariant* and *covariant components* of the position vector.

For the sake of brevity, we will often use the phrase "the vector x^{μ}" or "the vector x_{μ}" instead of the more accurate but cumbersome "the vector whose components are x^{μ} (or x_{μ}), with $\mu = 0, 1, 2, 3$."

The abbreviated phraseology blurs the distinction between a vector and its components, but there is not much harm in that, since a vector can be regarded as a set of components.

In terms of x^{μ} and x_{μ}, Eq. (2.13) becomes

$$ds^2 = dx^{\mu}dx_{\mu} \tag{2.16}$$

which resembles the familiar dot product of two vectors.

In general, tensor quantities with Greek superscripts are called *contravariant*, and quantities with subscripts are called *covariant*.

Thus, $\eta_{\mu\nu}$ is the covariant metric tensor.

If we invert the relation (2.14), we obtain

$$x^{\nu} = \eta^{\nu\mu} x_{\mu} \tag{2.17}$$

where $\eta^{\nu\mu}$ is the matrix inverse to $\eta_{\mu\nu}$.

This inverse is defined by

$$\eta^{\nu\mu}\eta_{\mu\alpha} = \delta^{\nu}_{\alpha} \tag{2.18}$$

where $\delta_{\alpha}{}^{\nu}$ is the Kronecker delta.

The matrix $\eta^{\nu\mu}$ is the contravariant metric tensor; the Kronecker delta $\delta_{\alpha}{}^{\nu}$ is a mixed tensor.

In terms of x^{μ}, the Lorentz transformation (2.9) can be written

$$\begin{aligned} x'^{0} &= \frac{x^0 - Vx^1}{\sqrt{1-V^2}} \\ x'^{1} &= \frac{x^1 - Vx^0}{\sqrt{1-V^2}} \\ x'^{2} &= x^2 \\ x'^{3} &= x^3 \end{aligned} \tag{2.19}$$

In these equations we have adopted the convention $c = 1$, which can always be achieved by a suitable choice of units.

For instance, we can measure time in seconds and distance in light-seconds; then $c = 1$ light-second per second = 1.

The convention $c = 1$ amounts to the same thing as omission of all factors of c.

The restoration of the missing factors of c is easy: Insert as many such factors as is necessary to make the equation dimensionally correct in cgs units.

For example, if we restore the factors of c, the equation for x'^0 becomes $x'^0 = (x^0 - Vx^1/c^2)/\sqrt{1 - V^2/c^2}$.

In numerical calculations, we will always use cgs units; that is, the factors of c are to be restored before inserting numerical values into any equations.

A neater way to write the Lorentz transformation is the following,

$$x'^{\mu} = a^{\mu}{}_{\nu} x^{\nu} \tag{2.20}$$

where $a^{\mu}{}_{\nu}$ is the Lorentz-transformation matrix

$$a^{\mu}{}_{\nu} = \begin{pmatrix} 1/\sqrt{1-V^2} & -V/\sqrt{1-V^2} & 0 & 0 \\ -V/\sqrt{1-V^2} & 1/\sqrt{1-V^2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.21}$$

Note that, as customary for matrices, the first index μ specifies the row, and the second index ν the column.

For some simple practice with index notation, let us evaluate (2.20) explicitly, by substituting (2.21). For $\mu = 0$, this gives $x'^0 = a^0{}_{\nu} x^{\nu} = a^0{}_0 x^0 + a^0{}_1 x^1 + a^0{}_2 x^2 + a^0{}_3 x^3 = (1/\sqrt{1 - V^2})x^0 - (V/\sqrt{1 - V^2})x^1 + 0 + 0$, which agrees with the usual Lorentz transformation equation for the time coordinate (likewise, the $\mu = 1$ component gives the Lorentz transformation for the space coordinate x^1).

An alternative method for the evaluation of Eq. (2.20) is to write the left side as a 4×1 matrix (that is, a column vector) and the right side as the matrix product of a 4×4 matrix and a 4×1 matrix, and then perform the usual row-by-column matrix multiplication:

$$\begin{pmatrix} x'^0 \\ x'^1 \\ x'^2 \\ x'^3 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{1-V^2} & -V/\sqrt{1-V^2} & 0 & 0 \\ -V/\sqrt{1-V^2} & 1/\sqrt{1-V^2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix}$$

$$= \begin{pmatrix} (x^0 - Vx^1)/\sqrt{1-V^2} \\ (x^1 - Vx^0)/\sqrt{1-V^2} \\ x^2 \\ x^3 \end{pmatrix}$$

However, when using this matrix method you must pay careful attention to the indices in the sum.

Row-by-column multiplication applies if the pair of dummy indices are adjacent [for example, the two ν indices in Eq. (2.20)].

But if they are not adjacent [for example, the two β indices in Eq. (2.39)], then column-by-column multiplication or, sometimes, row-by-row multiplication becomes necessary.

The Lorentz transformation (2.21) is a particularly simple kind of transformation, in that the velocity is entirely along the x-axis and therefore has only one component.

The matrix $a^{\mu}{}_{\nu}$ for a general Lorentz transformation depends on six parameters: three components of the relative velocity plus three angles that indicate how much the x', y', z'- axes are rotated with respect to the x, y, z-axes.

Lorentz transformations with zero rotation angles are called pure transformations, or boosts.

Lorentz transformations with zero velocity are merely ordinary rotations of the x, y, z-axes.

The matrix $a^{\mu}{}_{\nu}$ for a general Lorentz transformation must satisfy the equation

$$a^{\alpha}{}_{\mu}\eta_{\alpha\beta}a^{\beta}{}_{\nu} = \eta_{\mu\nu} \tag{2.22}$$

To see this, we substitute $x'^{\alpha} = a^{\alpha}{}_{\mu}x^{\mu}$ into the equation

$$x'^{\alpha}\eta_{\alpha\beta}x'^{\beta} = x^{\mu}\eta_{\mu\nu}x^{\nu} \tag{2.23}$$

that expresses the invariance of the (finite) spacetime interval between the origin and the point x^{μ}.

We then obtain

$$a^{\alpha}{}_{\mu}x^{\mu}\eta_{\alpha\beta}a^{\beta}{}_{\nu}x^{\nu} = x^{\mu}\eta_{\mu\nu}x^{\nu} \tag{2.24}$$

or

$$x^{\mu}x^{\nu}(a^{\alpha}{}_{\mu}\eta_{\alpha\beta}a^{\beta}{}_{\nu} - \eta_{\mu\nu}) = 0 \tag{2.25}$$

In Eq. (2.25), x^{μ} and x^{ν} are arbitrary.

Hence the coefficient of $x^{\mu}x^{\nu}$ must vanish; this implies Eq. (2.22).

We now give the general definitions of tensors.

The simplest tensor is a scalar (tensor of rank zero); this is a one-component object that remains unchanged under Lorentz transformations.

For example, any numerical constant is a scalar.

The spacetime interval ds^2 is a scalar, because the Lorentz transformation has been designed precisely so as to leave ds^2 invariant.

Another closely related scalar is the proper time interval $d\tau$ associated with the worldline of a particle.

This is defined as equal to ds along the worldline of the particle, that is,

$$d\tau = \sqrt{dx^\mu dx_\mu} = \sqrt{dt^2 - dx^2 - dy^2 - dz^2} = dt\sqrt{1 - v^2} \tag{2.26}$$

where $v = (dx^2 + dy^2 + dz^2)^{1/2}/dt$ is the particle's velocity.

In the particle's rest frame, $d\tau = dt$.

This gives us the physical interpretation for the proper time: It is the time interval measured by a clock that moves with the same velocity as the particle.

A *vector* (or tensor of rank one) is a four-component object A^μ that transforms under Lorentz transformations in exactly the same way as x^μ:

$$A'^\mu = a^\mu{}_\nu A^\nu \tag{2.27}$$

According to this definition, x^μ (and also dx^μ) serves as a prototype for all vectors.

If x^μ is the position vector of a particle, then the quantity

$$u^\mu \equiv \frac{dx^\mu}{d\tau} \tag{2.28}$$

is a vector since it consists of the product of a vector dx^μ by a scalar $1/d\tau$. This vector is called the four-velocity.

Written out explicitly, the components of u^μ are

$$u^\mu = \left(\frac{1}{\sqrt{1 - v^2}}, \frac{v_x}{\sqrt{1 - v^2}}, \frac{v_y}{\sqrt{1 - v^2}}, \frac{v_z}{\sqrt{1 - v^2}} \right) \tag{2.29}$$

Another important vector is the energy-momentum four-vector, defined as the product of the four-velocity and the particle's rest mass,

$$p^{\mu} = mu^{\mu} \tag{2.30}$$

This proportionality between p^{μ} and u^{μ} looks like the proportionality between Newtonian momentum and velocity, but each side of Eq. (2.30) is more complicated than in the Newtonian case.

Explicitly, the components of the energy-momentum four-vector are

$$p^{\mu} = \left(\frac{m}{\sqrt{1-v^2}}, \frac{mv_x}{\sqrt{1-v^2}}, \frac{mv_y}{\sqrt{1-v^2}}, \frac{mv_z}{\sqrt{1-v^2}} \right) \equiv (E, p_x, p_y, p_z) \tag{2.31}$$

These components are the energy (E) and the three components of the momentum (p_x , p_y , p_z) of the particle.

In Eq. (2.30) we constructed the relativistic momentum by a simple analogy with the Newtonian momentum.

For a deeper justification of this expression for momentum, we need to examine the conservation law for energy and momentum.

For instance, consider several particles engaged in elastic collisions among themselves.

If the conservation law for energy and momentum is to hold in the same way in all inertial reference frames, in accord with the principle of relativity, the total energy and momentum must be a four-vector, and therefore so must be the individual energy and momenta.

But the only four-vector we can construct with space components that are proportional to dx, dy, dz is $p^{\mu} \propto (dt/d\tau, dx/d\tau, dy/d\tau, dz/d\tau)$, and the constant of proportionality must be m, because at low speed, the relativistic momentum must agree with the Newtonian momentum.

This confirms Eq. (2.30).

In Chapter 3 we will see that this expression for the momentum also emerges from the Lagrangian formulation of relativistic mechanics (see Eq. 3.56).

The general definition of a (Lorentz)tensor is as follows:

A tensor of rank r is an object $A^{\mu\nu\ldots\varkappa}$ *with* 4^r *components that under a Lorentz transformation transforms according to*

$$A'^{\alpha\beta\ldots\gamma} = a^{\alpha}{}_{\mu} a^{\beta}{}_{\nu} \ldots a^{\gamma}{}_{\kappa} A^{\mu\nu\ldots\kappa} \tag{2.32}$$

Note that the object $A^{\mu\nu\ldots\varkappa}$ has *r* indices, and there is one Lorentz transformation matrix for each index.

As an example of a second-rank tensor, consider the 16-component object

$$x^{\mu} x^{\nu} \tag{2.33}$$

From the transformation law for x^{μ}, it is easy to see that this object does transform as a tensor.

In general, any object constructed by taking the "product" of two vectors is a second-rank tensor.

Thus, the objects

$$x^{\mu} p^{\nu} - x^{\nu} p^{\mu} \tag{2.34}$$

and

$$m u^{\mu} u^{\nu} \tag{2.35}$$

are such tensors or sums of such tensors.

The first is the angular-momentum tensor of a particle, and the second is related to the energy-momentum tensor (see later discussion).

We saw earlier that the position vector can be written either in the contravariant or in the covariant form. For an arbitrary contravariant vector A^{μ}, we define the covariant component A_{ν} by

$$A_\nu = \eta_{\nu\mu} A^\mu \tag{2.36}$$

This can be inverted to give

$$A^\mu = \eta^{\mu\nu} A_\nu \tag{2.37}$$

The operations performed with the metric tensor on the right sides of Eqs. (2.36) and (2.37) are called, respectively, lowering and raising an index.

The same procedure can be applied to one or more indices of a tensor of arbitrary rank.

For example,

$$\begin{aligned} A_\mu{}^\nu &= \eta_{\mu\alpha} A^{\alpha\nu} \\ A^\mu{}_\nu &= \eta_{\nu\alpha} A^{\mu\alpha} \\ A_{\mu\nu} &= \eta_{\mu\alpha}\eta_{\nu\beta} A^{\alpha\beta} \\ A^{\mu\nu} &= \eta^{\mu\alpha}\eta^{\nu\beta} A_{\alpha\beta} \end{aligned} \tag{2.38}$$

Our rules for raising and lowering indices can also be applied to the Lorentz-transformation matrix $a^{\mu}{}_{\nu}$. Thus, Eq. (2.22) can be written as

$$a_{\beta\mu} a^\beta{}_\nu = \eta_{\mu\nu} \tag{2.39}$$

If we raise the index μ, we obtain

$$a_\beta{}^\mu a^\beta{}_\nu = \eta^{\mu\alpha}\eta_{\alpha\nu}$$

But, by Eq. (2.18), the right side equals $\delta_{\mu}{}^{\nu}$, so

$$a_\beta{}^\mu a^\beta{}_\nu = \delta^\mu_\nu \tag{2.40}$$

This shows that $a_\beta{}^\mu$ is the inverse of $a^\beta{}_\nu$,

$$(a^{-1})^\mu{}_\beta = a_\beta{}^\mu \tag{2.41}$$

The matrix $a_\beta{}^\mu$ can be used to write the transformation law for covariant tensors in the form

$$A'_{\alpha\beta\ldots\gamma} = a_\alpha{}^\mu a_\beta{}^\nu \ldots a_\gamma{}^\kappa A_{\mu\nu\ldots\kappa} \tag{2.42}$$

This equation is obtained from (2.32) by lowering all upper indices and raising all lower indices.

In particular, for the covariant position vector, the transformation law (2.42) implies

$$x'_\mu = a_\mu{}^\nu x_\nu \tag{2.43}$$

Combining this with Eq. (2.41), we obtain

$$x_\nu = a^\mu{}_\nu x'_\mu \tag{2.44}$$

and this in turn gives the inverse Lorentz transformation

$$x^\nu = a_\mu{}^\nu x'^\mu \tag{2.45}$$

Differentiation of Eqs. (2.20) and (2.45) yields, respectively,

$$a^\mu{}_\nu = \frac{\partial x'^\mu}{\partial x^\nu} \tag{2.46}$$

$$a_\mu{}^\nu = \frac{\partial x^\nu}{\partial x'^\mu} \tag{2.47}$$

For any second-rank tensor $A^{\mu\nu}$, the quantity

$$A^\mu{}_\mu = a^0{}_0 + A^1{}_1 + A^2{}_2 + A^3{}_3 \tag{2.48}$$

is called the *trace* of the tensor.

This a one-component object that transforms as a scalar.

We will sometimes simply use the notation A for the trace of a tensor $A^{\mu\nu}$,

$$A \equiv A^{\mu}{}_{\mu} \tag{2.49}$$

$A^{\mu}{}_{\mu}$ remains invariant under Lorentz transformations, that is, $A'^{\mu}{}_{\mu} = A^{\mu}{}_{\mu}$ and also $A^{\mu}{}_{\mu} = A_{\mu}{}^{\mu}$.

Note that the trace of the metric tensor equals 4,

$$\eta_{\mu}{}^{\mu} = \eta^{\mu}{}_{\mu} = \delta^{\mu}_{\mu} = 4 \tag{2.50}$$

A tensor of rank higher than two has more than one trace. For example,

$$A^{\mu}{}_{\mu}{}^{\alpha}, A^{\mu\nu}{}_{\mu}, A^{\mu\alpha}{}_{\alpha} \tag{2.51}$$

are three different traces of a third-rank tensor $A^{\mu\nu\alpha}$.

The operation of taking the trace [as in Eqs. (2.48), (2.50), or (2.51)] is often called *contraction*.

Thus, $A^{\mu}{}_{\mu}{}^{\alpha}$ is the contraction of the first two indices of $A^{\mu\nu\alpha}$, and so on. It is obvious that contraction reduces the rank of a tensor by two.

For a tensor of the special type $B^{\mu}C^{\nu}$, the contraction gives the scalar

$$B^{\mu}C_{\mu} \tag{2.52}$$

This is sometimes simply written as $B\cdot C$ and called the scalar product of the vectors B^{μ} andC^{ν}.

For the four-velocity and the four-momentum, we have

$$u^{\mu}u_{\mu} = 1 \tag{2.53}$$

and

$$p^{\mu}p_{\mu} = m^2 \tag{2.54}$$

Note that in summations, such as those implied by the repeated dummy indices in Eqs. (2.51) and (2.52), one of the repeated dummy indices is always covariant, and one is contravariant.

If we were to form the sum $B^{\mu}C^{\mu}$, then the resulting object would not be a scalar – that is, the result would depend on the reference frame used in the evaluation; such an object is uninteresting.

We also note that $A^{\mu\mu\alpha}$ is not a vector.

A tensor $A^{\mu\nu}$ is said to be *symmetric* in μ, ν if

$$A^{\mu\nu} = A^{\nu\mu} \tag{2.55}$$

and it is said to be *antisymmetric* in μ, ν if

$$A^{\mu\nu} = -A^{\nu\mu} \tag{2.56}$$

Given an arbitrary tensor $B^{\mu\nu}$, it is always possible to write it as a linear combination of a symmetric and an antisymmetric part as follows:

$$B^{\mu\nu} = \frac{1}{2}(B^{\mu\nu} + B^{\nu\mu}) + \frac{1}{2}(B^{\mu\nu} - B^{\nu\mu}) \tag{2.57}$$

Obviously, the first term in parentheses is symmetric, and the second antisymmetric.

Equation (2.57) can also be written compactly as

$$B^{\mu\nu} = B^{(\mu\nu)} + B^{[\mu\nu]} \tag{2.58}$$

with the definitions

$$B^{(\mu\nu)} \equiv \frac{1}{2}(B^{\mu\nu} + B^{\nu\mu}) \tag{2.59}$$

$$B^{[\mu\nu]} = \frac{1}{2}(B^{\mu\nu} - B^{\nu\mu}) \tag{2.60}$$

Finally, given an arbitrary tensor $B^{\mu\nu}$ it is always possible to write it as a combination of a traceless tensor (that is, a tensor of zero trace) and $\eta^{\mu\nu}$ as follows:

$$B^{\mu\nu} = \left(B^{\mu\nu} - \tfrac{1}{4}\eta^{\mu\nu}B\right) + \tfrac{1}{4}\eta^{\mu\nu}B \tag{2.61}$$

Because $\eta_{\mu}{}^{\mu} = \delta_{\mu}{}^{\mu} = 4$, the term in parentheses is traceless.

2.3 Tensor fields

The definitions in the preceding section hold for tensors that are constant and also for tensors that vary in space and time; the latter are called tensor fields.

For example, a scalar field is a function $\phi(P)$ that has different values at different points *P* of spacetime.

At each given point of spacetime, the value of the scalar field remains invariant under Lorentz transformations:

$$\phi'(P) = \phi(P) \tag{2.62}$$

Instead of describing a given point of spacetime by the symbol *P* we can, of course, use coordinates.

If the point *P* has coordinates *x* and x', respectively, in the two coordinate systems, then Eq. (2.62) becomes

$$\phi'(x') = \phi(x) \tag{2.63}$$

Here, we use an abbreviated notation for spacetime variables, without superscripts.

In this abbreviated notation, *x* stands for x^0, x^1, x^2, x^3, and $\phi(x)$ stands for $\phi(x^0, x^1, x^2, x^3)$.

If the function $\phi(x)$ is given, then Eq. (2.63) determines the function $\phi'(x')$, because *x* can be expressed in terms of x' by the inverse Lorentz transformation, Eq. (2.45):

$$\phi'(x') = \phi(x(x')) \tag{2.64}$$

Note that $\phi'(x')$ is not the same function of x' as $\phi(x)$ is of x; for example, if $\phi(x) = (x^1)^2 + (x^2)^2$, then $\phi'(x') = (x'^1 + V x'^0)^2/(1 - V^2) + (x'^2)^2$, which is not $(x'^1)^2 + (x'^2)^2$.

The important thing is that each point of spacetime, $\phi'(x')$ and $\phi(x)$ take the same value.

This is what is meant by invariance, and the notation in Eq. (2.62) makes this clearer than Eq. (2.64).

We will often simply write $\phi' = \phi$, leaving it understood that the functions are to be evaluated at the same point.

In general, a tensor field is a function of spacetime such that the tensor transformation law Eq. (2.32) holds at each given point of spacetime.

It can be shown that all the fields that describe the state of (continuous) physical systems in classical physics must necessarily be tensor fields.

This important theorem rests on a group-theoretical analysis of the Lorentz transformations, and therefore we must forego the proof (Wigner).

The theorem tells us that fields with more complicated transformation laws are not admissible.

It follows that the only building blocks that we can use in the construction of a classical field theory are scalars, vectors, tensors, and so on.

We will rely heavily on this fact in the next chapter.

The gradient operator in spacetime is

$$\partial_\mu \equiv \left(\frac{\partial}{\partial x^0}, \frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2}, \frac{\partial}{\partial x^3} \right) \tag{2.65}$$

The chain rule for differentiation and Eq. (2.47) tell us that

$$\partial'_\mu = \frac{\partial}{\partial x'^\mu} = \frac{\partial x^\nu}{\partial x'^\mu}\frac{\partial}{\partial x^\nu} = a_\mu{}^\nu \partial_\nu \tag{2.66}$$

This shows that ∂_μ transforms as a covariant vector.

We can raise the index of ∂_μ to obtain

$$\partial^\mu \equiv \eta^{\mu\nu}\partial_\nu = \left(\frac{\partial}{\partial x^0}, -\frac{\partial}{\partial x^1}, -\frac{\partial}{\partial x^2}, -\frac{\partial}{\partial x^3}\right) \tag{2.67}$$

This operator has the transformation law of a contravariant vector,

$$\partial'^\mu = a^\mu{}_\nu \partial^\nu \tag{2.68}$$

The "Laplacian" operator in spacetime is

$$\partial^\mu\partial_\mu = \left(\frac{\partial}{\partial x^0}\right)^2 - \left(\frac{\partial}{\partial x^1}\right)^2 - \left(\frac{\partial}{\partial x^2}\right)^2 - \left(\frac{\partial}{\partial x^3}\right)^2 \tag{2.69}$$

This operator is invariant under Lorentz transformations:

$$\partial'^\mu\partial'_\mu = \partial^\mu\partial_\mu \tag{2.70}$$

As a consequence of Eq. (2.68), $\partial^\mu\phi$ is a vector whenever ϕ is a scalar.

It also follows that $\partial^\mu\partial^\nu\phi$ is a tensor and $\partial^\mu\partial_\mu\phi$ is a scalar.

In general, when ∂^μ operates on a tensor of rank r, the result is a tensor of rank r + 1.

$\partial^\mu A^{\alpha\beta\ldots\varkappa}$ is a tensor of rank $r + 1$ whenever $A^{\alpha\beta\ldots\varkappa}$ is a tensor of rank r

It is convenient to indicate derivatives by commas, in the style of the following examples:

$$
\begin{aligned}
\phi_{,\mu} &\equiv \partial_\mu \phi \\
\phi^{,\mu} &\equiv \partial^\mu \phi \\
\phi^{,\mu,\nu} &\equiv \partial^\nu \partial^\mu \phi \\
\phi^{,\mu}{}_{,\mu} &\equiv \partial_\mu \partial^\mu \phi \\
A^{\alpha\beta}{}_{,\mu} &\equiv \partial_\mu A^{\alpha\beta}
\end{aligned}
\tag{2.71}
$$

This concludes our discussion of the mathematical aspects of special relativity.

More on tensor analysis in 4-D spacetime can be found in textbooks on special relativity.

We now turn to some important applications of the formalism.

2.4 Energy-momentum tensor

In our search for a relativistic theory of gravitation the energy density will play an important role.

But it makes little sense to consider the energy density by itself, because what is energy density in one reference frame is some combination of energy density, energy flux density, and momentum flux density as seen from another reference frame.

Hence all these quantities form a single tensor object, and they must be considered together.

As a simple example of the construction of an energy-momentum tensor, consider a system consisting of a collection of noninteracting particles (a cloud of dust).

Suppose that in the neighborhood of a point in this cloud the density of particles is n per unit volume and their velocity is $\mathbf{v}$.

In this case the *energy density* can be expressed in terms of the density and the velocity of the particles:

$$T^{00} = \frac{nm}{\sqrt{1 - v^2/c^2}} \tag{2.72}$$

This expression follows the convention $c = 1$.

Restoration of factors of c would give $T^{00} = nmc^2/\sqrt{1 - v^2/c^2}$ in cgs units.

The *energy flux density* can be defined in the following way: The energy flux in the x-direction is the amount of energy transported in unit time across a unit y-z area, that is,

$$\frac{nmv_x}{\sqrt{1 - v^2/c^2}} \tag{2.73}$$

This is the product of the particle current density nv_x by the energy per particle.

In general, the energy flux density in the k direction is

$$T^{k0} = \frac{nmv^k}{\sqrt{1 - v^2/c^2}} \tag{2.74}$$

As an alternative notation in Eq. (2.74), we use T^{0k} so $T^{k0} = T^{0k}$.

Note that T^{k0} can also be regarded as the *density of momentum*.

Finally, let us define the *momentum flux density*.

The x-y momentum flux density is defined as the amount of x momentum that flows in the y direction per unit area and unit time.

Since the x-momentum density is

$$\frac{nmv_x}{\sqrt{1 - v^2}}$$

we must multiply this by the y-velocity to obtain the x-y momentum flux density,

$$\left(\frac{nmv_x}{\sqrt{1-v^2}}\right)v_y. \tag{2.75}$$

The general expression for the *k*-*l* momentum flux density is

$$T^{kl} = \frac{nmv^k v^l}{\sqrt{1-v^2}} \tag{2.76}$$

It is now easy to show that the 16-component object $T^{\mu\nu}$ given by Eqs. (2.72), (2.74), and (2.76) is a tensor under Lorentz transformations.

To show this, we write $T^{\mu\nu}$ as follows:

$$T^{\mu\nu} = n_0 m u^\mu u^\nu \tag{2.77}$$

where

$$n_0 \equiv n\sqrt{1-v^2} \tag{2.78}$$

The quantity n_0, called the *proper particle density*, is the particle density as measured in a reference frame that moves with the particles.

The relation between n_0 and n given by Eq. (2.78) is a consequence of the well-known volume-contraction effect of special relativity: A volume containing a given number of particles and moving along with them is contracted as measured in the laboratory frame; hence the laboratory density n is increased over the proper density n_0.

Since n_0 is a number measured in the local rest frame, it is a scalar.

Hence $n_0 m u^\mu u^\nu$ is the product of a scalar n_0 by the tensor $m u^\mu u^\nu$, and therefore $n_0 m u^\mu u^\nu$ is a tensor, that is, $T^{\mu\nu}$ is a tensor.

Note that this tensor is symmetric,

$$T^{\mu\nu} = T^{\nu\mu} \tag{2.79}$$

and note that it can also be expressed as

$$T^{\mu\nu} = \rho_0 u^\mu u^\nu \tag{2.80}$$

where

$$\rho_0 \equiv n_0 m \tag{2.81}$$

is the mass density as measured in the local rest frame of the particles, called the *proper mass density*.

The definition of the energy-momentum tensor $T^{\mu\nu}$ can be summarized as follows:

$$\begin{aligned} T^{00} &= \text{energy density} \\ T^{0k} &= T^{k0} = k\text{-momentum density } = \text{energy flux density in } k\text{-direction} \\ T^{kl} &= T^{lk} = k\text{-momentum flux density in } l\text{-direction} \end{aligned} \tag{2.82}$$

This definition applies in general.

For instance, if instead of a system of dust particles we have a system consisting of an ideal fluid, with a local flow velocity **v** and a pressure p, the energy-momentum tensor is

$$T^{\mu\nu} = (\rho_0 + p)u^\mu u^\nu - p\eta^{\mu\nu} \tag{2.83}$$

This differs from Eq. (2.80) by extra terms involving the pressure.

Whereas the dust particles produce an energy and momentum flux only by convective transport, the pressure of the fluid produces a momentum flux even if the fluid is at rest, because the internal pressure forces are a diagonal stress, which produces a flow of x-momentum in the x-direction, y-momentum in the y-direction, and z-momentum in the z-direction.

The expression (2.83) can be readily derived by beginning with the energy-momentum tensor in the rest frame of a small volume of fluid, where $T^{\mu\nu}$ is purely diagonal, with components

$$T^{00} = \rho_0, \quad T^{11} = T^{22} = T^{33} = p \tag{2.84}$$

Then perform a Lorentz transformation with velocity $-\mathbf{v}$, so as to give this small volume the appropriate flow velocity $\mathbf{v}$; this immediately leads to the result (2.83).

The definition (2.82) also applies to a system consisting of electromagnetic fields; we will deal with this case in the next section.

The energy-momentum tensor is usually a function of space and time.

For example, in Eq. (2.80) both the particle density and the velocity may vary in space and time.

The derivatives of the energy-momentum tensor of a complete system satisfy the (differential) conservation law

$$\partial_\nu T^{\mu\nu} = 0 \tag{2.85}$$

The proof of this conservation law is no harder than the well-known proof for the conservation law of charge.

Consider the $\mu = 1$ component of Eq. (2.85):

$$\frac{\partial}{\partial x^0} T^{10} + \frac{\partial}{\partial x^k} T^{1k} = 0 \tag{2.86}$$

To see how this equation arises from the conservation of momentum, consider a small volume in the system.

The x-momentum in the volume is $\int T^{10} d^3x$

and the rate of decrease of this x-momentum is

$$-\int \frac{\partial}{\partial x^0} T^{10} d^3x \tag{2.87}$$

This momentum decrease must be the result of x-momentum flowing out of the volume. The total outflow of x-momentum across the boundarv of the volume is

$$\int T^{1k} dS_k \tag{2.88}$$

We can apply Gauss's theorem to the three-component object T^{1k} ($k = 1, 2, 3$) and convert the surface integral (2.88) into a volume integral,

$$\int \frac{\partial}{\partial x^k} T^{1k} d^3x \tag{2.89}$$

By conservation of momentum, the terms (2.87) and (2.89) must be equal.

If the volume is small enough, the integrands are nearly constant and the integration can be replaced by a multiplication of integrand by volume:

$$\left(-\frac{\partial}{\partial x^0} T^{10}\right) \Delta x \Delta y \Delta z = \left(\frac{\partial}{\partial x^k} T^{1k}\right) \Delta x \Delta y \Delta z$$

This reduces to Eq. (2.86) if we cancel the volumes.

The proof for the other components of Eq. (2.85) is similar.

According to our definition of $T^{\mu\nu}$, the volume integrals of T^{00} and of T^{k0} give, respectively, the total energy and the total momentum of the system:

$$P^0 = \int T^{00} d^3x \tag{2.90}$$

$$P^k = \int T^{k0} d^3x \tag{2.91}$$

This four-component object P^{μ} ought to be a four-vector, but it is not at all obvious that the volume integrals appearing on the right side of Eqs. (2.90) and (2.91) transform correctly under Lorentz transformations, because these volume integrals must be evaluated at one instant of time, and the meaning of an instant of time is different in the two reference frames; that is, the hypersurfaces x^0 = constant and x'^0 = constant are different, so the "time slicing" is different.

The proof that the integrals (2.90) and (2.91) indeed transform as a four-vector hinges on a clever argument involving the conservation law (2.85).

We begin by multiplying this conservation law by an arbitrary *constant* vector B_ν , so

$$\frac{\partial}{\partial x^{\mu}}(T^{\nu\mu}B_{\nu}) = 0 \tag{2.92}$$

Integration over a 4-D volume then gives

$$\int \frac{\partial}{\partial x^{\mu}}(T^{\nu\mu}B_{\nu})dtd^3x = 0 \tag{2.93}$$

The quantity in parentheses in the integrand is a four-vector.

Accordingly, the left side of Eq. (2.93) is a 4-D volume integral of the divergence of a 4-D vector.

We can apply the 4-D analog of the familiar Gauss theorem to convert the volume integral into a "surface" integral,

$$\int \frac{\partial}{\partial x^{\mu}}(T^{\nu\mu}B_{\nu})dtd^3x = \int (T^{\nu\mu}B_{\nu})dS_{\mu} \tag{2.94}$$

Since the volume is 4-D, its "surface" is actually a 3-D hypersurface.

The vector dS_μ has a magnitude equal to the corresponding 3-D volume element and a direction perpendicular to the hypersurface; that is, $dS_\mu \delta x^\mu = 0$ whenever δx^μ is any displacement vector within the hypersurface.

For simplicity, we will assume that $T^{\nu\mu}$ differs from zero only in some finite region of space.

If we consider this finite region of space at successive instants of time, then in spacetime $T^{\nu\mu}$ is nonzero in a four-dimensional tube of infinite length in the time direction and of a finite cross section in the space directions.

Figure 2.5 shows this tube in the $x^0 - x^1$ plane; the two vertical lines form the boundaries of the tube, and $T^{\nu\mu}$ is different from zero only between these boundaries.

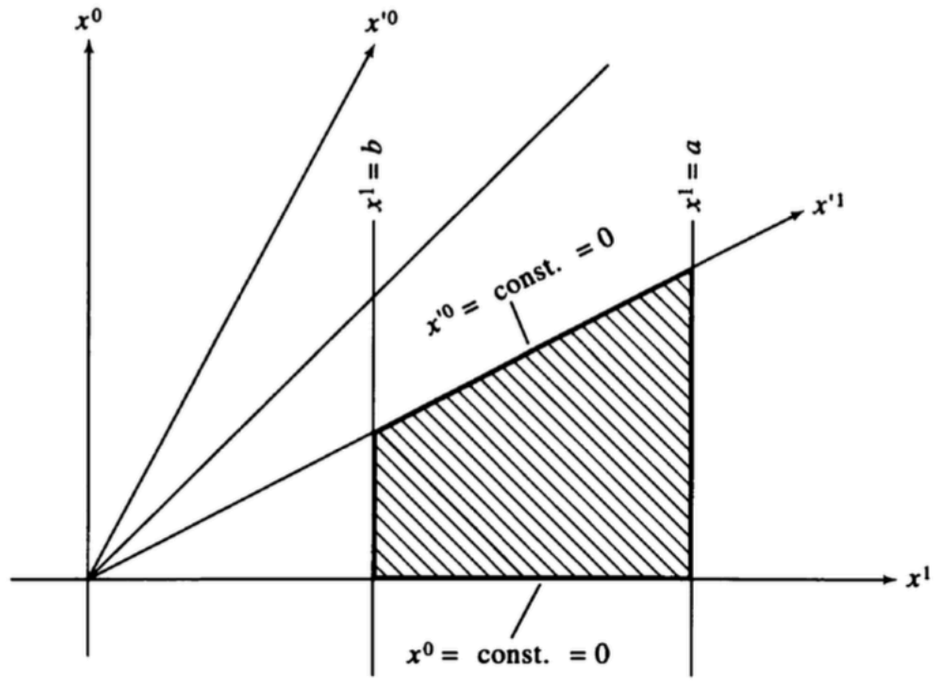


Fig. 2.5 The energy-momentum tensor is different from zero in the region included between the boundaries $x^1 = b$ and $x^1 = a$.

Consider now a Lorentz transformation that introduces new coordinates x'^{μ} according to Eq. (2.20). The equations

$$x^0 = \text{constant} \tag{2.95}$$

and

$$x'^0 = \text{constant} \tag{2.96}$$

define 3-D hypersurfaces in spacetime, that is, they define "time slices" in spacetime.

These surfaces are distinct; if we express Eq. (2.96) in terms of x coordinates, it becomes $x^0 + V x^1 =$ constant $\times \sqrt{1 - V^2}$, which differs from Eq. (2.95).

In Fig. 2.5, the hypersurfaces (2.95) and (2.96) are represented by the heavy lines that slice across the tube, and it has been assumed that the constants in Eqs. (2.95) and (2.96) are zero.

We now concentrate on that part of the tube included between these surfaces; in Fig. 2.5 this 4-D region is shown hatched.

If we apply Eqs. (2.93) and (2.94) to this region, we obtain

$$0 = \int T^{\nu\mu} B_\nu dS_\mu = \int_I T^{\nu\mu} B_\nu dS_\mu + \int_{II} T^{\nu\mu} B_\nu dS_\mu + \int_{III} T^{\nu\mu} B_\nu dS_\mu \tag{2.97}$$

where the surface *I* is at the top [Eq. (2.96)], the surface *II* is at the bottom [Eq. (2.95)], and the surface *III* is the "lateral" surface of the tube (see Fig. 2.5).

Since $T^{\nu\mu} = 0$ on the lateral surface, we can ignore the integral $\int_{III}$ and we obtain

$$\int_I T^{\nu\mu} B_\nu dS_\mu = -\int_{II} T^{\nu\mu} B_\nu dS_\mu \tag{2.98}$$

The surface element appearing on the right side is $dS_{..} = (-d^3x, 0, 0, 0)$, and hence the right side reduces to

$$\int_{II} T^{\nu 0} B_\nu d^3x \tag{2.99}$$

The surface element appearing on the left side of Eq. (2.98) is more complicated.

However, we can take advantage of the fact that the complete integrand is a scalar ($T^{\nu\mu}$ is a tensor, B_ν and dS_μ are vectors), so in terms of the new coordinates

$$\int_I T^{\nu\mu} B_\nu dS_\mu = \int_I T'^{\nu\mu} B'_\nu dS'_\mu \tag{2.100}$$

In the new coordinates $dS'_\mu = (d^3x', 0, 0, 0)$, and it follows that the left side of Eq. (2.98) is simply

$$\int_I T'^{\nu 0} B'_\nu d^3x' \tag{2.101}$$

We have therefore established the equality

$$B'_\nu \int_I T'^{\nu 0} d^3x' = B_\nu \int_{II} T^{\nu 0} d^3x \tag{2.102}$$

which shows that the quantity $B_\nu T^{\nu 0} d^3x$ is a scalar under Lorentz transformations.

Since, by hypothesis, B_ν is a covariant vector, we can rewrite Eq. (2.102) as

$$a_\mu{}^\nu B_\nu \int_I T'^{\mu 0} d^3x' = B_\nu \int_{II} T^{\nu 0} d^3x \tag{2.103}$$

And because the components B_ν are arbitrary, the coefficients of these components on both sides of this equation must be equal, so we can effectively cancel B_ν on both sides.

The resulting equation shows that $\int T^{\nu 0} d^3x$ is a contravariant vector.

This is what we wanted to prove.

As a corollary, we can now demonstrate Einstein's famous $E_0 = Mc^2$ relation between the energy in the rest frame of a system and the mass of that system.

The energy and momentum of a general system are given by the integrals involving the energy-momentum tensor, and we can then easily confirm that the $E_0 = Mc^2$ relation is a direct consequence of the four-vector character of the energy and momentum.

By definition, the momentum of the system is zero in its rest frame, so the energy-momentum four-vector is $P^{\mu} = (E_0, 0, 0, 0)$ in this frame.

A Lorentz transformation with velocity v in the, say, negative x-direction then gives the system a translational velocity and an energy-momentum vector

$$P'^{\mu} = a^{\mu}{}_{\nu} P^{\nu} = \left(\frac{E_0}{\sqrt{1 - v^2}}, \frac{E_0 v}{\sqrt{1 - v^2}}, 0, 0 \right) \tag{2.104}$$

Comparison with the energy-momentum vector of a particle [see Eq. (2.31)], shows that here E_0 plays the role of mass, with $E_0 = M$ or, in cgs units, $E_0 = Mc^2$.

Note that the difficulty in establishing this result is not in Eq. (2.104), which is trivial, but in the preliminary proof of the four-vector character of the energy and momentum of a general system, which requires sophisticated manipulations of 4-D integrals.

Apart from establishing Einstein's iconic result about energy and mass, this demonstration illustrates the crucial role in relativity of the differential conservation law for energy and momentum.

The conservation law $\partial_{\nu} T^{\mu\nu} = 0$ is, in essence, the relativistic version of the Newtonian conservation law for the total momentum of a system, $dP^k/dt = 0$, when external forces are absent.

The Newtonian conservation law hinges on the second law and also on the third law, because it is necessary to take into account the equal magnitudes and opposite directions of the action and reaction forces within the system, to ensure that the momentum gained by one part of the system is always compensated by momentum lost by another part.

Likewise, in relativity, the conservation law $\partial_{\nu} T^{\mu\nu} = 0$ takes into account that the momentum gained by one part of a system is compensated by momentum lost by another, adjacent part.

The relativistic and Newtonian momentum transfers differ in that the relativistic transfers are strictly local, between one place and an adjacent place.

This locality of momentum transfers is necessary to ensure that they occur simultaneously in all reference frames.

If the transfers were to occur between two distant places, then the relativity of simultaneity would make it impossible to guarantee that transfers that are simultaneous in one reference frame remain simultaneous in another – and momentum conservation would be violated in some reference frames, at least intermittently.

If the system is acted upon by an external force, then the relativistic dynamical equation that takes the place of Newton's second law is

$$\partial_\nu T^{\mu\nu} = f^\mu \tag{2.105}$$

Here f^μ is the force density, or the four-force per unit volume, arising from sources external to the system.

The space components f^1, f^2, f^3 of this four-force density are simply the components of the ordinary force per unit volume, and the time component f^0 is the power per unit volume delivered by the force.

Equation (2.105) is the fundamental dynamical equation for relativistic systems.

It is difficult to apply in practice, because the entire system must be treated as an elastic body, whose shape changes while it is being accelerated by the external force (rigid bodies cannot exist in relativity).

Furthermore, in the presence of an external force, the integrals of the energy and momentum densities [Eqs. (2.90) and (2.91)] do not have the simple Lorentz transformation properties we established earlier [Eq. (2.103)], and this leads to some ugly complications when we try to transform the equation of motion of the system from one reference frame to another.

2.5 Relativistic electrodynamics

In this section we examine the implications of the principle of relativity for Maxwell's equations.

In general, the principle of relativity demands the invariance of the laws of physics under Lorentz transformations.

It is obvious that if $A^{\mu\nu\ldots\varkappa}$ is a tensor of rank r, an equation of the form

$$A^{\mu\nu\ldots\kappa} = 0 \tag{2.106}$$

satisfies the principle of relativity. In fact, multiplication of Eq.(2.106) by $a^{\alpha}{}_{\mu}a^{\beta}{}_{\nu}\ldots a^{\gamma}{}_{\varkappa}$ yields $A'^{\alpha\beta\ldots\gamma} = 0$, which shows that the law (2.106) is invariant under the Lorentz transformation specified by the matrix $a^{\alpha}{}_{\mu}$.

Our next objective is to show that the equations of electrodynamics have the form of tensor equations, like Eq. (2.106).

In cgs Gaussian units, with the charge density and current density in esu Maxwell's equations of electrodynamics are

$$\nabla \cdot \mathbf{E} = 4\pi\rho \tag{2.107}$$

$$\nabla \times \mathbf{B} - \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} = \frac{4\pi}{c}\mathbf{j} \tag{2.108}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{2.109}$$

$$\nabla \times \mathbf{E} + \frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} = 0 \tag{2.110}$$

These equations do not seem to be of the type (2.106), and their Lorentz invariance is not at all obvious.

To discuss the Lorentz-transformation properties of Eqs. (2.107)–(2.110), we begin by rewriting them in 4-D component notation as follows:

$$\partial_\mu F^{\mu\nu} = \frac{4\pi}{c} j^\nu \tag{2.111}$$

$$\partial^\alpha F^{\mu\nu} + \partial^\mu F^{\nu\alpha} + \partial^\nu F^{\alpha\mu} = 0 \tag{2.112}$$

where the four-component current density j^ν is defined as

$$j^\nu \equiv (c\rho,\, j_x,\, j_y,\, j_z) \tag{2.113}$$

and the 16-component tensor $F^{\mu\nu}$ is defined as

$$F^{\mu\nu} = \begin{pmatrix} F^{00} & F^{01} & F^{02} & F^{03} \\ F^{10} & F^{11} & F^{12} & F^{13} \\ F^{20} & F^{21} & F^{22} & F^{23} \\ F^{30} & F^{31} & F^{32} & F^{33} \end{pmatrix} \equiv \begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{pmatrix} \tag{2.114}$$

The tensor $F^{\mu\nu}$ is called the *electromagnetic field tensor* (and we will establish its tensor transformation properties on the next several pages).

Note that this tensor is antisymmetric, so only 6 of its 16 components are independent.

With these definitions, it is a straightforward exercise to check that the components of Eq. (2.111) are equivalent to those of Eqs. (2.107) and (2.108) and that the components of Eq. (2.112) are equivalent to those of Eqs. (2.109) and (2.110).

For instance, if $\nu = 0$, Eq. (2.111) gives

$$\partial_0 F^{00} + \partial_1 F^{10} + \partial_2 F^{20} + \partial_3 F^{30} = \frac{4\pi}{c} j^0$$

According to Eqs. (2.113) and (2.114) this is simply

$$\frac{\partial}{\partial x} E_x + \frac{\partial}{\partial y} E_y + \frac{\partial}{\partial z} E_z = 4\pi\rho$$

which agrees with Eq. (2.107).

Written in the form (2.111) and (2.112), Maxwell's equations look like tensor equations [and they can be rewritten in the form of Eq. (2.106) by moving all terms to the left side].

However, appearances can be deceiving, and unless $F^{\mu\nu}$ is a second-rank tensor and j^{μ} is a four-vector, Maxwell's equations will not be tensor equations.

Before we jump to the conclusion that the equations are tensor equations, we must examine the Lorentz-transformation properties of j^{μ} and $F^{\mu\nu}$.

To establish that the object j^{μ} is a four-vector, we recall the definition of charge and current density.

Suppose we have a cloud of electrically charged particles (a fluid), all of the same kind.

In the neighborhood of any given point, we describe this distribution of particles by the number per unit volume and by the velocity.

Both the number density and the velocity are functions of space and time:

$$\begin{aligned} n &= n(x) \\ \mathbf{v} &= \mathbf{v}(x) \end{aligned} \tag{2.115}$$

The charge density is then defined by

$$\rho(x) = qn(x) \tag{2.116}$$

and the current density by

$$\mathbf{j}(x) = qn\mathbf{v}(x). \tag{2.117}$$

In these equations q is the charge of one particle.

For example, $q = -e$ for electrons, $q = +e$ for protons, and $q = 0$ for neutrons.

If several types of particles are present, then we must combine their charge and current densities additively.

Note that as a consequence of the conservation of the number of particles, the charge is conserved, so

$$\nabla \cdot \mathbf{j} + \frac{\partial \rho}{\partial t} = 0 \tag{2.118}$$

Conservation of the number of each kind of particle is sufficient, but not necessary, for the validity of Eq. (2.118).

Obviously all that is necessary is that in any reaction that destroys or creates particles, the net charge should be the same before and after.

In view of Eqs. (2.116) and (2.117), we can express the four-component object j^ν as

$$j^\nu = (c\rho,\, j_x,\, j_y,\, j_z) = (cqn,\, qnv_x,\ qnv_y,\ qnv_z) \tag{2.119}$$

We can also write this as

$$j^\nu = n_0 q u^\nu \tag{2.120}$$

or as

$$j^\nu = \rho_0 u^\nu \tag{2.121}$$

where

$$n_0 = n\sqrt{1 - v^2/c^2} \tag{2.122}$$

and

$$\rho_0 = q n_0 \tag{2.123}$$

Here n_0 is the proper particle density [see Eq. (2.78)] and ϱ_0 the proper charge density; these proper quantities are scalars.

Since j^ν consists of u^ν multiplied by a scalar, it must have the transformation law of a four-vector,

$$j'^\beta = a^\beta{}_\nu j^\nu \tag{2.124}$$

In terms of this current density four-vector, we can write the conservation law (2.118) as

$$\partial_\nu j^\nu = 0 \tag{2.125}$$

This equation is of the general form (2.106), and its Lorentz invariance is obvious.

Next, we want to establish the Lorentz-transformation properties of the 16-component object $F^{\mu\nu}$.

In contrast to j^ν, for which we could rely on the known Lorentz- transformation properties of the building blocks ϱ_0 and u^ν out of which j^ν is constructed, we now have no such building blocks.

Instead, we must appeal to the principle of relativity, which tells us that under a Lorentz transformation the field equation (2.111) must become

$$\partial'_\alpha F'^{\alpha\beta} = \frac{4\pi}{c} j'^\beta$$

Since $\partial'_\alpha = a_\alpha{}^\mu \partial_\mu$ and $j'^\beta = a^\beta{}_\sigma j^\sigma$, we can also express this as

$$a_\alpha{}^\mu \partial_\mu F'^{\alpha\beta} = \frac{4\pi}{c} a^\beta{}_\sigma j^\sigma$$

If we multiply this by $a_\beta{}^\nu$ and use Eq. (2.40), we obtain

$$a_\beta{}^\nu a_\alpha{}^\mu \partial_\mu F'^{\alpha\beta} = \frac{4\pi}{c} j^\nu \tag{2.126}$$

We now compare Eqs. (2.126) and (2.111). Agreement between these equations demands

$$a_\beta{}^\nu a_\alpha{}^\mu \partial_\mu F'^{\alpha\beta} = \partial_\mu F^{\mu\nu} \tag{2.127}$$

To extract the transformation properties of $F^{\mu\nu}$ from this equation, we note that, on physical grounds, the relationship between $F'^{\mu\nu}$ and $F^{\mu\nu}$ must be linear and homogeneous, that is, it must be of the form

$$F^{\mu\nu} = C^{\mu\nu}_{\alpha\beta} F'^{\alpha\beta} \tag{2.128}$$

where $C_{\alpha\beta}^{\mu\nu}$ is a set of constant numerical coefficients, independent of the values of the electromagnetic fields.

If we substitute Eq. (2.128) into Eq. (2.127), we obtain

$$a_\beta{}^\nu a_\alpha{}^\mu \partial_\mu F'^{\alpha\beta} = \partial_\mu C_{\alpha\beta}^{\mu\nu} F'^{\alpha\beta}$$

Because $\partial_\mu F'^{\alpha\beta}$ can attain any arbitrary value at a given point of spacetime, subject only to a restriction for antisymmetry in α, β, the coefficients multiplying $\partial_\mu F'^{\alpha\beta}$ on both sides of this equation must be equal to within an undetermined tensor $D^{\mu\nu}$ symmetric in α, β:

$$a_\beta{}^\nu a_\alpha{}^\mu = C_{\alpha\beta}^{\mu\nu} + D_{\alpha\beta}^{\mu\nu}$$

The undetermined tensor does not affect Eq. (2.128), which now becomes

$$F^{\mu\nu} = a_\beta{}^\nu a_\alpha{}^\mu F'^{\alpha\beta} \tag{2.129}$$

This confirms that $F^{\mu\nu}$ is a second-rank tensor and that Eqs. (2.111) and (2.112) are tensor equations of the general type of Eq. (2.106), which are in accord with the principle of relativity.

From the tensor-transformation equations for the field tensor $F^{\mu\nu}$ we can readily extract the transformation equations for the electric and magnetic fields, which are components of $F^{\mu\nu}$ [see Eq. (2.114)].

If one uses the tensor-transformation equation $F'^{\mu\nu} = a^\mu{}_\alpha a^\nu{}_\beta F^{\alpha\beta}$ and the Lorentz-transformation matrix (2.21) to obtain the transformation equations for E_x , E_y , E_z and B_x, B_y, B_z one can show that these transformation equations for the *x*-components (parallel to the velocity **V**) can be written as

$$\mathbf{E'}_{\parallel} = \mathbf{E}_{\parallel} \qquad \mathbf{B'}_{\parallel} = \mathbf{B}_{\parallel} \tag{2.130}$$

and for the y- and z-components (perpendicular to the velocity) they can be written as

$$\mathbf{E'}_{\perp} = \frac{1}{\sqrt{1-V^2}}(\mathbf{E}_{\perp} + \mathbf{V} \times \mathbf{B}) \qquad \mathbf{B'}_{\perp} = \frac{1}{\sqrt{1-V^2}}(\mathbf{B}_{\perp} - \mathbf{V} \times \mathbf{E}) \tag{2.131}$$

The equation of motion for a charged particle can be written in terms of the field tensor $F^{\mu\nu}$.

The equation of motion is

$$\frac{d\mathbf{p}}{dt} = q\mathbf{E} + \frac{q}{c}\mathbf{v} \times \mathbf{B} \tag{2.132}$$

where $\mathbf{p} = m\mathbf{v}/\sqrt{1 - v^2/c^2}$ is the relativistic momentum.

It is easy to check that Eq. (2.132) is equivalent to

$$\frac{dp^{\mu}}{d\tau} = \frac{q}{m} p_{\nu} F^{\mu\nu} \tag{2.133}$$

The Lorentz invariance of this equation is obvious.

In later chapters, we will need the energy-momentum tensor of the electromagnetic field (and of other systems).

The energy-momentum tensor of the electromagnetic field is

$$T^{\mu\nu} = -\frac{1}{4\pi}\left(F^{\mu\alpha}F^{\nu}{}_{\alpha} - \frac{1}{4}\eta^{\mu\nu}F^{\alpha\beta}F_{\alpha\beta}\right) \tag{2.134}$$

A complete derivation of this expression is given in Appendix A.1.

The T^{00} component is the familiar energy density of the electromagnetic field,

$$T^{00} = \frac{1}{8\pi}(\mathbf{E}^2 + \mathbf{B}^2) \tag{2.135}$$

and the T^{0k} components are the components of the Poynting vector,

$$T^{0k} = \frac{1}{4\pi}(\mathbf{E} \times \mathbf{B})^k \tag{2.136}$$

Note that the trace of $T^{\mu\nu}$ is zero:

$$T^{\mu}{}_{\mu} = 0 \tag{2.137}$$

If we take for granted that the energy density in the electromagnetic field is $(\mathbf{E}^2 + \mathbf{B}^2)/8\pi$, we can give a very simple derivation of the expression (2.134) for the energy- momentum tensor.

The argument is this: Suppose Eq. (2.135) for the energy density is true in every reference frame.

Since (2.134) agrees with (2.135), it follows that $T^{\mu\nu}$ can differ from the expression (2.134) only by a (symmetric) tensor $B^{\mu\nu}$ whose B^{00} component vanishes identically in every reference frame.

It is easy to see that in this case all components of $B^{\mu\nu}$ must vanish.

Finally, we present the relativistic formulation of the electromagnetic potentials. The usual connection between fields and potentials is

$$\mathbf{E} = -\nabla\phi - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} \tag{2.138}$$

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{2.139}$$

where ϕ is the "scalar" potential and $\mathbf{A}$ the "vector" potential (these are scalars and vectors only in the 3-D sense!).

If we define a four-component object

$$A^\mu = (\phi, A_x, A_y, A_z) \tag{2.140}$$

then we can write Eqs. (2.138) and (2.139) simply as

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu \tag{2.141}$$

Because the left side of Eq. (2.141) is a tensor, it follows that A^μ must be a four-vector, usually called the four-vector potential.

If we insert the expression (2.141) into the Maxwell equations (2.111) and (2.112), we find that the second of these is satisfied identically, whereas the first gives

$$\partial_\mu \partial^\mu A^\nu - \partial^\nu \partial_\mu A^\mu = \frac{4\pi}{c} j^\nu \tag{2.142}$$

This is the wave equation for the potential.

We will not attempt the solution of the relativistic Maxwell equations, because the application of these equations to electromagnetic problems is beyond the scope of our discussion.

We are interested in electrodynamics only as a mathematical example of a relativistic field theory.

In the next chapter we will see how the understanding of the electromagnetic field equations leads to an understanding of the gravitational field equations.

2.6 Differential forms and exterior calculus

In the preceding sections we have adopted the point of view that, in four dimensions, a vector or a tensor is an *n*-tuple of real numbers.

In three dimensions an alternative, "geometrical" point of view enjoys much popularity: A vector is thought of as a directed line segment, or an arrow, and it is designated by a vector symbol, such as $\mathbf{A}$ or $\vec{\mathrm{A}}$.

This point of view not only provides us with a nice intuitive mental picture of vectors but also proves very convenient when we attempt to formulate vector identities, especially identities involving the curl and those involving vector integrals.

For instance, the identities $\nabla\cdot(\nabla\times\mathbf{A}) = 0$, $\nabla\times(\nabla\phi) = 0$, and $\mathbf{B}\cdot d\mathbf{l} = (\nabla\times\mathbf{B})\cdot d\mathbf{S}$ (Stokes' theorem) have a very concise form in this geometrical vector language, but they look rather awkward when expressed in component language.

The connection between the "geometrical" description of a vector and the component description is given by

$$\mathbf{A} = A^1\mathbf{e}_1 + A^2\mathbf{e}_2 + A^3\mathbf{e}_3 \tag{2.143}$$

where $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ are the three unit vectors along the x-, y-, and z- axes, respectively.

Of course, in four dimensions, we could adopt an analogous picture of a vector as a directed line segment in spacetime, and we could express such a "geometric" vector as a superposition of four unit vectors $\mathbf{e}_0$, $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$.

But this point of view is of limited usefulness for several reasons.

First, much of its intuitive appeal is lost, because we are unable to visualize line segments in four dimensions.

More seriously, the naıve geometric picture of a vector as a directed line segment does not admit of generalization to *curved* spacetime and is therefore hardly worth bothering with.

We can see why curved space leads to trouble for the naive geometric picture of vectors in the following simple example of a curved 2-D space: Pretend that the surface of the Earth is a perfect sphere, and consider a displacement of 4,000 km due east along the surface of the sphere and a displacement of 4,000 km due north.

Can we regard these displacements as vectors, subject to the usual rules for vector addition?
We cannot, because the two displacements *do not commute* (see Fig. 2.6a).

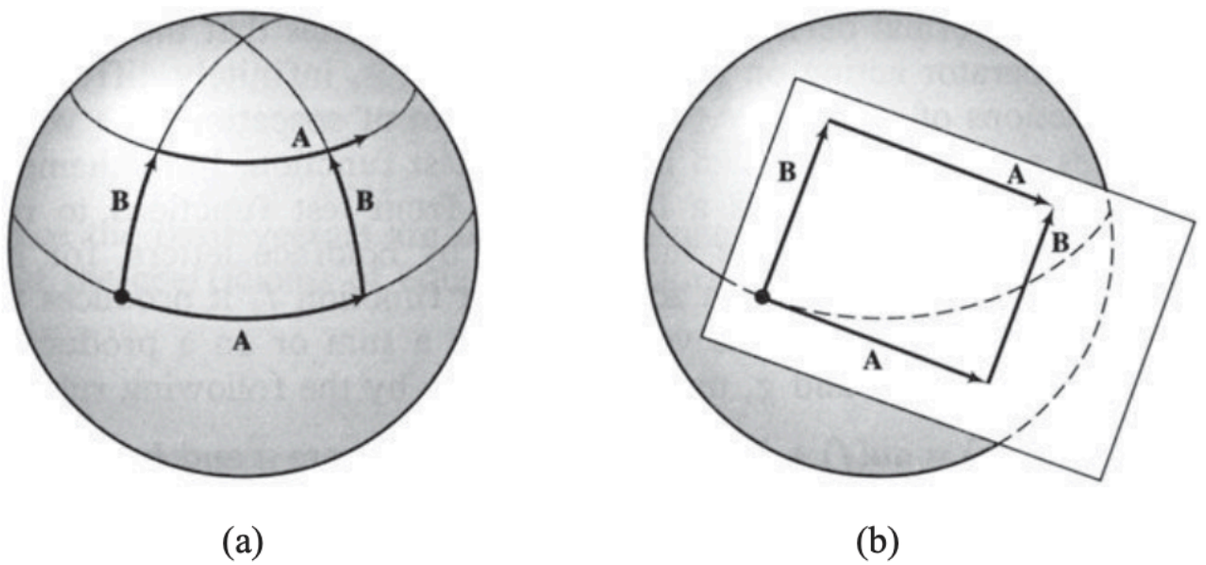


Fig. 2.6 (a) Addition of eastward (**A**) and northward (**B**) displacements on the surface of the Earth does not commute. (b) The displacements commute if the addition is performed on a flat plane tangent to the sphere.

There are two ways out of this difficulty.

We can postulate that the vector addition is to be performed on a flat plane tangent to the sphere at the starting point (see Fig. 2.6b).

This is, of course, how we normally deal with the addition of vectors on the surface of the Earth; but it has the serious drawback that such an addition yields a vector that lies outside of the curved 2-D surface of the sphere; that is, it yields a vector in (flat) 3-D space, rather than a vector in the (curved) 2-D space.

If we wanted to adopt this procedure for vector addition for a curved 4-D spacetime, we would first have to embed the curved spacetime in a flat space, a procedure that is awkward and requires the introduction of unphysical, unobservable extra dimensions of space and time.

Alternatively, we can postulate that vector addition merely means the addition of components.

For instance, if we adopt as coordinates on the surface of the sphere the usual polar angles $x^1 = \theta$ and $x^2 = \phi$, then a vector in this curved space has components (A^1, A^2), and the addition of this vector and a second vector (B^1, B^2) simply gives the vector sum $(A^1 + B^1, A^2 + B^2)$, which is obviously commutative.

However, such a vector in the curved space is nothing but a pair of numbers, and we cannot associate a mental picture with it.

Although it is mathematically consistent to think of a vector as nothing more than a set of components, this view seems to lack something.

We expect that if there are components, there should be some mathematical entity such that the components are the "pieces" of this mathematical entity.

Of course, we could say that the *n*-tuple is the entity of which the components are pieces, but this is a rather shallow and unsatisfying point of view, and we expect that it should be possible to construct some deeper, more interesting mathematical entity.

Such a construction is provided by the calculus of differential forms invented by E. Cartan.

In this intricate construction, vectors are viewed as linear operators acting on "test" functions, and tensors are viewed as operators acting on vectors.

Thus, in Cartan's calculus, everything is operators!

Vectors. Cartan's notion of vector hinges on the observation that, in flat space, such as 3-D space, displacement vectors are in one-to-one correspondence with directional differential operators.

Thus, with the displacement vector $\mathbf{A}$ we can associate the differential operator $\mathbf{A} \cdot \nabla$, or $A^k \partial/\partial x^k$, which is proportional to the component of the gradient operator along the direction of $\mathbf{A}$.

The algebra of such differential operators mimics the algebra of vectors – the differential operators can be added and multiplied by numbers in the same way as displacement vectors.

This means that the differential operators are "vectors," according to the general mathematical definition of an abstract vector space.

But the differential operators have a crucial advantage over displacement vectors in that they can be generalized to curved spacetime.

Vectors constructed from differential operators are called *tangent vectors*.

The formal definition of such a tangent vector says that it is a linear operator acting on test functions, that is, infinitely differentiable functions of position.

At any given point of spacetime, the vector produces a real number when it acts on a test function.

In mathematical language, the tangent vector is a linear map from test functions to real numbers.

We will designate such vectors by boldface letters; for instance, $\mathbf{u}$.

When the vector $\mathbf{u}$ acts on a test function f, it produces the real number $\mathbf{u}(f)$.

When the vector acts on a sum or on a product of two test functions f and g, the result is given by the following two rules:

$$(i)\ \ \mathbf{u}(af + bg) = a\mathbf{u}(f) + b\mathbf{u}(g), \text{ for any real numbers } a \text{ and } b.$$
$$(ii)\ \ \mathbf{u}(fg) = g\mathbf{u}(\mathrm{f}) + f\mathbf{u}(g).$$

The first of these rules merely expresses linearity, and the second rule is the familiar rule for the effect of a first-order differential operator on a product of functions.

Thus, the second rule characterizes the vector as a differential operator.

In component language, these rules can be shown to imply that **u** must be a superposition of the differential operators $\partial/\partial x^0$, $\partial/\partial x^1$, $\partial/\partial x^2$, $\partial/\partial x^3$:

$$\mathbf{u} = u^{\mu} \frac{\partial}{\partial x^{\mu}} \tag{2.144}$$

In this equation, we can regard the operators $\partial/\partial x^{\mu}$ as basis operators, or basis vectors.

Equation (2.144) is then analogous to Eq. (2.143), with $\partial/\partial x^{\mu}$ playing the role of the unit vectors $\mathbf{e}_k$.

The coefficients u^{μ} are the components of the vector **u** with respect to the basis vectors $\partial/\partial x^0$, $\partial/\partial x^1$, $\partial/\partial x^2$, $\partial/\partial x^3$ (note that in this context, $\partial/\partial x^0$, $\partial/\partial x^1$, $\partial/\partial x^2$, $\partial/\partial x^3$ are not thought of as components of the gradient operator $\partial/\partial x^{\mu}$, but as four separate entities).

The vector **u** is called "tangent" vector because the algebra of such vectors mimics the algebra we would obtain if we were to perform vector addition with ordinary displacement vectors on a flat plane tangent to the (possibly) curved spacetime at the given point.

The set of all possible vectors at the given point is called the *tangent space*.
The superscript notation adopted in Eq. (2.144) suggests that the components u^{μ} should be contravariant components, and we can verify that this is indeed the case.

If we perform a Lorentz transformation, the new components become u'^{μ} and the new basis vectors become $\partial/\partial x'^{\mu}$; however, the vector **u** *remains the same*:

$$\mathbf{u} = u^{\mu}\frac{\partial}{\partial x^{\mu}} = u'^{\nu}\frac{\partial}{\partial x'^{\nu}} \tag{2.145}$$

With Eq. (2.66), this reduces to

$$u^{\mu}\frac{\partial}{\partial x^{\mu}} = u'^{\nu}a_{\nu}{}^{\mu}\frac{\partial}{\partial x^{\mu}}$$

Since the basis vectors are linearly independent, this equation demands that the coefficients of equal basis vectors on both sides be equal, that is,

$$u^{\mu} = a_{\nu}{}^{\mu}u'^{\nu} \tag{2.146}$$

This indicates that the components u^{μ} have the contravariant transformation law, as expected.

As an example of a vector, consider the energy-momentum vector of a particle.

In component language, this vector consists of the components

$$p^{\mu} = (E, p_x, p_y, p_z)$$

In our new language, this vector becomes

$$\mathbf{p} = E\frac{\partial}{\partial x^0} + p_x\frac{\partial}{\partial x^1} + p_y\frac{\partial}{\partial x^2} + p_z\frac{\partial}{\partial x^3} \tag{2.147}$$

Note that in Cartan's formalism, **p** is a 4-D vector, not to be confused with the familiar 3-D vector **p** for which we use the same notation.

Vectors can be given a geometric interpretation in terms of directional derivatives along curves in spacetime.

Consider a curve specified by a parametric equation $x^{\mu}(\lambda)$, where λ is a parameter that varies smoothly along the curve.

Then the directional derivative of a test function f along the curve, or the rate of change of f along the curve, is

$$\mathbf{u}(f) = \frac{df}{d\lambda} = \frac{dx^\mu}{d\lambda}\frac{\partial f}{\partial x^\mu} \tag{2.148}$$

Thus, the components of the vector $\mathbf{u} = d/d\lambda$ are $dx^\mu/d\lambda$; these components agree with what, in component language, we would call the components of a vector tangent to the curve at the given point.

By choosing different curves, all passing through the same point but with different directions, we obtain the set of all possible vectors that can be defined at that point.

Any vector at this point can then be regarded as a directional derivative for one of these curves.

Note that the basis vectors $\partial/\partial x^0$, $\partial/\partial x^1$, $\partial/\partial x^2$, $\partial/\partial x^3$ are among the possible vectors we obtain for suitable choices of curves.

For instance, $\partial/\partial x^0$ is the directional derivative we obtain for a curve whose direction coincides with the x^0-axis and whose parameter is simply the time coordinate x^0.

The advantage of describing vectors in terms of directional derivatives along curves is that the curve and its parameter uniquely specify the vector (at a given point), without any need to invoke coordinates explicitly.

This is the rationale for the claim, often made by aficionados, that the Cartan definition of a vector is a "geometric" definition, independent of coordinates.

However, this claim is somewhat of an exaggeration.

Spacetime is featureless and amorphous; points and curves in spacetime cannot be identified and defined except by their coordinates.

We might be tempted to identify a point by describing an event that occurred at this point; for example, “this is the point where the small red Chinese firecracker exploded.”

But such an anecdotal identification is meaningful only to eyewitnesses of the event; the anecdotal identification lacks the universal communicability that is supposed to be a hallmark of scientific data.

If the eyewitnesses try to communicate the position of the event to an investigator or to make a written record for posterity, they find themselves reduced to describing the position by explicit or implicit coordinates: “The firecracker exploded five meters to the north of the fire hydrant on Mercer Street.”

In essence, coordinates give us an operational procedure for the identification of points in spacetime, which an anecdotal identification fails to do.

Furthermore, the anecdotal identification works only for those points of spacetime where there is an actual event (an actual happening).

Most of spacetime is empty, devoid of actual events.

The set of spacetime points is a continuous, uncountable set, whereas the set of actual events is a discrete, countable set; it is therefore not possible to regard these sets as identical.

Coordinates lie not only at the root of the definition of a manifold (see Section 2.1) and at the root of any identification of points or curves in spacetime but they are also essential whenever we want to compare the results of a calculation with experimental measurements.

Experimenters cannot measure a Cartan vector; they can only measure components.

Thus, the Cartan formalism is no more than a clever mathematical tool for the efficient and elegant derivation of some vector and tensor theorems.

It has no advantage in principle over the component formalism.

1-Forms. In the mathematics of vector spaces it is customary to define a dual vector as a linear map from vectors to real numbers, that is, the dual vector produces a number whenever it acts on a vector.

In the Cartan calculus, dual vectors are called *differential forms or 1-forms*.

Since a vector is a differential operator, the 1-form associates numbers with differential operators.

We will designate 1-forms by boldface Greek letters, such as $\boldsymbol{\alpha}$.

The 1-form $\boldsymbol{\alpha}$ produces a number $\boldsymbol{\alpha}(\mathbf{u})$ when it acts on the vector $\mathbf{u}$.

When acting on a superposition of two vectors, the 1-form obeys the rule

$$\boldsymbol{\alpha}(a\mathbf{u} + b\mathbf{w}) = a\boldsymbol{\alpha}(\mathbf{u}) + b\boldsymbol{\alpha}(\mathbf{w}) \tag{2.149}$$

As a basis for the space of dual vectors, we use the 1-forms $\boldsymbol{\omega}^0, \boldsymbol{\omega}^1, \boldsymbol{\omega}^2, \boldsymbol{\omega}^3$ defined by their action on the basis vectors $\partial/\partial x^0$, $\partial/\partial x^1$, $\partial/\partial x^2$, $\partial/\partial x^3$, as follows:

$$\boldsymbol{\omega}^\nu \left(\frac{\partial}{\partial x^\mu} \right) = \delta^\nu_\mu \tag{2.150}$$

For instance, $\boldsymbol{\omega}^0$ yields +1 when acting on $\partial/\partial x^0$, and zero when acting on any of the other basis vectors.

Any 1-form can then be expressed as a superposition of the basis 1-forms $\boldsymbol{\omega}^\mu$:

$$\boldsymbol{\alpha} = \alpha_\mu \boldsymbol{\omega}^\mu \tag{2.151}$$

Here, the coefficients α_μ are the components of the 1-form with respect to the basis $\boldsymbol{\omega}^\mu$.

When a 1-form with components α_μ acts on a vector with components u^μ, the result is

$$\boldsymbol{\alpha}(\mathbf{u}) = \alpha_\mu \boldsymbol{\omega}^\mu \left(u^\nu \frac{\partial}{\partial x^\nu} \right) = \alpha_\mu u^\nu \boldsymbol{\omega}^\mu \left(\frac{\partial}{\partial x^\nu} \right) = \alpha_\mu u^\nu \delta^\mu_\nu = \alpha_\mu u^\mu$$

In component language, we would call this the scalar product of the contravariant vector u^μ and the covariant vector α_μ (we will see later that the components α_μ are, indeed, covariant components).

For example, the energy-momentum can be expressed as a 1- form,

$$\boldsymbol{\pi} = E\boldsymbol{\omega}^0 + p_1\boldsymbol{\omega}^1 + p_2\boldsymbol{\omega}^2 + p_3\boldsymbol{\omega}^3 = E\boldsymbol{\omega}^0 - p_x\boldsymbol{\omega}^1 - p_y\boldsymbol{\omega}^2 - p_z\boldsymbol{\omega}^3$$

and if we let this 1-form energy-momentum act on the vector energy-momentum [see Eq. (2.147)], we obtain

$$\begin{aligned}\boldsymbol{\pi}(\mathbf{p}) &= (E\boldsymbol{\omega}^0 - p_x\boldsymbol{\omega}^1 - p_y\boldsymbol{\omega}^2 - p_z\boldsymbol{\omega}^3)\left(E\frac{\partial}{\partial x^0} + p_x\frac{\partial}{\partial x^1} + p_y\frac{\partial}{\partial x^2} + p_z\frac{\partial}{\partial x^3}\right)\\ &= (E^2 - p_x^2 - p_y^2 - p_z^2) = m^2\end{aligned}$$

Gradient. The gradient of a function is a special case of a 1-form.

The gradient **d**f of a function *f* is defined by what it does when it acts on some arbitrary tangent vector d/dλ:

$$\mathbf{d}f\left(\frac{d}{d\lambda}\right) = \frac{df}{d\lambda} \tag{2.152}$$

Thus, the number that the gradient produces when acting on the tangent vector d/dλ associated with a given curve is simply the derivate of the function along the curve.

As a simple example of this formula for the gradient, consider the function $f = x^0$, which depends linearly on the coordinate x^0, and suppose that the vector d/dλ is one of the basis vectors, $d/d\lambda = \partial/\partial x^\nu$.

Then Eq. (2.152) becomes

$$\mathbf{d}x^0\left(\frac{\partial}{\partial x^\nu}\right) = \frac{\partial x^0}{\partial x^\nu} = \delta^0_\nu \tag{2.153}$$

More generally, if $f = x^\mu$, so *f* depends linearly on the coordinate x^μ (and only on this one coordinate), we obtain

$$\mathbf{d}x^{\mu}\left(\frac{\partial}{\partial x^{\nu}}\right)=\frac{\partial x^{\mu}}{\partial x^{\nu}}=\delta^{\mu}_{\nu} \tag{2.154}$$

Comparing this with Eq. (2.150), we see that the gradients $\mathbf{d}x^0, \mathbf{d}x^1, \mathbf{d}x^2, \mathbf{d}x^3$ of the coordinates coincide with the basis 1-forms $\boldsymbol{\omega}^0, \boldsymbol{\omega}^1, \boldsymbol{\omega}^2, \boldsymbol{\omega}^3$, respectively:

$$\mathbf{d}x^{\mu}=\boldsymbol{\omega}^{\mu} \tag{2.155}$$

Thus, the component expression (2.151) for a 1-form can be written as

$$\boldsymbol{\alpha}=\alpha_{\mu}\mathbf{d}x^{\mu} \tag{2.156}$$

In particular, the gradient of a function *f* can be written as

$$\mathbf{d}f=\frac{\partial f}{\partial x^{\mu}}\mathbf{d}x^{\mu} \tag{2.157}$$

The components of this gradient are, of course, what we would normally expect from our experience with component language.

Keep in mind that although Eq. (2.157) looks very much like the familiar equation for a small change of a function engendered by a small displacement, it has nothing to do with that; $\mathbf{d}f$ and $\mathbf{d}x^{\mu}$ are not small changes, but 1-forms.

If we wanted to find the small change in the function *f* engendered by a small displacement, we would have to write

$$df=dx^{\mu}\frac{\partial f}{\partial x^{\mu}} \tag{2.158}$$

To interpret this equation in the language of differential forms, we observe that $dx^{\mu}\,\partial/\partial x^{\mu}$ is a vector $\boldsymbol{\delta}$ that acts on the function *f*; the small changes dx^{μ} are the components of this vector.

Equation (2.158) for the small change in the function f then becomes

$$df = \boldsymbol{\delta} f$$

Never confuse this df with **d**f !

In the preceding equations we adopted a subscript notation for the components of a 1-form.

This notation suggests that these components should obey the transformation law for covariant components,

$$\alpha_\nu = a^\mu{}_\nu \alpha'_\mu$$

By an argument similar to that given in connection with the components of a vector [see Eqs. (2.145) and (2.146)], it is easy to show that this is indeed the case.

Tensors. Tensors are defined as linear maps from 1-forms and vectors to real numbers, that is, tensors are operators that yield real numbers when acting on 1-forms and vectors.

For instance, a tensor **T** of third rank, contravariant in the first index and covariant in the next two indices [called a tensor of type $\binom{1}{2}$], produces a real number $\mathbf{T}(\alpha, \mathbf{u}, \mathbf{v})$ when it acts on one 1-form and two vectors.

The map is linear in the 1-form and in each of the vectors individually.

One way to construct a tensor of this type is by means of the "tensor" product of one vector and two 1-forms, such as

$$\mathbf{s} \otimes \boldsymbol{\zeta} \otimes \boldsymbol{\xi} \tag{2.159}$$

When this tensor acts on one 1-form α and on two vectors **u** and **w** it gives, by definition,

$$\mathbf{s} \otimes \boldsymbol{\zeta} \otimes \boldsymbol{\xi}(\boldsymbol{\alpha}, \mathbf{u}, \mathbf{w}) \equiv \boldsymbol{\alpha}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{u})\boldsymbol{\xi}(\mathbf{w}) \tag{2.160}$$

Here, the right side is simply a product of three real numbers. Note that the tensor-product symbol ⊗ merely indicates that the factors in Eq. (2.160) are to be treated independently when they act on 1-forms or vectors placed on the right, and only in the last step, after all the 1-forms and vectors have been acted upon and have been converted into numbers, is the product to be taken.

The most general tensor of the type $\binom{1}{2}$ is a superposition of tensor products of the kind displayed in Eq. (2.159), with different choices of $\mathbf{s}$, $\boldsymbol{\zeta}$, and $\boldsymbol{\xi}$.

If we construct tensor products with the basis vectors and basis 1-forms, we obtain basis tensors, such as

$$\frac{\partial}{\partial x^{\sigma}} \otimes \mathbf{d}x^{\mu} \otimes \mathbf{d}x^{\nu} \tag{2.161}$$

Note that when such a tensor acts on a vector, we must specify which one of the two basis 1-forms acts. In terms of these basis tensors, an arbitrary third-rank tensor of the type $\binom{1}{2}$ can be expressed as

$$\mathbf{T} = T^{\sigma}{}_{\mu\nu} \frac{\partial}{\partial x^{\sigma}} \otimes \mathbf{d}x^{\mu} \otimes \mathbf{d}x^{\nu} \tag{2.162}$$

The coefficients $T^{\sigma}{}_{\mu\nu}$ are the components of this tensor with respect to the basis tensors.

Vectors and 1-forms are themselves tensors of the first rank.

Thus, a 1-form is a first-rank tensor of the type $\binom{0}{1}$, which yields a real number when it acts on a vector.

And a vector is a first-rank tensor of the type $\binom{1}{0}$, which yields a real number when it acts on a 1-form.

The *contraction* of a tensor is accomplished by allowing it to act on $\mathbf{d}x^{\tau}$ and $\partial/\partial x^{\tau}$, with summation over τ.

For instance, the contraction of the third-rank tensor **T** in Eq. (2.162) results in

$$\overline{\mathbf{T}} = \mathbf{T}\left(\mathbf{d}x^{\tau}, \frac{\partial}{\partial x^{\tau}}, \right) \tag{2.163}$$

where the remaining empty slot in the tensor $\mathbf{T}$ indicates that it has now become a first-rank tensor, of type $\binom{1}{0}$, that acts on a tangent vector to produce a number.

In the language of differential forms, the metric tensor is a second-rank tensor of the type $\binom{0}{2}$, which can be written as

$$\boldsymbol{\eta} = \eta_{\mu\nu}\mathbf{d}x^{\mu} \otimes \mathbf{d}x^{\nu}. \tag{2.164}$$

The inverse of the metric tensor is a second-rank tensor of the type $\binom{2}{0}$,

$$\boldsymbol{\eta}^{-1} = \eta^{\mu\nu}\frac{\partial}{\partial x^{\mu}} \otimes \frac{\partial}{\partial x^{\nu}} \tag{2.165}$$

It is easy to verify that when η acts on η^{-1}, the result is the identity operator:

$$\boldsymbol{\eta}\boldsymbol{\eta}^{-1} = \mathbf{1} \tag{2.166}$$

When the metric tensor (2.164) acts on two vectors, it produces a real number:

$$\begin{aligned}\boldsymbol{\eta}(\mathbf{u}, \mathbf{w}) &= \eta_{\mu\nu}\mathbf{d}x^{\mu} \otimes \mathbf{d}x^{\nu}(\mathbf{u}, \mathbf{w}) = \eta_{\mu\nu}\mathbf{d}x^{\mu}(\mathbf{u})\mathbf{d}x^{\nu}(\mathbf{w}) \\ &= \eta_{\mu\nu}\mathbf{d}x^{\mu}\left(u^{\sigma}\frac{\partial}{\partial x^{\sigma}}\right)\mathbf{d}x^{\nu}\left(w^{\tau}\frac{\partial}{\partial x^{\tau}}\right) \\ &= \eta_{\mu\nu}u^{\mu}w^{\nu}\end{aligned} \tag{2.167}$$

When the metric tensor acts on a single vector, it produces a 1-form:

$$\begin{aligned}\boldsymbol{\eta}(\mathbf{u}) &= \eta_{\mu\nu}\mathbf{d}x^{\mu} \otimes \mathbf{d}x^{\nu}(\mathbf{u}) = \eta_{\mu\nu}\mathbf{d}x^{\mu}(\mathbf{u})\mathbf{d}x^{\nu} \\ &= \eta_{\mu\nu}\mathbf{d}x^{\mu}\left(u^{\sigma}\frac{\partial}{\partial x^{\sigma}}\right)\mathbf{d}x^{\nu} \\ &= \eta_{\mu\nu}u^{\nu}\mathbf{d}x^{\mu} = u_{\mu}\mathbf{d}x^{\mu}\end{aligned} \tag{2.168}$$

This action of the metric tensor on a single vector replaces the original vector and its contravariant components u^σ by a 1-form and its covariant components u_μ.

Thus, the action of the metric tensor on a vector is effectively equivalent to the operation of lowering an index we encountered in component language.

In fact, all the index manipulations and all the equations in component language can be translated into the language of differential forms.

Sometimes the result of such a translation is rather cumbersome; for instance, Eq. (2.168) is more messy than the simple and direct equation $u_\nu = \eta_{\nu\sigma}\, u^\sigma$ of component language.

But sometimes the translation into the language of differential forms yields elegant and concise results, especially when dealing with equations containing derivatives, as we will see later.

2-Forms. The exterior product, or the wedge product, of 1-forms is an antisymmetrized tensor product.

The exterior product of two 1-forms is called a *2-form*.

This exterior product is defined by

$$\boldsymbol{\alpha} \wedge \boldsymbol{\beta} = \boldsymbol{\alpha} \otimes \boldsymbol{\beta} - \boldsymbol{\alpha} \otimes \boldsymbol{\beta} \tag{2.169}$$

Note that the exterior product is antisymmetric under the exchange of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, that is, $\boldsymbol{\alpha} \wedge \boldsymbol{\beta} = -\boldsymbol{\beta} \wedge \boldsymbol{\alpha}$.

As a consequence of this asymmetry, the exterior product of any 1-form with itself vanishes identically, $\boldsymbol{\alpha} \wedge \boldsymbol{\alpha} = 0$.

In terms of components, the 2-form $\boldsymbol{\alpha} \wedge \boldsymbol{\beta}$ becomes

$$\boldsymbol{\alpha} \wedge \boldsymbol{\beta} = \alpha_\mu \mathbf{d}x^\mu \wedge \beta_\nu \mathbf{d}x^\nu \tag{2.170}$$

We can rewrite the right side of this equation as

$$\begin{aligned}\boldsymbol{\alpha} \wedge \boldsymbol{\beta} &= \frac{1}{2}(\alpha_\mu \beta_\nu \mathbf{d}x^\mu \wedge \mathbf{d}x^\nu - \alpha_\mu \beta_\nu \mathbf{d}x^\nu \wedge \mathbf{d}x^\mu) \\ &= \frac{1}{2}(\alpha_\mu \beta_\nu - \alpha_\nu \beta_\mu)\mathbf{d}x^\mu \wedge \mathbf{d}x^\nu\end{aligned} \tag{2.171}$$

where the second equation merely involves an exchange of the dummy indices μ, ν.

We can regard the coefficients $^1/_2\,(\alpha_\mu\beta_\nu - \alpha_\nu\beta_\mu)$ as the components of the 2-form; Eq. (2.171) shows that the components are antisymmetric in μ and ν.

More generally, given a set of coefficients $\alpha_{\mu\nu}$ antisymmetric in μ and ν, we can define the 2-form $^1/_2\,\alpha_{\mu\nu}\,\mathbf{d}x^\mu \wedge \mathbf{d}x^\nu$.

An example of such a 2-form is the electromagnetic field tensor **F** defined by

$$\mathbf{F} = \frac{1}{2}F_{\mu\nu}\mathbf{d}x^\mu \wedge \mathbf{d}x^\nu \tag{2.172}$$

where the components $F_{\mu\nu}$ have the values

$$F_{\mu\nu} = \begin{pmatrix} 0 & E_x & E_y & E_z \\ -E_x & 0 & -B_z & B_y \\ -E_y & B_z & 0 & -B_x \\ -E_z & -B_y & B_x & 0 \end{pmatrix} \tag{2.173}$$

Note that the covariant tensor $F_{\mu\nu}$ differs from the contravariant tensor $F^{\mu\nu}$ of Eq. (2.114) by sign reversal of the first row and first column.

If we write Eq. (2.172) out in full, it becomes

$$\begin{aligned}\mathbf{F} = {} & E_x \mathbf{d}x^0 \wedge \mathbf{d}x^1 + E_y \mathbf{d}x^0 \wedge \mathbf{d}x^2 + E_z \mathbf{d}x^0 \wedge \mathbf{d}x^3 \\ & - B_z \mathbf{d}x^1 \wedge \mathbf{d}x^2 + B_y \mathbf{d}x^1 \wedge \mathbf{d}x^3 - B_x \mathbf{d}x^2 \wedge \mathbf{d}x^3\end{aligned} \tag{2.174}$$

Here, terms with 1-forms in opposite order have been combined.

With this electromagnetic field tensor we can write the equation of motion of an electric charge as follows:

$$\frac{d\mathbf{p}}{d\tau} = \frac{q}{m}\mathbf{F}(\mathbf{p}) \tag{2.175}$$

where $\mathbf{p}$ is the momentum vector,

$$\mathbf{p} = p^\mu \frac{\partial}{\partial x^\mu} \tag{2.176}$$

Here it is assumed that, as in Eq. (2.168), the first 1-form in each term in $\mathbf{F}$ acts on $\mathbf{p}$, while the second 1-form remains inactive.

Equation (2.175) looks very simple, and this is one of the advantages of the language of differential forms.

But, unfortunately, we cannot do anything much with this simple equation of motion.

Whenever we want to solve the equation of motion for a particle in a given electromagnetic field with some given initial conditions, we have to disassemble this equation into components, and compute the (numerical) values of these components one by one.

p-Forms. The exterior product can also be defined for more than two 1-forms.

The exterior product of p 1-forms is called a p-form.

This is an antisymmetrized tensor product, which is antisymmetric in each adjacent pair of factors.

For instance, with three 1-forms we can define the 3-form

$$\begin{aligned}\boldsymbol{\alpha}\wedge\boldsymbol{\beta}\wedge\boldsymbol{\zeta} = {} & \boldsymbol{\alpha}\otimes\boldsymbol{\beta}\otimes\boldsymbol{\zeta}+\boldsymbol{\beta}\otimes\boldsymbol{\zeta}\otimes\boldsymbol{\alpha}+\boldsymbol{\zeta}\otimes\boldsymbol{\alpha}\otimes\boldsymbol{\beta} \\ & -\boldsymbol{\beta}\otimes\boldsymbol{\alpha}\otimes\boldsymbol{\zeta}-\boldsymbol{\alpha}\otimes\boldsymbol{\zeta}\otimes\boldsymbol{\beta}-\boldsymbol{\zeta}\otimes\boldsymbol{\beta}\otimes\boldsymbol{\alpha}\end{aligned} \tag{2.177}$$

If we express the 1-forms in components, we can rewrite this as

$$\begin{aligned}\boldsymbol{\alpha}\wedge\boldsymbol{\beta}\wedge\boldsymbol{\zeta} = {} & \frac{1}{3!}(\alpha_\mu\beta_\nu\zeta_\sigma+\alpha_\nu\beta_\sigma\zeta_\mu+\alpha_\sigma\beta_\mu\zeta_\nu \\ & -\alpha_\nu\beta_\mu\zeta_\sigma-\alpha_\mu\beta_\sigma\zeta_\nu-\alpha_\sigma\beta_\nu\zeta_\mu)\mathbf{d}x^\mu\wedge\mathbf{d}x^\nu\wedge\mathbf{d}x^\sigma\end{aligned} \tag{2.178}$$

With some arbitrary set of coefficients $\alpha_{\mu\nu\sigma}$, antisymmetric in pairs of adjacent indices, we can construct a general 3-form,

$$\frac{1}{3!}\alpha_{\mu\nu\sigma}\mathbf{d}x^\mu\wedge\mathbf{d}x^\nu\wedge\mathbf{d}x^\sigma \tag{2.179}$$

Note that in 4-D spacetime, any 5-form, or any 6-form, or any higher p-form necessarily vanishes,

$$\boldsymbol{\alpha}\wedge\boldsymbol{\beta}\wedge\boldsymbol{\zeta}\wedge\boldsymbol{\xi}\wedge\boldsymbol{\chi}=0, \text{ etc.} \tag{2.180}$$

This is so because, when written with the basis 1-forms $\mathbf{d}x^\mu$, the left side of Eq. (2.180) contains five factors $\mathbf{d}x^\mu$, which means that at least one of the basis 1-forms occurs twice – but the exterior product of equal 1-forms is zero.

Also, note that, to within an overall numerical factor, there is only one 4-form,

$$\boldsymbol{\varepsilon}=\mathbf{d}x^0\wedge\mathbf{d}x^1\wedge\mathbf{d}x^2\wedge\mathbf{d}x^3 \tag{2.181}$$

This 4-form can also be written as

$$\boldsymbol{\varepsilon}=\frac{1}{4!}\varepsilon_{\mu\nu\sigma\tau}\mathbf{d}x^\mu\wedge\mathbf{d}x^\nu\wedge\mathbf{d}x^\sigma\wedge\mathbf{d}x^\tau \tag{2.182}$$

where $\varepsilon_{\mu\nu\sigma\tau}$ is the "alternating" symbol, antisymmetric under the exchange of any two adjacent indices, whose components are zero unless all the indices are different, in which case $\varepsilon_{0123}=\varepsilon_{1032}=\varepsilon_{1320}=\ldots=1$ and $\varepsilon_{1023}=\varepsilon_{0132}=\varepsilon_{0321}=\ldots=-1$.

Exterior Derivative. The exterior derivative of a p-form is a $(p + 1)$-form.

If the p-form is

$$\boldsymbol{\alpha} = \frac{1}{p!}\alpha_{\mu\nu\ldots}\mathbf{d}x^{\mu} \wedge \mathbf{d}x^{\nu} \wedge \ldots \tag{2.183}$$

then the exterior derivative is defined by

$$\mathbf{d}\boldsymbol{\alpha} = \frac{1}{p!}\mathbf{d}\alpha_{\mu\nu\ldots} \wedge \mathbf{d}x^{\mu} \wedge \mathbf{d}x^{\nu} \wedge \ldots \tag{2.184}$$

Note that the extra 1-form $\mathbf{d}\alpha_{\mu\nu\ldots}$ that makes its appearance on the right side of Eq. (2.184) is simply the gradient of the function $\alpha_{\mu\nu\ldots}$.

With the usual formula (2.157) for the gradient, Eq. (2.184) becomes

$$\mathbf{d}\boldsymbol{\alpha} = \frac{1}{p!}\left(\frac{\partial}{\partial x^{\sigma}}\alpha_{\mu\nu\ldots}\right)\mathbf{d}x^{\sigma} \wedge \mathbf{d}x^{\mu} \wedge \mathbf{d}x^{\nu} \wedge \ldots \tag{2.185}$$

The exterior derivative of an exterior derivative is always zero.

This is an immediate consequence of Eq. (2.157):

$$\mathbf{dd}\boldsymbol{\alpha} = \frac{1}{p!}\left(\frac{\partial}{\partial x^{\tau}}\frac{\partial}{\partial x^{\sigma}}\alpha_{\mu\nu\ldots}\right)\mathbf{d}x^{\tau} \wedge \mathbf{d}x^{\sigma} \wedge \mathbf{d}x^{\mu} \wedge \mathbf{d}x^{\nu} \wedge \ldots \tag{2.186}$$

Here, the second derivative with respect to x^{τ} and x^{σ} is symmetric in τ and σ, but the exterior product $\mathbf{d}x^{\tau} \wedge \mathbf{d}x^{\sigma}$ multiplying it is antisymmetric; hence the summation over τ and σ necessarily yields zero.

The identity $\mathbf{dd}\,\alpha = 0$ can be regarded as a generalization of the familiar 3-D identity $\nabla \cdot (\nabla \times \mathbf{A}) = 0$.

Note that the rule that two exterior derivatives in succession give zero also applies to the gradient of a function, for which we readily obtain the identity $\mathbf{dd}f = 0$.

This identity is a generalization of the familiar 3-D identity $\nabla \times (\nabla f) = 0$.

Exterior derivatives permit us to express Maxwell's equations in an elegant way.

It is easy to check that the Maxwell equation (2.112) simply says that the exterior derivative of the electromagnetic field tensor $\mathbf{F}$ is zero:

$$\mathbf{dF} = 0 \tag{2.187}$$

The other Maxwell equation, Eq. (2.111), is somewhat more difficult to translate into the language of differential forms.

A clever way to accomplish the translation is to start with the contravariant components of the electromagnetic field tensor and of the current density and then define a 2-form $^*\mathbf{F}$ and a 3-form $^*\mathbf{j}$ as follows:

$$^*\mathbf{F} \equiv \frac{1}{4}\varepsilon_{\mu\nu\sigma\tau} F^{\sigma\tau} \mathbf{d}x^\mu \wedge \mathbf{d}x^\nu \tag{2.188}$$

$$^*\mathbf{j} \equiv \frac{1}{3!}\varepsilon_{\mu\nu\sigma\tau} j^\tau \mathbf{d}x^\mu \wedge \mathbf{d}x^\nu \wedge \mathbf{d}x^\sigma \tag{2.189}$$

These are called, respectively, the dual electromagnetic field tensor and the dual current density (in this context, the word *dual* has a new, different meaning from that in dual vector space).

The Maxwell equation (2.111) then says that the exterior derivative of $^*\mathbf{F}$ is proportional to $^*\mathbf{j}$:

$$\mathbf{d}^*\mathbf{F} = \frac{4\pi}{c}{}^*\mathbf{j} \tag{2.190}$$

Note that if we take the exterior derivative of Eq. (2.190), we find that the left side is zero, according to the usual rule for the exterior derivative of an exterior derivative.

Hence

$$0 = \mathbf{d}^*\mathbf{j} \tag{2.191}$$

This is the conservation law for electric charge, expressed in the language of differential forms.

Finally, in the language of differential forms, the vector potential is a 1-form,

$$\mathbf{A} = A_\mu \mathbf{d}x^\mu \tag{2.192}$$

and the connection between the vector potential and the electromagnetic field tensor is simply

$$\mathbf{F} = \mathbf{dA} \tag{2.193}$$

Equation (2.187) is then an identity,

$$\mathbf{dF} = \mathbf{ddA} = 0 \tag{2.194}$$

Although all the equations of electrodynamics look deceptively simple when expressed in the language of differential forms, we must keep in mind that the simple notation hides a great many complications. All these complications reemerge when we attempt any numerical calculation, where we must necessarily use components.

These examples of the application of differential forms to some equations of physics give us a glimpse of the elegance of the language of differential forms.

But to grasp the full power of this language, we have to look at how it deals with the integration of tensor quantities over regions of spacetime.

With a suitable definition of the integral of a differential form, it can be shown that Gauss's theorem, Stokes' theorem, and other integral theorems of this kind are corollaries of a single general theorem that relates the integral of a differential form over a region of spacetime to the integral over the boundary of this region.

These results can be found in treatises on differential forms.

3 - The Linear Approximation

Un poco di vero fa creder tutta la bugia.
[A little truth makes the whole lie believable.]

Traditional Italian proverb

To discover the relativistic field equations for gravitation we begin with a linear approximation for the gravitational field, that is, we neglect the effects of the gravitational field on itself.Of course, if Newton's principle of equivalence ($m_I = m_G$) is to hold as an *exact* statement, gravitational energy must gravitate, and the exact field equations must be nonlinear.

Although it is true that some of the most spectacular results of gravitational theory depend in a crucial way on the nonlinearity of the field equations, almost all of the results that have been the subject of experimental investigation can be described by the linear approximation.

For example, the deflection of light, the time delay of light, gravitational time dilation, gravitational lensing, and gravitational radiation emerge from the linear approximation.

Furthermore, this approximation applies to all phenomena that lie in the region of overlap between Newton's and Einstein's theories.

Most discussions of gravitational theory begin by postulating that spacetime is curved and from there proceed to formulate the nonlinear equations that govern the curved spacetime geometry.

The linear approximation then arises in the end from the full nonlinear equations.

The great disadvantage of this approach is that it never makes clear just why anybody would entertain the preposterous notion that our beautiful flat spacetime should be curved, bent, and deformed.

The reason why we believe in curved spacetime can best be understood by beginning with flat spacetime and seeking the field equations that describe the gravitational field (in linear approximation) in this flat spacetime.

As we will see, the behavior of particles in the gravitational field and the behavior of "clocks" and "meter sticks" are then such that measurements of spacetime indicate that the real geometry is curved.

Thus, the flat spacetime with which we begin turns out to be an unobservable, fictitious geometry.

But, nevertheless, flat spacetime is a useful starting point in the case of weak fields.

3.1 The example of electromagnetism

Since electromagnetism is the linear theory par excellence, it will be instructive to play a game.

Let us pretend that we know that there is a long-range interaction between charges, but that we do not know the field equations (Maxwell's equations) that govern this interaction.

How could we go about finding the equations?

We will, of course, demand that the equations that describe the interaction between charges obey the principle of relativity; accordingly, we demand that the equations be tensor equations.

This immediately tells us that an interaction that involves only the charge density is impossible.

The charge density ϱ and the current density j^{μ} taken together form a four-vector [see Eq. (2.119)], and any tensor equation that involves one component of this four-vector must also involve the other components.

For example, suppose we have a reference frame in which there is a charge density, but no current.

In a new, moving reference frame there will then be both a charge density $j'^0 = a^0{}_0 j^0$ and a current density $j'^k = a^k{}_0 j^k$, so the charge density in the first reference frame becomes a combination of the charge density and current density.

If the charge density in the first reference frame participates in interactions, an observer in the second reference frame will perceive this as an interaction of both charges and currents.

Any law that pretends to be valid in all reference frames must therefore include both the interactions of charge densities and of current densities – if the charge density j^0 generates some kind of force field, so will the current density j^k.

To put it another way, the four-vector j^μ is the source of the electromagnetic field.

The field equations are supposed to establish the link between the field and its source j^μ.

As a first step in our search for the field equations, we must decide what kind of field is to be used to describe the interaction.

Our choices are a scalar, a vector, and a tensor of higher rank.

Because the source of the field is the four-vector j^μ, the most natural choice is a vector field, and this is what we will try.

This choice of a vector field can be made plausible by an analogy with Lagrangian mechanics.

The equation of motion for a generalized coordinate q^k under the influence of a generalized driving force Q^k is typically of the form

$$\frac{d^2 q^k}{dt^2} \propto Q^k \tag{3.1}$$

In our case, the field plays the role of the generalized coordinate, and the current density plays the role of the driving force.

The equation of motion should then have the form

$$\frac{\partial^2 A}{\partial t^2} \propto j^\nu \tag{3.2}$$

Here the field variable has simply been written as A, without any index.

This quantity A could be a scalar, a vector, or a tensor of higher rank.

One obvious defect of our tentative equation of motion is that whereas a Lorentz transformation changes A into A' and j^ν into j'^ν, it does not change $\partial^2/\partial t^2$ into $\partial^2/\partial t'^2$; hence the equation is not a tensor equation and fails to satisfy the principle of special relativity.

To achieve invariance, we recall that $\partial^2/\partial t^2$ is only one part of the operator $\partial_\mu \partial^\mu$.

If we include the complete operator $\partial_\mu \partial^\mu$, then a Lorentz transformation changes it into $\partial_\mu{}' \, \partial'^\mu$, as desired.

We therefore arrive at a tentative equation of motion

$$\partial_\mu \partial^\mu A \propto j^\nu \tag{3.3}$$

Since the right side of this equation is a four-vector, it follows that A must also be a four-vector.

We designate this four-vector by A^ν or by $A^\nu(x)$.

We now look for the most general second-order linear differential equation for $A^\nu(x)$, with $j^\nu(x)$ acting as source.

This means that the equation should have the form of Eq. (3.2): on the left side there should appear a linear second-order differential operator or a combination of such operators acting on the field, and on the right side should appear the current.

For example,

$$\partial_\mu \partial^\mu A^\nu + b\partial^\nu \partial_\mu A^\mu + aA^\nu = 4\pi j^\nu \tag{3.4}$$

where a and b are constants, is an acceptable equation.

In fact, Eq. (3.4) is the only acceptable equation, if we insist that the differential order should be no more than two.

To see this, note that Lorentz invariance demands that the left side be a four-vector.

But the only four-vectors linear in A^ν that can be constructed out of A^ν and no more than two of the differential operators ∂^μ are

$$\partial_\mu \partial^\mu A^\nu, \quad \partial^\nu \partial_\mu A^\mu, \quad A^\nu$$

Of course, we would generalize Eq. (3.4) slightly by inserting an extra arbitrary constant in front of the first term on the left side, but this amounts to no more than a redefinition of A^ν and is therefore not needed.

Likewise, an extra multiplicative constant (besides 4π) that could be inserted on the right side can always be absorbed in the definition of the current.

In fact, the current already contains an adjustable constant q, which represents the charge on a particle [see the definition given by Eqs. (2.119) and (2.120)].

This is a coupling constant; it characterizes the strength of the coupling between matter and field.

The value of this constant q must be determined from experiment, which can be done only after we discover the field equations.

If we do not know Maxwell's equations and the exact form of the law of force between charges, then we cannot give an operational procedure for measuring the constant q that appears in Eqs. (2.119) and (2.120).

But we can write down the expressions (2.119) and (2.120), leaving q provisionally as a unknown parameter.

In contrast, the values of the constants a and b are completely determined by the following arguments.

As concerns the constant a, we can immediately say that it must be positive or zero.

This can best be seen in the special case of vacuum solutions of Eq. (3.4).

If $j^\nu = 0$, we have

$$\partial_\mu \partial^\mu A^\nu + b\partial^\nu \partial_\mu A^\mu + aA^\nu = 0 \tag{3.5}$$

This is a wave equation describing the propagation of the vector field.

For propagation in the z-direction,

$$A^\nu = (0, \cos(kz - \omega t), 0, 0) \tag{3.6}$$

with

$$\omega = \sqrt{k^2 + a} \tag{3.7}$$

is a possible solution of Eq. (3.5).

The phase and the group velocities of the wave (3.6) are given by, respectively,

$$v_p = \frac{\omega}{k} = \frac{\sqrt{k^2 + a}}{k} \quad \text{and} \quad v_g = \frac{d\omega}{dk} = \frac{k}{\sqrt{k^2 + a}} \tag{3.8}$$

From this we see that a must be positive, because if a were negative, the group velocity would exceed the speed of light (assumed to be $c = 1$ in our units).

We can compare the expression for the group velocity with the expression for the velocity of a particle of momentum $p = \hbar k$:

$$v = \frac{\hbar k}{\sqrt{(\hbar k)^2 + m^2}} = \frac{k}{\sqrt{k^2 + (m/\hbar)^2}}$$

Accordingly, a is proportional to m^2, which is positive.

A negative value of a would correspond to an imaginary mass and propagation with speeds larger than the speed of light (tachyons).

To decide what (positive) value of a we should adopt, let us look at another special solution of Eq. (3.4), with a time-independent, spherically symmetric charge distribution.

If we consider the $\nu = 0$ component of Eq. (3.4) and omit all the terms containing time derivatives, we obtain a differential equation that relates A^0 to the charge density ϱ:

$$-\nabla^2 A^0 + aA^0 = 4\pi\rho \tag{3.9}$$

In the region exterior to the charge distribution, this has the solution

$$A^0 \propto \frac{e^{\pm\sqrt{a}r}}{r} \tag{3.10}$$

Only the solution with the negative sign is relevant, and this solution shows that the field decays exponentially with distance outside of the charge distribution.

By the rules of our game, we do not know Maxwell's equations, but we do know that the interaction is long range; we therefore will take $a = 0$ (see Section 1.2 for experimental limits on a).

Alternatively, we could introduce the assumption that the speed of electromagnetic waves should be equal to the speed of light; Eq. (3.8) would then lead to $a = 0$.

Our field equation then reduces to

$$\partial_\mu \partial^\mu A^\nu + b\partial^\nu \partial_\mu A^\mu = 4\pi j^\nu \tag{3.11}$$

The only unknown constant remaining in our field equation is b.

To determine this constant, we need to consider the conservation equation for the electric charge, $\partial_\nu j^\nu = 0$.

Historically, this conservation of charge is an empirical law of nature, accurately confirmed by many experiments.

However, we can also regard it as a consequence of the equations of electrodynamics, rather than as a separate empirical statement.

The advantage of this point of view is that it gives us an explanation for current conservation, so we do not have to accept conservation as an independent law of nature, imposed on us without any obvious reason.

If the charge-conservation equation $\partial_\nu j^\nu = 0$ is to be a consequence of the field equation (3.11), the left side of this equation must reduce to zero identically when we apply the operator ∂_ν :

$$\partial_\nu(\partial_\mu \partial^\mu A^\nu + b\partial^\nu \partial_\mu A^\mu) = 0 \tag{3.12}$$

If we exchange the (dummy) μ and ν indices in the second term on the left side, we obtain

$$(1+b)\partial_\mu \partial^\mu \partial_\nu A^\nu = 0 \tag{3.13}$$

Evidently, this equation will be identically true if $b = -1$, and we therefore adopt this value of b.

With $b = -1$, the final form of our field equation becomes

$$\partial_\mu \partial^\mu A^\nu - \partial^\nu \partial_\mu A^\mu = 4\pi j^\nu \tag{3.14}$$

This equation has the same form as the Maxwell equation for the potential that we wrote down in Section 2.5 [see Eq. (2.142)].

Obviously, our vector field A^ν is to be identified with the four-vector potential.

To obtain the Maxwell equations for the fields, we define the tensor $F^{\mu\nu}$ as

$$F^{\mu\nu} \equiv \partial^\mu A^\nu - \partial^\nu A^\mu \tag{3.15}$$

and then Eq. (3.14) becomes

$$\partial_\mu F^{\mu\nu} = 4\pi j^\nu \tag{3.16}$$

Furthermore, the tensor (3.15) identically satisfies the equation

$$\partial^\alpha F^{\mu\nu} + \partial^\mu F^{\nu\alpha} + \partial^\nu F^{\alpha\mu} = 0 \tag{3.17}$$

Equations (3.16) and (3.17) are, of course, exactly the Maxwell equations (2.111) and (2.112) that we formulated in Section 2.5, except that in Section 2.5 we retained the speed of light in these equations, whereas here we assume $c = 1$.

Although in some sense the preceding arguments may be regarded as a derivation of Maxwell's equations, we should keep the limitations of this approach in mind.

Clearly, we have had to make various assumptions to reach Eq. (3.14).

The goal of our game with electrodynamics was to obtain a recipe for finding the field equations in the hope that an analogous recipe will lead us to the field equations for gravitation.

The number of ingredients in this recipe need not concern us, as long as each ingredient in the electromagnetic case has an analogous ingredient in the gravitational case.

As we will see in the next section, the preceding arguments can indeed be reproduced step-by-step in the gravitational case and lead to a reasonable result.

Before we proceed to the gravitational case, we want to call attention to a remarkable property of the electromagnetic field equation (3.14):

Whenever A^ν *satisfies Eq.* (3.14), *then so does* $A^\nu + \partial^\nu \Lambda$, *where* Λ *is any arbitrary scalar function*.

This is so because in the equation

$$\partial_\mu \partial^\mu (A^\nu + \partial^\nu \Lambda) - \partial^\nu \partial_\mu (A^\mu + \partial^\mu \Lambda) = 4\pi j^\nu \tag{3.18}$$

the terms involving Λ cancel identically, and therefore the differential equation is left unchanged.

The transformation

$$A^\nu \rightarrow A^\nu + \partial^\nu \Lambda \tag{3.19}$$

is called a *gauge transformation*.

According to Eq. (3.18), the field equation is invariant under this gauge transformation.

In general, whenever an object is left unchanged (invariant) by some operation, the operation is said to be a symmetry of the object.

Thus, gauge invariance is a symmetry of the field equation.

The gauge function Λ is completely arbitrary; this means that our equations do not determine the field A^ν uniquely.

There is no harm in this, because an ambiguity of the form $\partial^\nu \Lambda$ in the field A^ν leads to no observable consequences.

The gauge transformations of A^ν leave the field tensor $F^{\mu\nu}$ unchanged, which implies that the observable quantities **E** and **B** are gauge independent.

For the purposes of this chapter we adopt the conservation law for the current as the motivation for the form of the field equation (3.14), with $b = -1$, so gauge invariance emerges as a consequence of the conservation law.

In Chapter 5 we will see that gauge invariance actually plays a more fundamental role and takes precedence – both the form of the field equation and the conservation law are consequences of gauge invariance, and the latter is motivated by the spin angular momentum and the zero mass of the electromagnetic field.

In short, spin 1 and mass zero of the electromagnetic field imply gauge invariance, and gauge invariance implies charge conservation.

The study of field theory reveals that, in general, gauge symmetries imply conservation laws; this is Noether's theorem.

However, it is helpful to make a particular choice of the gauge function and eliminate some of the ambiguity in A^{ν}.

We can do this by imposing the *Lorenz gauge condition* on the field,

$$\partial_\mu A^\mu = 0 \tag{3.20}$$

To see how this works, suppose that A^{μ} does not satisfy the condition (3.20).

Then we can always construct a new field $\bar{A}^\mu$ that does.

We simply take

$$\overline{A}^\mu = A^\mu + \partial^\mu \Lambda \tag{3.21}$$

and choose Λ in such a way that $\partial_\mu \bar{A}^\mu = 0$:

$$0 = \partial_\mu \overline{A}^\mu = \partial_\mu A^\mu + \partial_\mu \partial^\mu \Lambda \tag{3.22}$$

that is,

$$\partial_\mu \partial^\mu \Lambda = -\partial_\mu A^\mu \tag{3.23}$$

This is a differential equation which can be solved for Λ .

Equation (3.23) always has a solution.

In fact, it has several solutions, because to any solution Λ we can always add a solution of the homogeneous equation $\partial_\mu \partial^\mu \Lambda = 0$.

Accordingly, the Lorenz condition does not remove the ambiguity in the field A^μ entirely, and we can make further, supplementary, gauge adjustments within the constraint $\partial_\mu \partial^\mu \Lambda = 0$.

We will from now on assume that the Lorenz condition holds.

Note that with this condition the field equation reduces to

$$\partial_\mu \partial^\mu A^\nu = 4\pi j^\nu \tag{3.24}$$

This form of the field equation is often the most convenient for finding solutions for a given current.

In vacuo, where $j^\nu = 0$, Eq. (3.24) is a simple wave equation with a propagation speed equal to the speed of light.

For a wave propagating in the z-direction, Eqs. (3.24) and (3.20), with $j^\nu = 0$, have the following linearly independent solutions:

$$A^\nu = \varepsilon_{(n)}{}^\nu \cos(\omega z - \omega t) \tag{3.25}$$

where ε^ν is one of the following four-vectors:

$$\varepsilon_{(1)}{}^\nu = (0, 1, 0, 0), \quad \varepsilon_{(2)}{}^\nu = (0, 0, 1, 0), \quad \varepsilon_{(3)}{}^\nu = (1, 0, 0, 1) \tag{3.26}$$

The subscript n has been enclosed in parentheses to indicate that it is not a tensor index; it is only a label that distinguishes between the four-vectors in Eq. (3.26).

To complete our construction of electromagnetic theory, we need to find the equation of motion of a charged particle in an electromagnetic field.

This is best done by exploiting the principle of extremum action, because this principle incorporates the relativistic version of Newton's third law and ensures that energy and momentum changes of the particle are equal and opposite to energy and momentum changes of the field that acts on the particle, so the total energy and momentum of the particle plus field are always conserved.

We will deal with the derivation of the equation of motion from the action principle in Section 3.3.

3.2 Linear field equations for gravitation

To discover the field equations of gravitation, we proceed in a manner entirely analogous to what we did for electromagnetism.

This means that we assume that our spacetime is the flat spacetime of special relativity, and we look for a suitable field equation for the gravitational field.

The first thing we must do is decide what the source of gravitation is going to be; that is, we must ask what quantity plays a role analogous to that of the electromagnetic source j^{μ}.

The answer is given by Newton's principle of equivalence, according to which the (total) inertial mass of a system gravitates, that is, the energy gravitates.

The source of gravitation must therefore be the energy density.

However, it is impossible to construct a Lorentz-invariant theory of gravitation in which the energy density is the only source of gravitation.

We recall that the energy density T^{00} is only one component of the second-rank energy-momentum tensor $T^{\mu\nu}$.

What is pure T^{00} in one reference frame becomes $T'^{\mu\nu} = a^{\mu}{}_{0} a^{\nu}{}_{0} T^{00}$ in another reference frame, and in general, this is not pure T^{00}.

Thus, what is an energy density in one reference frame is some combination of energy density, energy flux density, and momentum flux density as seen in another reference frame, and an observer in this other reference frame will refute the claim that the energy density is the only source of gravitation.

If the laws are to have the same form in all Lorentz frames, then all these quantities must be sources of gravitation; that is, the complete tensor $T^{\mu\nu}$ must be the source of gravitation.

Before we proceed, a remark on a possible alternative.

We could try to use the invariant trace of the energy-momentum tensor as the source of gravitation.

Since the trace equals

$$T_{\mu}{}^{\mu} = T_0{}^0 + T_1{}^1 + T_2{}^2 + T_3{}^3 \tag{3.27}$$

we see that $T_{\mu}{}^{\mu}$ contains $T_0{}^0$, and therefore it would seem that the expression (3.27) satisfies the requirement that the energy density be the main source of gravitation.

The trouble with (3.27) is that for electromagnetic fields, $T_{\mu}{}^{\mu} = 0$ [see Eq. (2.137)].

Hence electromagnetic energy would not gravitate, in contradiction with the principle of equivalence.

The absence of gravitational interactions for electromagnetic energy would mean that light is not deflected by gravitational fields and also that the electrostatic energy within nuclei does not contribute to the gravitational mass, which disagrees with the Eotvos experiments.

We must now decide what type of field we will use as carrier of the interaction.

Because the source of gravitation is a second-rank, symmetric tensor, arguments analogous to those given in the electromagnetic case suggest that the obvious and natural choice is a second-rank, symmetric tensor field.

We will use the notation $h^{\mu\nu}$ for the gravitational field tensor.

As an alternative to this second-rank tensor, we might try a third-rank tensor. This gives results equivalent to those obtained by using a second-rank tensor.

It corresponds to the possibility of writing Einstein's equations in terms of $\Gamma^{\mu}{}_{\alpha\beta}$ rather than the usual $g_{\mu\nu}$; see Chapter 7.

The most general field equation that is linear in $h^{\mu\nu}$, is of second differential order, and contains $T^{\mu\nu}$ as source has the form

$$\begin{aligned}\partial_\lambda\partial^\lambda h^{\mu\nu} + a\partial^\mu\partial^\nu h_\sigma{}^\sigma - a'(\partial_\lambda\partial^\nu h^{\mu\lambda} + \partial_\lambda\partial^\mu h^{\nu\lambda}) + b\eta^{\mu\nu}\partial_\lambda\partial^\lambda h_\sigma{}^\sigma + b'\eta^{\mu\nu}\partial_\lambda\partial_\sigma h^{\lambda\sigma} \\ = -\kappa T^{\mu\nu}\end{aligned} \tag{3.28}$$

where a, a′, b, b′, and $\varkappa$ are constants.

We arrive at this form by writing down all possible symmetric tensors that can be constructed out of $h^{\mu\nu}$ and no more than two gradient operators ∂^σ.

The only such tensors, linear in $h^{\mu\nu}$ are the following:

$$\partial_\lambda\partial^\lambda h^{\mu\nu}, \quad \partial^\mu\partial^\nu h_\sigma{}^\sigma, \quad \partial_\lambda\partial^\nu h^{\mu\lambda}, \quad \partial_\lambda\partial^\mu h^{\nu\lambda}, \quad \eta^{\mu\nu}\partial_\lambda\partial^\lambda h_\sigma{}^\sigma,$$
$$\eta^{\mu\nu}\partial_\lambda\partial_\sigma h^{\lambda\sigma}, \quad h^{\mu\nu}, \quad \eta^{\mu\nu}h_\sigma{}^\sigma$$

The left side of Eq. (3.28) is a general linear combination of these tensors, except that we have omitted the last two, $h^{\mu\nu}$ and $\eta^{\mu\nu}h_\sigma{}^\sigma$.

The presence of such terms in Eq. (3.28) would produce an exponential decrease with distance in the static gravitational field [as in Eq. (3.10)]; we will assume that the interaction is long range and that therefore such terms are undesirable.

Incidentally, we could attempt to generalize (3.28) by using $\kappa T^{\mu\nu} + \kappa' \eta^{\mu\nu} T$ on the right side.

But this amounts to no more than a redefinition of the constants a, a′, b, b′ on the left side.

Our next task will be to determine the values of the constants a, a′, b, b′; we will see that our theory determines all these values unambiguously.

The constant κ characterizes the strength of the coupling between matter and gravitation, and its value must be determined by experiment.

This constant plays a role analogous to that of the electric charge.

The only difference is that the electric charge q is included in the definition of j^μ, whereas the constant κ is written separately from $T^{\mu\nu}$; this difference in notation is sanctified by tradition.

It is convenient to introduce the notation h for the trace of $h^{\mu\nu}$:

$$h \equiv h_\sigma{}^\sigma \tag{3.29}$$

With this, Eq. (3.28) becomes

$$\partial_\lambda \partial^\lambda h^{\mu\nu} + a\partial^\mu \partial^\nu h - a'(\partial_\lambda \partial^\nu h^{\mu\lambda} + \partial_\lambda \partial^\mu h^{\nu\lambda}) + b\eta^{\mu\nu} \partial_\lambda \partial^\lambda h + b'\eta^{\mu\nu} \partial_\lambda \partial_\sigma h^{\lambda\sigma} = -\kappa T^{\mu\nu} \tag{3.30}$$

The conservation law for the energy and momentum of matter is $\partial_\mu T^{\mu\nu} = 0$, which is analogous to the conservation law for electric charge, $\partial_\mu j^\mu = 0$.

By analogy with the case of electromagnetism in the preceding section, we want the conservation law $\partial_\mu T^{\mu\nu} = 0$ to be a direct consequence of the gravitational field equation.

We therefore require that when we operate with ∂_μ on both sides of Eq. (3.30), the left side must reduce to zero identically:

$$\begin{aligned}\partial_\lambda\partial^\lambda\partial_\mu h^{\mu\nu} + a\partial_\mu\partial^\mu\partial^\nu h - a'(\partial_\lambda\partial^\nu\partial_\mu h^{\mu\lambda} + \partial_\lambda\partial^\mu\partial_\mu h^{\nu\lambda}) \\ + b\partial^\nu\partial_\lambda\partial^\lambda h + b'\partial^\nu\partial_\lambda\partial_\sigma h^{\lambda\sigma} = 0\end{aligned} \tag{3.31}$$

To achieve this, the constants in this equation must satisfy the conditions

$$\begin{aligned}-a' + b' &= 0 \\ 1 - a' &= 0 \\ a + b &= 0\end{aligned} \tag{3.32}$$

Hence $a' = 1$, $b' = 1$, and $b = -a$.

Our field equation then reduces to

$$\partial_\lambda\partial^\lambda h^{\mu\nu} + a\partial^\mu\partial^\nu h - (\partial_\lambda\partial^\nu h^{\mu\lambda} + \partial_\lambda\partial^\mu h^{\nu\lambda}) - a\eta^{\mu\nu}\partial_\lambda\partial^\lambda h + \eta^{\mu\nu}\partial_\lambda\partial_\sigma h^{\lambda\sigma} = -\kappa T^{\mu\nu} \tag{3.33}$$

The value of the constant a remains undetermined, but it is of no physical relevance.

Different values of a merely correspond to different ways of describing the same physical gravitational field by means of different, but related, tensors.

Thus, instead of using the tensor $h^{\mu\nu}$ to describe the gravitational field, we could use a new tensor $\bar{h}^{\mu\nu}$ related to the former by

$$h^{\mu\nu} = \bar{h}^{\mu\nu} - C\eta^{\mu\nu}\bar{h} \tag{3.34}$$

where C is a constant.

Then the field equation for $\bar{h}^{\mu\nu}$ would have exactly the same form as Eq. (3.33), but a would be replaced by $\bar{a}$:

$$\bar{a} = a(1 - C) + 2C \tag{3.35}$$

A particularly convenient choice for C is

$$C = \frac{1-a}{2-4a} \tag{3.36}$$

This makes $\bar{a} = 1$, so we obtain the field equation

$$\partial_\lambda \partial^\lambda \bar{h}^{\mu\nu} + \partial^\mu \partial^\nu \bar{h} - (\partial_\lambda \partial^\nu \bar{h}^{\mu\lambda} + \partial_\lambda \partial^\mu \bar{h}^{\nu\lambda}) - \eta^{\mu\nu} \partial_\lambda \partial^\lambda \bar{h} + \eta^{\mu\nu} \partial_\lambda \partial_\sigma \bar{h}^{\lambda\sigma} = -\kappa T^{\mu\nu} \tag{3.37}$$

Note that $a = 1/2$ leads to trouble in Eq. (3.36).

However, this troublesome value of a must in any case be excluded, because for $a = 1/2$ our field equation (3.33) contains insufficient information for determining the field.

The choice $\bar{a} = 1$ has the advantage that it leads to a simple expression for gauge invariance (see later discussion); other choices of a make gauge invariance more messy.

Since the overbars in Eq. (3.37) serve no further purpose, we omit them and write our final equation as

$$\partial_\lambda \partial^\lambda h^{\mu\nu} + \partial^\mu \partial^\nu h - (\partial_\lambda \partial^\nu h^{\mu\lambda} + \partial_\lambda \partial^\mu h^{\nu\lambda}) - \eta^{\mu\nu} \partial_\lambda \partial^\lambda h + \eta^{\mu\nu} \partial_\lambda \partial_\sigma h^{\lambda\sigma} = -\kappa T^{\mu\nu} \tag{3.38}$$

This is the relativistic field equation for gravitation in the linear approximation.

It plays a role analogous to that of the Maxwell equations: Eq. (3.38) determines the gravitational field generated by a given matter distribution.

The analogy with electrodynamics is summarized in Table 3.1.

Since the gravitational field $h^{\mu\nu}$ is the analog of the four-vector potential A^μ, it would perhaps be more appropriate to call $h^{\mu\nu}$ the gravitational tensor potential; but for convenience we will continue to call it simply the gravitational field because, in the linear approximation, it is the only field quantity that we need.

Table 3.1 Analogy between the electromagnetic and gravitational field theories

Description	Electromagnetism	Gravitation (Linear Approximation)
Source of field	j^ν	$T^{\mu\nu}$
Conservation law	$\partial_\nu j^\nu = 0$	$\partial_\nu T^{\mu\nu} = 0$
Field	A^ν	$h^{\mu\nu}$
Field equation	$\partial_\mu \partial^\mu A^\nu - \partial^\nu \partial_\mu A^\mu = 4\pi j^\nu$	$\partial_\lambda \partial^\lambda h^{\mu\nu} + \partial^\mu \partial^\nu h - (\partial_\lambda \partial^\nu h^{\mu\lambda} + \partial_\lambda \partial^\mu h^{\nu\lambda}) - \eta^{\mu\nu} \partial_\lambda \partial^\lambda h + \eta^{\mu\nu} \partial_\lambda \partial_\sigma h^{\lambda\sigma} = -\kappa T^{\mu\nu}$
Gauge transformation	$A^\mu \rightarrow A^\mu + \partial^\mu \Lambda$	$h^{\mu\nu} \rightarrow h^{\mu\nu} + \partial^\mu \Lambda^\nu + \partial^\nu \Lambda^\mu$
Gauge condition	$\partial_\mu A^\mu = 0$	$\partial_\mu (h^{\mu\nu} - \frac{1}{2}\eta^{\mu\nu} h) = 0$
Field equation in this gauge	$\partial_\mu \partial^\mu A^\nu = 4\pi j^\nu$	$\partial_\lambda \partial^\lambda (h^{\mu\nu} - \frac{1}{2}\eta^{\mu\nu} h) = -\kappa T^{\mu\nu}$
Particle equation of motion	$m\, du_\mu/d\tau - q F_{\mu\nu} u^\nu = 0$	$du_\mu/d\tau + \kappa h_{\mu\alpha,\beta} u^\alpha u^\beta - \frac{1}{2}\kappa h_{\alpha\beta,\mu} u^\alpha u^\beta = 0$
Proper time interval	$d\tau^2 = \eta_{\alpha\beta} dx^\alpha dx^\beta$	$d\tau^2 = (\eta_{\alpha\beta} + \kappa h_{\alpha\beta}) dx^\alpha dx^\beta$

The assumption that the gravitational field is entirely described by a single tensor field entered in a crucial way into our construction of the field equation.

This is the simplest of all possible choices for the gravitational field.

All the available experimental data are in agreement with this tensor theory of gravitation, and therefore a more complicated choice would be contrary to Newton's First Rule for Reasoning in Philosophy (also known as Occam's razor): "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances."

To the best of our knowledge, one gravitational field is sufficient to explain all gravitational effects, and more fields would be a wretched excess.

If, nevertheless, we were to hypothesize that besides the tensor field there exist one or several extra gravitational scalar, vector, or tensor fields, then we could construct no end of endlessly complicated theories.

Among these alternatives the least complicated is the scalar-tensor theory of Jordan (1951) and Brans and Dicke (1961), which contains an extra scalar field; qualitatively, this scalar field has the effect of making the gravitational constant dependent on position.

The experimental evidence (see Section 4.3) speaks against the scalar-tensor theory, and we will therefore not bother with it.

In a region free of matter, the field equation (3.38) becomes

$$\partial_\lambda \partial^\lambda h^{\mu\nu} + \partial^\mu \partial^\nu h - (\partial_\lambda \partial^\nu h^{\mu\lambda} + \partial_\lambda \partial^\mu h^{\nu\lambda}) - \eta^{\mu\nu} \partial_\lambda \partial^\lambda h + \eta^{\mu\nu} \partial_\lambda \partial_\sigma h^{\lambda\sigma} = 0$$

This equation describes the propagation of free gravitational waves in empty space; we will investigate such waves in Chapter 5.

This equation coincides with the wave equation for a spin 2 particle of zero mass, the graviton.

Thus, we could have obtained our field equation for the gravitational tensor field by postulating that the gravitational interaction is carried by quanta of spin 2 and mass zero, just as the electromagnetic interaction is carried by quanta of spin 1 and mass zero, that is, photons.

However, we wanted to formulate the equation in a purely classical context, by relying on the analogy with electromagnetism.

In the electromagnetic case, we were able to obtain a simplified field equation [Eq. (3.24)] by taking advantage of gauge invariance.

We can also do this in the gravitational case. It is easy to show that Eq. (3.38) is invariant under the gauge transformation

$$h^{\mu\nu} \rightarrow h^{\mu\nu} + \partial^\mu \Lambda^\nu + \partial^\nu \Lambda^\mu \tag{3.39}$$

where $\Lambda^\mu(x)$ is an arbitrary vector field.

Exactly as in the electromagnetic case, gauge invariance implies that the field equations do not determine the field uniquely, but the ambiguity in the field is not serious and leads to no observable consequences.

We find it convenient to eliminate the ambiguity (partially) by imposing the following gauge condition on the fields:

$$\partial_\mu \left(h^{\mu\nu} - \tfrac{1}{2}\eta^{\mu\nu} h \right) = 0 \tag{3.40}$$

This is known as the *Hilbert gauge condition*.

If we are given a field $h^{\mu\nu}$ that does not satisfy Eq. (3.40), then, by a gauge transformation, we can always construct a new field $\bar{h}^{\mu\nu}$,

$$\bar{h}^{\mu\nu} = h^{\mu\nu} + \partial^\mu \Lambda^\nu + \partial^\nu \Lambda^\mu \tag{3.41}$$

such that the latter does satisfy Eq. (3.40).

Show that if Λ^ν is a solution of the differential equation

$$\partial_\mu \partial^\mu \Lambda^\nu = -\partial_\mu h^{\mu\nu} + \tfrac{1}{2}\partial^\nu h \tag{3.42}$$

then $\bar{h}^{\mu\nu}$ satisfies Eq. (3.40).

We now assume that the gauge condition (3.40) holds.

Then the field equation simplifies to

$$\partial_\lambda \partial^\lambda \left(h^{\mu\nu} - \tfrac{1}{2}\eta^{\mu\nu} h \right) = -\kappa T^{\mu\nu} \tag{3.43}$$

Here it is convenient to define a new field variable:

$$\phi^{\mu\nu} \equiv h^{\mu\nu} - \tfrac{1}{2}\eta^{\mu\nu} h \tag{3.44}$$

In terms of $\phi^{\mu\nu}$, the field equation then becomes

$$\partial_\lambda \partial^\lambda \phi^{\mu\nu} = -\kappa T^{\mu\nu} \tag{3.45}$$

and the gauge condition becomes

$$\partial_\mu \phi^{\mu\nu} = 0 \tag{3.46}$$

Eq. (3.44) implies

$$h^{\mu\nu} = \phi^{\mu\nu} - \tfrac{1}{2}\eta^{\mu\nu}\phi \tag{3.47}$$

As in the case of the electromagnetism, the gauge condition (3.46) does not remove the ambiguity in the field entirely, because we can always add to the gauge function calculated from the differential equation (3.42) another gauge function that satisfies the homogeneous equation

$$\partial_\mu \partial^\mu \Lambda^\nu = 0 \tag{3.48}$$

The solutions of this equation are free waves (that is, waves in vacuum, without interactions), and it is therefore not surprising that the supplementary gauge transformations constructed with solutions of Eq. (3.48) play a special role in the study gravitational waves.

We will explore this topic in Chapter 5.

Equation (3.45) expresses the field equation in a particularly simple and convenient form.

The remainder of this chapter and the next chapter are largely concerned with the investigation of different solutions of Eq. (3.45).

Thus, in Section 4.1 we will find the field surrounding a static, spherically symmetric mass, and in Section 4.7 the field surrounding a rotating mass.

In Chapter 5, we will look at the wave solutions of our field equation and investigate the emission of gravitational radiation.

However, it does us little good to calculate the field $h^{\mu\nu}$ (or $\phi^{\mu\nu}$) generated by some matter distribution if we do not know how this field acts on particles that are placed in it.

Our next task is to discover the equation of motion of a particle in a given gravitational field.

3.3 Variational principle and equation of motion

Hamilton's variational principle for the "action" gives us a convenient and elegant way to derive the equation of motion of a particle in electromagnetic or gravitational fields.

According to this principle, the motion of a particle is such as to make the action stationary whenever the particle moves from one given spacetime point P_1 to another spacetime point P_2.

Thus, the action evaluated along the actual worldline of the particle is equal to the action evaluated along any infinitesimally nearby worldline with the same endpoints P_1, P_2 (see Fig. 3.1).

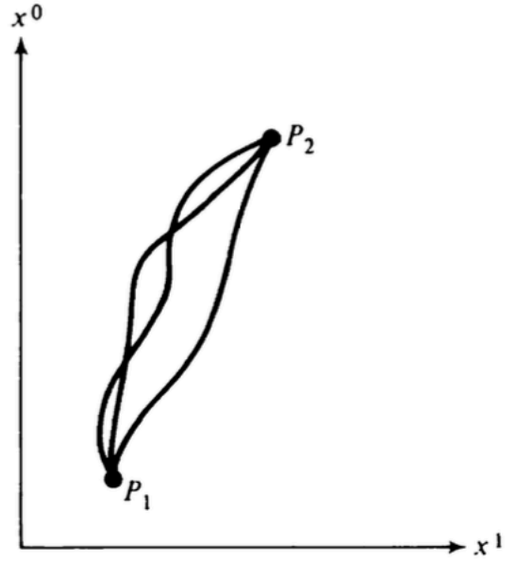


Fig. 3.1 Worldlines between two spacetime points P_1, P_2. One of these worldlines, say, the lowest, represents the actual worldline of the particle; the others are obtained by variation. For small variations, the action along any such worldline is equal to the action along the actual worldline.

To formulate the condition of stationary action, we need to introduce a parameter that labels the points along the worldlines.

In the relativistic case, it is expedient to use the proper time of the actual worldline as the parameter; we will use the usual notation τ for this parameter.

It is to be emphasized that, although τ is the proper time measured along the actual worldline, it is not the proper time along any of the other nearby worldlines that are considered in the variational principle.

Along these other worldlines, τ is to be regarded as a purely mathematical parameter, without any special physical interpretation.

The reason why we cannot parametrize the other worldlines by their own proper time is that the parameter used in the formulation of the variational equation must range over exactly the same set of values along all the nearby worldlines between P_1 and P_2, and this cannot be achieved if each worldline is parametrized with its own proper time.

To avoid this complication, the ordinary time coordinate t is often chosen as the parameter, instead of τ ; but our choice of τ will prove essential for dealing with the geodesic equation [see Eq. (3.116)], for which the proper time is the preferred parameter, and the choice we are making now anticipates this.

The action can then be written as an integral over this proper time,

$$I = \int_{\tau_1}^{\tau_2} L(x^\mu, u^\mu) d\tau \tag{3.49}$$

where the Lagrangian L is a function of the position x^μ and the corresponding velocities $u^\mu = dx^\mu/d\tau$.

The details of the functional dependence on coordinates and velocities depend, of course, on the dynamical system.

In general, the Lagrangian depends not only on the velocity but also on the presence of electromagnetic, gravitational, and other fields that act on the particle.

If we vary the actual worldline at the point $x^\mu(\tau)$ by an amount $\delta x^\mu(\tau)$, then the velocity will vary by $\delta u^\mu = (d/d\tau)\delta x^\mu$.

The change δI in the action will therefore be

$$\delta I = \int \left(\frac{\partial L}{\partial x^\mu} \delta x^\mu + \frac{\partial L}{\partial u^\mu} \delta u^\mu \right) d\tau = \int \left(\frac{\partial L}{\partial x^\mu} \delta x^\mu + \frac{\partial L}{\partial u^\mu} \frac{d}{d\tau} \delta x^\mu \right) d\tau$$

The second term in the integrand can be integrated by parts.

Because $\delta x^\mu = 0$ at the endpoints of the integration interval, we obtain

$$\delta I = \int \left(\frac{\partial L}{\partial x^\mu} - \frac{d}{d\tau} \frac{\partial L}{\partial u^\mu} \right) \delta x^\mu d\tau$$

To satisfy the condition $\delta I = 0$ of stationary action for arbitrary variations δx^μ, we need

$$\frac{\partial L}{\partial x^\mu} - \frac{d}{d\tau} \frac{\partial L}{\partial u^\mu} = 0 \tag{3.50}$$

This is the *Euler-Lagrange equation*.

We can also write this equation as

$$\frac{d}{d\tau} P_\mu - \frac{\partial L}{\partial x^\mu} = 0 \tag{3.51}$$

where

$$P_\mu \equiv \frac{\partial L}{\partial u^\mu} \tag{3.52}$$

is the momentum canonically conjugate to the coordinate x^μ, usually called the canonical momentum.

By way of illustration, let us consider the case of a free particle.

The equation of motion for a free particle is

$$\frac{d}{d\tau}(mu^\mu) = 0 \tag{3.53}$$

This equation can be obtained from the action

$$I = \int_{\tau_1}^{\tau_2} \frac{1}{2} m \frac{dx^\mu}{d\tau} \frac{dx_\mu}{d\tau} d\tau \tag{3.54}$$

with the Lagrangian

$$L_0 = \frac{1}{2} m \frac{dx^\mu}{d\tau} \frac{dx_\mu}{d\tau} = \frac{1}{2} m u^\mu u_\mu = \frac{1}{2} m \eta_{\mu\nu} u^\mu u^\nu \tag{3.55}$$

This Lagrangian gives a canonical momentum

$$P_\mu = \frac{\partial}{\partial u^\mu} \left(\frac{1}{2} m \eta_{\mu\nu} u^\mu u^\nu \right) = m u_\mu \tag{3.56}$$

and it gives the expected equation of motion (3.53).

The momentum mu_μ is the relativistic momentum that we already adopted in Chapter 2 on the basis of arguments related to momentum conservation.

Note that the condition $(dx^\mu/d\tau)(dx_\mu/d\tau) = 1$ can be inserted into the equation of motion, but it must *not* be inserted into Eq. (3.55) before obtaining the equation of motion.

This condition is valid only along the actual worldline of the particle where the parameter τ coincides with the proper time.

The other nearby worldlines that are considered in the variational principle have their own proper time distinct from τ , and hence along them $(dx^\mu/d\tau)(dx_\mu/d\tau) = 1$.

If we carelessly insert $(dx^\mu/d\tau)(dx_\mu/d\tau) = 1$ into Eq. (3.55), we obtain $L_0 = {}^1/_2\, m$, and then no equation of motion emerges.

The equation of motion for a free particle can also be obtained from a somewhat different Lagrangian,

$$L_0 = m\sqrt{u^\mu u_\mu} \tag{3.57}$$

with the action

$$I = \int_{\tau_1}^{\tau_2} m\sqrt{\frac{dx^\mu}{d\tau}\frac{dx_\mu}{d\tau}}d\tau \tag{3.58}$$

The Euler-Lagrange equation is then

$$\frac{d}{d\tau}\frac{mu^\mu}{\sqrt{u^\nu u_\nu}} = 0 \tag{3.59}$$

Since this equation refers to the actual worldline, it is valid to insert the condition $u^\nu u_\nu = 1$ into it.

We then recover the equation of motion (3.53) of a free particle.

The integral (3.58) can be expressed in the concise form

$$I = \int_{P_1}^{P_2} m\sqrt{dx^\mu dx_\mu} \tag{3.60}$$

which, except for the factor of m, is equal to the proper time along a worldline connecting the initial and final spacetime points (again, keep in mind that although $\sqrt{dx^\mu dx_\mu}$ is the proper time interval along any arbitrary worldline, the notation $d\tau$ is reserved for the proper time interval along the actual worldline determined by the equation of motion).

Thus, the principle of stationary action for Eq. (3.60) tells us that along the worldline of a free particle the proper time is stationary.

It is easy to show that the proper time is actually at a maximum.

This concludes our review of the variational principle for free particles.

We next consider the formulation of the equations of motion of a particle in electromagnetic or gravitational fields.

As in the earlier sections of this chapter, we begin with the formulation of the equation of motion of a charged particle in an electromagnetic field, as a preliminary exercise that we can then imitate in the gravitational case.

It is desirable to formulate the dynamics of the joint system consisting of a particle (or particles) and fields by an action principle with a complete Lagrangian that describes the particle, the fields, and their interaction.

This approach ensures that the joint energy and momentum of the particle and fields will be conserved, which amounts to a relativistic implementation of Newton's third law about action and reaction.

The complete Lagrangian consists of three terms: one term that contains only the particle variables, another term that contains only the field variables, and a third term that contains products of particle and field variables.

The third term is the interaction term; it describes the interaction between the particle and the fields.

We need to find this interaction term, so we can calculate what contribution it makes to the Euler-Lagrange equations for the motion of the particle when under the influence of the field.

To find this interaction term, we can exploit what we know about the electromagnetic field equation (3.14): the right side of the field equation represents the contribution that the particle makes to the Euler-Lagrange equation for the fields; thus, the right side of the field equations supplies information about the interaction.

We can extract additional information from the balance of actions and reactions.

The particle acts on the fields while the fields act on the particle, and, in some sense, actions and reactions are "equal and opposite," so if we know one of them we can discover the other.

However, because we are dealing with a particle and a field, both of which must be treated relativistically, the situation is not quite as simple as in the case of two particles exerting forces on each other in Newtonian physics, where the simple statement about equal and opposite forces in the third law suffices to implement the balance of actions and reactions.

Since the field variable A^μ plays the role of a coordinate that represents the state of the electromagnetic field, the contribution of the interaction term to the Euler-Lagrange equation for the field will involve the derivative $\partial L_{int}/\partial A^\mu$ of the interaction Lagrangian, and this derivative is what we see on the right side of Eq. (3.14), that is, $\partial L_{int}/\partial A^\mu \propto j_\mu$.

Accordingly, L_{int} must be of the form

$$L_{\text{int}} \propto j_\mu A^\mu \tag{3.61}$$

Here we have ignored the factor of 4π on the right side of Eq. (3.14); effectively, this means that we have elected to move this factor to the left side of Eq. (3.14), as a factor of $1/4\pi$.

Such rearrangements in the placement of constants in field equations are closely related to the question of choice of units for the fields and the electric charge, and we know, from painful experience with SI units vs. cgs, that there is much freedom of choice in these units.

For our present purposes, we elect to use Eq. (3.61) as is, without any extra factor.

When we now attempt to treat Eq. (3.61) as the Lagrangian for the motion of a particle, we encounter a tricky complication, because the action I_{int} for fields must be expressed as a volume integral of the Lagrangian, whereas the action for the particle must be an integral along the worldline [as in Eq. (3.58)].

So the action I_{int} must actually be both a volume integral (in regard to fields) and a line integral (in regard to the particle).

The correct combination requires a delicate use of delta functions and is a bit complicated.

But even without these details, it is intuitively obvious that in the Lagrangian of a single particle, the current density j_μ in Eq. (3.61) will have to be replaced by the current qu_μ of the particle along its worldline.

So the interaction term in the particle Lagrangian that produces the force on a single particle must be $qu_\mu A^\mu$ instead of $j_\mu A^\mu$, and if we add this interaction term to the free-particle Lagrangian (3.55), we obtain

$$L = \frac{1}{2} m u^\mu u_\mu + q u_\mu A^\mu \tag{3.62}$$

This immediately gives us an Euler-Lagrange equation

$$\frac{d}{d\tau}(m u_\mu + q A_\mu) - q u_\nu \frac{\partial A^\nu}{\partial x^\mu} = 0 \tag{3.63}$$

Note that here the canonical momentum $P_\mu = mu_\mu + qA_\mu$ differs from the ordinary ("kinetic") momentum mu_μ.

With $(d/d\tau)A_\mu = (dx^\nu/d\tau)(\partial/\partial x^\nu)A_\mu = u^\nu A_{\mu,\nu}$, Eq. (3.63) is equivalent to

$$\frac{d}{d\tau}(m u_\mu) = q u^\nu (A_{\nu,\mu} - A_{\mu,\nu}) \tag{3.64}$$

which agrees with Eq. (2.133).

The equation of motion of a particle in a gravitational field can be obtained by analogous methods.

The field variable $h_{\mu\nu}$ plays the role of a coordinate representing the state of the gravitational field, and by inspection of the right side of the gravitational field equation (3.38) we see that the derivative $\partial L_{int}/\partial h_{\mu\nu}$ of the interaction Lagrangian L_{int} must be

$$\frac{\partial L_{\text{int}}}{\partial h_{\mu\nu}} \propto T^{\mu\nu} \tag{3.65}$$

Here, as in the electromagnetic case, we have ignored the constant on the right side of Eq. (3.38). In the gravitational case, the conventional choice of constant in Eq. (3.65) is a factor of $\kappa/2$, so

$$\frac{\partial L_{\text{int}}}{\partial h_{\mu\nu}} = \frac{\kappa}{2} T^{\mu\nu} \tag{3.66}$$

We will adopt this choice, which means we implicitly adopt a choice of units for $h_{\mu\nu}$ consistent with Eq. (3.66). (Ultimately, the numerical value of κ will have to be determined by comparison of the field equation and equation of motion with the equations of Newtonian gravitation, and the insertion of the extra factor of $1/2$ merely changes how κ is related to the gravitational constant G.)

From Eq. (3.66) we conclude

$$L_{\text{int}} = \frac{1}{2}\kappa h_{\mu\nu} T^{\mu\nu} \tag{3.67}$$

For a cloud of particles, with an energy-momentum tensor $T^{\mu\nu} = \varrho_0 u^\mu u^\nu$ [see Eq. (2.80)], this interaction Lagrangian becomes $L_{int} = 1/2\ \kappa\, h_{\mu\nu}\, \varrho_0 u^\mu u^\nu$, where ϱ_0 is the rest-mass density.

For a single particle, integration of ϱ_0 over the volume of the particle then yields a term $1/2\ \kappa\, h_{\mu\nu}\, m u^\mu u^\nu$, and therefore this must be the interaction Lagrangian that produces the gravitational force.

If we add this term to the free-particle Lagrangian (3.55), we find

$$L = \tfrac{1}{2} m u^\mu u_\mu + \tfrac{1}{2} m \kappa h_{\mu\nu} u^\mu u^\nu \tag{3.68}$$

From this Lagrangian, we obtain the canonical momentum P_μ according to Eq. (3.52):

$$P_\mu = \frac{\partial L}{\partial u^\mu} = m u_\mu + m \kappa h_{\mu\alpha} u^\alpha \tag{3.69}$$

and the Euler-Lagrange equation

$$\frac{d}{d\tau}(u_\mu + \kappa h_{\mu\alpha} u^\alpha) - \frac{\kappa}{2} h_{\alpha\beta,\mu} u^\alpha u^\beta = 0 \tag{3.70}$$

This is the equation of motion for a particle in a (weak) gravitational field.

Note that in this equation the mass m has canceled, because each term in the equation has the same mass factor.

Thus, the motion of a particle in a gravitational field is independent of the mass of the particle, in agreement with the Galileo principle.

We can rewrite Eq. (3.70) in several ways.

For example, we can use

$$\frac{d}{d\tau} \kappa h_{\mu\alpha} u^\alpha = \kappa h_{\mu\alpha} \frac{d}{d\tau} u^\alpha + \kappa h_{\mu\alpha,\beta} u^\alpha u^\beta \tag{3.71}$$

Since the acceleration of the particle is of order κ h, the first term on the right side of Eq. (3.71) is of order $(\kappa \mathrm{h})^2$, and it can be neglected.

Therefore Eq. (3.70) is approximately equivalent to

$$\frac{du_\mu}{d\tau} + \left(\kappa h_{\mu\alpha,\beta} u^\alpha u^\beta - \frac{\kappa}{2} h_{\alpha\beta,\mu} u^\alpha u^\beta\right) = 0 \tag{3.72}$$

The $\mu = 0$ component of the canonical momentum P_μ is the kinetic plus potential energy of the particle in the gravitational field:

$$P_0 = mu_0 + m\kappa h_{0\alpha} u^\alpha \tag{3.73}$$

According to Eq. (3.70), this energy is a constant of the motion provided the gravitational field is time independent, so $h_{\alpha\beta,0} = 0$.

But if the gravitational field is time dependent, then the particle energy is not constant.

However, it is a general theorem of Lagrangian mechanics that

$$u^\mu \frac{\partial L}{\partial u^\mu} - L \tag{3.74}$$

is a constant of the motion, provided that L has no explicit dependence on the proper time parameter τ .

In our case, this integral of the motion takes the form

$$\frac{1}{2} m(u^\mu u_\mu + \kappa h_{\mu\nu} u^\mu u^\nu) = \text{constant} \tag{3.75}$$

The value of the constant on the right side of Eq. (3.75) is $^1/_2$ m.

This becomes obvious if we suppose that the particle was initially outside of the gravitational field (where $h_{\mu\nu} = 0$) or, equivalently, that the gravitational field was initially zero and reached its present value by gradually changing over time.

Hence Eq. (3.75) becomes

$$(\eta_{\mu\nu} + \kappa h_{\mu\nu}) u^\mu u^\nu = 1 \tag{3.76}$$

Alternatively, we can express this integral of the motion in terms of the (ordinary) momentum $p^\mu = mu^\mu$:

$$(\eta_{\mu\nu} + \kappa h_{\mu\nu}) p^\mu p^\nu = m^2 \tag{3.77}$$

Note that with $u^\mu = dx^\mu/d\tau$, Eq.(3.76) gives us a new relation between the parameter $d\tau$ and dx^μ:

$$d\tau^2 = (\eta_{\mu\nu} + \kappa h_{\mu\nu}) dx^\mu dx^\nu \tag{3.78}$$

This equation differs from the relation $d\tau^2 = \eta_{\mu\nu} dx^\mu dx^\nu$, which is valid in the absence of gravitation.

We will examine the implications of Eq. (3.78) in Section 3.5.

For now, we remark only that in the evaluation of Eq. (3.73) we must use the expression (3.78) for $d\tau$, rather than the familiar $d\tau^2 = \eta_{\mu\nu} dx^\mu dx^\nu$.

From Eq. (3.78) it follows that

$$u_0 = \frac{dt}{d\tau} = \frac{dt}{\left(1 - v^2 + \kappa h_{\mu\nu} \dfrac{dx^\mu}{d\tau} \dfrac{dx^\nu}{d\tau}\right)^{1/2} dt}$$

$$\cong \frac{1}{\sqrt{1 - v^2}} - \frac{1}{2} \frac{1}{(1 - v^2)^{3/2}} \kappa h_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau}$$

and

$$P_0 \cong \frac{m}{\sqrt{1 - v^2}} - \frac{1}{2} \frac{m}{(1 - v^2)^{3/2}} \kappa h_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau} + m\kappa h_{0\alpha} u^\alpha \tag{3.79}$$

The first term on the right side of this equation has the form of the usual sum of rest-mass and kinetic energy; the other terms represent a velocity-dependent gravitational interaction energy (potential energy).

3.4 Nonrelativistic limit and Newton's theory

Our linear gravitational theory will be a good approximation if the gravitational field is weak.

We can express this condition more precisely if we compare the linear and nonlinear terms in Eq. (3.70).

For instance, in Eq. (3.71) we neglected a nonlinear term $\kappa\, h_{\mu\alpha} du^{\alpha}/d\tau$.

Since $du^{\alpha}/d\tau$ is of the order of $\varkappa h$ (apart from tensor indices, derivatives, and factors of u^{μ}), this nonlinear term is of the order of $\kappa\, ^2h^2$.

In contrast, the linear terms are of the order $\varkappa h$, so these two kinds of terms typically differ by a factor $\kappa\, h$.

The nonlinearities are therefore insignificant if $\kappa\, h^{\mu\nu} \ll 1$.

As we will see in Section 4.1, the field surrounding a static mass M is of the order of $h^{\mu\nu} \cong GM/\kappa\, rc^2$ (see Eq. [4.9]), and hence the condition for the validity of the linear approximation becomes

$$\frac{GM}{rc^2} \ll 1 \tag{3.80}$$

This is the same as the condition for the validity of the Newtonian theory (see Section 1.1); that is, both the Newtonian and the linear relativistic theories hinge on the assumption that the field is weak.

However, Newtonian theory also requires the assumption that the speeds are low ($v \ll c$, or $v \ll 1$); this assumption is not required in the relativistic theory.

Thus the relativistic theory is an improvement because we can treat the motion of relativistic particles in (weak) gravitational fields, and we can treat time-dependent fields.

For instance, we can calculate the motion of light (photons) in the field of the Sun; we cannot do this in Newtonian theory.

We will now prove that, in the special case of static fields and low speeds, Newton's gravitational theory emerges, as it should, from the relativistic theory.

In the non- relativistic limit, $v^k \equiv dx^k/dt \ll 1$, so we can neglect terms of order v^2.

We must also decide what to do with terms of order $\kappa\, h \times v$.

For a particle orbiting a static mass M, these terms are of order $(GM/r) \times \sqrt{GM/r}$, which is somewhat larger than the terms of order $v^2 \cong (GM/r)$.

In the lowest approximation, we neglect both of these kinds of terms. In the next approximation, we would neglect the terms of order v^2, but we would retain the terms of order $\kappa\, h \times v$.

We will see that the lowest approximation corresponds to Newton's gravitational theory.

In the context of this lowest approximation, we can substitute $du^k/d\tau \cong dv^k/dt$ in the first term in Eq. (3.72), and we can substitute $u^0 \cong 1$, $u^k \cong 0$ in the term in parentheses.

This leads to a simplified approximate equation of motion

$$h_{00}\frac{dv^k}{d\tau} = -\kappa\left(h^k{}_{0,0} - \frac{1}{2}h_{00}{}^{,k}\right) = 0 \tag{3.81}$$

or, if we assume that the field is time independent,

$$\frac{dv^k}{d\tau} = \frac{1}{2}\kappa h_{00}{}^{,k} \tag{3.82}$$

We can compare this with the equation of motion given by Newton's theory of gravitation [see Eq. (1.31)]:

$$\frac{dv^k}{dt} = \Phi^{,k} \tag{3.83}$$

From this we see that Newton's equation of motion is a consequence of our theory in the nonrelativistic limit, provided that the Newtonian potential is proportional to h_{00}:

$$\Phi = \frac{1}{2}\kappa h_{00} \tag{3.84}$$

To complete the proof of the correspondence between Newton's theory and relativistic theory, we must show that the Newtonian field equation

$$\nabla^2\Phi = 4\pi G\rho_0 \tag{3.85}$$

is a consequence of our relativistic field equations.

In the nonrelativistic limit, the dominant component of the energy-momentum tensor is T^{00}, because this is the component that contains the rest-mass density.

We therefore have

$$T^{00} = \rho_0, \quad T^{k\mu} \cong 0 \tag{3.86}$$

The 00 component of the field equation (3.38) is

$$\partial_\lambda\partial^\lambda h^{00} + \partial^0\partial^0 h - 2\partial_\lambda\partial^0 h^{0\lambda} - \partial_\lambda\partial^\lambda h + \partial_\lambda\partial_\sigma h^{\lambda\sigma} = -\kappa T^{00} \tag{3.87}$$

To eliminate the last two terms on the left side, we take the trace of the field equation (3.38), which gives

$$-2\partial_\lambda\partial^\lambda h + 2\partial_\lambda\partial_\sigma h^{\lambda\sigma} = -\kappa T \tag{3.88}$$

We then subtract one-half times Eq. (3.88) from Eq. (3.87), with the result

$$\partial_\lambda \partial^\lambda h^{00} + \partial^0 \partial^0 h - 2\partial_\lambda \partial^0 h^{0\lambda} = -\kappa\left(T^{00} - \tfrac{1}{2}T\right) \tag{3.89}$$

For a time-independent field, this reduces to

$$-\nabla^2 h^{00} = -\kappa\left(T^{00} - \tfrac{1}{2}T\right) \tag{3.90}$$

With the energy-momentum tensor given by Eq. (3.86), we obtain $(T^{00} - 1/2\, T) \cong 1/2\, \varrho_0$, and therefore

$$\nabla^2 h^{00} = \tfrac{1}{2}\kappa\rho_0 \tag{3.91}$$

Equation (3.91) has exactly the same form as Eq. (3.85) and shows that the Newtonian field equation is a consequence of the relativistic field equation.

However, note that if some of the other components of the energy-momentum tensor are large, then we will obtain a field equation that differs from the Newtonian equation.

For instance, if the energy-momentum tensor is that of a fluid with a large pressure, with $T_1^{\,1} = T_2^{\,2} = T_3^{\,3} = -p$ [see Eq. (2.84)], then $T^{00} - 1/2\, T = \varrho_0 - 1/2\,(\varrho_0 - 3p) = 1/2\,(\varrho_0 + 3p)$, and

$$\nabla^2 h^{00} = \tfrac{1}{2}\kappa(\rho_0 + 3p) \tag{3.92}$$

This shows that pressure acts like an enhancement of the effective mass density, that is, pressure is an extra source of gravity.

When seeking the values of a, a′, b, b′ in Section 3.2, we could have saved ourselves some labor by demanding that our equations reduce to those of Newtonian theory in the nonrelativistic limit.

However, it is very satisfying that this correspondence should have emerged spontaneously.

Since, in the nonrelativistic limit, the equations of our theory are the same as those of Newtonian theory, the signs of the gravitational force acting between two particles are the same in both theories.

Interaction by means of a tensor field automatically gives an attractive force, just as interaction by a vector field automatically gives a repulsive force (between electric charges of the same sign).

To obtain the numerical value of the constant κ, substitute Eq. (3.84) into (3.91):

$$\frac{2}{\kappa}\nabla^2\Phi = \frac{1}{2}\kappa\rho_0 \tag{3.93}$$

Comparison of this with Eq. (3.85) then gives

$$\kappa = \sqrt{16\pi G} \tag{3.94}$$

In cgs units this becomes

$$\kappa = \sqrt{16\pi G/c^4} = 2.04 \times 10^{-24}\text{s}/(\text{g}\cdot\text{cm})^{1/2} \tag{3.95}$$

To examine the equation of motion and the field equation in the next approximation, we must retain the terms of order $\kappa h \times v$.

This means that in such terms we can substitute $u^0 \cong 1$, but we need to take $u^k \cong v^k \neq 0$, except in quadratic terms, where $u^k u^l \cong 0$. In a time-independent field, with $h_{\alpha\beta,0} = 0$, the equation of motion (3.72) then becomes

$$\frac{du_\mu}{d\tau} + \kappa(h_{\mu 0,l} - h_{0l,\mu})v^l - \frac{\kappa}{2}h_{00,\mu} = 0 \tag{3.96}$$

Here, it is convenient to define a new field quantity:

$$f_{\alpha\beta} = \frac{\kappa}{2}(h_{0\beta,\alpha} - h_{0\alpha,\beta}) \tag{3.97}$$

This quantity is a 4×4 antisymmetric matrix; note that $f_{\alpha\beta}$ is not a genuine tensor quantity; its Lorentz transformation properties are not those of a second-rank tensor, but those of particular 0 components of a third-rank tensor, as indicated by Eq. (3.97).

We can write the matrix $f_{\alpha\beta}$ as

$$f_{\alpha\beta} = \begin{pmatrix} 0 & g_x & g_y & g_z \\ -g_x & 0 & -b_z/2 & b_y/2 \\ -g_y & b_z/2 & 0 & -b_x 2 \\ -g_z & -b_y/2 & b_x/2 & 0 \end{pmatrix} \tag{3.98}$$

where $g_x = f_{01} = -(\kappa/2)h_{00,1}$, $b_x = 2f_{32} = \kappa(h_{02,3} - h_{03,2})$, and so on.

The three- component quantities **g** and **b** appearing here are three-dimensional vectors (but they are not vectors with respect to Lorentz transformations).

In terms of $f_{\alpha\beta}$, the equation of motion (3.96) becomes

$$\frac{dv_k}{dt} = 2f_{kl}v^l - f_{0k} \tag{3.99}$$

or, in vector notation,

$$\frac{d\mathbf{v}}{dt} = \mathbf{g} + \mathbf{v} \times \mathbf{b} \tag{3.100}$$

This equation is analogous to the usual equation for the Lorentz force in electrodynamics.

The gravitational field **g**, which is simply the Newtonian gravitational field [see Eq. (3.98)], is analogous to the electric field **E**; and the field **b** is analogous to the magnetic field **B**.

Because of this analogy, the field **b** is called the *gravitomagnetic field*.

Of course, the notation for **g** and **b** used in Eq. (3.98) was chosen in anticipation of this analogy.

The analogy between gravitational and electromagnetic quantities also applies to the field equations for **g** and **b** (Braginsky).

As a consequence of Eq. (3.97), the quantity $f_{\alpha\beta}$ identically satisfies the differential equation

$$\partial_\alpha f_{\mu\beta} + \partial_\mu f_{\beta\alpha} + \partial_\beta f_{\alpha\mu} = 0 \tag{3.101}$$

Note that this equation is analogous to Eq. (2.112) for the electromagnetic field.

Because Eq. (3.98) is analogous to Eq. (2.114), except for factors of $^1/_2$, we immediately recognize that the vectors **g** and **b** must satisfy equations analogous to the last two Maxwell equations, Eqs. (2.109) and (2.110):

$$\nabla \cdot \mathbf{b} = 0 \tag{3.102}$$

$$\nabla \times \mathbf{g} + \frac{1}{2}\frac{\partial \mathbf{b}}{\partial t} = 0 \tag{3.103}$$

It is also easy to verify these equations by direct substitution of Eq. (3.98) into (3.101).

The other two "Maxwell equations" emerge from the gravitational field equation, Eq. (3.38).

As we saw in Eq. (3.90), for a time-independent field, the 00 component of the field equation reduces to

$$-\nabla^2 h^{00} = -\kappa\left(T^{00} - \tfrac{1}{2}T\right) \tag{3.104}$$

In terms of $\mathbf{g} = -(\kappa/2)\nabla h^{00}$, this becomes

$$\nabla \cdot \mathbf{g} = -4\pi G(2T^{00} - T) \tag{3.105}$$

Likewise, for a time-independent field, the $0k$ component of the field equation (3.38) is

$$-\nabla^2 h^{0k} - \partial_l \partial^k h^{0l} = -\kappa T^{0k} \tag{3.106}$$

It is easy to check that this equation is equivalent to the *k* component of the following equation for the curl of **b**:

$$\nabla \times \mathbf{b} = -16\pi G\mathbf{S} \tag{3.107}$$

where **S** is the momentum density, or the energy flux density (the Poynting vector), with components $S^k = T^{0k}$.

Equations (3.105) and (3.107) are analogous to the Maxwell equations (2.107) and (2.108).

According to this analogy, $2T^{00} - T$ plays the role of a gravitational charge density (source of the gravitational field **g**), and the momentum density T^{0k} plays the role of a gravitational current density (source of the gravitomagnetic field **b**).

Note that in Eq. (3.107), the gravitational current density appears with a factor of 16π, in contrast to the factor of 4π that appears in the analogous Maxwell equation.

Thus, although the analogy between static gravitational fields and static electrodynamic fields is surprisingly close, it is not perfect.

The analogy breaks down altogether if the fields are not static.

The gravitational equations then contain several terms involving time derivatives that are not present in the electromagnetic equations.

We could further improve our approximate relativistic gravitational theory by including additional terms in our equations.

The systematic procedure for developing the equation of motion (and also the field equations, including their nonlinearities) in a series of terms involving powers of the speed and powers of $h_{\mu\nu}$ is called the post-Newtonian formalism.

The gravitational field is of the order GM/r, and the speed of an orbiting particle is of the order $\sqrt{GM/r}$; hence all the terms of the series are powers of $\sqrt{GM/r}$ or, in cgs units, $\sqrt{GM/rc^2}$, which is a dimensionless quantity.

The post-Newtonian formalism is used for accurate predictions of the motion of the planets [currently, in the most accurate numerical calculations, terms up to order $(\sqrt{GM/rc^2})^2$ are retained].

The post-Newtonian formalism is also often used to compare different gravitational theories with each other and with experiment. For this purpose, adjustable parameters are included with the coefficients of the powers of GM/rc^2, so that different theories can be characterized by different values of these parameters.

This parametrized post-Newtonian formalism (PPN) leads to very general, but very messy equations for the motion and for the fields (see Will).

Since we will deal exclusively with Einstein's gravitational theory, we will have no occasion to use the PPN formalism.

3.5 Geometric interpretation; curved spacetime

According to Eq. (3.78), the proper time interval for a particle moving in a gravitational field is given by

$$d\tau^2 = (\eta_{\mu\nu} + \kappa h_{\mu\nu})dx^\mu dx^\nu \tag{3.108}$$

This differs from the expression

$$d\tau^2 = \eta_{\mu\nu}dx^\mu dx^\nu \tag{3.109}$$

that is valid in the (flat) relativistic spacetime.

In fact, (3.108) is the spacetime interval for a curved spacetime geometry, and in Chapter 6 we will begin a full discussion of such geometries.

For now we only state briefly that, in a curved spacetime geometry, the constant Minkowski tensor $\eta_{\mu\nu}$ is replaced by a function $g_{\mu\nu}(x)$ of space and time, and the spacetime interval becomes

$$d\tau^2 = g_{\mu\nu}(x)dx^\mu dx^\nu \tag{3.110}$$

This expression determines the spacetime distances.

For example, a coordinate displacement dx^1 along the *x*-axis has an actual length $\sqrt{-g_{11}}\,dx^1$; that is, the measured distance differs from dx^1 by a factor $\sqrt{-g_{11}}$.

Likewise, a coordinate time (*t*-time) displacement dx^0 has a duration $\sqrt{g_{00}}\,dx^0$ when measured by the proper time of a clock at rest.

Because $g_{\mu\nu}$ determines the spacetime distances, it is called the *metric tensor* of the curved spacetime.

We will establish the precise connection between $g_{\mu\nu}$ and the curvature of the space-time in Chapter 6.

Under exceptional circumstances, the spacetime interval (3.110) may merely represent flat spacetime in curvilinear coordinates.

However, there is a simple test for discriminating between flat and curved spacetime: if the spacetime interval (3.110) represents flat spacetime in curvilinear coordinates, then a coordinate transformation to a suitable set of rectangular coordinates will transform $g_{\mu\nu}$ into $\eta_{\mu\nu}$.

If the space-time interval represents a genuine curved spacetime, then there exists no coordinate transformation that accomplishes this throughout the entire spacetime.

A coordinate transformation of a curved spacetime can, at the most, approximately transform $g_{\mu\nu}$ into $\eta_{\mu\nu}$ in a small neighborhood, but it cannot accomplish this transformation exactly and everywhere (curved spacetime is locally approximately flat, but not globally flat).

Comparison with Eq. (3.110) shows that the metric tensor corresponding to the spacetime interval (3.108) is

$$g_{\mu\nu} = \eta_{\mu\nu} + \kappa h_{\mu\nu} \tag{3.111}$$

This suggests a geometric interpretation of gravitation: a gravitating mass distorts space-time, giving it a curvature.

Furthermore, the motion of a particle in the field of a gravitating mass can be interpreted as a free motion along the geodesics, or the straightest possible worldlines, in the curved spacetime geometry.

Such straightest possible worldlines are defined by the condition that the proper time along the worldline between two points remains stationary with respect to arbitrary (but small) variations of the worldline, a condition that can be expressed by the variational principle

$$\delta \int_{P_1}^{P_2} \sqrt{g_{\mu\nu} dx^\mu dx^\nu} = 0 \tag{3.112}$$

or

$$\delta \int_{\tau_1}^{\tau_2} \sqrt{g_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau}} d\tau = 0 \tag{3.113}$$

This definition is motivated by what we know about the worldlines of a free particle in flat spacetime: it is the curve of extremum proper time connecting two points [see Eq. (3.60)].

We can verify that the resulting equation for a geodesic in curved spacetime agrees with our equation of motion for a particle in the gravitational field, that is, Eq. (3.70).

To obtain the explicit differential equation for the geodesic, we note that the Lagrangian is

$$L = \sqrt{g_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau}} \tag{3.114}$$

so the Euler-Lagrange equation is

$$\frac{d}{d\tau}\frac{\partial L}{\partial(dx^\mu/d\tau)} - \frac{\partial L}{\partial x^\mu} = \frac{d}{d\tau}\left(\frac{1}{L}g_{\mu\nu}\frac{dx^\nu}{d\tau}\right) - \frac{1}{2L}g_{\alpha\beta,\mu}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0 \tag{3.115}$$

Along the geodesic, τ is the proper time, and we can therefore insert the numerical value $g_{\mu\nu}(dx^\mu/d\tau)(dx^\nu/d\tau) = 1$ and obtain

$$\frac{d}{d\tau}\left(g_{\mu\nu}\frac{dx^\nu}{d\tau}\right) - \frac{1}{2}g_{\alpha\beta,\mu}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0 \tag{3.116}$$

This is called the *geodesic equation*.

With $g_{\mu\nu} = \eta_{\mu\nu} + \kappa\, h_{\mu\nu}$, we see that the geodesic equation takes the form

$$\frac{d}{d\tau}\left(\eta_{\mu\nu}\frac{dx^\nu}{d\tau} + \kappa h_{\mu\nu}\frac{dx^\nu}{d\tau}\right) - \frac{\kappa}{2}h_{\alpha\beta,\mu}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0 \tag{3.117}$$

which is the same as Eq. (3.70).

Our equation of motion for a particle in a gravitational field therefore says that the particle moves along a geodesic in a curved spacetime with a metric tensor $g_{\mu\nu} = \eta_{\mu\nu} + \kappa\, h_{\mu\nu}$.

This means that the effects of the gravitational field and the gravitational force can be regarded as purely geometric effects associated with motion in a curved geometry.

This geometric interpretation of the gravitational field hinges on the physical interpretation of the parameter τ defined by Eq. (3.108).

In the equation of motion, this parameter certainly seems to play the role of proper time.

We therefore expect that it actually is the proper time measured by a clock carried along by the particle.

Our expectation is based on what we know to be the case in Newtonian physics: the parameter t that appears in the Newtonian equation of motion does indeed agree with the time measured by clocks.

However, relativistic physics is more complicated.

We know that in special relativity, the parameter τ in the equation of motion does not agree with the time measured by clocks, but only with the proper time, that is, the time measured by a preferred clock that accompanies the particle during its motion.

And it is conceivable that, in the presence of gravitation, the situation is even more complicated and that the parameter τ has some new meaning involving some recalibration of clocks by the gravitational field.

In a purely mathematical sense, the variable τ enters the equation of motion (3.117) merely as a parameter that labels points along the worldline, and Eq. (3.108) relates this parameter to spacetime displacements.

Once the equation of motion has been solved, the parameter τ can be eliminated from the solution, leaving $\mathbf{x}$ as a function of t.

Thus, the value of the parameter does not affect the spacetime trajectory of the particle.

It might, therefore, be questioned whether we necessarily have to regard τ as the proper time of the particle.

Could it be that the variable τ given by Eq. (3.108) is a purely mathematical device having nothing to do with proper time?

If so, then there would be no need to regard spacetime as curved.

To confirm the physical interpretation of Eq. (3.108), we must check that a clock placed in a gravitational field shows a proper time $\sqrt{g_{\mu\nu}dx^\mu dx^\nu}$ rather than $\sqrt{\eta_{\mu\nu}dx^\mu dx^\nu}$.

For this, we must analyze the behavior of clocks in a gravitational field.

The essential ingredient in such an analysis is the principle of equivalence, which tells us that a clock in free fall in a gravitational field will behave in exactly the same way as though the gravitational field were absent.

Whether the clock is a mechanical clock (regulated by an oscillating balance wheel), an electronic clock (regulated by a vibrating quartz crystal), or an atomic clock (regulated by the vibrations of atoms) makes no difference – in any clock in free fall in a gravitational field, all the parts of the clock fall with the same acceleration, and hence the gravitational field cannot exert any direct effect on the rate of the clock.

The only direct effect of the gravitational field on a clock arises from tidal forces; but for most clocks, these are insignificant if the clock is of sufficiently small size.

Although the gravitational field cannot exert any direct effect on the rate of a clock, it exerts an indirect effect on the relative rates of clocks placed at different positions in this gravitational field.

To discover this indirect effect, we must compare the rates of clocks at different positions.

Since the behavior of freely falling clocks is completely predictable from the principle of equivalence, we will use freely falling clocks for all our measurements in the gravitational field, even measurements at a fixed position, for instance, a measurement at a fixed position on the surface of the Earth.

For this purpose, we use a freely falling clock, instantaneously at rest at the fixed position.

As soon as this clock has fallen too far from our fixed position and acquired too much speed, we must replace it by a new clock, instantaneously at rest.

Whenever we speak of the time as measured by “a clock located at a fixed position” in a gravitational field, this phrase must be understood as shorthand for a complicated measurement procedure, involving many freely falling, disposable clocks, used in succession.

We are now ready to compare the rates of two clocks at different fixed positions.

Suppose that clock 1 is located at position P_1 and clock 2 at position P_2, and suppose that the gravitational field is time independent (if the gravitational field is time dependent, then the comparison of clock rates is ambiguous – the result depends on how and when the comparison is made).

We can use either of two methods for the comparison of the rates of the clocks at these different fixed positions: we can send signals from P_1 to P_2, or we can transport a third, auxiliary clock from clock 1 to clock 2 and check each of the two clocks against this auxiliary clock.

Since we understand the behavior of freely falling clocks, it will be best to use the method of transport of an auxiliary clock and perform this transport by free fall.

Suppose that the auxiliary clock has a speed v_1 when it passes clock 1 and a speed v_2 when it passes clock 2.

The rest frame of the auxiliary clock is a (local) inertial frame, and relative to this reference frame the first clock has a speed v_1 and the second clock has a speed v_2.

Thus, relative to the rest frame of the auxiliary clock, both of the other clocks will suffer a time dilation.

Accordingly, when clock 1 passes by the auxiliary clock, we find that the rate of the clock 1 differs from that of the auxiliary clock by a factor

$$\frac{d\tau_1}{d\tau_{\text{aux}}} = \sqrt{1 - {v_1}^2} \tag{3.118}$$

Likewise, when clock 2 passes by the auxiliary clock, we find that

$$\frac{d\tau_2}{d\tau_{\text{aux}}} = \sqrt{1 - {v_2}^2} \tag{3.119}$$

The ratio of these two rates gives us the rate of clock 1 relative to the rate of clock 2:

$$\frac{d\tau_2}{d\tau_1} = \sqrt{\frac{1 - v_2{}^2}{1 - v_1{}^2}} \tag{3.120}$$

Although this result is exact, it becomes more useful if we make some approximations.

The speeds v_1 and v_2 are reckoned in a local inertial reference frame, and in a curved spacetime such a local speed is not simply equal to the rate of change calculated from the coordinate and time displacements dx^k and dt.

However, to lowest order in the gravitational constant, we can ignore this fine distinction, and we can approximate the speed by the ratio of space and time coordinate changes, $v = \sqrt{dx^2 + dy^2 + dz^2}/dt$.

Furthermore, for a freely falling body, we can approximate the difference between $v_1{}^2$ and $v_2{}^2$ by its Newtonian value, $v_2{}^2 - v_1{}^2 \sim= 2\Phi_1 - 2\Phi_2$.

With these approximations we obtain

$$\frac{d\tau_2}{d\tau_1} \cong 1 - \frac{1}{2}\left(v_2{}^2 - v_1{}^2\right) \cong 1 - (\Phi_1 - \Phi_2) \tag{3.121}$$

This shows that if clock 2 is at a lower potential than clock 1, so $\Phi_2 < \Phi_1$, then $d\tau_2$ is smaller than $d\tau_1$.

Thus, the clock that is deeper in the gravitational potential runs slower.

Equation (3.121) is the gravitational time-dilation formula, or redshift formula; we will discuss it further in Section 4.2.

The equation is valid to lowest order in the gravitational field strength.

If clock 1 is outside of the gravitational potential, at infinity, where $\Phi_1 = 0$, then its proper time coincides with the conventional time coordinate t, and $d\tau_1 = dt$.

Then Eq. (3.121) becomes

$$d\tau_2 \cong (1 + \Phi_2)dt \tag{3.122}$$

If we express this in terms of h_{00} and omit the subscript, it reduces to

$$d\tau \cong (1 + \tfrac{1}{2}\kappa h_{00})dt \tag{3.123}$$

This result shows that the proper time measured by a clock at rest in a gravitational field differs from *dt* by a factor $(1 + {}^1/_2\ \kappa h_{00})$.

Since $(1 + {}^1/_2\ \kappa h_{00}) = \sqrt{1 + \frac{1}{2}\kappa h_{00}} = \sqrt{g_{00}}$, this time-dilation factor agrees with the factor that determines the measured duration of coordinate time interval *dt* according to Eq. (3.110), $d\tau = \sqrt{g_{00}}\ dt$.

The agreement between the time-dilation factors obtained from our analysis of measurements by clocks and the factor given by Eq. (3.110) verifies the physical interpretation of $(\eta_{\mu\nu} + \kappa\ h_{\mu\nu})$ as the metric tensor of a curved spacetime, at least as concerns the 00 component of this tensor.

To verify the physical interpretation of the other components of the metric tensor, we must investigate the measurement of space intervals.

Such measurements can be conveniently performed by the "radar-ranging" procedure described in Section 2.1.

For a spatial displacement, the general procedure described in Section 2.1 reduces to ordinary radar ranging.

This means we place a clock at one end of the spatial displacement, and we then send a light pulse from this end to the other end, where we reflect the light pulse back to the starting point (see Fig. 3.2).

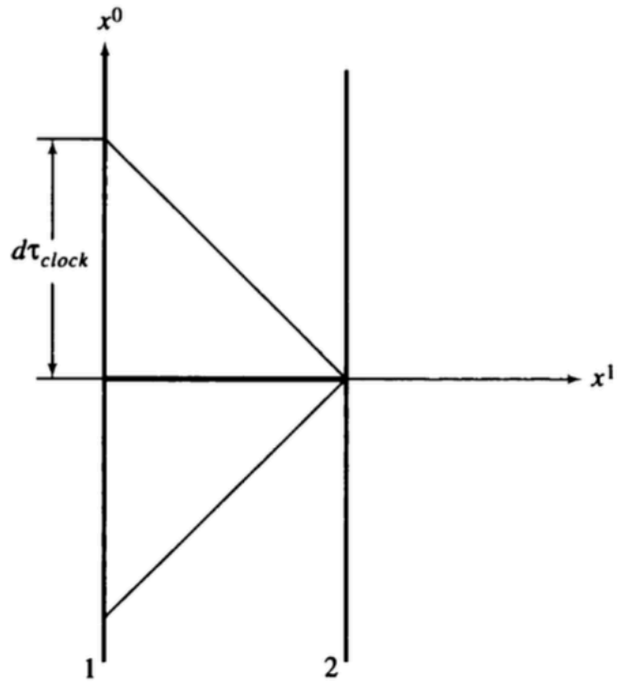


Fig. 3.2 Measurement of the length of a spatial displacement by radar ranging. The light signal starts at one end of the spatial displacement, reflects off the other end, and returns to the first end.

The clock tells us the travel time of the light pulse, that is, it tells us the proper time registered by the clock.

If we designate $^1/_2$ the travel time by $d\tau_{1/2}$, then the distance from one end of the displacement to the other is $dl = d\tau_{1/2}$.

Note that the radar-ranging procedure implicitly relies on the principle of equivalence: in the local inertial reference frame of the freely falling clock, light propagates as in the absence of the gravitational field, at its standard speed.

Because we understand the behavior of clocks in the gravitational field, we are able to relate the result of such a distance measurement to the field tensor $h_{\mu\nu}$.

First we need to examine what effect the field tensor $h_{\mu\nu}$ has on the propagation of light.

We already know the equation of motion of particles in the gravitational field, and we therefore find it convenient to think of light as consisting of particles, or photons, of zero mass.

For the case of zero mass, Eq. (3.77) becomes

$$(\eta_{\mu\nu} + \kappa h_{\mu\nu}) p^\mu p^\nu = 0 \tag{3.124}$$

The direction of propagation of a photon coincides with the direction of its momentum.

Hence, along the worldline of the photon, $dx^\mu \propto p^\mu$, and therefore

$$(\eta_{\mu\nu} + \kappa h_{\mu\nu}) dx^\mu dx^\nu = 0 \tag{3.125}$$

This establishes that, along the worldline of a light signal, ds^2 is zero,

$$ds^2 = (\eta_{\mu\nu} + \kappa h_{\mu\nu}) dx^\mu dx^\nu = 0 \tag{3.126}$$

Of course, this result is the obvious generalization of what we know to be true in the absence of gravitation, where $ds^2 = \eta_{\mu\nu} dx^\mu dx^\nu = 0$ for a light signal.

Equation (3.126) tells us how the increments dx^μ of the conventional space and time coordinates are related along the worldline of a light signal.

To exploit this equation in our radar-ranging procedure, suppose that $h_{0k} = 0$ for the sake of simplicity.

[If $h_{0k} \neq 0$, then it is awkward to express the radar-ranging procedure in terms of dx^k, because the light signal takes different amounts of *t* time going out and coming back.

The condition $h_{0k} = 0$ can always be achieved by a gauge transformation; it therefore does not really restrict the gravitational field in any serious way.]

Equation (3.126) then tells us

$$\sqrt{1 + \kappa h_{00}}\, dt = \sqrt{-(\eta_{kl} + \kappa h_{kl}) dx^k dx^l} \tag{3.127}$$

From our analysis of the behavior of clocks, we know that $\sqrt{1+\kappa h_{00}}dt$ is the proper time registered by a clock placed at a fixed position in the gravitational field.

If this clock is used in the radar-ranging procedure, the measured length of the spatial displacement (dx, dy, dz) is equal to half the proper time registered by the clock.

Thus, dl = $d\tau_{1/2}$ = $\sqrt{1+\kappa h_{00}}dt$, and according to Eq. (3.127) we can then conclude that

$$dl = \sqrt{-(\eta_{kl} + \kappa h_{kl})dx^k dx^l} \tag{3.128}$$

or

$$dl^2 = -(\eta_{kl} + \kappa h_{kl})dx^k dx^l \tag{3.129}$$

This equation expresses the measured length in terms of the coordinate increments, and it verifies the physical interpretation of the spatial components of the tensor ($\eta_{\mu\nu} + \kappa\ h_{\mu\nu}$) as components of the metric tensor of a curved spacetime.

If we introduce the notation $g_{kl} = (\eta_{kl} + \kappa\ h_{kl})$, we can write Eq. (3.129) as

$$dl^2 = -g_{kl}dx^k dx^l \tag{3.130}$$

We can then combine our results for the measurement of space and time intervals to find a general expression for an arbitrary spacetime interval measured by means of clocks placed in a gravitational field:

$$ds^2 = g_{00}dt^2 - dl^2 = g_{00}dt^2 + g_{kl}dx^k dx^l = g_{\mu\nu}dx^\mu dx^\nu \tag{3.131}$$

This result is in complete agreement with the expression for ds^2 that we found from our examination of the equation of motion.

Thus, all measurements based on clocks indicate that the metric of spacetime is $\eta_{\mu\nu} + \kappa\ h_{\mu\nu}$, rather than $\eta_{\mu\nu}$, and we therefore conclude that the real observable geometry is curved.

The flat spacetime geometry that we took as the starting point of our calculations in this chapter is unobservable, except, of course, if gravitational fields are absent.

It is to be emphasized that the geometric interpretation of gravity is not optional.

The arguments concerning measurements with clocks in the gravitational field compel us to accept this interpretation.

From these arguments we see that the formalism of the tensor theory of gravity implicitly carries within it its own physical interpretation.

For our further discussion of the linear approximation, we will take for granted the geometric interpretation, with the metric tensor $g_{\mu\nu} = \eta_{\mu\nu} + \kappa h_{\mu\nu}$.

For some parts of our discussion this is not really essential, because some experiments involve only measurements with clocks placed outside of the gravitational field, where we do not directly detect the curved geometry.

The deflection of light by the gravitational field of the Sun (see Section 4.3) is a typical experiment of this kind; the deflection angle is measured in a region that is far away from the deflecting body.

Note the crucial role played by the equivalence principle in several stages of the earlier arguments.

Both the comparison of clocks at different positions and the radar-ranging procedure hinge on the principle of equivalence.

This means that any deviation from the principle of equivalence – such as the deviations produced by tidal forces – would have serious consequences.

For instance, we might ask, can the tidal forces disturb an atomic clock?

A simple estimate shows that not even the strong tidal forces found on the surface of a neutron star can change the frequency of atomic vibrations by very much; the internal forces of the atom dominate and resist the tidal distortions.

In practice, we need to consider not only tidal forces but also accelerations.

Atomic clocks are usually not found in free fall.

They are usually at rest in the gravitational field (for instance, at rest on the surface of the Earth), and this means that they are accelerated relative to freely falling clocks.

We might worry about the disturbances caused by this acceleration.

These disturbances can be estimated, and their effect on the atomic vibrations can be seen to be negligible.

Incidentally, there is direct experimental evidence that confirms that even very large accelerations have an insignificant effect on the frequency of a well-built clock.

In the most severe experimental test of the effects of acceleration, nuclei were used as clocks.

The nuclear vibration frequency, as revealed by the emission of gamma rays via the Mossbauer effect, played the role of clock frequency.

When a sample of nuclei was placed in an ultracentrifuge and whirled around with an acceleration of 10^8 cm/s^2, the nuclear frequency was unaffected, except for the ordinary time dilation predicted by special relativity.

4 - Applications of the Linear Approximation

Twinkle, twinkle, little star
How I wonder where you are.
"1.75 seconds of arc from where I seem to be
For $ds^2 = (1-2GM/r)dt^2 - (1+2GM/r)(dx^2 + dy^2 + dz^2)$." *Source unknown*

In this chapter we will obtain some simple time-independent solutions of the linear field equations, such as the solution for the field produced by a static spherical mass and by a steadily rotating mass.

For a given solution of the field equations, the equation of motion then permits us to predict the trajectories of particles or of light signals, which we can compare with experimental or observational data.

Most of the experimental or observational tests that have so far been performed on the relativistic theory of gravitation involve only the linear approximation.

For instance, the gravitational time dilation predicted by the linear approximation has been tested and confirmed by experiments with clocks in the gravitational field of the Earth.

Some other tests exploit the motion of light signals in the gravitational field of the Sun.

Since light signals are necessarily relativistic, their motion provides a direct test of the relativistic features of our equations.

The gravitational field equations and the equation of motion predict a deflection of light rays and also a time delay of light signals passing through a gravitational field.

The observational data on the deflection and on the time delay of light by the gravitational field of the Sun confirm the predictions of the linear approximation.

The deflection of light was one of the early predictions of Einstein, and the observational confirmation of this effect in 1919, just after the Great War, was viewed as a spectacular success of his theory of general relativity.

By now, the deflection of light has become a tool for astronomers, who use it in the study of the multiple images produced when rays of light from a distant galaxy are deflected by the gravitational field of a foreground galaxy, which acts as a "gravitational lens."

4.1 Field of a spherical mass

The gravitational field surrounding a spherically symmetric, static mass distribution is as important in the relativistic theory as it is in Newton's theory.

In the exterior of the mass $T^{\mu\nu} = 0$, and we must therefore construct the spherically symmetric, time-independent solution of the field equations (3.45) and (3.46):

$$\partial_\lambda \partial^\lambda \phi^{\mu\nu} = 0 \tag{4.1}$$

$$\partial_\mu \phi^{\mu\nu} = 0 \tag{4.2}$$

This special solution describes the gravitational field produced by the Sun or by the Earth, provided we ignore the rotational motion and any equatorial bulge of these bodies.

With the assumption that the field is time independent, Eq. (4.1) reduces to

$$-\nabla^2 \phi^{\mu\nu} = 0 \tag{4.3}$$

The spherically symmetric solution of this equation is

$$\phi^{\mu\nu} = \frac{C^{\mu\nu}}{r} \tag{4.4}$$

where $C^{\mu\nu}$ is a constant tensor.

We could also include an additive constant in the solution (4.4), but we will assume that $\phi^{\mu\nu} \to 0$ as $r \to \infty$, so there is no such additive constant.

Substitution of Eq. (4.4) into Eq. (4.2) yields

$$C^{k\nu}\partial_k\left(\frac{1}{r}\right) = 0$$

that is,

$$C^{k\nu}\left(\frac{-x^k}{r^3}\right) = 0 \tag{4.5}$$

This can hold for all values of x^k only if $C^{k\nu} = 0$.

Hence the only nonzero component of $C^{\mu\nu}$ is C^{00}.

It then follows that

$$h_{00} = \phi_{00} - \frac{1}{2}\eta_{00}\phi = \frac{C_{00}}{r} - \frac{1}{2}\frac{C_{00}}{r} = \frac{C_{00}}{2r} \tag{4.6}$$

This determines the value of C_{00}, because we already know that h_{00} is proportional to the Newtonian potential [see Eq. (3.84)]:

$$h_{00} = \frac{2}{\kappa}\Phi = \frac{2}{\kappa}\left(-\frac{GM}{r}\right) \tag{4.7}$$

Hence $C_{00} = -4GM/\kappa$ and

$$\phi_{\mu\nu} = \begin{pmatrix} -4GM/\kappa r & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{4.8}$$

The corresponding expression for $h_{\mu\nu}$ is

$$h_{\mu\nu} = \begin{pmatrix} -2GM/\kappa r & 0 & 0 & 0 \\ 0 & -2GM/\kappa r & 0 & 0 \\ 0 & 0 & -2GM/\kappa r & 0 \\ 0 & 0 & 0 & -2GM/\kappa r \end{pmatrix} \tag{4.9}$$

Note that since the Newtonian potential of the mass M is $\Phi = -GM/r$, Eq. (4.9) can be written as

$$h_{\mu\nu} = \begin{pmatrix} 2\Phi/\kappa & 0 & 0 & 0 \\ 0 & 2\Phi/\kappa & 0 & 0 \\ 0 & 0 & 2\Phi/\kappa & 0 \\ 0 & 0 & 0 & 2\Phi/\kappa \end{pmatrix} \tag{4.10}$$

In this form, the equation is also valid for mass distributions other than spherical.

A general mass distribution can be regarded as a collection of point masses, and the linearity of our equations then tells us that the net field $h_{\mu\nu}$ of the mass distribution is the sum of the fields of the point masses.

Therefore the net field $h_{\mu\nu}$ of the mass distribution is related to the Newtonian potential of this mass distribution by Eq. (4.10).

For instance, if we want to find the field $h_{\mu\nu}$ in the interior of the Sun, we can simply substitute the Sun's Newtonian potential into Eq. (4.10).

We could use Eq. (4.9) and our relativistic equation of motion to study the motion of planets in the gravitational field of the Sun.

However, we will not bother with this because the relativistic deviations from the Newtonian orbits are extremely small.

The most noticeable deviation between the Newtonian and the relativistic orbital motions is the perihelion precession.

The only other deviation that we might hope to detect is a difference in the predicted instantaneous velocities along the orbit, which cannot be detected by ordinary telescopic observation.

With the development of radar astronomy, it has become possible to measure the instantaneous velocity of the (inner) planets to 1 part in 10^6 by means of the Doppler shift suffered by radar signals sent to these planets and reflected by them back to Earth.

But not even these radar- echo experiments are quite precise enough to detect the velocity deviations between Newtonian and relativistic gravitation.

At present, the only available test of the relativistic equation of motion of massive particles comes from the perihelion precession.

Although we could calculate the perihelion precession from our linear theory, it turns out that this precession receives comparable contributions from the linear approximation for $h_{\mu\nu}$ and from nonlinear corrections to $h_{\mu\nu}$ [note that the right side of Eq. (4.9) is of order GM; the nonlinear corrections are of order $(GM)^2$].

Such corrections come from the nonlinearities of the exact field equations that we will explore in Chapter 7.

Let us try to understand how these nonlinearities affect the motion of planets.

Consider the equation of motion, Eq. (3.70):

$$\frac{1}{m}\frac{d}{d\tau}P_\mu = \frac{\kappa}{2}h_{\alpha\beta,\mu}u^\alpha u^\beta$$

The right side of this equation may be regarded as a force per unit mass, that is, an acceleration.

From $u^0 \cong 1$ and $u^k \cong v^k$, we see the right side contains terms of the types $\kappa\, h_{00,\mu}$ and $\kappa\, h_{kl,\mu} v^k v^l$.

If we use the linear approximation (4.9), then the term $\kappa h_{00,\mu}$ is simply the Newtonian acceleration; this yields no perihelion precession.

The term $\kappa\, h_{kl,\mu} v^k v^l$ represents a deviation from the Newtonian acceleration and produces a perihelion precession.

Because the speed of a planet in orbit is roughly $v \cong \sqrt{GM/r}$, this term is of order $(GM)^2$.

We now notice that nonlinear corrections will change $\kappa\, h_{00,\mu}$ by a term of order $(GM)^2$.

This nonlinear correction is roughly of the same importance as $\kappa h_{kl,\mu} v^k v^l$.

Hence, for the evaluation of the perihelion precession, we need the terms of order $(GM)^2$ in $\kappa\, h_{00}$, although we need only the terms of order GM in $\kappa\, h_{kl}$.

We actually could calculate the required corrections by examining the contribution that gravitational field energy makes to the energy-momentum tensor and how this changes the field equation and its solution; we could then find the correct perihelion precession.

But it is preferable to set the whole question aside until after we formulate the exact nonlinear Einstein equations and obtain their exact solution in Chapter 7.

However, there are three interesting observable effects that we can calculate from the linear approximation: the gravitational time dilation, the deflection of light by gravitational fields, and the time delay of light signals.

4.2 Gravitational time dilation

According to Section 3.5, the tensor $\eta_{\mu\nu} + \kappa h_{\mu\nu}$ we obtained in the study of the equation of motion is the metric tensor $g_{\mu\nu}$ measured by means of clocks.

In the particular case of a spherically symmetric mass distribution, the metric tensor is therefore approximately

$$\begin{aligned} g_{\mu\nu} &= \eta_{\mu\nu} + \kappa h_{\mu\nu} \\ &= \begin{pmatrix} 1-2GM/r & 0 & 0 & 0 \\ 0 & -1-2GM/r & 0 & 0 \\ 0 & 0 & -1-2GM/r & 0 \\ 0 & 0 & 0 & -1-2GM/r \end{pmatrix} \end{aligned} \tag{4.11}$$

This interval corresponds to a curved spacetime; we will calculate the curvature in Chapter 6.

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = \left(1 - \frac{2GM}{r}\right)dt^2 - \left(1 + \frac{2GM}{r}\right)(dx^2 + dy^2 + dz^2) \tag{4.12}$$

Although the arguments of Section 3.5 establish that Eq. (4.12) describes the geometry surrounding, say, the Earth and that therefore we are living in a curved spacetime, it is desirable to have some direct experimental evidence for this curvature.

In 1821 Gauss attempted to find a curvature of space by laying out a triangle between mountain peaks, about 100 km apart. Gauss laid out this triangle with light rays, and he used surveying instruments to measure the sum of interior angles.

He found no deviation from 180°, to within his experimental error of 0.7 arcsec.

But even if he had found a deviation, could it have been interpreted as a simple bending of the path of light in the Earth's gravitational field with no geometrical implications?

In view of our definition of c as the local standard of velocity, the answer must be in the negative.

A deflection of light implies that the effective wave velocity defined by reference to standards of length and time kept outside of the gravitational field is a function of position.

Hence this effective velocity disagrees with the velocity measured locally in a freely falling reference frame, a paradox that can be resolved only if spacetime is curved (this is discussed in more detail in Section 4.4).

The gravitational time dilation, or slowing down of clocks in a gravitational field, serves as an even more direct test of the curvature of spacetime.

We already obtained the time-dilation formula from our general discussion of measurement procedures in Section 3.5.

There, we found a time dilation when we compared two clocks at different positions by means of an auxiliary clock allowed to fall freely from one position to the other.

In practice, most time-dilation experiments have been performed with light or radio signals sent from one clock to the other.

If a regular succession of identical signals (messenger rockets, cannonballs, periodic wavecrests of light, or whatever) is sent from one fixed position to the other, then each signal takes exactly the same amount of Δt time to complete this trip.

This is an immediate consequence of the time independence of the gravitational field and the repetitive character of the signals – the worldlines of successive signals are copies of each other, merely shifted forward in t-time (see Fig. 4.1).

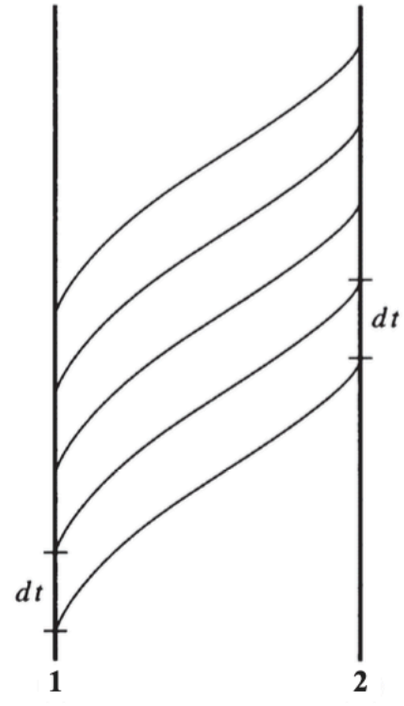


Fig. 4.1 Worldlines of successive, identical signals plotted in *r*, *t* coordinates. The *t*-time difference between departure of one signal and the next at the radial position 1 is *dt*, and the *t*-time difference between the arrival of one signal and the next at the radial position 2 is also *dt*.

But if each signal takes the same amount of time to complete the trip, and if the *t*-time difference between one signal and the next is *dt* at the point of departure, then the *t*-time difference must also be *dt* at the point of arrival.

However, *t*-time is not the time measured by clocks placed in the gravitational field.

Rather, *t*-time is the time indicated by a clock placed at a large distance from the gravitating body.

For such a distant clock, $r \to \infty$ and $g_{00} \to 1$; hence, the proper time indicated by such a clock is

$d\tau = \sqrt{g_{00}}\ dt = dt.$

In contrast, for clocks 1 and 2 placed at finite distances r_1 and r_2 the proper times are, respectively,

$$d\tau_1 = \sqrt{g_{00}(1)}dt \quad \text{and} \quad d\tau_2 = \sqrt{g_{00}(2)}dt \tag{4.13}$$

Given that the delay *dt* between successive signals traveling from r_1 to r_2 is the same at both positions, the ratio for the proper times that the two clocks indicate between such signals is simply

$$\frac{d\tau_2}{d\tau_1} = \frac{\sqrt{g_{00}(2)}}{\sqrt{g_{00}(1)}} \tag{4.14}$$

$$= \frac{\sqrt{1-2GM/r_2}}{\sqrt{1-2GM/r_1}} \cong 1 - \frac{GM}{r_2} + \frac{GM}{r_1} \tag{4.15}$$

Since $GM/r_1 - GM/r_2 = -\Phi_1 + \Phi_2$, the ratio $d\tau_2/d\tau_1$ for signals sent from one clock to the other agrees with the ratio we obtained in Section 3.5 by transport of an auxiliary clock from one to the other – both calculations give us the same value for the gravitational time dilation.

Equation (4.14) indicates that the experiments on gravitational time dilation may be regarded as direct measurements of the g_{00} component of the metric tensor.

The experiments show that in the vicinity of a massive body the metric tensor does not have the flat spacetime value $g_{00} = 1$, but rather the value $g_{00} = 1 - 2GM/r$.

We are constrained to accept this purely geometric interpretation of the time-dilation effect because, according to the principle of equivalence, it is impossible for the gravitational field to influence the rate of a (freely falling) clock in any other way.

A modification of the clock rate by a change in the spacetime geometry is the only possible alternative.

Most gravitational time-dilation experiments were performed with electromagnetic waves sent from one clock to the other.

In that case, the clocks measure signal frequencies; that is, the number of signals (or wavecrests) emitted per unit clock time at 1 and received per unit clock time at 2.

The frequencies measured by the clocks at 1 and at 2 are then equal to the inverses of the proper time intervals, $\nu_1 = 1/d\tau_1$ and $\nu_2 = 1/d\tau_2$, respectively, so Eq. (4.15) becomes

$$\frac{\nu_1}{\nu_2} = 1 - \frac{GM}{r_2} + \frac{GM}{r_1} \tag{4.16}$$

This gives us the fractional difference between the signal frequencies measured by clocks 1 and 2:

$$\frac{\Delta\nu}{\nu} \cong \frac{\nu_2 - \nu_1}{\nu_2} = -GM\left(\frac{1}{r_1} - \frac{1}{r_2}\right) \tag{4.17}$$

Here we have written the fractional frequency difference as $\Delta\nu/\nu$, without specifying the choice of ν_1 or ν_2 in the denominator, because this choice alters the result by only a small, insignificant amount of order $(GM)^2$.

If $r_2 > r_1$, then $\Delta\nu < 0$, and hence periodic signals sent from the clock at the smaller radius r_1 will display a reduced frequency when they arrive at, and are measured with, the clock at the larger radius r_2.

In particular, the spectral lines in the light emitted by an atom placed deep in a gravitational potential will display a *redshift* when compared with the spectral lines emitted by a similar atom placed outside of the potential.

If the clocks are near the surface of the Earth, we can make the approximation

$$GM\left(\frac{1}{r_1} - \frac{1}{r_2}\right) = GM\left(\frac{1}{R} - \frac{1}{R + \Delta r}\right) \cong GM\frac{\Delta r}{R^2} = g\Delta r \tag{4.18}$$

where Δr is the difference in height between the clocks and R is the radius of the Earth.

With this approximation Eq. (4.17) becomes, in cgs units,

$$\frac{\Delta\nu}{\nu} = -\frac{g\Delta r}{c^2} \tag{4.19}$$

For a height difference of 10 or 20 m, such as might be used in a laboratory experiment, the frequency shift is extremely small, of the order of $\Delta\nu/\nu \approx -10^{-15}$.

Even for a height difference of 10 km, which is the typical height attained by an aircraft, the frequency shift is only $\Delta\nu/\nu \approx -10^{-12}$.

The first laboratory measurement of the gravitational time dilation was performed at Harvard by Pound and Rebka (1960), who exploited iron nuclei as high-precision "nuclear" clocks, whose frequency can be compared by emission and absorption of γ rays.

They placed an emitter of γ rays (^{57}Fe) at ground level and detected the γ rays with an absorber placed on top of a tower, 22.6 m above the emitter (see Fig. 4.2).

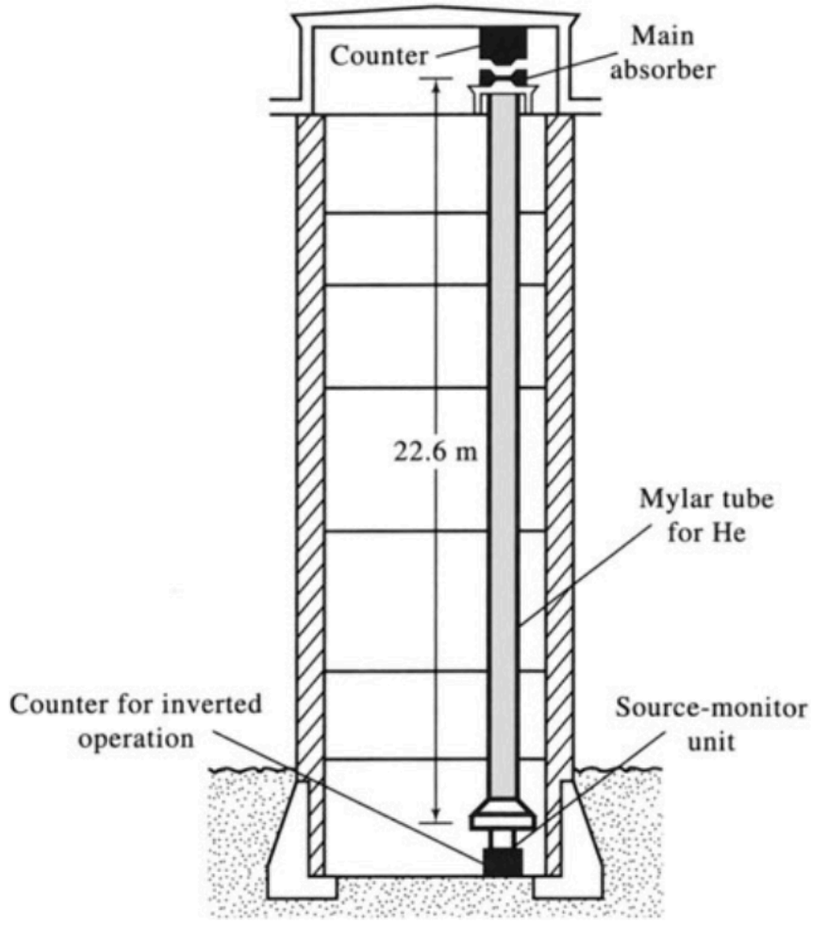


Fig. 4.2 Arrangement of source and absorber of γ rays used by Pound and Rebka (1960).

For that height, the predicted frequency shift is $\Delta \nu/\nu = -2.46 \times 10^{-15}$, but the measurement of this small frequency shift is possible because resonant absorption of γ rays in an absorber of ^{57}Fe is very sensitive to the frequency of the incident γ rays.

The experiment relied on the Mossbauer effect to prevent loss of any of the γ-ray energy by recoil of the nucleus during emission or absorption.

The most precise version of this experiment (by Pound and Snider, 1964, 1965) agreed with the predicted value to within the experimental error, about 1%.

The gravitational time dilation has also been tested with atomic clocks.

For a direct test of the time dilation, Hafele and Keating (1972) carried cesium-beam clocks on flights around the world in commercial aircraft.

On each flight, they carried four cesium clocks, and they intercompared these continuously, so that small spontaneous jumps in rate, to which these atomic clocks are prone, could be identified and corrected for. The clocks carried to high altitude (about 10 km) by the aircraft were found to have gained of the order of 100 nanoseconds when brought back to a laboratory clock that stayed on the ground.

To be precise, the clocks were found to have gained time after a kinematic correction was subtracted to take into account the special-relativistic time dilation.

The kinematic correction and the predicted redshift were calculated from the flight data of the aircraft. The experiment verified the gravitational time dilation to within about 10%.

A similar, but more precise test was performed by Alley (1979), who flew an aircraft equipped with atomic clocks in a holding pattern above a ground station, also equipped with atomic clocks.

The clocks in the aircraft and on the ground were compared during flight by laser signals, and the aircraft was tracked continuously by radar, to determine height and speed.

The development of highly stable hydrogen-maser clocks has permitted the most precise test to date. Vessot et al. used a Scout rocket to launch a hydrogen-maser clock to an altitude of about 10,000 km.

Radar signals were used to compare the rate of the clock on the rocket with a similar clock on the ground.

Furthermore, separate radar signals sent to the rocket and returned by a transponder were used to monitor the height and instantaneous speed of the rocket, by the Doppler shift (if a radar signal travels up and down in the gravitational field, it would return to the ground with unchanged frequency if the rocket were instantaneously stationary; any frequency shift of such radar signal is a Doppler shift, which reveals the speed of the rocket).

The data from this experiment confirmed the prediction for the gravitational time dilation to within 2 parts in 10^4.

Gravitational redshifts have also been measured for light emitted by atoms on the surface of stars.

For example, relative to oscillations of atoms at a large distance, oscillations of atoms on the surface of the Sun have a frequency shift of

$$\frac{\Delta\nu}{\nu} = -\frac{GM_\odot}{R_\odot c^2} = -2.12 \times 10^{-6} \tag{4.20}$$

It has been difficult to obtain clear-cut experimental results for this redshift because there are strong convection currents in the solar atmosphere, and the spectral lines are subjected to Doppler shifts arising from the motion of the gas.

The best results were obtained by observations at the limb of the Sun, where the Doppler shifts from convection currents are minimal and the observed redshift was found to be in agreement with the value given by Eq. (4.20), to within the observational uncertainty of about 2%.

The gravitational time dilation was also measured and confirmed by the Voyager and Galileo spacecraft while in the gravitational field of Saturn and the Sun, respectively, to within an observational uncertainty smaller than 1%.

Table 4.1 lists the best tests of the time dilation; the last column of this table gives the ratio of experimental results to theoretical predictions.

Table 4.1 Tests of the gravitational time dilation

Experimenter(s)	Year	Method	$\Delta\nu_{\text{exp}}/\Delta\nu_{\text{theory}}$
Pound and Rebka	1960	redshift of γ rays	1.05 ± 0.1
Pound and Snider	1964	redshift of γ rays	1.00 ± 0.01
Häfele and Keating	1972	cesium clocks on aircraft	0.9 ± 0.1
Alley	1979	cesium and rubidium clocks on aircraft	1 ± 0.02
Vessot and Levine	1979	hydrogen-maser clock on rocket	1 ± 0.0002
Krisher et al.	1990	Voyager spacecraft in Saturn's field	1 ± 0.01
Krisher et al.	1993	Galileo spacecraft in Sun's field	1 ± 0.005
LoPresto and Schrader	1991	redshift of O lines on Sun	0.99 ± 0.02

Time-dilation measurements play a crucial role in our understanding of gravitation, because they offer direct empirical evidence that spacetime in a gravitational field is curved.

This evidence is independent of the theoretical considerations that led us to the curved spacetime described by Eq. (4.10).

If spacetime were flat, a clock slowly transported from a point P_1 in a weak gravitational field to a point P_2 in a stronger field and then back to its starting point would not show any deviation from a similar clock held stationary at the starting point.

In flat spacetime, such slow-clock transport is a standard technique for transferring synchronization from one location to another.

The flat spacetime interval $d\tau^2 = dt^2 - dx^2 - dy^2 - dz^2 = (1 - V^2)dt^2$ ensures that, in the limit of zero speed, the elapsed proper time shown by the transported clock agrees with ordinary time, because

$$\Delta\tau = \int \sqrt{1 - V^2}dt \cong \int (1 - \tfrac{1}{2}V^2)dt = \Delta t - \tfrac{1}{2}V \int dl \to \Delta t \text{ for } V \to 0 \tag{4.21}$$

Any deviation between the readings of the slowly transported clock and the stationary clock is therefore conclusive evidence that $d\tau^2$ differs from $dt^2 - dx^2 - dy^2 - dz^2$, which implies that spacetime is not flat, but curved.

The Hafele-Keating experiment gives us the most direct empirical evidence of this kind.

The low precision of the experiment does not affect its evidentiary value – any deviation between the clocks, even if small, is empirical proof for curved spacetime.

The other time-dilation experiments listed in Table 4.1 involved clock comparisons made by signals instead of direct comparison by a round-trip clock transport.

But even such an indirect comparison of clocks is sufficient for our purposes, because the worldlines of successive signals used in these experiments are identical (see Fig. 4.1), and if spacetime were flat, the proper time intervals between successive signals emitted and received by two stationary clocks would be equal, $d\tau_1 = dt$ and $d\tau_2 = dt$, and there could be no redshift.

The observed redshift contradicts this result and establishes that the proper times differ; that is, the relationship between proper time $d\tau$ and coordinate time dt is position dependent, which, again, implies that spacetime must be curved.

The gravitational time dilation finds a practical application in the operation of the Global Positioning System (GPS), which calculates the position of a radio receiver on the ground by means of the measured travel times of radio signals transmitted from several Earth satellites equipped with highly accurate, synchronized clocks.

Calculation of the position to within a few meters requires synchronization of the clocks to within about 10 nanoseconds, which can be achieved only if the gravitational time dilation of the clocks is taken into account.

Recently, a pair of optical atomic clocks (that is, clocks operating at optical frequencies) was used in a laboratory experiment that detected the gravitational time dilation within a height difference of less than a meter.

This is an impressive demonstration of the high precision of atomic clocks, but because of the rather large experimental errors, it does not give us an improved test of the gravitational time dilation.

However, the experimenters suggested that with further developments of this technique the gravitational time dilation might be used for the measurement of small height differences of a few centimeters, in practical applications in geodesy and geophysics.

Finally, we discuss a simple derivation of the redshift of historical and conceptual interest, which was formulated by Einstein in 1911 before he recognized that spacetime is curved.

Einstein directly exploited the equivalence principle by means of a Gedanken-experiment that replaces the gravitational field by the pseudo-force field generated in an accelerated reference frame.

He simulated the effect of a uniform gravitational field g by placing the two clocks in a reference frame ("elevator") accelerating upward at the rate g (in the absence of any gravitational field; see Fig. 4.3).

An inertial observer then describes the situation as follows: Clocks 1 and 2, separated by a vertical distance Δr, are both accelerating upward with acceleration g.

Clock 1 is emitting light pulses at the rate ν_1 signals per second.

In the time $\cong \Delta r/c$ that these light pulses take to reach clock 2, the velocity of clock 2 increases by $\Delta V = g\,\Delta r/c$.

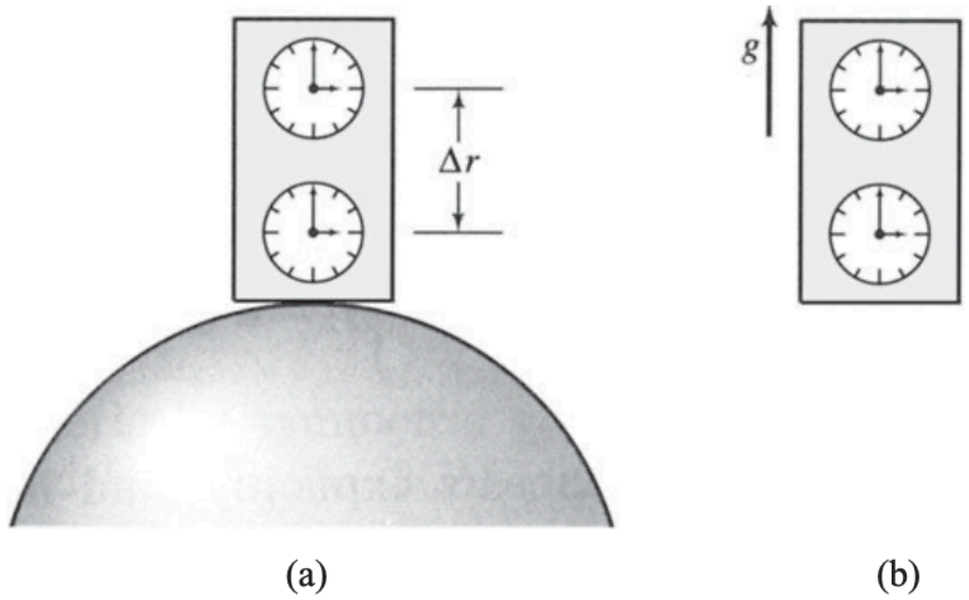


Fig. 4.3 (a) Two clocks at rest in the gravitational field of the Earth. (b) Two clocks in an accelerated reference frame.

This means that the rate ν_2 at which pulses are received by clock 2 is decreased by the Doppler shift, so,

$$\nu_2 \cong \nu_1 \left(1 - \frac{\Delta V}{c}\right) = \nu_1 \left(1 - \frac{g\Delta r}{c}\right) \tag{4.22}$$

which gives us a result

$$\frac{\Delta \nu}{\nu} \cong \frac{\nu_2 - \nu_1}{\nu_1} \cong -\frac{g\Delta r}{c^2} \tag{4.23}$$

in agreement with Eq. (4.19).

But there is something incongruous about this deduction of the gravitational redshift.

The calculation seems to rely on what is no more than simple Newtonian physics, that is, the Doppler-shift formula and the equivalence of gravitational and inertial effects, both of which seem innocuous and trivial.

In Newtonian physics, the approximate Doppler-shift formula in Eq. (4.22) is an obvious consequence of the velocity-addition law, and it is also obvious that a constant gravitational field can be replaced by an accelerated reference frame, because the gravitational force in the former is identical to the acceleration pseudo-force in the latter.

But it is puzzling that a time dilation of clocks could emerge from these straightforward ideas – time dilation is a relativistic effect, and it should not emerge from Newtonian physics.

The answer to this conundrum is that Einstein's calculation actually introduces the special-relativistic time dilation implicitly, by the assumption of a constant, universal speed of light signals, unaffected by the speed of the elevator.

We can understand the crucial role of relativistic physics in Einstein's derivation of the elevator redshift more clearly if we replace the light pulses by sound pulses propagating in the air contained in the elevator (alternatively, we could replace the light pulses by an evenly spaced sequence of BB pellets fired from a BB gun).

Each pulse of sound has the same speed relative to the elevator, but between emission of one pulse and the next, the elevator increases its speed; therefore, relative to the inertial reference frame with respect to which the elevator accelerates, each pulse has a slightly higher speed than the preceding one.

By taking into account the Newtonian velocity-addition law, we then readily find that the frequency shift is reduced to zero – there is no frequency shift at all.

And, by the equivalence principle, we would then conclude that there is no frequency shift for a sequence of sound pulses in a gravitational field

This makes it evident that to obtain the gravitational redshift we need to consider relativistic corrections to the Newtonian calculation (for another hint in that direction, we might recall that the derivation of the gravitational time dilation in Section 3.5, by means of a freely falling clock, relied directly on the time-dilation effect of special relativity).

And, indeed, if instead of the Newtonian addition law for velocity we use the special-relativistic velocity-addition law, then the elevator Gedanken-experiment with sound pulses yields the expected redshift, exactly as in Einstein's calculation with light waves.

For the sake of simplicity, assume that the speed u of sound pulses is much larger than the increment of speed of the elevator during the travel time of the signal.

The first pulse is launched at time $t = 0$ and reaches the ceiling at time $t_a \cong \Delta r/u$.

The next pulse is launched at time $t = 1/\nu_1$ and reaches the ceiling at time $t_b \cong 1/\nu_1 + \Delta r/(u' - g/\nu_1)$, where g/ν_1 is the speed of the elevator at launch and u' is the pulse speed relative to our original reference frame.

According to the relativistic combination law for velocities, $u' = (u + g/\nu_1)/(1 + ug/\nu_1 c^2)$.

Ignoring terms of order g^2, we then find

$$t_b = \frac{1}{\nu_1} + \frac{\Delta r}{u} + \frac{g\Delta r}{\nu_1 c^2} \tag{4.24}$$

The frequency received at 2 is therefore given by

$$\frac{1}{\nu_2} = t_b - t_a = \frac{1}{\nu_1} + \frac{g\Delta r}{\nu_1 c^2} \tag{4.25}$$

and the fractional redshift is $(\nu_2 - \nu_1)/\nu_2 = -g\ \Delta r/c^2$, in agreement with the usual result (4.23).

The preceding calculation establishes an important point about Einstein's derivation of the gravitational redshift: It involves an inconsistent mix of Newtonian physics for the equivalence principle and relativistic physics for the velocity-addition law, applied either to light or to sound pulses or pellets, or any other signaling method.

Newtonian physics does not suffice to derive the gravitational redshift via the elevator argument.

Relativistic physics plays an essential role, and this means we must somehow adapt special relativity and the equivalence principle to each other.

As we will see in later chapters, this adaptation is achieved by general relativity and curved spacetime.

4.3 Deflection of light

To find out how light propagates in a gravitational field, we adopt the view that light consists of particles (photons), because this permits us to rely on our relativistic equation of motion for a particle in the gravitational field.

However, because the mass of the photon is zero, we must first rewrite the equation of motion somewhat.

For this purpose, we multiply the Eq. (3.72) by mdτ:

$$d(mu_\mu) + \left(\kappa h_{\mu\alpha,\beta} m u^\alpha u^\beta d\tau - \frac{\kappa}{2} h_{\alpha\beta,\mu} m u^\alpha u^\beta d\tau\right) = 0 \tag{4.26}$$

If we substitute $p_\mu = mu_\mu$ and $dx^\beta = u^\beta d\tau$, we obtain

$$dp_\mu + \left(\kappa h_{\mu\alpha,\beta} p^\alpha - \frac{\kappa}{2} h_{\alpha\beta,\mu} p^\alpha\right) dx^\beta = 0 \tag{4.27}$$

In this equation neither *m* nor dτ appears explicitly, and hence the equation can be applied to a photon.

It tells us how the direction of propagation of the photon is changed by the gravitational field and permits us to calculate the deflection of a ray of light passing through the gravitational field.

For such a calculation we can treat the factor p^α in the second term in Eq. (4.27) as constant, because any nonconstant portion in p^α is of order κ and would make a negligible contribution of order κ^2 in this second term.

The net change in p_μ produced while the photon passes through the gravitational field is therefore given by a simple integral:

$$\Delta p_\mu = -\kappa p^\alpha \int_{-\infty}^{\infty} \left(h_{\mu\alpha,\beta} - \frac{1}{2} h_{\alpha\beta,\mu} \right) dx^\beta \tag{4.28}$$

where it is understood that the integration is performed along the approximately straight trajectory of the photon.

The first term on the right side of Eq. (4.28) can be omitted; it equals $\int h_{\mu\alpha,\beta} dx^\beta = h_{\mu\alpha}(\infty) - h_{\mu\alpha}(-\infty)$, which is zero because the field $h_{\mu\alpha}$ is zero at infinite distance from the gravitating mass.

Thus,

$$\Delta p_\mu = \frac{\kappa}{2} p^\alpha \int_{-\infty}^{\infty} h_{\alpha\beta,\mu} dx^\beta \tag{4.29}$$

Consider a ray of light passing by the Sun with impact parameter b (see Fig. 4.4).

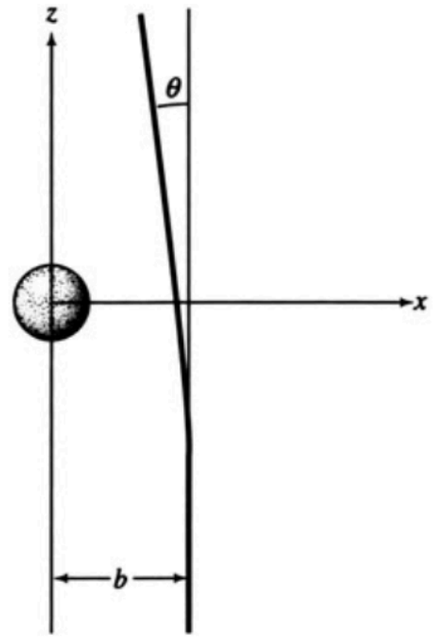


Fig. 4.4 Path of a ray of light in the gravitational field of the Sun. The deflection is exaggerated.

Taking the z -axis along the direction of incidence, we see that along this (approximately) straight path the displacement dx^β and the vector p^α are, respectively,

$$dx^\beta \cong (dt, 0, 0, dz) = (dz, 0, 0, dz) \tag{4.30}$$

and

$$p^\alpha \cong (p^3,\ 0,\ 0,\ p^3) \tag{4.31}$$

Inserting these expressions into Eq. (4.29), we obtain

$$\Delta p_1 = \frac{\kappa}{2} p^3 \int_{-\infty}^{\infty} (h_{00,1} + h_{03,1} + h_{30,1} + h_{33,1}) dz \tag{4.32}$$

For a spherically symmetric gravitational field, the terms h_{03} and h_{30} vanish, and $h_{00} = h_{33} = -2GM_\odot / \kappa r$.

Hence

$$\Delta p_1 = \frac{\kappa}{2} p^3 \int_{-\infty}^{\infty} \frac{\partial}{\partial x} (h_{00} + h_{33}) dz \tag{4.33}$$

$$= -\frac{\kappa}{2} p^3 \int_{-\infty}^{\infty} \frac{\partial}{\partial x} \frac{4GM_\odot}{\kappa \sqrt{x^2 + z^2}} \bigg|_{x=b} dz = 2GM_\odot p^3 \int_{-\infty}^{\infty} \frac{b}{(z^2 + b^2)^{3/2}} dz$$

$$= \frac{4GM_\odot}{b} p^3 \tag{4.34}$$

The deflection angle for the light ray is therefore

$$\theta \cong \frac{\Delta p_x}{p_z} = -\frac{\Delta p_1}{p^3} = -\frac{4GM_\odot}{b} \tag{4.35}$$

or, expressed in cgs units,

$$\theta = -\frac{4GM_\odot}{bc^2} \tag{4.36}$$

Note that Eq. (4.33) can be expressed in terms of the Newtonian potential $\Phi = -GM_\odot/r$ of the gravitating mass,

$$\Delta p_x = -\Delta p_1 = -2p^3 \int_{-\infty}^{\infty} \left. \frac{\partial \Phi}{\partial x} \right|_{x=b} dz \tag{4.37}$$

Written in this form, the equation is valid for the deflection produced by some general mass distribution with some general potential Φ, because any such general mass distribution can be regarded as a collection of point masses, which produce additive contributions to the potential Φ and to the deflection angle.

Accordingly, the deflection, in the *x*-direction, for a ray passing by (or through) a general mass distribution is

$$\theta = -2 \int_{-\infty}^{\infty} \frac{\partial \Phi}{\partial x} dz \tag{4.38}$$

where it is understood that the derivative is to be evaluated at the impact parameter of the ray.

If the potential has a nonzero gradient in the *y*-direction, then there will also be a deflection in the *y*-direction, with $\partial \Phi / \partial y$ in the integrand of Eq. (4.38).

This general result will be useful later in this chapter.

Figure 4.5 shows a plot of deflection angle versus the impact parameter for rays passing by the Sun.

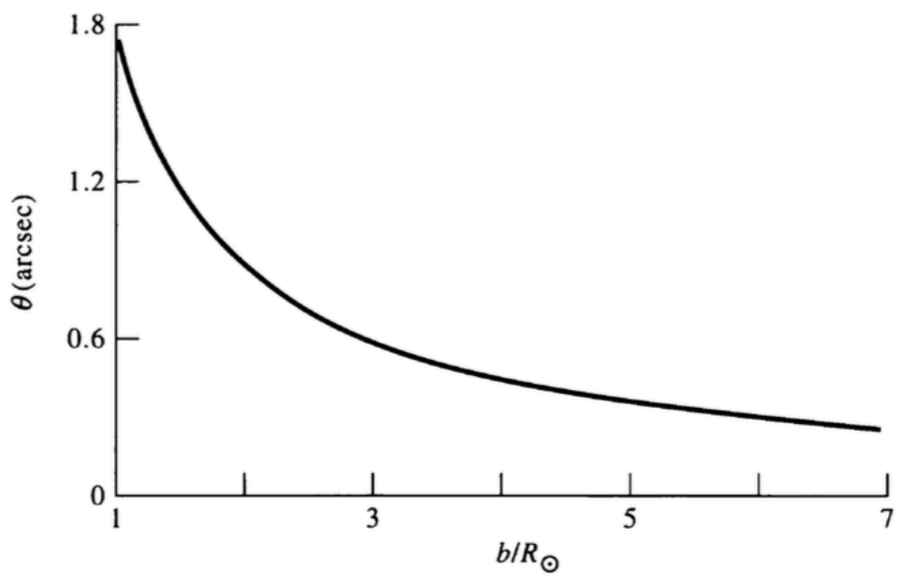


Fig. 4.5 Deflection angle vs. impact parameter.

For the case of a light ray grazing the Sun, we must take $b = R_\odot = 6.96 \times 10^{10}$ cm and $M_\odot = 1.99 \times 10^{33}$ g, which gives

$$\theta = -\frac{4GM_\odot}{R_\odot c^2} = -8.48 \times 10^{-6}\,\text{radian} = -1.75\,\text{arcsec} \tag{4.39}$$

Such a deflection of the light reaching the Earth from a star results in an (apparent) shift in the position of the star as seen from the Earth (see Fig. 4.6).

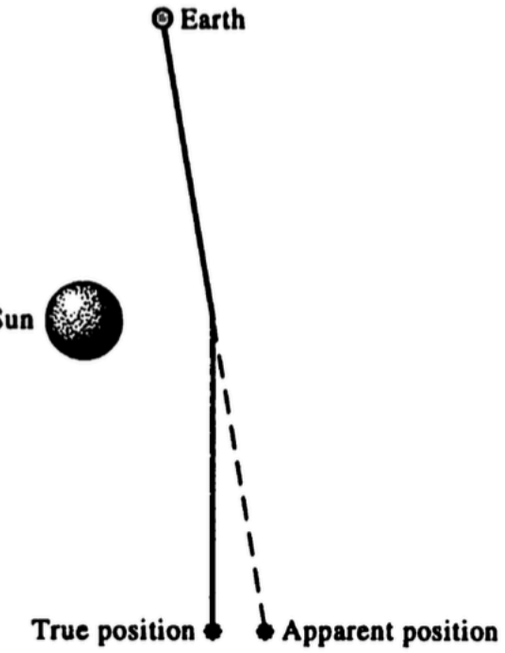


Fig. 4.6 Apparent shift in the position of the star.

It is difficult to perform optical observations of this effect.

The stars near the Sun are visible only during a total eclipse of the Sun, and even then the brightness of the solar corona restricts observations to impact parameters $b > 2R_\odot$.

The observational procedure involves taking a photograph of the star field surrounding the eclipsed Sun and comparing this photograph with another one taken at night several months before (or after) when the Sun is not in this star field.

The technical difficulties are that the eclipse photograph must be taken in the very short time interval of totality; furthermore, disturbances to the instrument must be prevented (or corrected for) during the waiting period between the eclipse photograph and the night photograph.

In a temporary field station set up in the path of the eclipse in some remote part of the world, all of this is hard to achieve.

In response to Einstein's prediction of light deflection, Eddington and Dyson organized two expeditions to observe the eclipse of May 29, 1919, at Sobral (Brazil) and on the island of Principe (in the Gulf of Guinea).

Figure 4.7 shows the instruments installed at Sobral.

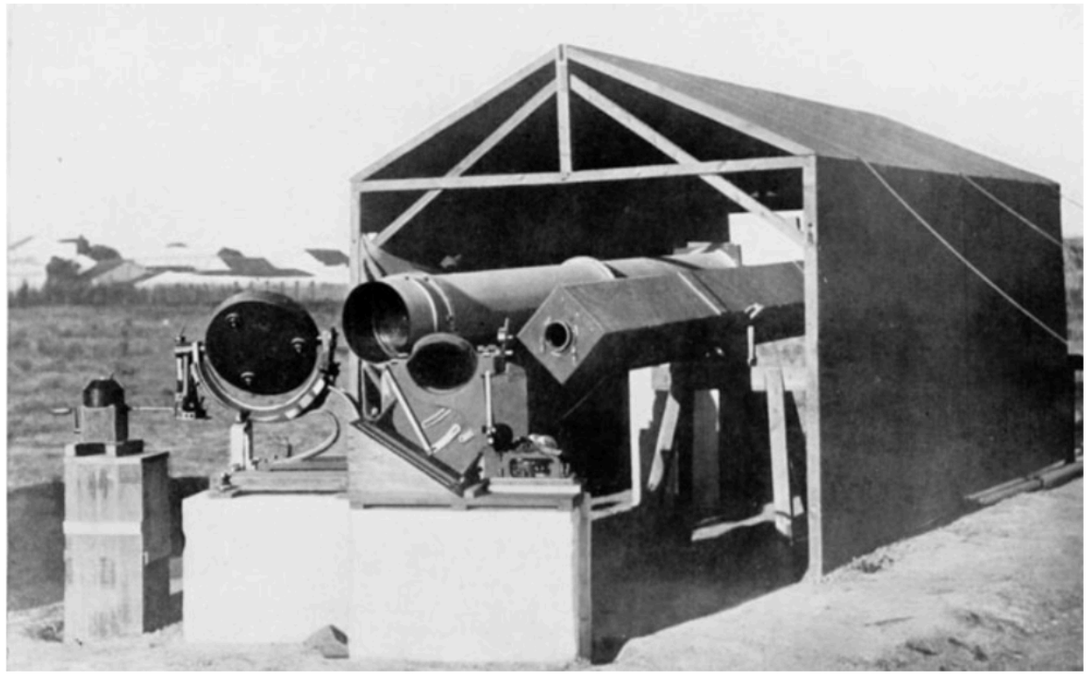


Fig. 4.7 Eclipse instruments at Sobral. The two separate telescopes are mounted horizontally and mirrors (center left) are used to throw the Sun's image into them. (From Eddington, 1920.)

The experimental results obtained by these expeditions were judged to be in reasonable agreement with the prediction, to within ±10%, but the analysis of the photographs was somewhat arbitrary and questionable.

The announcements of the results created a sensation and made Einstein a worldwide celebrity.

Since then, observations have been carried out during a dozen other eclipses.

Unfortunately, all the eclipse observations are afflicted with large experimental errors, and despite the best efforts of astronomers over a span of a half-century, the accuracy of the observations has barely improved over that of the first observations.

All we can say is that the observed deflection is approximately 2 arcsec, but a precise test of the predicted deflection is not possible by optical methods.

More precise results have been obtained by the use of radio waves.

In this case, it is not necessary to wait for an eclipse; rather one must wait for the Sun's limb to approach some radio source or group of radio sources in the sky.

For instance, the quasistellar source 3C279 is occulted by the Sun on October 8 of each year, and radio interferometers, consisting of two or more antennas, can be used to measure the apparent shift of the position of this source.

The shift is measured relative to the quasistellar sources 3C273 and 3C48, which are about 10° farther away and therefore suffer a much smaller shift.

The deflection of the radio waves must be corrected for refraction by the plasma in the solar corona.

This is accomplished by measuring the deflections at two or more different radio frequencies; the slight differences between the deflections at different frequencies permit a calculation of the plasma density.

The first results on the deflection of radio waves were obtained in the 1970s, and they confirmed the theoretical prediction to within about 1%.

More precise results were obtained later, by very long baseline interferometry (VLBI), which combines measurements taken simultaneously with several radio telescopes separated by transcontinental or intercontinental distances.

The signals were recorded electronically and tagged with precise time markers provided by synchronized atomic clocks at each radio telescope.

The recordings were later played back in precise synchronization, so the superposition of their signals gave an interference pattern, as though the separate radio telescopes had operated as a single radio interferometer.

According to Rayleigh's criterion, the resolving power of a radio interferometer is limited by the width of the diffraction peak:

$$\Delta\theta \cong \frac{\text{wavelength}}{\text{baseline}} = \frac{\lambda}{d} \tag{4.40}$$

With an intercontinental baseline of ≈ 10,000 km and a wavelength of 3 cm, this theoretical diffraction limit is $\Delta\theta \cong 0.006$ arcsec.

The uncertainty of ±0.0001 arcsec in the best available result for the deflection angle is actually below this diffraction limit because radio astronomers can plot the shape of the diffraction peak and determine the location of its center with an accuracy that is not limited by its width.

The observational results are listed in Table 4.2; they are expressed in terms of a ratio of the observed deflection and the predicted deflection.

Table 4.2 Some tests of the deflection of radio waves

Radio Telescope	Observer(s)	Baseline	$\theta_{exp}/\theta_{theory}$
Owens Valley	Seielstadt et al. (1970)	1.07 km	1.01 ± 0.11
National RAO	Sramek (1971)	2.7	0.9 ± 0.05
National RAO	Fomalont and Sramek (1976)	35	1.007 ± 0.009
VLBI	Robertson et al. (1991)	≈10,000	1.000 ± 0.002
VLBI	Lebach et al. (1995)	4,000	0.9998 ± 0.0008
VLBI	Shapiro et al. (2004)	≈10,000	0.9999 ± 0.0002

4.4 Time delay of light

Another observable effect that we can obtain from the linear approximation is the time delay suffered by a radar signal sent from the Earth to a planet and reflected back to Earth.

As was first recognized by Shapiro, the gravitational field of the Sun contributes a measurable increment to this time delay, because it reduces the speed of propagation of light signals.

The speed of propagation can be deduced from Eq. (4.27). Consider a light signal that moves along a straight line connecting the Earth and another planet (see Fig. 4.8; we take the z-axis parallel to this line).

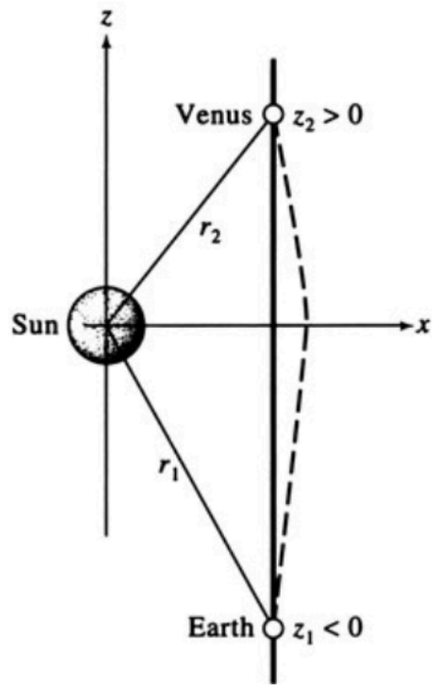


Fig. 4.8 Path of a light ray between the Earth (at z_1) and a planet (at z_2). The true path is indicated by the dashed line. The straight line is an approximation.

Of course, because of the deflection of the path of light, the true path will not be exactly straight.

But this deviation has next to no effect on the travel time because the difference in length between the straight and curved paths shown in Fig. 4.8 is only of order θ^2 (where θ is the deflection angle); this difference is therefore a second-order correction, which we can ignore.

Along the straight line, the change in p_3 given by Eqs. (4.27), (4.30), and (4.31) is

$$dp_3 = -\kappa h_{33,3} p^3 dz + \frac{\kappa}{2} h_{33,3} p^3 dz + \frac{\kappa}{2} h_{00,3} p^3 dz \tag{4.41}$$

With $h_{00} = h_{33} = -2GM_\odot/$ r, the terms on the right side of this equation cancel; thus $dp_3 = 0$, and p_3 is constant.

The change in p_0 is

$$dp_0 = -\kappa h_{00,3} p^0 dz \cong -\kappa h_{00,3} p^3 dz \tag{4.42}$$

which leads to

$$p_0 = \text{constant} - \kappa h_{00} p^3 = (1 - \kappa h_{00}) p^3 \tag{4.43}$$

From $p^\mu = m dx^\mu/d\tau$, we find $p^\mu/p^\nu = dx^\mu/dx^\nu$, and this remains valid in the limit of zero mass.

Accordingly, in terms of the coordinates z and t, the velocity of the photon can be expressed as

$$\frac{dz}{dt} = \frac{dx^3}{dx^0} = \frac{p^3}{p_0} = \frac{p^3}{(1 - \kappa h_{00}) p^3} \cong (1 + \kappa h_{00}) \tag{4.44}$$

or as

$$\frac{dz}{dt} = 1 - \frac{2GM_\odot}{r} \tag{4.45}$$

Note that this velocity is independent of the frequency; hence there is no dispersion, and the group velocity (signal velocity) coincides with the phase velocity.

The result (4.45) shows that the gravitational field of the Sun decreases the speed of propagation of light.

In fact, the deflection of light, which we calculated earlier, can be regarded as a refraction of the wavefront: The part of the wavefront nearer the Sun is slowed down more than the part farther away, and hence the direction of advance of the wave is deflected toward the Sun.

Figure 4.8 shows the Earth (at $z_1 < 0$) and the target planet (at $z_2 > 0$) in the gravitational field of the Sun.

The travel time for a light or radio signal from z_1 to z_2 is

$$\Delta t = \int_{z_1}^{z_2} \frac{z}{dz/dt} \cong \int \left(1 + \frac{2GM_\odot}{r}\right) dz = \int \left(1 + \frac{2GM_\odot}{\sqrt{z^2 + b^2}}\right) dz$$
$$= z_2 + |z_1| + 2GM_\odot \ln \frac{\sqrt{{z_2}^2 + b^2} + z_2}{\sqrt{{z_1}^2 + b^2} - |z_1|} \tag{4.46}$$

The logarithmic term on the right side of Eq. (4.46) represents the extra time delay produced by the gravitational field.

The extra delay is large when the Earth and planet are on nearly opposite sides of the Sun ("superior conjunction").

Then b is small and the extra time delay, in cgs units, becomes approximately

$$\Delta t = \frac{2GM_\odot}{c^3} \ln \frac{4\,|z_1 z_2|}{b^2} \tag{4.47}$$

Clearly, b cannot be smaller than the radius of the Sun (grazing light ray).

If we set $b = R_\odot = 6.96\times10^{10}$ cm, $|z_1|$ = radius of Earth orbit $\cong 14.9\times10^{12}$ cm, and z_2 = radius of Mercury orbit = 5.8×10^2 cm, we find from Eq. (4.47) that the maximum possible extra time delay is $\cong 1.1 \times 10^{-4}$ s, and the extra delay for the signal to go to Mercury and return is twice as large.

Before we can compare this with experiment, we must make a correction because the clocks on the Earth used to measure the time delay give the proper time rather than the t time that appears in Eq. (4.47).

According to Section 4.2, the connection between proper time and t time for a clock at rest in the gravitational field of the Sun is

$$\Delta\tau = \sqrt{g_{00}}\Delta t = \sqrt{1+\kappa h_{00}}\Delta t \cong \left(1 - \frac{GM_\odot}{r}\right)\Delta t \tag{4.48}$$

To convert the expression (4.47) to proper time at the Earth, we must therefore multiply it by $1 - GM_\odot/r$, where r is the distance between Earth and Sun.

Because the Earth is moving, the time-dilation effect of special relativity must also be taken into account.

If the orbit of the Earth were circular, these corrections would merely add a constant, uninteresting term to the time delay; but because the orbit of the Earth is elliptical, with a variable distance and speed, these corrections add a variable term to the time delay, which alters the measured time delay given by Eq. (4.46) by as much as 10% for the case of Mercury.

Further (nongravitational) corrections must be made for the change in the velocity of propagation of the radio signal caused by the solar corona and interplanetary plasma, and also for the distortion of the return signal caused by topographic features on the reflecting surface of the target planet.

The delays of radar echoes from Mercury and Venus were measured by Shapiro et al., using the Haystack and Arecibo radio telescopes.

To separate the extra time delay caused by gravitation from the total measured delay, it is necessary to know the position of the reflecting surface very precisely.

This means that both the position and the size of the reflecting planet must be known with an accuracy that is beyond what can be achieved by ordinary optical astronomy.

This difficulty was resolved by performing a large number of time-delay measurements over a period of several years.

The orbital elements of the planets, their sizes, topographical features, masses, and finally the gravitational delay were then obtained from these data.

The curve in Fig. 4.9 is a plot of the predicted extra time delay for Venus produced by gravitational effects as a function of time.

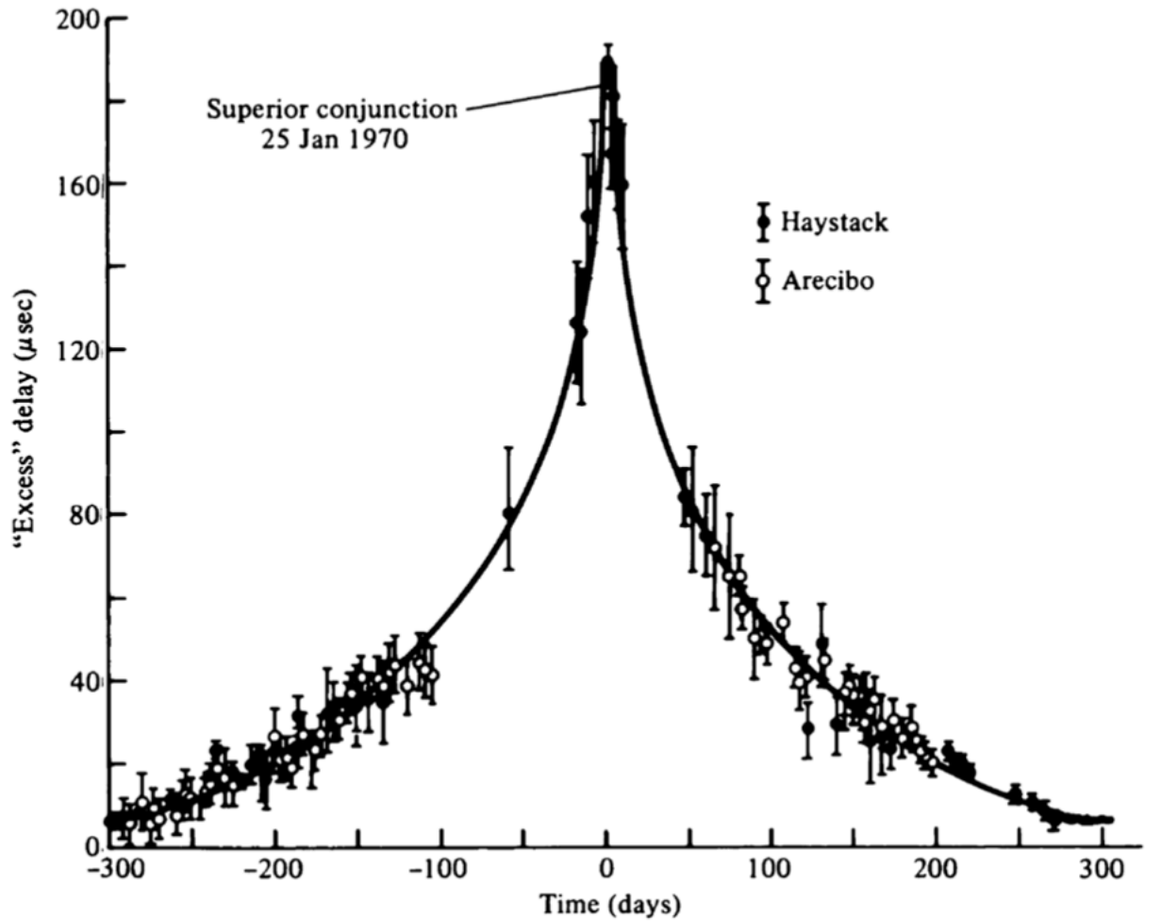


Fig. 4.9 Results of Earth-Venus time-delay measurements. The solid curve gives the theoretical prediction. (From Shapiro et al., 1971.)

The data points obviously fit this curve very well – the theoretical and experimental values agree to within 2%.

Time delays were also measured by Anderson et al. for signals transmitted to the Mariner 6 and 7 spacecraft in orbit around the Sun and retransmitted to the Earth by a transponder on the spacecraft.

The experimental delays were found to be in excellent agreement with the theoretical values.

Unfortunately, these spacecraft are subject to unknown random forces produced by outgassing of the attitude-control systems and by fluctuations in the solar radiation pressure.

These unknown perturbations introduce uncertainties in the positions and consequent uncertainties in the delays; a realistic estimate of the total uncertainty is 3%.

Results of higher precision were obtained with signals transmitted to the Viking orbiters and landers in orbit around Mars and on the ground on Mars.

Under these conditions, the spacecraft are "anchored" to Mars, and the uncertainties in their positions are minimized.

Analysis of the Viking data has confirmed the theoretical value of the time delay to within 0.1%.

The best result was obtained in 2002 with signals transmitted to the Cassini spacecraft when it passed behind the Sun while on its voyage to Saturn.

Instead of measuring the round-trip time delay for a radio pulse, Bertotti et al. measured the Doppler shift for a round trip of a continuous radio wave.

The fractional Doppler shift equals the rate of change of the round-trip travel time of the wave, for which Eq. (4.47) gives

$$-\frac{\Delta\nu}{\nu} = \frac{d(2\Delta t)}{dt} = -4\frac{GM_\odot}{c^3 b}\frac{db}{dt} \tag{4.49}$$

The factor 1/b in this expression leads to a strong peak in the Doppler shift when b approaches zero at conjunction (besides this Doppler shift generated by the rate of change of b, there is of course also a Doppler shift associated with the outward radial speeds dz_1/dt and dz_2/dt of the spacecraft, but these do not have a 1/b dependence and remain nearly constant and small during conjunction).

Thus, the measurement of the Doppler shift serves as a test of the time delay, and the Cassini data confirmed the predicted time delay to 2 parts in 10^5.

Table 4.3 summarizes the results of time-delay measurements.

Table 4.3 Some tests of the time delay of light

Target	Observer(s)	$\Delta\tau_{exp}/\Delta\tau_{theory}$
Mercury, Venus	Shapiro et al. (1968, 1971)	1.02 ± 0.05
Mariner 6, 7	Anderson et al. (1975)	1.00 ± 0.03
Mariner 9 (orbiting Mars)	Reasenberg and Shapiro (1976)	1.00 ± 0.02
Viking (on Mars)	Reasenberg et al. (1979)	1.000 ± 0.001
Cassini spacecraft	Bertotti et al. (2003)	1.00002 ± 0.00002
PSR J0737−3039	Kramer et al. (2006)	1.0000 ± 0.0005

Time delays have also been measured in binary pulsar systems, where a pulsar orbits around a star or another pulsar.

Pulsars are neutron stars rotating at high speeds, typically ranging from a few revolutions per second to hundreds of revolutions per second.

They have a strong magnetic field whose axis is inclined relative to the rotation axis, and they emit radio waves along the magnetic axis.

The regularly repeating radio pulses observed at Earth are due to a light-house effect when the rotating radio beam sweeps across our line of sight.

The radio pulses serve as a highly accurate clock, which permits the detection of the delay in the travel time of the radio beam if the pulsar comes near superior conjunction with its companion.

The most precise such pulsar time delay has been measured in the binary system PSR J0737−3039, which consists of two pulsars whose orbital plane is almost tangent to our line of sight.

The measured time delay agrees with the theoretical prediction to 5 parts in 10^4.

Although this result is somewhat less precise than for the Cassini time delay, it is nevertheless exceptionally interesting because it involves propagation of radio waves in the strong gravitational fields of the two pulsars, so the calculation of the predicted delay requires the nonlinear theory of general relativity (in contrast, all the other time delays listed in Table 4.3 involve weak gravitational fields for which the linear theory suffices).

In the calculation leading to the velocity formula (4.45) we adopted the view that light is made of particles.

Of course, we can also adopt the view that light is a wave, and we expect that the wave speed matches the particle speed.

In this context, Eq. (4.43) poses somewhat of a puzzle, because the familiar proportionality $\hbar\,\omega = p_0$ between frequency and energy of a photon suggests that the frequency of the wave varies during propagation.

But such a variation of the frequency of the wave is in conflict with a general property of wave propagation in a static medium – in such a medium, the wavelength of the wave varies, but the frequency always remains constant.

The answer to this puzzle is that the familiar proportionality of ω and p_0 fails.

The frequency ω of the wave is not proportional to the "kinetic" energy p_0, but rather to the "canonical" energy P_0,

$$\hbar\omega = P_0 = p_0 + \kappa h_{0\alpha} p^{\alpha} = (1 + \kappa h_{00}) p_0 \tag{4.50}$$

(More generally, the wave vector k_μ is proportional to the canonical energy-momentum P_μ.)

Equation (4.50) simply means that the frequency is proportional to the *total* energy of the photon in the gravitational field.

According to Eq. (4.43), $p_0 = (1 - \kappa\, h_{00})p^3$, and therefore to within terms of order κ^2,

$$\hbar\omega = (1+\kappa h_{00})p_0 = (1+\kappa h_{00})(1-\kappa h_{00})p^3 \cong p^3 = \text{constant} \tag{4.51}$$

So ω is indeed constant as the wave propagates through the gravitational field, that is, the light wave propagates without change in frequency.

This constant frequency is consistent with what we learned in Section 4.2: The redshift of light is due to the time dilation of clocks when placed in a gravitational field, whereas the frequency of light, in *t*-time, remains constant.

Note that we could have obtained the velocity formula (4.45) directly from the curved spacetime interval (4.12) by imposing the usual condition that the propagation of light corresponds to a lightlike interval, so $ds^2 = g_{\mu\nu}dx^\mu dx^\nu = 0$.

Equation (4.12) then yields

$$\left(1-\frac{2GM_\odot}{r}\right)dt^2 - \left(1-\frac{2GM_\odot}{r}\right)(dx^2+dy^2+dz^2) = 0 \tag{4.52}$$

and the speed of propagation of the light signal therefore is

$$v = \frac{\sqrt{dx^2+dy^2+dz^2}}{dt} = \sqrt{\frac{1-2GM_\odot/r}{1+2GM_\odot/r}} \cong 1-\frac{2GM_\odot}{r} \tag{4.53}$$

This agrees with Eq. (4.45).

A comment on just what is meant by the "slowing down" of the speed of propagation of light may be helpful.

Although a light signal that passes close to the Sun suffers a delay, this does not mean that an observer who measures the speed of the signal at a point near the Sun will obtain a result different from the usual speed of light.

Observers who use "good" meter sticks and "good" clocks will always find a speed $c = 3 \times 10^{10}$cm/s.

At a point in the gravitational field of the Sun, the time measured by a clock at rest is $d\tau = (1-2GM_\odot/r)^{1/2}dt$, and the distance measured by a meter stick is $dl = (1-2GM_\odot/r)^{1/2}(dx^2+dy^2+dz^2)^{1/2}$; therefore Eq. (4.52) merely says $d\tau^2 - dl^2 = 0$, which means that the locally measured speed of light is $dl/d\tau = 1$, or 3×10^{10} cm/s in cgs units.

The slowed speed given by Eq. (4.53) is not a locally measured speed, but an effective speed measured with standards of time and length that are far away, outside of the gravitational field.

An idealized situation is shown in Fig. 4.10.

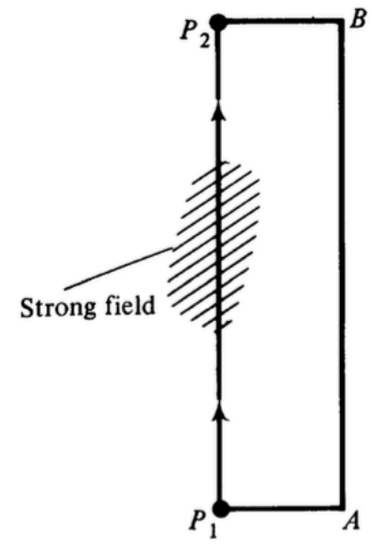


Fig. 4.10 Light path and survey used for the measurement of the effective speed.

We have two points P_1, P_2 connected by a light signal.

These points P_1, P_2 are outside of the gravitational field (alternatively, the field is so weak at these points that it can be neglected for our purposes), but the light ray connecting them passes through a region where the field is strong.

To measure the actual distance and the speed of the light signal along P_1P_2 we place our meter sticks along P_1P_2.

To establish the coordinate distance (that is, the change in y along P_1P_2), we use surveying instruments to lay out the lines P_1A, AB, BP_2, where all the angles are right angles.

All these lines are outside of the gravitational field, their angles and lengths are given by Euclidean geometry, so the changes in x, y, z along these lines are equal to changes of length.

We assign to P_1P_2 a change in the coordinates x, y, z equal to that for AB.

Furthermore, clocks at P_1, A, B, P_1 can be synchronized by the procedure familiar from special relativity.

We can then say that the effective speed along P_1P_2 is the distance measured along AB divided by the time elapsed between departure from P_1 and arrival at P_2.

It is this effective speed that is smaller than c.

The discrepancy between the speed c measured locally by meter sticks and clocks placed in the gravitational field and the smaller speed found by measurements with survey instruments and clocks placed outside of the gravitational field is clear evidence for curved spacetime – this discrepancy would not exist in flat spacetime.

Note that, by itself, the observation of a slowing of the speed of light in a gravitational field is inconclusive.

It could be explained by assigning an index of refraction to the gravitational field.

But it becomes conclusive evidence for curved spacetime when combined with the conflicting observation of an unchanged local value of the speed.

The slowdown of the speed of light given by Eq. (4.53) has a paradoxical consequence for the motion of high-speed particles in a gravitational field.

If a particle enters a gravitational field at a speed near the speed of light, say, on a radial line, then this particle must slow down as it moves downward, because the speed of light is the ultimate speed limit for particles, and if this speed limit decreases, so must the speed of the particle.

Thus, in contrast to low-speed particles, which always accelerate downward with the standard acceleration $GM_\odot/r^2$, high-speed particles *decelerate*.

This does not mean that gravity becomes repulsive at high speed.

The speed limit (4.53) is not a locally measured speed, and the deceleration of a high-speed particle is not a locally measured deceleration.

If we measure the acceleration of a fast-moving particle locally, according to its own proper time interval $d\tau$, the proper speed $|dr/d\tau|$ always increases.

But the coordinate speed $|dr/dt|$ differs from this proper speed by a factor $d\tau/dt$ that is smaller than 1 and decreasing.

For a high-speed particle, the decrease of this factor can overwhelm the increase of $|dr/d\tau|$ and lead to a decrease of $|dr/dt|$, that is, a deceleration. It can be shown that the switchover from acceleration to deceleration occurs at a critical speed $|dr/dt| = c/\sqrt{3}$.

The deceleration of high-speed particles would seem to be a violation of the principle of equivalence – high-speed particles and low-speed particles have different accelerations in a gravitational field.

However, in Chapter 6 we will see that this discrepancy is an artifact of our choice of coordinates.

We will examine other ways of introducing coordinates in curved spacetime, and we will find that the discrepancy between the accelerations of particles of different speeds actually disappears when we adopt a “better” choice of coordinates (local geodesic coordinates).

Direct experimental measurements of the speed of ultrarelativistic particles would serve as a valuable test of the relativistic theory or gravitation, but such measurements are difficult to perform with any precision.

A recent test of the speed of bunches of neutrinos propagating from the CERN laboratory at Geneva to the Gran Sasso laboratory in Italy, 730 km away, at first seemed to indicate a speed exceeding the speed of light.

But this proved to be a mistake, arising from an instrumental malfunction that caused an error in the synchronization of the clocks at the starting point and the end point.

The best available evidence concerning the speed of neutrinos in a gravitational field comes from serendipitous observations of the arrival of a bursts of neutrinos and of light from a supernova explosion in the Large Magellanic Cloud in 1987 (SN 1987A).

These bursts arrived within 3 h of each other, after a journey of about 160,000 yr, so the average speeds of these bursts agreed to within 2 parts in 10^9.

The average speed of light includes a reduction of about 1 part in 10^7 due to the gravitational field of our Galaxy.

From this we can conclude that the gravitational effects on the speeds of neutrinos and light agree to within 0.2% or better.

4.5 Gravitational lenses

When rays of light from a distant source are deflected by the gravitational field of a mass, the rays that pass on opposite sides of the mass intersect at some large distance beyond (see Fig. 4.11).

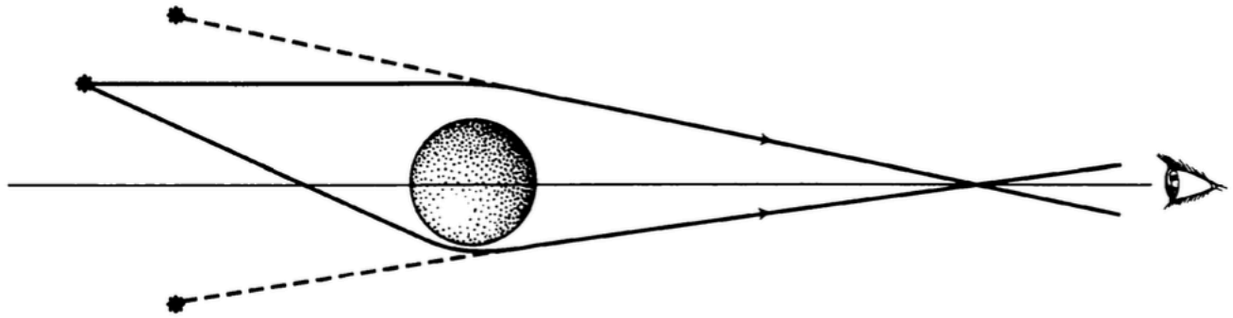


Fig. 4.11 The observer sees the source at two distinct locations.

An observer placed at such a large distance will simultaneously see the source at two locations in the sky; that is, she will see two images of the source.

For rays deflected by the Sun, the nearest point at which these rays intersect and two images become visible, is about 50 light-years.

Hence, on the Earth, we cannot see any of the multiple images produced by light deflection in the gravitational field of the Sun.

However, we can expect to see the multiple images produced by light deflection in the gravitational field of distant stars or distant galaxies.

Stars or galaxies that produce multiple images are called gravitational lenses.

Such “lenses” do not have the optical properties of genuine lenses.

The magnitude of the deflection angle of a ray passing through a genuine lens increases in direct proportion to the impact parameter [with the sign convention used in Eq. (4.36), the deflection angle is negative for a convex lens, and positive for a concave lens].

When a bundle of (almost) parallel rays from a very distant source is incident on a genuine lens, the rays either converge to a point (convex lens, real image), or they diverge so their backward extrapolations appear to come from a point (concave lens, virtual image).

But, according to Eq. (4.36), the magnitude of the deflection angle of a ray passing through the gravitational field of a star decreases with the impact parameter, and therefore such a gravitational lens has no well-defined focal length and cannot produce genuine images, real or virtual.

When a bundle of parallel rays passes through this gravitational lens, the rays diverge, but their backward extrapolation has no unique point of intersection (see Fig. 4.12), and the rays do not form a well-defined image point in the sky.

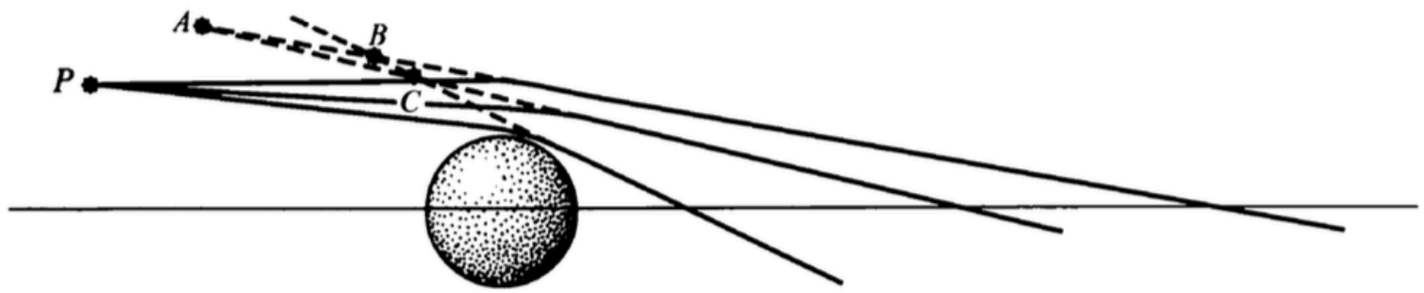


Fig. 4.12 Backward extrapolation of rays from a source *P* fails to give a unique intersection. The points *A, B, C* indicate different intersections for different pairs of rays.

Thus, the multiple "images" generated by gravitational lenses are merely multiple directions of incidence of rays on the eye or on the telescope of the observer.

The "images" are like the rainbow produced by raindrops, which also corresponds merely to a direction of incidence of light on the observer, not a genuine image in the sky.

If the observer shifts her position, the apparent position of the rainbow also shifts.

For raindrops, as for gravitational lenses, the only genuine images are those produced by the lens of the observer's eye on her retina or by a camera on a photographic film, where each of the multiple directions of incidence of rays registers as an image.

If the source of light, the deflecting star, and the observer are aligned exactly, all the rays that pass at the appropriate impact parameter around the deflecting star, at any azimuth, reach the position of the observer.

Under these exceptional circumstances, the observer sees an infinite number of images, which form a ring around the deflecting star, called the *Einstein ring*.

Figure 4.13 shows the positions of the source and the observer at distances $D1$ and $D2$, respectively, from the lens.

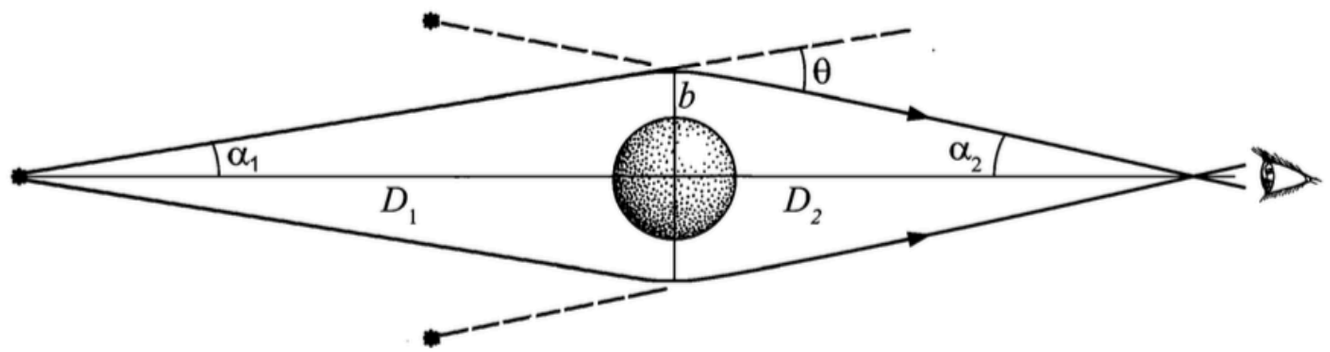


Fig. 4.13 Alignment of source, lens, and observer results in an Einstein ring.

The ray propagating from source to observer makes angles α_1 and α_2 with the centerline.

From the diagram, we see that in terms of the distances and the impact parameter b these angles are $\alpha_1 \cong b/D_1$ and $\alpha_2 \cong b/D_2$, respectively.

The sum of these two angles equals the deflection angle, Eq. (4.36),

$$\frac{b}{D_1} + \frac{b}{D_2} = \frac{4GM}{b} \tag{4.54}$$

This determines the value of the impact parameter, $b = [4GMD_1D_2/(D_1 + D_2)]^{1/2}$.

Accordingly, the angular radius of the Einstein ring is

$$\alpha_2 = \frac{b}{D_2} = \sqrt{\frac{4GMD_1}{D_2(D_1 + D_2)}} \tag{4.55}$$

For a typical star in our Galaxy, with $M \cong M_\odot$ and $D_1 \cong D_2 \cong 10^4$ light-years, the angular radius of the Einstein ring is $\alpha_2 \cong 5 \times 10^{-9}$ radian $\cong 10^{-3}$ arcsec; for a star in a distant galaxy, at a cosmological distance of $D_1 \cong D_2 \cong 10^{10}$ light-years, $\alpha_2 \cong 10^{-6}$ arcsec.

These angles characterize not only the size of the Einstein ring but also the angular separation between multiple images, as well as the maximum permissible deviation from alignment if several images are to be seen.

The small angular size of Einstein rings produced by individual stars makes observation extremely difficult and requires extremely precise (and rare) alignment.

An angle of 10^{-3} arcsec is beyond the limit of resolution of the existing optical telescopes.

Thus, the only observable effect of the alignment of the Earth, a deflecting star, and a pointlike source – such as another, more distant star or a quasar in another galaxy – would be an increase in the intensity that we would perceive when the light from all the azimuthal directions is deflected toward us.

But the chances for a sufficiently close alignment of a source and a foreground star are not good.

Only a few events of this kind have been observed (see the discussion of “microlensing” at the end of this section).

A much better opportunity for observation of Einstein rings and multiple images is provided by entire galaxies acting as gravitational lenses.

Light or radio waves passing outside of a galaxy are deflected in much the same way as light passing by an individual star.

In fact, the deflection produced by a galaxy can be calculated by performing a (vector) sum of the deflections produced by all the stars or mass elements in the galaxy (see the next section).

Since the mass distribution of the galaxy is not spherical, the details of the deflection of light are somewhat more complicated than for the spherically symmetric gravitational field of a single star.

However, the simple formula (4.55) gives us an order-of-magnitude estimate for the deflection.

For a typical galaxy, with $M \cong 10^{11} M_\odot$ and $D_1 \cong D_2 \cong 10^{10}$ light-years, the deflection angle is $\theta_2 \cong 1$ arcsec.

Such an angle is within the limit of resolution attainable by radio telescopes and also by the Hubble Space Telescope.

Furthermore, since an alignment to within a few arcseconds suffices for the production of multiple images, the probabilities for alignment for a galaxy lens are much better than for a star lens.

With the current estimates for the overall density of galaxies in the universe, it is expected that at least 1 galaxy in 10 is sufficiently well aligned with another background galaxy to produce multiple images of the latter.

Lensing phenomena produced by galaxies are called *macrolensing*, in contrast to such phenomena produced by individual stars, which are called *microlensing*.

When a galaxy acts as a gravitational lens, it produces deflections not only of the rays passing outside it but also of rays passing through the mass distribution of the galaxy.

Because the distances between the stars in a galaxy are much larger than the diameters of the stars, the chance of interception of a ray of light by the body of a star is negligible, and the galaxy is effectively transparent to light rays.

It is therefore possible for rays to reach the observer by several paths, passing outside or inside the galaxy.

The observer then sees multiple images of the source.
The number of images depends on the details of the mass distribution; we will investigate the optical properties of gravitational lenses with arbitrary, complicated mass distributions in the next section.

But in all cases, the number of images produced by a transparent mass distribution is an odd number.

This general theorem is most easily proved by appealing to wave optics.

When the wavefronts pass through the gravitational field, they are deformed, because of the slowing down of the speed of light in the field.

The deformed wavefront subsequently folds on itself (see Fig. 4.14).

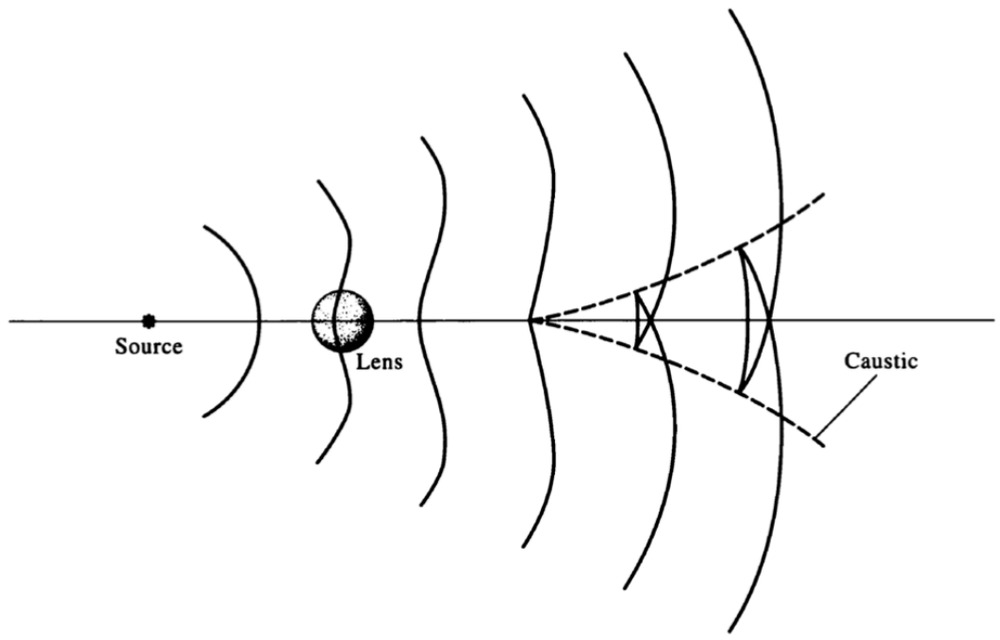


Fig. 4.14 Wavefronts deformed by the gravitational field fold upon themselves. The number of layers of any of the folded wavefronts passing over the observer is odd. The edge of the fold of the wavefront traces out a surface in space (dashed lines).

When the folded wavefront reaches the observer, it gives rise to the multiple images.

Each layer in the folded wavefront has a different direction of propagation (different wave vector), and when the folded, layered wavefront passes over the observer, each layer is perceived as arriving from a different direction, so the observer sees images in different directions.

Because the portions of the wavefront far from the gravitational field are more or less undisturbed, the number of layers in each folded wavefront is necessarily odd.

You can easily convince yourself of this by crumpling and folding the central part of a (stretchy) bedsheet while keeping the edges of the sheet more or less fixed.

The observer therefore sees an odd number of images.

Note that this theorem does not apply to the images produced by an individual star, because the star is not transparent and blocks out some rays (that is, the star cuts holes in the wavefront, and we then cannot make any self- evident assertion about the number of layers in the wavefront).

However, the theorem applies to images produced by galaxies, except perhaps a galaxy with heavy dust clouds or a dense galactic nucleus that might cut holes into the wavefront.

The edge of the fold of the traveling wavefront traces out a singular surface in space, called the *caustic surface* (see Fig. 4.14).

The layers of the wavefront are tangent to each other at the fold – that is, their rays are parallel – and they are tangent to the caustic surface.

Hence, the caustic surface is a surface of tangency of rays; that is, it is an envelope of rays.

On the caustic surface, the rays are concentrated and the intensity of light is exceptionally high.

Figure 4.15, illustrates this concentration of rays.

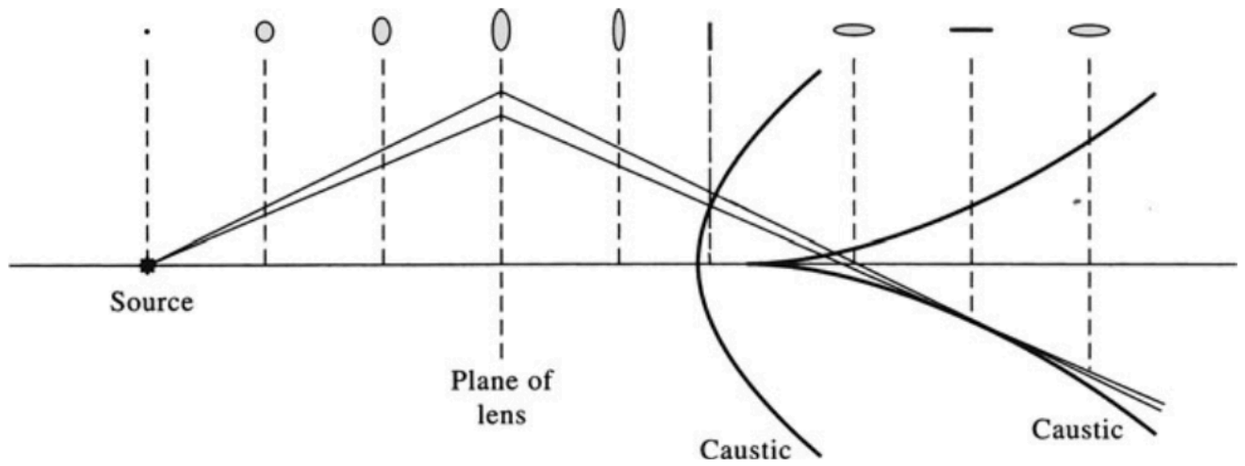


Fig. 4.15 A narrow conical bundle of rays is emitted by the source on the left. In the case illustrated here, the gravitational lens has two distinct caustic surfaces (for a more complete 3-D view of the shape of the caustic surfaces of a typical gravitational lens, see Fig. 4.19). The bundle of rays first converges tangentially, forming a radial (vertical) ribbon at the first caustic surface; it then converges radially, forming a tangential (horizontal) ribbon at the second caustic surface. The diagrams at the top indicate the cross sections of the bundle at different locations (Blandford and Narayan, 1992).

Consider a narrow conical bundle of rays emitted by the source on the left.

Initially, the cross section of this bundle is circular, but it stretches into an ellipse as the bundle propagates through the lens.

The ellipse continues to stretch until it reaches a caustic surface, where the ellipse collapses into a line segment, or a ribbon, lying flat on the surface.

Thus, at each caustic surface, the bundle of rays becomes a ribbon, tangent to the surface.

Note that at each caustic surface, the gravitational lens achieves this concentration of rays for only one of the transverse dimensions of the bundle.

In contrast, a genuine lens, at its real image point, achieves simultaneous concentration for both of the transverse dimensions of such a bundle.

The real image point formed by a genuine lens is a caustic point, which can be regarded as an extreme, degenerate form of a caustic surface, that is, a caustic surface that has contracted to a point.

The caustic surfaces of a gravitational lens separate the regions of space in which the observer sees different numbers of images; whenever the observer crosses a caustic surface, the number of images she sees increases or decreases by two.

For instance, if the observer approaches the caustic surface illustrated in Fig. 4.14 from inside, she will see that two of the three images merge, become exceptionally bright on the caustic, and then disappear.

The most favorable configuration for the observation of multiple images is the alignment of a quasar with a galaxy.

Quasars are intense, almost pointlike sources, and they are therefore likely to yield sharp, nonoverlapping multiple images.

The first such multiple image was discovered by Walsh, who noticed that the two quasars Q0957+561 A and B had identical optical and radio spectra, with the same cosmological redshifts (this redshift is a Doppler shift of the light, resulting from the motion of expansion of the universe; a large redshift means a large distance; for more about this, see Chapter 9.)

Walsh et al. proposed that these "twins" were two images of a single quasar formed by some invisible foreground galaxy acting as a gravitational lens (see Fig. 4.16a).

Subsequently, a faint galaxy was discovered in roughly the right position for gravitational lensing.

This galaxy is part of a cluster of more than 100 galaxies, all of which also contribute to the gravitational-lens effect of the "twins."

The positions and the intensities of the images can be accounted for with a plausible theoretical model for the mass distribution of the galaxies.

The theoretical model predicts that a third image should also be present.

But this third image is faint and overlaps the lensing galaxy; it is therefore not readily detectable.

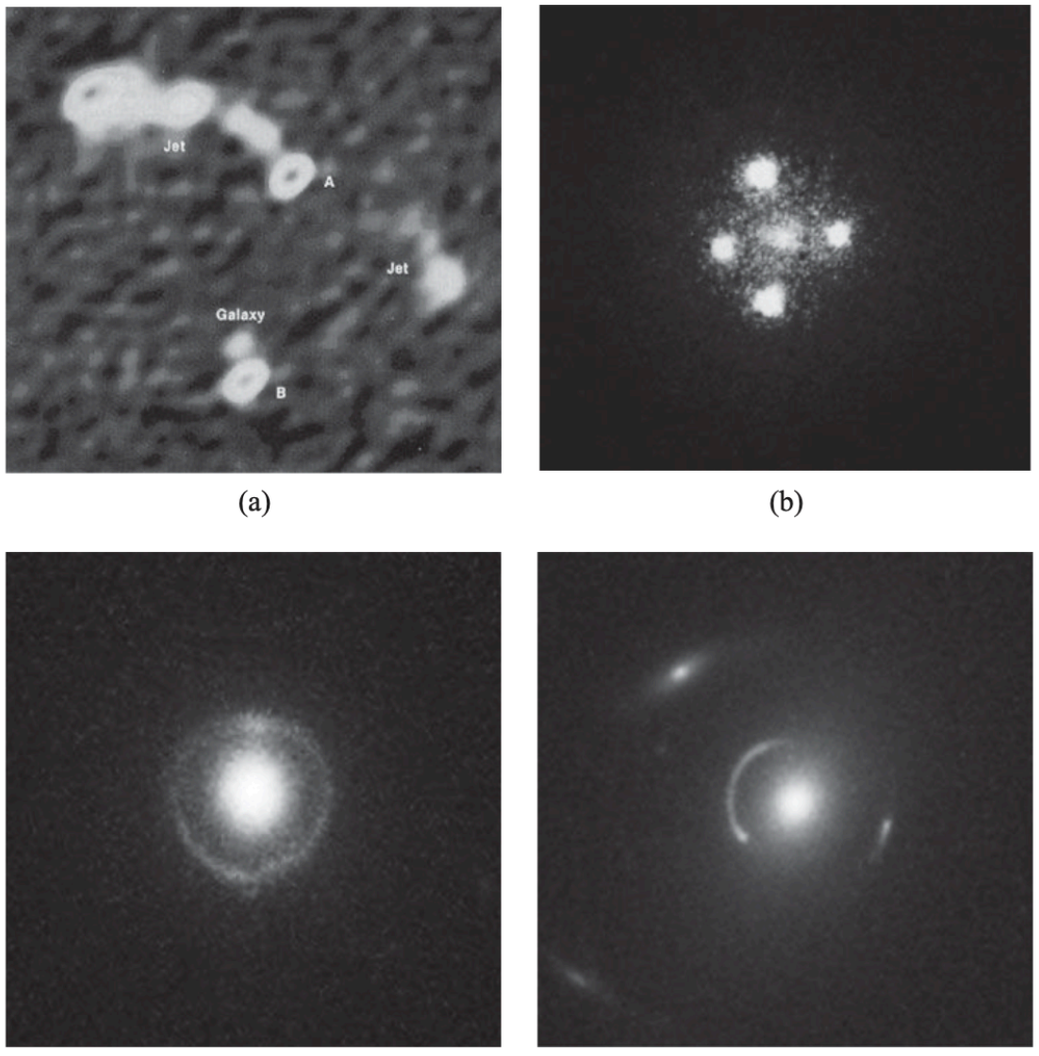
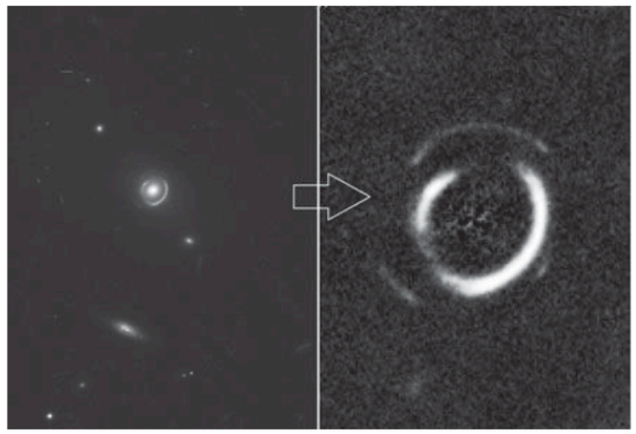
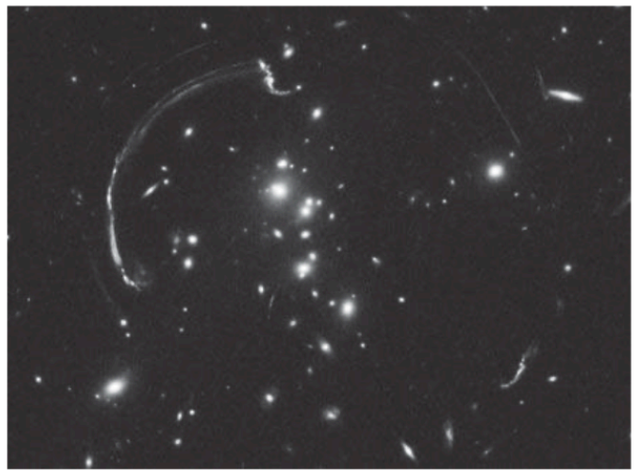


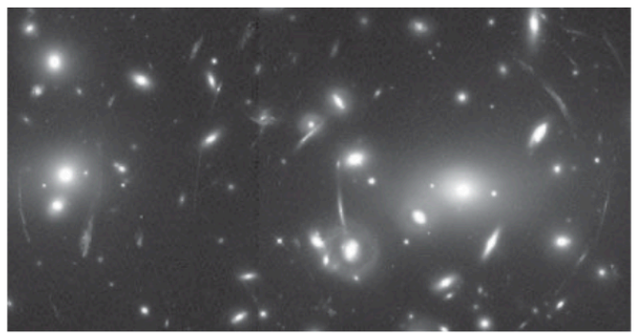
Fig. 4.16 Examples of gravitational lensing.
(a) Q0957+561 A and B, the "Twins." This picture, obtained with a radio telescope, shows two images (A, B) of a quasar. The foreground galaxy can be seen faintly just above the lower image B. The jet that extends outward from the upper image fails to produce a twin jet because it is too far from the lensing galaxy. (Credit: J. N. Hewitt, Massachusetts Institute of Technology)
(b) G2237+0305, the "Einstein Cross." This photograph made with the Hubble Space Telescope shows a quadruplet of images of a very distant quasar. These images are produced by the lensing action of the nearer galaxy seen at the center of the cross. The quasar is at a distance of 8 billion ly and the lensing galaxy is at 400 million ly. The angular separation between the upper and lower quasar images is 1.6 arcseconds. According to the general theorem for gravitational lenses, there also should be a fifth image, but this image is faint, and probably lies somewhere within the glare of the lensing galaxy. (Credit: NASA, ESA, and STScI)
(c) J162746–005357 and (d) J120540+491029. These are two examples of Einstein rings discovered by the Hubble Space Telescope. In each of these images, the central blob is a giant elliptical galaxy roughly 2 to 4 billion ly away, which acts as a gravitational lens for the light from a source galaxy about twice as far away. In (c) the source is perfectly aligned with the lens; in (d) it is slightly off-center. (Credit: NASA, ESA, and the SLACS team)



(e)



(f)



(g)

Fig. 4.16 (*continued*) (e) SDSSJ0946+1006. This shows a double Einstein ring, produced by an elliptical galaxy in the foreground (at 3 billion ly); in the enlarged photograph, the elliptical galaxy has been erased. This elliptical galaxy acts as a gravitational lens for the light from two galaxies aligned exactly behind it, at 6 billion ly and at 11 billion ly. The larger ring belongs to the more distant of these galaxies. (Credit: NASA, ESA, and the SLACS team) (f) RCS2 032727−132623. The galaxy cluster shown here acts as a gravitational lens for the light of a more distant background galaxy. The large arc at the left is a magnified and amplified (and distorted) image of the background galaxy. The galaxy cluster is at a distance of 5 billion ly, and the background galaxy is at 10 billion ly. The image is 20 times larger and more than 3 times brighter than typical images of lensed galaxies. The cluster effectively acts as a "zoom" lens – one of the most striking examples of gravitational lensing. [Credit: NASA, ESA, J. Rigby (NASA Goddard Space Flight Center), and K. Sharon (Kavli Institute for Cosmological Physics, University of Chicago).] (g) Numerous small arcs encircle the core of the rich cluster Abell 2218. These arclets are the distorted images of much more distant galaxies located behind the cluster. The abundance of lensing features can be exploited to

Many more instances of gravitational lensing of quasar images have been discovered, most as a result of systematic search programs with optical and radio telescopes.

We now know of about 200 confirmed cases of double and multiple images.

Good optical images have been obtained with the Hubble Space Telescope.

Figure 4.16 shows a selection of photographs of such multiple images and arc images (that is, partial rings) of quasars and of distant galaxies.

If the source is not a pointlike quasar, but a galaxy or a radio source with a disk of noticeable extent, the gravitational lens distorts its shape and size.

The distortion becomes severe near the caustic surface.

For instance, a disklike source will be distorted into one or several arcs, and if the alignment is exceptional, the arcs spread and merge to form an annular image, that is, an Einstein ring.

Note that in almost all of these cases of gravitational lensing, the number of observed images is even, in contradiction to the general theorem that the number should be odd.

There are several possible explanations for this discrepancy: the missing image or images might be too faint to be seen, or an image might merge with another image (when near a caustic), or an image might be blocked out by a cloud of dust in the lensing galaxy.
Long before the discovery of any gravitational lenses, Zwicky speculated on their possible applications to investigations of distant galaxies.

Today, gravitational lenses have become an important tool for the investigation of astronomical and astrophysical problems, such as the determination of the dark mass distribution in the outer reaches of galaxies, exploration of the geometry and the kinematics of our expanding universe, and the investigation of galaxies, galactic nuclei, and black holes that are too distant and too faint to be resolved with our available telescopes.

Galaxies and clusters of galaxies contain a large amount of dark mass, different from luminous stars and from ordinary clouds of dust and gas.

This dark mass is thought to be made of particles other than baryons, and it can be detected only by its gravitational effects, such as the orbital speed it gives to stars orbiting around the galaxy.

Measurements of stellar orbital speeds suggest that the amount of dark mass in galaxies and in clusters of galaxies is quite large, several times as large as the visible, luminous mass in stars.

The dark mass would contribute to the gravitational lensing generated by a galaxy or a cluster of galaxies, and this gives us a tool for investigating the distribution of dark mass.

From the observed lensing effects produced by a galaxy, we can determine the mass distribution by constructing a model that reproduces the locations, shapes, and brightnesses of the images.

A study of a large sample of galactic gravitational lenses found that the dark mass fraction gradually increases with distance from the galactic center, which is in agreement with dark-mass determinations from stellar orbital speeds.

We have similar observational evidence for dark mass from the study of the swarms of “arclets” seen when a cluster of galaxies acts as a gravitational lens.

These arclets are distorted images of background galaxies located at much larger distances behind the cluster (see Fig. 4.16g).

From the analysis of the amount of distortion as a function of radial distance from the center of the cluster, it is possible to calculate the cluster mass.

The mass deduced in this way is larger than the sum of the masses of the luminous galaxies in the cluster; thus, there must be dark mass in the cluster.

When galaxies act as gravitational lenses, the image of the source is typically magnified by a factor of 10 compared with the source itself.

This means that these gravitational lenses can be used as “telescopes” for the inspection of faint, distant sources.

Astronomers are currently collecting lensing data for the study of such sources.

An important cosmological application of the lensing phenomenon exploits time-variable sources, such as quasars or supernovas.

The multiple images produced by the lens will then vary in brightness, but with time delays between different images.

For instance, the light from the “twin” quasar Q0957+561 travels to the Earth along two different paths, of different lengths for each image; furthermore, where the paths penetrate through the gravitational field of the lens, the effective speeds of light are different.

Hence the travel times for the light arriving from the two images are different.

The observed time delay between the images A and B in the “twins” is about 540 days, with image A leading image B.

If the overall geometry of the light rays is assumed known from a theoretical model of the mass distribution of the lensing galaxies, then the difference in travel times fixes the overall scale of the distances to the lensing galaxies and to the source quasar.

The scale of distances is the essential ingredient for the calculation of the Hubble constant, that is, the ratio of the speed of recession of a galaxy or a quasar divided by its distance.

The speed of recession is easily determined from the cosmological redshift, but the distance can only be determined indirectly (see Chapter 9).

The value of the Hubble constant found from such time-delay measurements is 70 km/s per Mpc, with an uncertainty of about 5%.

The image of a compact source, such as a quasar, formed by a ray passing through a galaxy is expected to exhibit intensity fluctuations when stars moving within this galaxy pass close to the ray and produce extra, time-dependent, deflections in excess of the deflection contributed by the average, smoothed mass distribution of the galaxy.

Such microlensing events have been observed in images of distant stars and quasars, where intensity enhancements, typically lasting 10 to 100 days, were produced by a temporary alignment with a foreground star.

Microlensing is strongly suppressed if the angular size of the disk of the source exceeds the size of the Einstein ring of the microlensing star, because in that case the light merely gets rearranged within the disk, with no net loss of light from the disk.

Hence, from the observation of a fluctuation of intensity produced by a microlensing star, astronomers can determine the size of the quasar.

For instance, a microlensing observation in one of the multiple images of the Einstein Cross (Fig. 4.16b) established that this quasar source has a size no larger than 10^{15} cm, about the same as the size of the Solar System.

The large energy flux of this quasar requires that it has a large mass – at least 10^8 $M_\odot$ – and such a large mass confined to a relatively small region indicates the presence of a black hole.

Besides giving us information about the properties of the source, microlensing events also give us information about the mass of the lens.

Thus, microlensing can be exploited for the investigation of massive compact halo objects (MACHOS), such as brown dwarfs, that is, stars of very low mass that are too small to generate light by nuclear reactions.

This has led to the detection of some MACHOS, but not in sufficient abundance to solve the dark-mass problem.

Microlensing has been used with considerable success in searches for exoplanets.

If a microlensing star has a nearby planet, then this planet can become aligned with the source and add an extra enhancement of the intensity of the source; that is, a small "blip" of intensity added to the larger and longer lasting intensity enhancement generated by the microlensing star.

More than a dozen exoplanets have been discovered by such composite microlensing events.

The masses of these planets range from 1 to 10^3 Earth masses.

The advantage of this method of detection of exoplanets over other methods – such as measurement of a star's periodic motion by Doppler shifts or direct observation of a planetary transit across a star's face – is that microlensing events can be observed even for planets of small mass, or planets of large orbital radii, or planets in galaxies other than our own.

But the disadvantage is that we must monitor millions of stars to find the rare, accidental alignments of a source and a microlensing star and planet.

Besides, these microlensing events are ephemeral – the alignment of a particular microlensing planet lasts for a few hours or a few days and does not repeat again, so we cannot carry out any further investigations on the planet.

Accumulation of statistical data about exoplanet microlensing events permits us to draw statistical conclusions about the prevalence of planets in star systems.

The somewhat small sample of exoplanet events available today already indicates that, on average, every star is accompanied by at least one or two planets.

4.6 Optics of gravitational lenses

To determine the optical properties of gravitational lenses we must examine the trajectories of rays that start at a source and pass by or through some mass distribution placed between the source and the observer.

In optics, the determination of the trajectory of a ray is called ray tracing; its aim is to find the direction of arrival of the final ray at the position of the observer.

Suppose that the deflecting mass distribution is located at $z = 0$, the source at $z = D_1$, and the observer at $z = -D_2$ (see Fig. 4.17).

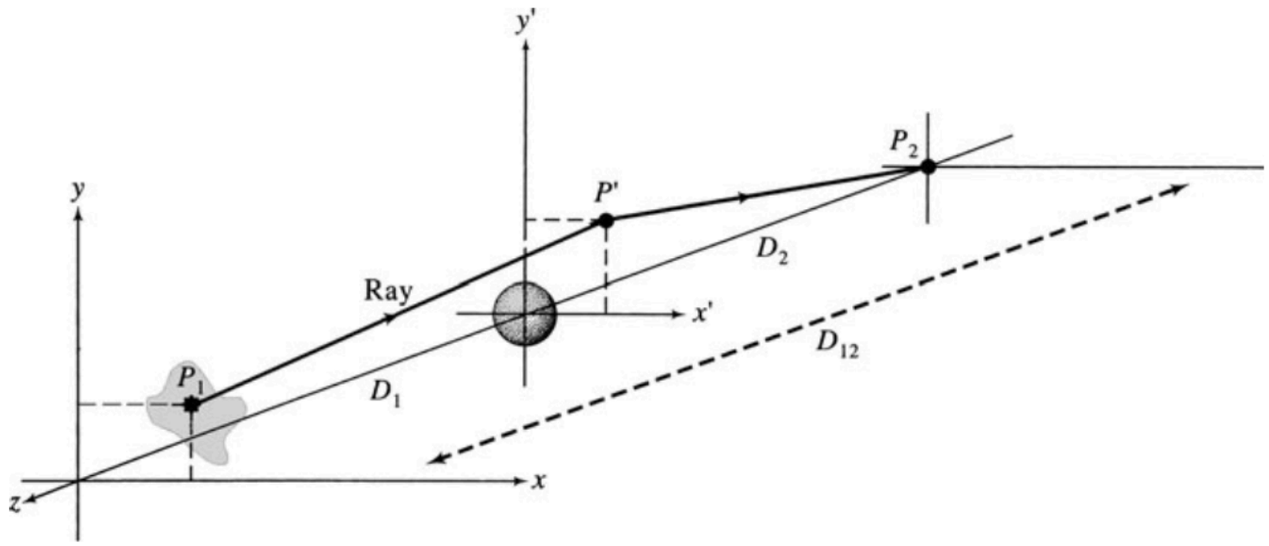


Fig. 4.17 Source, deflector, and observer with z-coordinates $z = D_1$, $z = 0$, and $z = -D_2$, respectively. The observer is at P_2, on the z-axis. The ray starts at some point P_1 within the source, with coordinates $x \ll D_1$, $y \ll D_1$.

The transverse coordinates of the ray are x, y in the plane of the source and x', y' in the plane of the deflecting mass.

The ray suffers simultaneous deflections α_x and α_y in the x and the y directions.

These deflections can be expressed in terms of the gradients of the Newtonian potential, as in Eq. (4.38):

$$\alpha_x = -2\int_{-\infty}^{\infty} \frac{\partial \Phi}{\partial x'} dz', \quad \alpha_y = -2\int_{-\infty}^{\infty} \frac{\partial \Phi}{\partial y'} dz' \tag{4.56}$$

For convenience, we combine these two components into a complex angle:

$$\alpha = \alpha_x + i\alpha_y \tag{4.57}$$

If we then define a complex scattering function I(x′,y′) as

$$I(x', y') \equiv \mathrm{Re}(I) + i\mathrm{Im}(I) = \int_{-\infty}^{\infty} \left(\frac{\partial \Phi}{\partial x'} - i\frac{\partial \Phi}{\partial y'} \right) dz' \tag{4.58}$$

we obtain a simple formula for the complex angle α:

$$\alpha = -2[\mathrm{Re}(I) - i\mathrm{Im}(I)] = -2I^* \tag{4.59}$$

The scattering function I(x′,y′) can be expressed as an integral over the density of the deflecting mass. From the familiar formula for the Newtonian potential,

$$\Phi(\mathbf{x}') = -G \int \frac{\rho(\xi, \zeta, \chi)}{[(x' - \xi)^2 + (y' - \zeta)^2 + (z' - \chi)^2]^{1/2}} d\xi d\zeta d\chi \tag{4.60}$$

we obtain

$$\frac{\partial \Phi}{\partial x'} = G \int \frac{(x' - \xi)\rho(\xi, \zeta, \chi)}{[(x' - \xi)^2 + (y' - \zeta)^2 + (z' - \chi)^2]^{3/2}} d\xi d\zeta d\chi \tag{4.61}$$

and when we integrate this over z′, we find

$$\int_{-\infty}^{\infty} \frac{\partial \Phi}{\partial x'} dz' = 2G \int \frac{(x' - \xi)\rho(\xi, \zeta, \chi)}{(x' - \xi)^2 + (y' - \zeta)^2} d\xi d\zeta d\chi$$

Furthermore, when we integrate the mass density ϱ(ξ,ζ,χ) over χ, we obtain the projected surface mass density σ (ξ , ζ) in the plane of the deflecting mass:

$$\int \rho(\xi, \zeta, \chi) d\chi = \sigma(\xi, \zeta)$$

Therefore

$$\int_{-\infty}^{\infty} \frac{\partial \Phi}{\partial x'} dz' = 2G \int \frac{(x' - \xi)\sigma(\xi, \zeta)}{(x' - \xi)^2 + (y' - \zeta)^2} d\xi d\zeta$$

With a similar expression for $\int (\partial \Phi / \partial y') dz'$, we arrive at the following result for the scattering function I:

$$I(x', y') = 2G \int \frac{(x' - \xi) - i(y' - \zeta)}{(x' - \xi)^2 + (y' - \zeta)^2} \sigma(\xi, \zeta) d\xi d\zeta \tag{4.62}$$

$$= 2G \int \frac{\sigma(\xi, \zeta)}{(x' - \xi) + i(y' - \zeta)} d\xi d\zeta \tag{4.63}$$

For a given projected surface mass density, the evaluation of this integral is straightforward in principle, but tedious in practice.

Explicit formulas for I can be obtained for some fairly simple mass distributions with spheroidal symmetry.

Numerical integration of Eq. (4.63) is often the most convenient approach, but it is necessary to be careful about handling the real and imaginary quantities in the integration [real and imaginary parts are easily separated in Eq. (4.62)].

From the deflection angle α, we can calculate the angular displacement of the image relative to the source.

According to Fig. 4.17, the deflection angle is the difference between the angles of the lines P_1P' and P_2P':

$$\alpha_x = \frac{x - x'}{D_1} - \frac{x'}{D_2} = \frac{x}{D_1} - \frac{x' D_{12}}{D_1 D_2}$$

where $D_{12} \equiv D_1 + D_2$.

This gives

$$\frac{x'}{D_2} = \frac{x}{D_{12}} - \alpha_x \frac{D_1}{D_{12}} \tag{4.64}$$

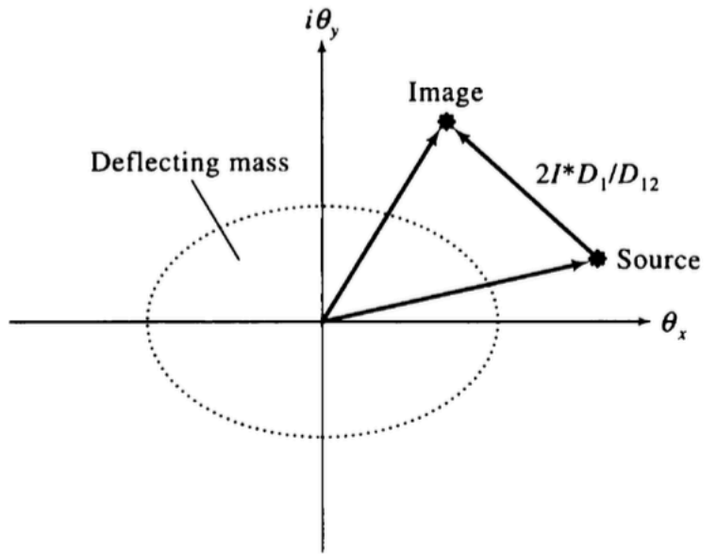


Fig. 4.18 In this diagram, the angular positions of the (undeflected) source point and the image point in the sky are represented by complex angles $\theta_x + i\,\theta_y$. The angular displacement of the image relative to the source is $-\alpha D_1/D_{12}$, or $2I^*D_1/D_{12}$.

and, likewise,

$$\frac{y'}{D_2} = \frac{y}{D_{12}} - \alpha_y \frac{D_1}{D_{12}} \tag{4.65}$$

Here x/D_{12}, y/D_{12}, and x'/D_2, y'/D_2 are the angular positions in the sky at which the observer would locate the (undeflected) source and the image, respectively.

Thus, in the sky, the complex angular displacement of the image point relative to the source point is $-\alpha D_1/D_{12}$, which, in terms of the scattering function, equals $2I^*D_1/D_{12}$.

Figure 4.18 illustrates this angular displacement.

The scattering function completely determines the direction of arrival of the final ray relative to the direction of the source.

The scattering function tells us not only the position of the image relative to the source but also the brightness of the image relative to the brightness of the source.

Suppose that when the gravitational lens is absent, the source is a small disk whose surface brightness (energy per unit solid angle and unit time as seen from the position of the observer) is uniform.

Suppose that the displacement from the center of this disk to a point at its edge is δx, δy. For small displacements, Eqs. (4.64) and (4.65) give linear relations among δx, δy and $\delta x'$, $\delta y'$:

$$\begin{aligned}\frac{\delta x'}{D_2} &= \frac{\delta x}{D_{12}} - \delta x' \frac{\partial \alpha_x}{\partial x'} \frac{D_1}{D_{12}} - \delta y' \frac{\partial \alpha_x}{\partial y'} \frac{D_1}{D_{12}} \\ &= \frac{\delta x}{D_{12}} + 2\delta x' \frac{\partial}{\partial x'} \mathrm{Re}(I) \frac{D_1}{D_{12}} + 2\delta y' \frac{\partial}{\partial y'} \mathrm{Re}(I) \frac{D_1}{D_{12}}\end{aligned} \tag{4.66}$$

$$\begin{aligned}\frac{\delta y'}{D_2} &= \frac{\delta y}{D_{12}} - \delta x' \frac{\partial \alpha_y}{\partial x'} \frac{D_1}{D_{12}} - \delta y' \frac{\partial \alpha_y}{\partial y'} \frac{D_1}{D_{12}} \\ &= \frac{\delta y}{D_{12}} - 2\delta x' \frac{\partial}{\partial x'} \mathrm{Im}(I) \frac{D_1}{D_{12}} - 2\delta y' \frac{\partial}{\partial y'} \mathrm{Im}(I) \frac{D_1}{D_{12}}\end{aligned} \tag{4.67}$$

We can solve these equations for the angular displacements $\delta x'/D_2$, $\delta y'/D_2$ of the edge of the image relative to its center.

In matrix form, the result is

$$\begin{pmatrix} \delta x'/D_2 \\ \delta y'/D_2 \end{pmatrix} = \frac{1}{{F_1}^2 - |F_2|^2} \begin{pmatrix} F_1 + \mathrm{Re}(F_2) & -\mathrm{Im}(F_2) \\ -\mathrm{Im}(F_2) & F_1 - \mathrm{Re}(F_2) \end{pmatrix} \begin{pmatrix} \delta x/D_{12} \\ \delta y/D_{12} \end{pmatrix} \tag{4.68}$$

where

$$F_1 \equiv 1 - \left(\frac{\partial I}{\partial x'} + i \frac{\partial I}{\partial y'} \right) \frac{D_1 D_2}{D_{12}}$$
$$F_2 \equiv \left(\frac{\partial I}{\partial x'} - i \frac{\partial I}{\partial y'} \right) \frac{D_1 D_2}{D_{12}} \tag{4.69}$$

The eigenvalues of the matrix appearing on the right side of Eq. (4.68) tell us by what factor the semiaxes of the image ellipse are enlarged relative to the radius of the source disk.

These eigenvalues are

$$\frac{1}{F_1 - |F_2|} \quad \text{and} \quad \frac{1}{F_1 + |F_2|} \tag{4.70}$$

The product of these eigenvalues is proportional to the solid angle subtended by the image.

Thus, the solid angle subtended by the image is enlarged relative to the solid angle subtended by the source by a factor of

$$\frac{1}{{F_1}^2 - |F_2|^2} \tag{4.71}$$

The surface brightness (energy per unit solid angle and unit time) of the image is the same as the surface brightness of the source.

Hence the enhancement (4.71) of solid angle leads to a corresponding enhancement of the brightness of the image relative to the brightness of the source.

When we observe quasars, the sources are almost pointlike, and so are the images produced by the gravitational lens.

The distortions of the shapes of the images are therefore not directly observable; however, the sizes of the images affect their relative brightnesses, which are observable.

If the denominator in Eq. (4.71) vanishes, the intensity predicted by this equation becomes infinite.

The equation $F_1^2 - |F_2|^2 = 0$ determines when the observer is on a caustic surface.

Although our simple calculation, based on geometrical optics, predicts an infinite intensity on the caustic surface, a more careful calculation, based on wave optics, establishes that the intensity on the caustic surface is limited by diffraction effects, and it remains finite.

In the calculations of this section we have assumed that the observation point is fixed on the z-axis, whereas the source point has variable coordinates x, y, and $z = D_1$.

Hence the equation $F_1^2 - |F_2|^2 = 0$ actually gives the locus of those source points that place the (fixed) observation point on a caustic.

We will call the locus of such source points the *conjugate caustic surface*.

The conjugate caustic is actually the caustic that the lens would form if we were to place a light source at the position of the observer.

This follows from optical reversibility – if the gravitational field is time independent, we can retrace the rays in Fig. 4.17 from left to right, and the rays that diverge from the position of the observer on the old caustic will then reconverge at the position of the source.

Thus, we have a reciprocal relationship: If the observer is on the caustic generated by the source, then the source is on the caustic generated by a source placed at the position of the observer (this is analogous to the reciprocal relationship between object and image points for lenses in optics: if we place a source at the image point, the lens forms an image at the object point).

If the source and the observer are symmetrically located in relation to the lens, then the conjugate caustic surface has the same shape and size as the caustic surface – the equations of the two surfaces are related by the simple exchange $z \leftrightarrow -z$.

But if the source and observer are not symmetrically located, then the shapes and sizes of the caustic and the conjugate caustic surfaces are different.

In discussions of gravitational lenses, it has become customary to examine the conjugate caustic, rather than the caustic.

Unfortunately, it has also become customary to call this conjugate caustic simply the "caustic," which can be confusing.

The shape of the conjugate caustic surface or surfaces can be calculated from the mass distribution in the lens and the equation $F_1^2 - |F_2|^2 = 0$.

In general, these surfaces are complicated, with several intersecting smooth curved pieces.

Figure 4.19 illustrates the conjugate caustic surfaces associated with an ellipsoidal mass distribution of uniform density (such a mass distribution is a simple, but crude, model for a galaxy).

For this mass distribution there are two caustic surfaces.

At the far left, the larger of these surfaces is roughly an ellipse in cross section; the smaller surface is an astroid with four spikes.

These two caustic surfaces correspond to the two ways that a bundle of rays can be focused: tangentially or radially (compare Fig. 4.15).

The result (4.71) for the enhancement of the intensity is valid only for a small (infinitesimal) source.

For an extended source, the shape of the image must be calculated by mapping the boundary of the source point by point into the image plane, by means of Eqs. (4.64) and (4.65).

If the source is at or near a conjugate caustic, the image suffers a severe distortion, and its solid angle becomes much larger than that subtended by the source.

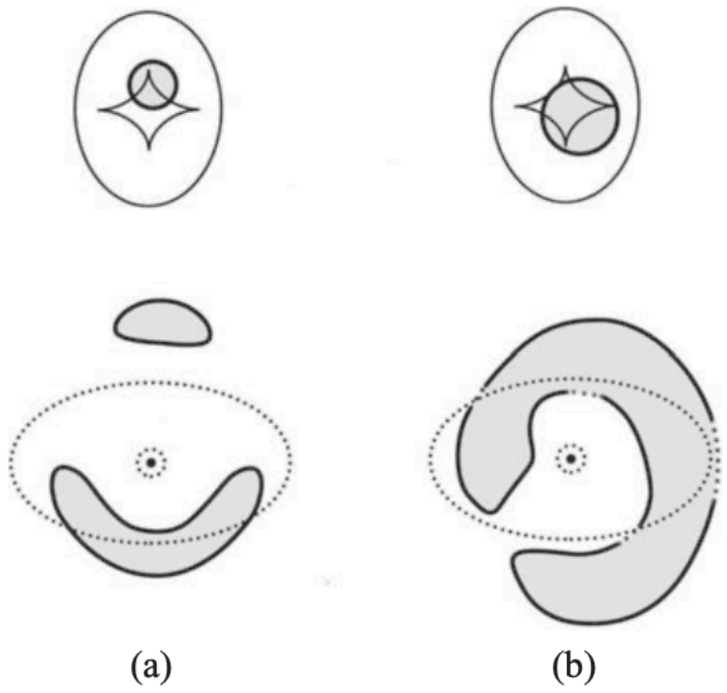


Fig. 4.20 Diagrams at the top show the position of the source disk in relation to the conjugate caustic. Diagrams at the bottom show the corresponding images. (a) If a disklike source is near the cusp of the conjugate caustic, the image is an elongated arc. (b) If the source is almost centered on the conjugate caustic, the image is an almost complete ring (Blandford and Narayan, 1992).

For example, Fig. 4.20a shows the image seen when the source, in the shape of a disk, overlaps the cusp of the conjugate caustic; the image displays one small and one large arc.

Figure 4.20b shows the image seen when the source is almost centered on the conjugate caustic; the image then displays an almost complete Einstein ring.

4.7 Field of a rotating mass; Lense-Thirring effect

The gravitational field surrounding a rotating mass differs from that surrounding a non-rotating mass.

We can understand this by analogy with the case of a rotating, uniformly charged sphere; such a sphere produces both electric and magnetic fields, whereas a nonrotating sphere produces only an electric field.

Since the rotations of the Sun and of the Earth are quite slow (nonrelativistic), the effects of the rotation on the gravitational fields and on the motion of particles in these fields are very small.

However, the Stanford Gravity Probe B experiment, a high-precision test of general relativity completed in 2010, has detected the effect of the rotation of the Earth on gyroscopes placed in a satellite orbiting the Earth.

To study this effect, we need the solution for the gravitational field surrounding a nonrelativistic rotating mass.

The general solution with relativistic rotation will be given in Section 8.4.

We assume that the rotation is steady, so the field is time independent.

It is convenient to begin with the general solution of the gravitational field equation.

In the time-independent case, the field equation (3.45) reads

$$-\nabla^2\phi^{\mu\nu}(\mathbf{x}) = -\kappa T^{\mu\nu} \tag{4.72}$$

and it has the solution

$$\phi^{\mu\nu}(\mathbf{x}) = -\frac{\kappa}{4\pi}\int \frac{T^{\mu\nu}(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}d^3x' \tag{4.73}$$

where the integration extends over the entire region in which $T^{\mu\nu}$ is nonzero.

To evaluate the integral appearing in Eq. (4.73), we use the Taylor-series expansion for $1/|\mathbf{x}-\mathbf{x}'|$ [(see Eq. (1.23)] and keep only the first two terms:

$$\frac{1}{|\mathbf{x}-\mathbf{x}'|} \cong \frac{1}{r} + \frac{x^k x'^k}{r^3} \tag{4.74}$$

This gives

$$\phi^{\mu\nu}(\mathbf{x}) = -\frac{\kappa}{4\pi r}\int T^{\mu\nu}(\mathbf{x}')d^3x' - \frac{\kappa x^k}{4\pi r^3}\int x'^k T^{\mu\nu}(\mathbf{x}')d^3x' \tag{4.75}$$

The integral in the first term of Eq. (4.75) has the values

$$\int T^{00}(\mathbf{x}')d^3x' = M \tag{4.76}$$

$$\int T^{k\mu}(\mathbf{x}')d^3x' = 0 \tag{4.77}$$

where M is the mass of the system.

Equation (4.76) is obvious and requires no comment.

To derive Eq. (4.77), we begin with the conservation law

$$\frac{\partial}{\partial x'^0} T^{0\mu}(\mathbf{x}') + \frac{\partial}{\partial x'^l} T^{l\mu}(\mathbf{x}') = 0 \tag{4.78}$$

Since the matter distribution is assumed to be time independent, this reduces to

$$\frac{\partial}{\partial x'^l} T^{l\mu}(\mathbf{x}') = 0. \tag{4.79}$$

Next, we multiply Eq. (4.79) by x'^k and integrate over the volume of the system:

$$\int x'^k \frac{\partial}{\partial x'^l} T^{l\mu}(\mathbf{x}') d^3x' = 0 \tag{4.80}$$

An integration by parts then gives Eq. (4.77).

The integral appearing in the second term in Eq. (4.75) has the values

$$\int x'^k T^{00}(\mathbf{x}') d^3x' = 0 \tag{4.81}$$

$$\int x'^k T^{l0}(\mathbf{x}') d^3x' = \frac{1}{2} \varepsilon^{kln} S^n \tag{4.82}$$

$$\int x'^k T^{ln}(\mathbf{x}') d^3x' = 0 \tag{4.83}$$

where ε^{kln} is the object defined in Eq. (1.67), and S^n is the spin angular momentum of the system.

Equation (4.81) simply states that the origin of coordinates is at the center of mass; this is an assumption that we make for the sake of simplicity.

The derivation of Eq. (4.82) proceeds as follows: We multiply Eq. (4.79) by $x'^k x'^n$ and integrate over the volume of the system:

$$\int x'^k x'^n \frac{\partial}{\partial x'^l} T^{l\mu}(\mathbf{x}') d^3x' = 0 \tag{4.84}$$

With an integration by parts, we obtain the identity

$$-\int x'^n T^{k\mu} d^3x' - \int x'^k T^{n\mu} d^3x' = 0 \tag{4.85}$$

The spin angular momentum is the integral of the cross product of the position vector and the density of momentum.

Thus, the x-component of the spin angular momentum is defined as

$$S^1 = \int (x'^2 T^{30} - x'^3 T^{20}) d^3x' \tag{4.86}$$

The other components of the spin angular momentum are defined similarly.

In general we can show

$$S^n = \int \varepsilon^{nkl} x'^k T^{l0} d^3x' \tag{4.87}$$

In view of the identity (4.85), each of the two terms in the integrand of Eq. (4.86) contributes the same amount, and therefore

$$S^1 = 2 \int x'^2 T^{30} d^3x' \tag{4.88}$$

This equation confirms the $k = 2, l = 3$ component of Eq. (4.82).

The other components can be checked in the same way.

Finally, to derive Eq. (4.83) we begin with the identity

$$T^{ln}(\mathbf{x}') = \frac{1}{2}\frac{\partial}{\partial x'^m}[x'^n T^{lm}(\mathbf{x}') + x'^l T^{nm}(\mathbf{x}')] \tag{4.89}$$

Hence, with an integration by parts

$$\begin{aligned}\int x'^k T^{ln} d^3x' &= \frac{1}{2}\int x'^k \frac{\partial}{\partial x'^m}(x'^n T^{lm} + x'^l T^{nm}) d^3x' \\ &= -\frac{1}{2}\int \delta^k_m (x'^n T^{lm} + x'^l T^{nm}) d^3x' = -\frac{1}{2}\int (x'^n T^{lk} + x'^l T^{nk}) d^3x'\end{aligned} \tag{4.90}$$

The last expression vanishes as a consequence of the identity (4.85).

If we insert the results (4.76), (4.77), and (4.81)–(4.83) into Eq. (4.75), we obtain

$$\begin{aligned}\phi^{00}(\mathbf{x}) &= -\frac{\kappa M}{4\pi r} \\ \phi^{l0}(\mathbf{x}) &= \phi^{0l}(\mathbf{x}) = -\frac{\kappa}{8\pi r^3}\varepsilon^{kln} x^k S^n \\ \phi^{ln}(\mathbf{x}) &= 0\end{aligned} \tag{4.91}$$

The corresponding expressions for $h_{\mu\nu}$ are

$$\begin{aligned}h_{00}(\mathbf{x}) &= -\frac{2GM}{\kappa r} \\ h_{l0}(\mathbf{x}) &= h_{0l}(\mathbf{x}) = \frac{2G}{\kappa r^3}\varepsilon^{kln} x^k S^n \\ h_{ln}(\mathbf{x}) &= -\frac{2GM}{\kappa r}\delta^n_l\end{aligned} \tag{4.92}$$

If the direction of the spin angular momentum coincides with the z-axis of our coordinate system, we obtain

$$h_{\mu\nu} = \begin{pmatrix} -2GM/\kappa r & -2GS_z y/\kappa r^3 & 2GS_z x/\kappa r^3 & 0 \\ -2GS_z y/\kappa r^3 & -2GM/\kappa r & 0 & 0 \\ 2GS_z x/\kappa r^3 & 0 & -2GM/\kappa r & 0 \\ 0 & 0 & 0 & -2GM/\kappa r \end{pmatrix} \tag{4.93}$$

Comparing this with Eq. (4.9), we see that the field of a rotating mass differs from that of a nonrotating mass by the presence of off-diagonal components h_{0l}.

The spins of the Sun and Earth are fairly small, and the effects of the off-diagonal components of the gravitational field on the orbits of planets and ordinary satellites are difficult to detect.

However, in the case of the Earth, artificial satellites of a special design have made it possible to detect these effects.

There are two different ways in which the off-diagonal components reveal themselves: They produce (small) precessions in the orientation of the orbital plane and the perigee of satellites orbiting the Earth, and they produce precessions in the orientation of spinning bodies (gyroscopes) orbiting the Earth.

The first of these effects can be thought of as a spin-orbit coupling and the second as a spin-spin coupling.

Both of these precessions are called *Lense-Thirring effects*.

The Lense-Thirring precession of satellite orbits can be calculated by examining the equation of motion of a particle in the gravitational field (4.93).

However, the calculations are rather complicated and here we will merely calculate the precession of the orbital plane by means of the approximate gravitomagnetic equation of motion Eq. (3.100),

$$m\frac{d\mathbf{v}}{dt} = m\mathbf{g} + m\mathbf{v} \times \mathbf{b} \tag{4.94}$$

Here **b** is the gravitomagnetic field, calculated from Eq. (4.93).

By means of Eqs. (3.97) and (3.98), we find that the gravitomagnetic field generated by the spin of the Earth is

$$\mathbf{b} = -4G\frac{3\mathbf{n}\,\mathbf{n}\cdot\mathbf{S} - \mathbf{S}}{2r^3} \tag{4.95}$$

where **n** is a unit vector in the radial direction, $\mathbf{n} \equiv \mathbf{r}/r$.

This gravitomagnetic field is analogous to the magnetic field generated by the magnetic moment **m** of the nucleus of an atom.

Such a magnetic field causes a precession of the orbital plane of an electron, and the gravitomagnetic field causes a similar precession of the orbital plane of an Earth satellite.

The orientation of the orbital plane is characterized by the orbital angular momentum **L**, which is constant if the only force acting on the satellite is the radial Newtonian gravitational force $-GMm/r^2$.

However, the gravitomagnetic force $m\mathbf{v} \times \mathbf{b}$ in Eq. (4.94) is not radial and it exerts a torque that changes the direction of the orbital angular momentum and the orientation of the orbital plane.

For the sake of simplicity, we consider a polar circular orbit, initially in the *x*-*z* plane (see Fig. 4.21), so its orbital angular momentum is initially in the *y*-direction.

For this orbit $\mathbf{r} \cdot \mathbf{v} = 0$, so the torque produced by the gravitomagnetic force is

$$\boldsymbol{\tau} = m\mathbf{r} \times (\mathbf{v} \times \mathbf{b}) = m\mathbf{v}\,\mathbf{r} \cdot \mathbf{b} \tag{4.96}$$

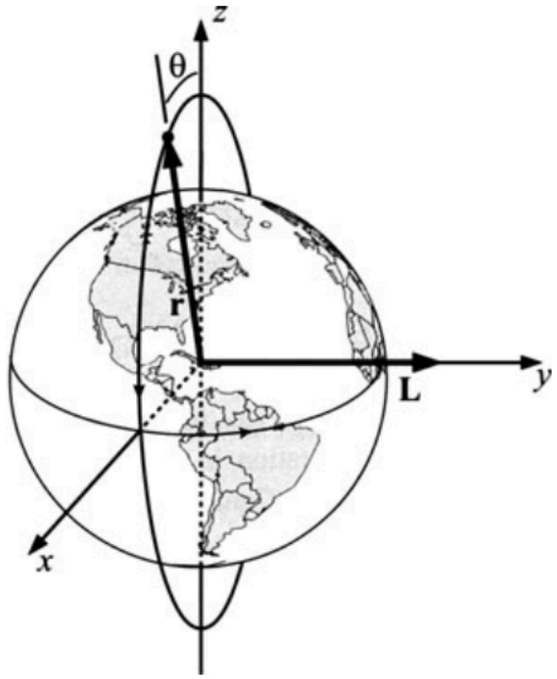


Fig. 4.21 A polar circular orbit. The orbit is initially in the *x*-*z* plane, and the orbital angular momentum vector is initially in the *y*-direction. The satellite has an instantaneous angular position θ.

With Eq. (4.95), the torque becomes

$$\boldsymbol{\tau} = m\mathbf{v}\left(-\frac{4G}{2r^3}\right)(3\mathbf{r}\cdot\mathbf{n}\,\mathbf{n}\cdot\mathbf{S} - \mathbf{r}\cdot\mathbf{S}) = -m\mathbf{v}\frac{4GS}{r^2}\cos\theta \tag{4.97}$$

The instantaneous velocity is $\mathbf{v} = v\,\hat{\mathbf{x}}\cos\theta - v\,\hat{\mathbf{z}}\sin\theta$, and if we substitute this into Eq. (4.97) and average over the circular orbit (that is, over θ), we find an average torque

$$\langle\boldsymbol{\tau}\rangle = -mv\hat{\mathbf{x}}\frac{2GS}{r^2} \tag{4.98}$$

This average torque is in the *x*-direction, which is perpendicular to **L**.

It therefore produces a change of direction of **L**, but no change of magnitude; that is, it produces a precession of the orbital angular momentum and of the orbital plane.

The rate of precession, in radians per second, is given by the ratio $(1/|\mathbf{L}|)(|d\mathbf{L}/dt|)$,

$$\frac{1}{|\mathbf{L}|}\left|\frac{d\mathbf{L}}{dt}\right| = \frac{1}{mvr}|\langle\boldsymbol{\tau}\rangle| = \frac{2GS}{r^3} \tag{4.99}$$

The direction of this precession is eastward, so we can write the precession angular velocity in vector notation (and in cgs units) as

$$\boldsymbol{\Omega}_{LT} = \frac{2G}{r^3 c^2}\mathbf{S} \tag{4.100}$$

It is easy to verify that the same formula applies to the Lense-Thirring precession of any circular orbit, of any inclination.

However, for an elliptical orbit, with eccentricity e and semimajor axis a, the formula is modified to $\boldsymbol{\Omega}_{LT} = 2G\mathbf{S}/a^3(1-e^2)^{3/2}c^2$.

Because the Lense-Thirring precession of the orbital plane is eastward, in the same direction as the rotation of the Earth, this result is often interpreted as a "frame-dragging" effect – supposedly, the rotation of the Earth drags the surrounding space along and makes it (partially) participate in the rotation of the Earth.

However, this intuitive interpretation is far too simplistic.

Examination of the precession of the perigee of an equatorial elliptical orbit reveals that the perigee has a precession of about the same magnitude as given by Eq. (4.100), but in the *westward* direction.

Furthermore, for an equatorial circular orbit, examination of the gravitomagnetic effect reveals that the orbital speed decreases if the orbital motion is eastward, and increases if the orbital motion is westward.

This contradicts the naıve notion of frame dragging.

If the polar orbit is of low altitude ($r \cong r_E$), the magnitude of the Lense-Thirring precession amounts to about 0.05 arcsec per year.

Unfortunately, satellite orbits are subject to much larger precessions due to the quadrupole moment of the Earth.

The quadrupole precession vanishes if the orbit is exactly polar, but this does not help – there are practical limitations on our ability to measure the angle of the plane of the orbit with ground-based instrumentation, and we cannot be sure that the orbit is exactly polar.

As a clever way out of this difficulty, Ciufolini proposed to use an existing satellite already in orbit (LAGEOS, or *L*aser *GEO*dynamics *S*atellite) and place a second satellite (LAGEOS II) in a similar orbit of opposite inclination (see Fig. 4.22).

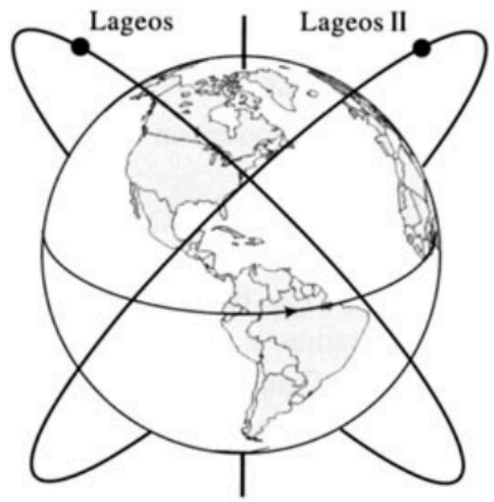


Fig. 4.22 Orbits of the LAGEOS satellite and of the LAGEOS II satellite of opposite inclination.

The quadrupole moment of the Earth then produces precessions of opposite directions for the two satellites, whereas the Lense-Thirring effect produces an eastward precession for both.

In principle, the average precession of the two satellites then measures the Lense-Thirring precession.

In practice, when LAGEOS II was launched in 1992, it was placed in an orbit of unequal inclination and shape, so the Lense-Thirring and quadrupole precession values had to be extracted from a full analysis of the measured orbital data for both satellites.

The LAGEOS satellites are covered with light reflectors, which permit very accurate laser ranging of their positions with pulses of laser light, to within a few millimeters.

Four years of data yielded a value for the Lense-Thirring precession in agreement with the theoretical prediction, with an uncertainty of ±10 %.

Although the Lense-Thirring effect was originally conceived as an orbital precession, a similar precession occurs when the gravitomagnetic field of the Earth acts on a rotating mass, that is, a gyroscope.

The spin $\mathbf{S}'$ of such a gyroscope then couples to the spin $\mathbf{S}$ of the Earth via the gravitomagnetic field.

The gyroscope experiences a torque, and if it is suspended in frictionless mountings, it will precess.

We can regard this kind of Lense-Thirring precession as analogous to the precession of a spinning particle of magnetic moment $\mathbf{m}'$ placed in the magnetic field $\mathbf{B}$ of a spinning particle of magnetic moment $\mathbf{m}$.

The torque exerted by the magnetic field $\mathbf{B}$ on the magnetic moment $\mathbf{m}'$ is $\boldsymbol{\tau} = \mathbf{m}' \times \mathbf{B}$, and this torque causes $\mathbf{m}'$ to precess, with an angular velocity proportional to the strength of the magnetic field.

Similarly, the gravitomagnetic field (4.95) of the Earth exerts a torque on the spin $\mathbf{S}'$ of a gyroscope,

$$\boldsymbol{\tau} = \frac{\mathbf{S}'}{2} \times \mathbf{b} = -G\mathbf{S}' \times \frac{3\mathbf{n}\,\mathbf{n}\cdot\mathbf{S} - \mathbf{S}}{r^3} \tag{4.101}$$

of the magnetic moment $\mathbf{m}'$ is $\mathbf{S}'/2$, rather than $\mathbf{S}'$ (the definition $\mathbf{m} = \frac{1}{2}\int \mathbf{x} \times \mathbf{j} d^3x$ in electromagnetism includes a factor of $1/2$, whereas the definition of $\mathbf{S}$, Eq. (4.87), includes no such factor).

Accordingly, the rate of change of the spin $\mathbf{S}'$ is

$$\frac{d\mathbf{S}'}{dt} = -G\mathbf{S}' \times \frac{3\mathbf{n}\,\mathbf{n}\cdot\mathbf{S} - \mathbf{S}}{r^3} \tag{4.102}$$

From this we can immediately identify the precession angular velocity (in cgs units):

$$\boldsymbol{\Omega} = \frac{G}{c^2}\frac{3\mathbf{n}\,\mathbf{n}\cdot\mathbf{S} - \mathbf{S}}{r^3} \tag{4.103}$$

For a circular or nearly circular polar orbit, the average value $\langle 3\mathbf{n}\,\mathbf{n}\cdot\mathbf{S} - \mathbf{S}\rangle$ is $^3/_2\,\mathbf{S} - \mathbf{S} = {}^1/_2\,\mathbf{S}$, and for an equatorial orbit it is $-\mathbf{S}$.

For a gyroscope placed on the surface of the Earth, the friction in the mountings is always much larger than the gravitomagnetic torque.

But the Gravity Probe B experiment, to be further discussed in Section 7.7, has detected this torque with a gyroscope placed in a satellite in orbit, where the gyroscope is weightless and requires no suspension.

However, such an orbiting gyroscope is subject to an additional precession effect, arising from its motion through the gravitational field of the Earth.

This additional precession, called the geodetic precession, or the de Sitter precession, is about 100 times larger than the Lense-Thirring precession.

These two kinds of precession can be separated experimentally because they differ in their dependences on the direction of the axis of the gyroscope.

The Gravity Probe B experiment was designed to measure both of these precession effects.

This concludes our discussion of the time-independent solutions of the linear field equations.

In the next chapter we will investigate the time-dependent fields in the radiation zone, that is, the fields at large distance from some time-dependent mass distribution.

We could also investigate the fields in the convection zone, that is, the fields in the immediate vicinity of a time-dependent mass distribution.

In fact, we could work out a set of “Lienard-Wiechert” potentials for gravitation entirely in analogy with electrodynamics.

Although the close similarity between linear gravitation and electrodynamics may tempt us to solve the same mathematical problems, we must keep in mind that there are few gravitational problems that are accessible to experiment.

We cannot manufacture measurable gravitational fields as easily as we manufacture electric and magnetic fields.

With the exception of the Cavendish experiment, the gravitational fields that we use in our experiments are generated by processes that are beyond our control.

5 - Gravitational Waves

> If you ask me whether there are gravitational waves or not,
> I must answer that I do not know.
> But it is a highly interesting problem.
>
> Albert Einstein

Gravitational effects cannot propagate with infinite speed.

This is obvious both from the lack of Lorentz invariance of infinite speed and from the causality violations that are associated with signal speeds in excess of the speed of light.

Since the speed of light is the only Lorentz-invariant speed, we expect that gravitational effects propagate in the form of waves at the speed of light.

As a concrete example, consider an apple that hangs on a tree.

At some time, the stem of the apple breaks and the apple falls to the ground, which means there is a sudden change in the terrestrial mass distribution.

The gravitational field surrounding the Earth must then adapt itself to this new mass distribution.

The change in the field will not occur simultaneously throughout the universe – at any given point of space the change will be delayed by a time equal to the time needed for a light signal to travel from the Earth to that point.

Hence the disturbance in the gravitational field propagates outward at the speed of light.

Such a propagating disturbance is a gravitational wave.

The existence of gravitational waves is an immediate consequence of special relativity, and, to some extent, the experimental discovery of gravitational waves would merely confirm the obvious.

Although the existence of waves is ensured by general arguments, the strength and type of wave depend on the details of the gravitational theory, and hence the experimental investigation of the properties of the waves would serve as a test of the theory.

Even more important, gravitational-wave astronomy would be a useful adjunct to optical, radio, and X-ray astronomy.

Gravitational waves would permit us to "look" into the very cores of quasars and other regions of strong gravitational fields.

The energy, pulse shape, and polarization of bursts of gravitational radiation could reveal to us the astrophysical processes by which these bursts are generated.

5.1 Plane waves

According to Eqs. (3.45) and (3.46), the linear field equation in vacuo is

$$\partial_\lambda \partial^\lambda \phi^{\mu\nu} = 0 \tag{5.1}$$

with the gauge condition

$$\partial_\mu \phi^{\mu\nu} = 0 \tag{5.2}$$

We will look for plane-wave solutions of the form

$$\phi^{\mu\nu} = \varepsilon^{\mu\nu} \cos k_\alpha x^\alpha \tag{5.3}$$

Here $\varepsilon^{\mu\nu}$ is a constant tensor, and k_α is a constant vector; these are called, respectively, the polarization tensor and the wave vector.

This plane-wave solution satisfies Eqs. (5.1) and (5.2) provided that

$$k^\alpha k_\alpha = 0 \tag{5.4}$$

$$k_\mu \varepsilon^{\mu\nu} = 0 \tag{5.5}$$

The frequency of the wave is $\omega = k^0$.

According to Eq. (5.4),

$$k^0 = \left(k_x{}^2 + k_y{}^2 + k_z{}^2\right)^{1/2} = |\mathbf{k}| \tag{5.6}$$

and hence $\omega/|\mathbf{k}| = 1$, so the speed of the wave is that of light.

The tensor $\varepsilon^{\mu\nu}$ must be symmetric, in accord with the general requirement that $h^{\mu\nu}$ and $\phi^{\mu\nu}$ be symmetric.

A symmetric second-rank tensor has 10 independent components, but Eq. (5.5) imposes an extra four conditions on these 10 components.

Hence $\varepsilon^{\mu\nu}$ has only six independent components, that is, there are six linearly independent tensors that solve Eq. (5.5).

Let us suppose that the wave propagates in the z-direction, with

$$k^\alpha = (\omega, 0, 0, \omega) \tag{5.7}$$

and

$$\phi^{\mu\nu} = \varepsilon^{\mu\nu} \cos(\omega t - \omega z) \tag{5.8}$$

For this special case, the linearly independent solutions of Eq. (5.5) may be taken to be the following:

$$\varepsilon_{(1)}{}^\mu \varepsilon_{(1)}{}^\nu - \varepsilon_{(2)}{}^\mu \varepsilon_{(2)}{}^\nu \tag{5.9}$$

$$\varepsilon_{(1)}{}^\mu \varepsilon_{(2)}{}^\nu + \varepsilon_{(1)}{}^\nu \varepsilon_{(2)}{}^\mu \tag{5.10}$$

$$\varepsilon_{(1)}{}^\mu k^\nu + \varepsilon_{(1)}{}^\nu k^\mu \tag{5.11}$$

$$\varepsilon_{(2)}{}^\mu k^\nu + \varepsilon_{(2)}{}^\nu k^\mu \tag{5.12}$$

$$k^\mu k^\nu \tag{5.13}$$

$$\varepsilon_{(1)}{}^\mu \varepsilon_{(1)}{}^\nu + \varepsilon_{(2)}{}^\mu \varepsilon_{(2)}{}^\nu \tag{5.14}$$

where the vectors $\varepsilon_{(1)}{}^\mu$ and $\varepsilon_{(2)}{}^\mu$ are defined as follows:

$$\varepsilon_{(1)}{}^{\mu} = (0, 1, 0, 0) \tag{5.15}$$

$$\varepsilon_{(2)}{}^{\mu} = (0, 0, 1, 0) \tag{5.16}$$

The polarization tensors (5.9) and (5.10) are called *transverse*.

They have the form, respectively,

$$\varepsilon_{\oplus}{}^{\mu\nu} = \varepsilon_{(1)}{}^{\mu}\varepsilon_{(1)}{}^{\nu} - \varepsilon_{(2)}{}^{\mu}\varepsilon_{(2)}{}^{\nu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{5.17}$$

$$\varepsilon_{\otimes}{}^{\mu\nu} = \varepsilon_{(1)}{}^{\mu}\varepsilon_{(2)}{}^{\nu} + \varepsilon_{(1)}{}^{\nu}\varepsilon_{(2)}{}^{\mu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{5.18}$$

These are the only polarizations that correspond to physical gravitational waves.
The other four polarizations, (5.11)–(5.14), carry no energy and no momentum.

In fact, waves of the other four types can always be completely canceled by an extra, supplementary, gauge transformation.

As mentioned in Chapter 3, the Hilbert gauge condition does not determine the field $h^{\mu\nu}$ uniquely, because we can still add to the gauge function Λ^{μ} that implements the gauge condition (5.2) a solution of the homogeneous wave equation (3.48),

$$\partial_{\lambda}\partial^{\lambda}\Lambda^{\mu} = 0 \tag{5.19}$$

Such supplementary gauge functions can be used to cancel the waves of type (5.11)–(5.14).

For instance, consider a wave of the type (5.11),

$$\phi^{\mu\nu} = \left(\varepsilon_{(1)}{}^{\mu} k^{\nu} + \varepsilon_{(1)}{}^{\nu} k^{\mu}\right) \cos k_{\alpha} x^{\alpha} \tag{5.20}$$

The gauge transformation (3.39) for $h^{\mu\nu}$ implies a corresponding gauge transformation for $\phi^{\mu\nu}$:

$$\phi^{\mu\nu} \to \bar{\phi}^{\mu\nu} = \phi^{\mu\nu} + \partial^{\nu}\Lambda^{\mu} + \partial^{\mu}\Lambda^{\nu} - \eta^{\mu\nu}\partial_{\alpha}\Lambda^{\alpha} \tag{5.21}$$

It is easy to see that this gives $\bar{\phi}^{\mu\nu} = 0$ if

$$\Lambda^{\mu} = -\varepsilon_{(1)}{}^{\mu} \sin k_{\alpha} x^{\alpha} \tag{5.22}$$

This is a permissible supplementary gauge function; it obeys the homogeneous wave equation (5.19) and is consistent with the primary gauge condition (5.2).

Similar supplementary gauge functions can be used to cancel all the other waves of the type (5.11)–(5.14).

Thus, all these waves merely represent a "wrong" choice of gauge function; they can all be canceled by adopting a better choice of gauge function.

These waves are merely "gauge waves," not physical waves.

In contrast, the two polarizations (5.17) and (5.18) cannot be canceled by a gauge transformation – they represent physical waves.

Like transverse electromagnetic waves, transverse gravitational waves with the polarization tensors $\varepsilon_{\oplus}{}^{\mu\nu}$ and $\varepsilon_{\oplus}{}^{\mu\nu}$ carry energy and momentum.

To calculate their energy and momentum, we need to examine the energy-momentum tensor for the gravitational field.

The derivation of this energy-momentum tensor is somewhat complicated; it is given in Appendix A.2.

The expression for the energy-momentum tensor of a free gravitational field obeying the field equation (5.1) and the gauge condition (5.2) is given by Eq. (A.39):

$$t_{(0)}{}^{\mu\nu} = \frac{1}{4}\left[2\phi^{\alpha\beta,\nu}\phi_{\alpha\beta}{}^{,\mu} - \phi^{,\nu}\phi^{,\mu} - \frac{1}{2}\eta^{\mu\nu}\left(2\phi^{\alpha\beta,\sigma}\phi_{\alpha\beta,\sigma} - \phi^{,\sigma}\phi_{,\sigma}\right)\right] \tag{5.23}$$

Here the subscript (0) serves to indicate that $t_{(0)}{}^{\mu\nu}$ is only a first approximation to the exact energy-momentum tensor for the full, nonlinear gravitational field. Note that the expression (5.23) has the desirable properties of an energy-momentum tensor: it is quadratic in the field variables, it is symmetric, and, in the static case, the $t_{(0)}{}^{00}$ component agrees with the energy density $(\nabla\ \Phi\ )^2/8\pi G$ given by the Newtonian approximation [see Eq. (1.21)].

If we substitute the wave solution (5.3) into the energy-momentum tensor $t_{(0)}{}^{\mu\nu}$ of Eq. (5.23), several terms drop out and we obtain

$$t_{(0)}{}^{\mu\nu} = \frac{1}{2}\left(\varepsilon^{\alpha\beta}\varepsilon_{\alpha\beta} - \frac{1}{2}\varepsilon_{\alpha}{}^{\alpha}\varepsilon_{\beta}{}^{\beta}\right)k^{\mu}k^{\nu}\sin^2 k_\sigma x^\sigma \tag{5.24}$$

The transverse waves, with amplitudes $A_\oplus$ and $A_\otimes$,

$$\phi_{\oplus}{}^{\mu\nu} = A_\oplus\varepsilon_{\oplus}{}^{\mu\nu}\cos k_\alpha x^\alpha \tag{5.25}$$

$$\phi_{\otimes}{}^{\mu\nu} = A_\otimes\varepsilon_{\otimes}{}^{\mu\nu}\cos k_\alpha x^\alpha \tag{5.26}$$

give, respectively, energy-momentum tensors

$$t_{(0)}{}^{\mu\nu} = (A_\oplus)^2 k^\mu k^\nu \sin^2 k_\alpha x^\alpha \tag{5.27}$$

$$t_{(0)}{}^{\mu\nu} = (A_\otimes)^2 k^\mu k^\nu \sin^2 k_\alpha x^\alpha \tag{5.28}$$

The $t_{(0)}{}^{03}$ component of these tensors gives the energy flux in the z-direction.

For the average energy flux, we must replace $\sin^2 k_\alpha x^\alpha$ by $1/2$, and we obtain, respectively,

$$\langle t_{(0)}{}^{03} \rangle = \frac{1}{2c}(A_\oplus)^2 \omega^2 \tag{5.29}$$

$$\langle t_{(0)}{}^{03} \rangle = \frac{1}{2c}(A_\otimes)^2 \omega^2 \tag{5.30}$$

where we have inserted a factor of $1/c$, to express the energy flux in cgs units.

Each of the "gauge waves" of types (5.11)–(5.14) has $t_{(0)}{}^{\mu\nu} = 0$.

Furthermore, if $\varepsilon^{\mu\nu}$ in Eq. (5.24) is an arbitrary superposition of the polarization tensors (5.9)–(5.14), then the value of $t_{(0)}{}^{\mu\nu}$ is exactly what it is if the terms involving the polarizations (5.11)–(5.14) are omitted from the superposition; that is, the value of $t_{(0)}{}^{\mu\nu}$ depends on the amplitudes $A_\oplus$ and $A_\otimes$ only.

Thus, the gauge waves are phantom waves – they are devoid of material substance and they are merely figments of the mathematical formalism.

It is often convenient to make use of the complex waves

$$\varepsilon_\oplus{}^{\mu\nu} e^{ik_\alpha x^\alpha} \tag{5.31}$$

$$\varepsilon_\otimes{}^{\mu\nu} e^{ik_\alpha x^\alpha} \tag{5.32}$$

in which case it is understood that the real part is to be regarded as the physical wave.

We can also define circularly polarized waves as

$$(\varepsilon_\oplus{}^{\mu\nu} - i\varepsilon_\otimes{}^{\mu\nu})\, e^{ik_\alpha x^\alpha} \tag{5.33}$$

$$(\varepsilon_\oplus{}^{\mu\nu} + i\varepsilon_\otimes{}^{\mu\nu})\, e^{ik_\alpha x^\alpha} \tag{5.34}$$

It can be shown that such circularly polarized waves carry angular momentum.

The amount of angular momentum is proportional to the amount of energy carried by the wave:

$$\text{angular momentum in wave} = \frac{2}{\omega} \times \text{energy in wave} \tag{5.35}$$

Waves with the polarization (5.33) have angular momentum parallel to the energy flux and are said to have *positive helicity*; those with the polarization (5.34) have angular momentum antiparallel to the energy flux and are said to have *negative helicity*.

The result (5.35) cannot be obtained directly from our solution (5.3); this solution ignores the boundaries of the wave in the transverse direction, and it is precisely the boundary region that is crucial for the transport of angular momentum.

The quantum-mechanical interpretation of Eq. (5.35) is that the quanta of the gravitational field, or gravitons, have spin $2\hbar$.

The ratio of angular momentum to energy per quantum is

$$\frac{2\hbar}{\hbar\omega} = \frac{2}{\omega} \tag{5.36}$$

in agreement with Eq. (5.35).

The result that there exist only two physical polarization states for a gravitational wave of given momentum corresponds to the well-known result of relativistic quantum theory that a particle of mass zero must have its spin along the direction of motion (positive helicity) or opposite to the direction of motion (negative helicity); any other direction for the spin is forbidden.

This gives us a new insight into the reason for gauge symmetry.

We can start the construction of any kind of relativistic field theory by deciding on the choice of the fundamental building blocks; that is, the spins and masses of the fields to that are to be included.

Each field of given spin and mass constitutes an irreducible representation of the Lorentz group; that is, a family of states such that each state is related to all the others by Lorentz transformations (including boosts, rotations, and translations).

Fields of different spins and/or masses constitute mutually exclusive families, not related by Lorentz transformations.

The electromagnetic fields (spin 1, mass zero) and our linear gravitational fields (spin 2, mass zero) are examples of such irreducible representations that constitute separate families of states.

Obviously, the simplest possible field theory involves only one family of a given spin and mass.

More complicated theories involve two or more families, which might be coupled to sources in different ways, and may also interact with each other.

For instance, the Brans-Dicke theory proposes to describe gravitation by a combination of a spin 2 tensor field and a spin 0 scalar field, both of mass zero.

Other proposed gravitational theories add more scalar, vector, or tensor fields.

All these more complicated theories contain a mix of intermingled families of fields.

Such theories with mixed ingredients are not necessarily wrong, but – as recommended by Newton's First Rule for Reasoning in Philosophy – it is desirable to avoid such a mix and its extra complications whenever possible.

Our construction of the linear gravitational field theory in Chapter 3 rested on the implicit assumption that we are dealing with a pure tensor field, of spin 2 and mass zero.

The mass is revealed by the value of the invariant "length" of the wave vector, $k_\alpha k^\alpha \propto m^2$, or, more precisely, $\hbar^2 k_\alpha k^\alpha = m^2$.

The spin can be identified by the factor of 2 that appears on the right side of Eq. (5.35).

However, our symmetric tensor field $h^{\mu\nu}$ or $\phi^{\mu\nu}$ has 10 components, whereas a plane wave of spin 2, mass zero is supposed to have only two linearly independent physical polarization states.

This is where gauge invariance comes into play – by adopting a field equation with gauge invariance, we can exploit gauge transformations to eliminate the superfluous, unphysical components of the gravitational wave field, and we can reduce the number of independent polarization states to two.

As we saw earlier, elimination of the unphysical polarizations involves repeated use of gauge transformations: First select a gauge function to achieve the gauge condition (5.2), which reduces the possible wave solutions to the six polarizations (5.9)–(5.14); then use a supplementary gauge function obeying the homogeneous wave equation (5.19) to eliminate four of these polarizations and reduce the possible wave solutions to the two transverse waves (5.25), (5.26).

Accordingly, we see that the conservation law for energy-momentum which we took as the motivation for gauge invariance in Chapter 3 is not the fundamental reason for the imposition of gauge invariance, but merely a byproduct.

The real fundamental reason is the spin 2, mass zero, character of the tensor field.

We can summarize this by the following scheme:

$$\text{tensor field of spin 2, mass zero} \Rightarrow \text{gauge invariance} \Rightarrow \text{conservation law for } T^{\mu\nu}$$

A similar scheme applies in electrodynamics, where the spin 1, mass-zero vector field A^μ has wave solutions with unphysical polarizations, which need to be eliminated by gauge transformations.

The conservation law for the electric current then emerges as a consequence of the gauge invariance.

5.2 Interaction of particles with a gravitational wave

To find the effect of a gravitational wave on a particle, we need to examine the equation of motion, Eq. (3.72):

$$\frac{du_\mu}{d\tau} = -\kappa\left(h_{\mu\alpha,\beta} - \tfrac{1}{2}h_{\alpha\beta,\mu}\right)u^\alpha u^\beta \tag{5.37}$$

For a particle initially at rest, $u_0 = 1$ and $u_k = 0$, so

$$\frac{du_\mu}{d\tau} = -\kappa\left(h_{\mu 0,0} - \tfrac{1}{2}h_{00,\mu}\right) \tag{5.38}$$

But for each of the physical gravitational waves (5.25) and (5.26) we have $h_{\mu 0} = 0$, so the right side of Eq. (5.38) vanishes.

Thus, a particle initially at rest remains at rest.

Although this means that the position of the particle, as measured by our coordinates, does not change, we must be careful not to jump to the conclusion that there is no physical effect at all.

Consider two particles placed on the *x*-axis, one at $x = \Delta x$ and the other at $x = -\Delta x$, and suppose they are subjected to a gravitational wave of the type (5.25).

According to Eq. (5.38) the coordinate interval between these particles remains fixed at the value $2\Delta x$.

But the physical distance, as measured by a meter stick or by radar ranging, depends not only on Δx but also on the metric tensor of spacetime.

We already saw in Section 3.5 that the metric tensor is not $\eta_{\mu\nu}$, but rather

$$g_{\mu\nu} = \eta_{\mu\nu} + \kappa h_{\mu\nu} \tag{5.39}$$

If so, then the measured distance between the two particles is

$$\Delta l^2 = -g_{11}(2\Delta x)^2 = (1 - \kappa h_{11})(2\Delta x)^2 = (1 - \kappa A_{\oplus} \cos \omega t)(2\Delta x)^2$$

In the context of the linear approximation, the amplitude $A_{\oplus}$ is small, and hence

$$\Delta l \cong \left(1 - \frac{\kappa}{2} A_{\oplus} \cos \omega t\right)(2\Delta x) \tag{5.40}$$

Thus, the distance between the particles oscillates.

Expressed in another way, the particles remain at rest relative to the coordinates, but the coordinates oscillate relative to our meter sticks, and hence our measurements tell us that the particles move.

Similarly, the distance between two particles on the *y*-axis, at $y = \pm \Delta y$, and exposed to the same gravitational wave varies as follows:

$$\Delta l \cong \left(1 + \frac{\kappa}{2} A_{\oplus} \cos \omega t\right)(2\Delta y) \tag{5.41}$$

Comparison of Eqs. (5.40) and (5.41) shows that the distance between the first pair of particles is at minimum when the distance between the second pair is at maximum, and vice versa.

Figure 5.1 displays the deformation produced by the gravitational wave (5.25) on a "chain" of free particles originally placed in a circle in the *x*-*y* plane.

Figure 5.2 shows the corresponding result for the gravitational wave (5.26).

It is obvious that the only difference between these figures is a rotation by 45°.

In fact, the polarization tensor (5.17) can be obtained from (5.18) by a 45° rotation.

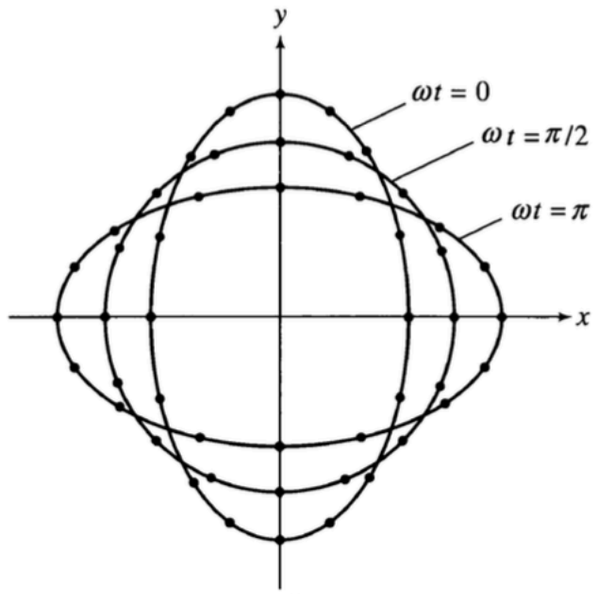


Fig. 5.1 Deformation of a circular "chain" of particles by a gravitational wave with polarization of type $\oplus$.

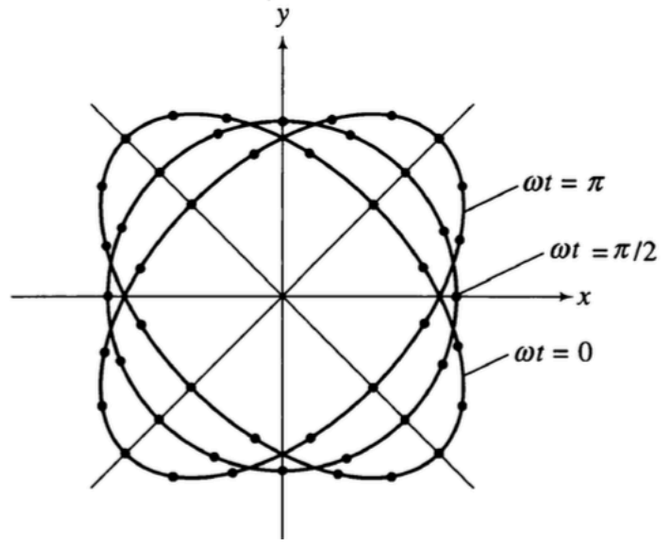


Fig. 5.2 Deformation of a circular "chain" of particles by a gravitational wave with polarization of type $\otimes$.

Finally, Figures 5.3 and 5.4 show the deformations produced by the circularly polarized waves (5.33) and (5.34).

In these two cases, only the deformation, or bulge, rotates in the direction shown; the particles on the chain do not rotate, but only oscillate in and out around their initial positions.

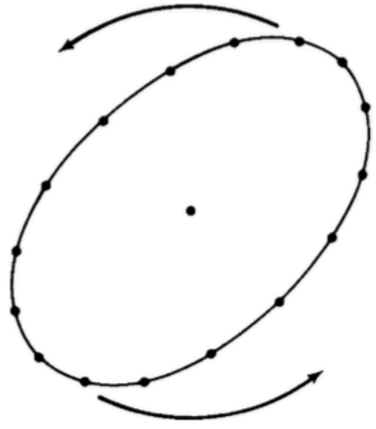


Fig. 5.3 Deformation produced by a wave of positive helicity. The deformation rotates in the counterclockwise direction if the wave is propagating out of the plane of the page, toward you.

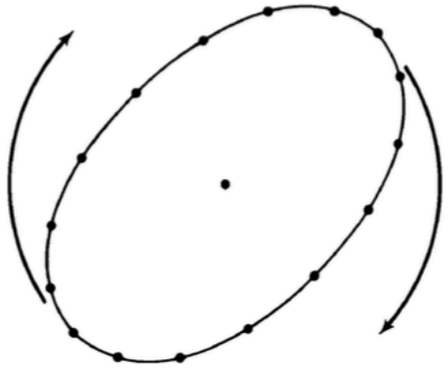


Fig. 5.4 Deformation produced by a wave of negative helicity. The deformation rotates in the clockwise direction.

The deformations shown in all these figures are essentially tidal effects, similar to those of Section 1.8, but the effects are time dependent rather than static.

This suggests we should calculate the tidal force needed to produce the motion given in Eqs. (5.40) and (5.41).

For a particle placed at Δ x on the *x*-axis, the acceleration away from the origin is

$$\frac{d^2}{dx^2}\frac{\Delta l}{2} = -\frac{\kappa}{2}A_\oplus \Delta x \frac{d^2}{dt^2}\cos\omega t = \frac{\kappa}{2}A_\oplus \Delta x\, \omega^2 \cos\omega t \tag{5.42}$$

If the mass of this particle is m, the corresponding tidal force is

$$f_x = \left(m\frac{\kappa}{2}A_\oplus \omega^2 \cos\omega t\right)\Delta x \tag{5.43}$$

Likewise, the tidal force on a particle on the *y*-axis is

$$f_y = -\left(m\frac{\kappa}{2}A_\oplus \omega^2 \cos\omega t\right)\Delta y \tag{5.44}$$

These expressions should be compared with our earlier results [see Eqs. (1.47), (1.48)] for the tidal force in a static gravitational field.

In both cases the strength of the tidal force increases directly with the distance.

Note that in Eqs. (5.43) and (5.44), Δ x and Δ y may be regarded, to a sufficient approximation, as the distances of the particles from the origin.

The difference between the true distance Δl and the coordinate interval Δ x is crucial in Eq. (5.40), but it is not important in Eqs. (5.43) and (5.44), since there it leads only to higher order corrections.

Incidentally, the forces f_x and f_y associated with a gravitational wave can be written in the form

$$f_x = F(t)\Delta x \qquad f_y = F(t)\Delta y$$

where we are now assuming that both Δx and Δy are nonzero, so both the forces (5.43) and (5.44) are present simultaneously.

From this it is easy to see that

$$\frac{\partial f_x}{\partial(\Delta x)} + \frac{\partial f_y}{\partial(\Delta y)} = 0 \tag{5.45}$$

Hence the divergence of the force **f** is zero, which tells us that the tidal force field can be represented graphically by field lines.

The field lines for a gravitational wave of type $\oplus$ are drawn in Figure 5.5, for $t = 0$.

Obviously, such a force field will tend to produce the deformations shown in Figure 5.1.

The field lines for a gravitational wave of type $\otimes$ are drawn in Figure 5.6.

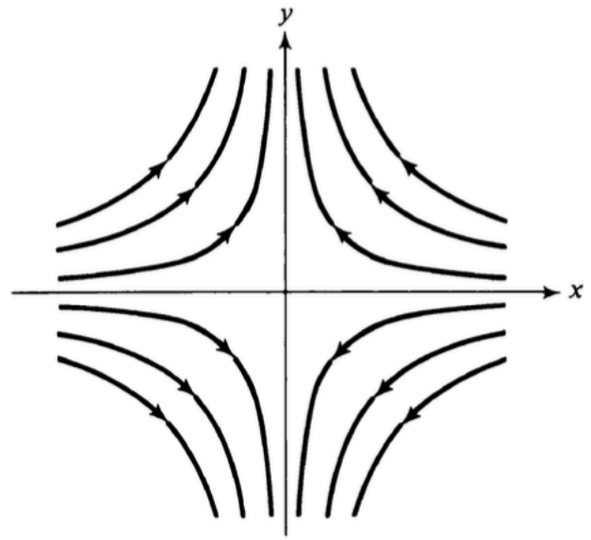


Fig. 5.5 The tidal field lines for a gravitational wave of type $\oplus$. The field lines reverse direction every half-period.

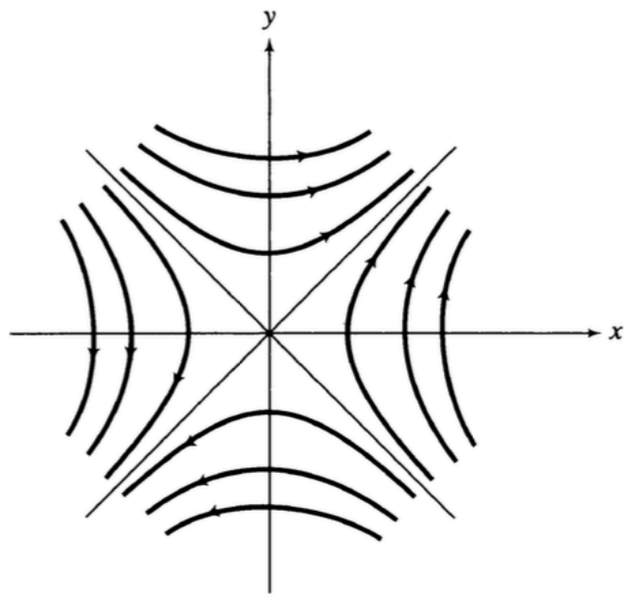


Fig. 5.6 The tidal field lines for a gravitational wave of type $\otimes$.

It can be shown that the radial component of the tidal force is

$$f_r = \left(m\frac{\kappa}{2}A_{\oplus}\omega^2\cos\omega t\right) r_0 \cos 2\phi \tag{5.46}$$

where $r_0 = (\Delta x^2 + \Delta y^2)^{1/2}$ is the radial coordinate of the particle, and ϕ the azimuthal coordinate.

Note that our results are valid only for weak gravitational waves.

To be precise, the wave must have a small amplitude, so $\kappa A \ll 1$.

As an illustration of these results, consider the effect of a gravitational wave on two free particles separated by a distance of several km, such as the test masses used in modern interferometric detectors of gravitational waves, for instance, the LIGO detector, where the test masses are mirrors hanging in a pendulum suspension, which provides a negligible restoring force, so the test masses are nearly free.

In response to the gravitational wave, incident transversely, the distance between the two masses will oscillate as indicated by Eq. (5.40).

The quantity $2\ \Delta x$ may be interpreted as the distance measured when the wave is absent, which is 4 km in the LIGO detector.

For a strong astrophysical source of gravitational radiation, say, a supernova at a distance of 3×10^4 light-years, the gravitational wave might have an amplitude of about $\kappa h_{11} = \kappa A_{\oplus} \cong 10^{-18}$ (see Table 5.3).

This means that the distance between the particles would vary by about 1 part in 10^{18}, or about $4\ \text{km} \times 10^{-18} = 4 \times 10^{-13}$ cm.

LIGO and other similar interferometric detectors now in operation are capable of detecting small displacements of this magnitude (see Section 5.7).

5.3 Emission of gravitational radiation

Gravitational waves will be emitted by almost any system with an energy-momentum tensor that changes in time.

The calculation of the radiation fields in the linear approximation proceeds in much the same way as the calculation of the electromagnetic radiation fields associated with a time-dependent charge distribution.

We begin with the linear field equation (3.45),

$$\partial_\lambda \partial^\lambda \phi^{\mu\nu}(t, \mathbf{x}) = -\kappa T^{\mu\nu}(t, \mathbf{x}) \tag{5.47}$$

where $T^{\mu\nu}(t, \mathbf{x})$ is the energy-momentum tensor of a time-dependent matter distribution.

The retarded solution of Eq. (5.47) is

$$\phi^{\mu\nu}(t, \mathbf{x}) = -\frac{\kappa}{4\pi} \int \frac{T^{\mu\nu}(t - |\mathbf{x} - \mathbf{x}'|, \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} d^3x' \tag{5.48}$$

We will assume that the field point $\mathbf{x}$ is in the radiation zone, far away from the matter system (see Fig. 5.7); this means that we can replace $|\mathbf{x} - \mathbf{x}'|$ by $|\mathbf{x}|$ in the denominator of the integrand in Eq. (5.48).

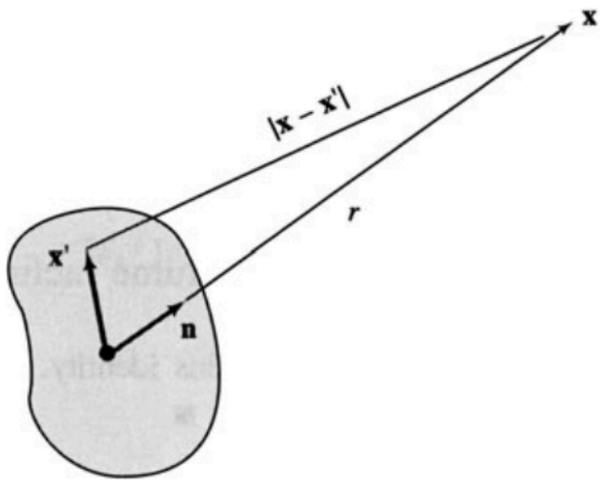


Fig. 5.7 The field point **x** is at a large distance from the region (shaded) containing the distribution of energy momentum.

We will further assume that $T^{\mu\nu}$ does not change very quickly in time; it will then be a good approximation to replace $t - |\mathbf{x} - \mathbf{x}'|$ by $t - |\mathbf{x}|$ in the numerator of the integrand in Eq. (5.48).

With $r \equiv |\mathbf{x}|$, we then obtain

$$\phi^{\mu\nu}(t, \mathbf{x}) = -\frac{\kappa}{4\pi r} \int T^{\mu\nu}(t - r, \mathbf{x}')\, d^3x' \tag{5.49}$$

In the linear approximation, $T^{\mu\nu}$ satisfies the conservation law

$$\partial_\nu T^{\mu\nu} = 0 \tag{5.50}$$

We can separate Eq. (5.50) into space and time components, respectively:

$$\frac{\partial}{\partial t}T^{k0} = -\partial_n T^{kn} \tag{5.51}$$

$$\frac{\partial}{\partial t}T^{00} = -\partial_n T^{0n} \tag{5.52}$$

In consequence of Eq. (5.51), we then obtain the following identity:

$$\int T^{kl}d^3x = \frac{1}{2}\frac{\partial}{\partial t}\int (T^{k0}x^l + T^{l0}x^k)d^3x \tag{5.53}$$

where the integration volume includes all of the matter system.
Similarly, in consequence of Eq. (5.52), we obtain the identity

$$\int (T^{k0}x^l + T^{l0}x^k)d^3x = \frac{\partial}{\partial t}\int T^{00}x^k x^l d^3x \tag{5.54}$$

If we combine Eqs. (5.53) and (5.54), we find that

$$\int T^{kl}d^3x = \frac{1}{2}\frac{\partial^2}{\partial t^2}\int T^{00}x^k x^l d^3x \tag{5.55}$$

which shows that most of the integrals appearing on the right side of (5.49) can be expressed in terms of T^{00}.

In essence, the reason why we can express the integral over T^{kl} in terms of derivatives of the integral over T^{00} is that, whenever the momentum flux density T^{kl} is not zero, the momentum density must accumulate somewhere in the volume, that is, T^{k0} must change.
And whenever the energy flux density T^{k0} is nonzero, the energy density must accumulate somewhere in the volume, that is, T^{00} must change, which gives rise to the time derivatives on the right side of Eq (5.55).

For nonrelativistic matter, we can approximate

$$T^{00} = \text{energy density} \cong \text{rest-mass density} = \rho_0 \tag{5.56}$$

and we obtain

$$\phi^{kl}(t, \mathbf{x}) = -\left[\frac{\kappa}{8\pi r}\frac{\partial^2}{\partial t^2}\int \rho_0(t, \mathbf{x}')x'^k x'^l d^3x'\right]_{t-r} \tag{5.57}$$

This integral can also be expressed in terms of the quadrupole-moment tensor [see Eq. (1.27)],

$$Q^{kl} = \int \left(3x'^k x'^l - r'^2\delta^l_k\right)\rho_0(t, \mathbf{x}')d^3x' \tag{5.58}$$

in terms of which

$$\phi^{kl}(t, \mathbf{x}) = -\frac{\kappa}{8\pi r}\frac{1}{3}\left[\frac{\partial^2}{\partial t^2}Q^{kl} + \delta^l_k\frac{\partial^2}{\partial t^2}\int r'^2\rho_0(t, \mathbf{x}')d^3x'\right]_{t-r} \tag{5.59}$$

For the calculation of the energy flux, we can omit the term proportional to $\delta_k{}^l$ in the brackets in Eq. (5.59), because this part of the wave carries away no energy.

To see this, note that if the distance r is large enough, then ϕ^{kl} can be regarded, to a sufficient approximation, as a plane wave in the vicinity of the point $\mathbf{x}$.

According to Section 5.1, the only plane-wave polarizations that carry energy are those of polarizations $\varepsilon_\oplus{}^{kl}$ and $\varepsilon_\otimes{}^{kl}$ given by Eqs. (5.17) and (5.18); obviously, $\delta_k{}^l$ is not a polarization of this type.

If we omit the $\delta_k{}^l$ term, then

$$\phi^{kl}(t, \mathbf{x}) = -\frac{\kappa}{8\pi r}\frac{1}{3}\ddot{Q}^{kl} \tag{5.60}$$

where the dots above the letter Q stand for time derivatives and it is understood that the right side is to be evaluated at the retarded time t − r.

The energy flux in the radial direction can be evaluated from the $t_{(0)}{}^{0s}$ components of the energy-momentum tensor (5.23) for the gravitational field.

The unit vector in the radial direction is **n** (see Fig. 5.7), and its components are $n^s = (n_x, n_y, n_z) = (x/r, y/r, z/r)$.

Therefore the energy flux in that direction is

$$t_{(0)}{}^{0s} n^s = \frac{n^s}{4}\left(2\phi^{\alpha\beta,0}\phi_{\alpha\beta}{}^{,s} - \phi^{,0}\phi^{,s}\right) \tag{5.61}$$

Let us separate $\phi^{\alpha\beta}$ into space and time parts, so

$$\begin{aligned} t_{(0)}{}^{0s} n^s &= \frac{n^s}{4}\left(2\phi^{kl,0}\phi^{kl,s} - 4\phi^{k0,0}\phi^{k0,s} + 2\phi^{00,0}\phi^{00,s} - \phi^{kk,0}\phi^{ll,s}\right. \\ &\quad \left. + \phi^{kk,0}\phi^{00,s} + \phi^{00,0}\phi^{kk,s} - \phi^{00,0}\phi^{00,s}\right) \\ &= \frac{n^s}{4}\left(2\phi^{kl,0}\phi^{kl,s} - 4\phi^{k0,0}\phi^{k0,s} + \phi^{00,0}\phi^{00,s}\right) \end{aligned} \tag{5.62}$$

where we have taken into account that, according to Eqs. (5.58) and (5.60), $\phi^{kk} \propto \ddot{Q}^{k\,k} = 0$.

In the evaluation of $\phi^{\alpha\beta,s}$ we can neglect the derivative of the factor $1/r$ appearing in Eq. (5.59), because we are assuming that r is very large.

The only other place where a dependence on r occurs is in the retarded time, $t - r$; this means that r and t appear only in the combination $t - r$, and therefore

$$\phi^{\alpha\beta}{}_{,s} = \frac{\partial\phi^{\alpha\beta}}{\partial r}\frac{\partial r}{\partial x^s} = -\phi^{\alpha\beta}{}_{,0}\frac{\partial r}{\partial x^s} = -\phi^{\alpha\beta}{}_{,0} n^s \tag{5.63}$$

Accordingly, the derivative appearing in the first term in the parentheses in Eq. (5.62) becomes

$$\phi^{kl,s} = \phi^{kl,0} n^s \tag{5.64}$$

We can obtain convenient expressions for the other terms appearing in Eq. (5.62) by beginning with our familiar gauge condition:

$$\phi^{\mu 0}{}_{,0} = -\phi^{\mu l}{}_{,l} \tag{5.65}$$

This gives

$$\phi^{k0}{}_{,0} = -\phi^{kl}{}_{,l} = \phi^{kl}{}_{,0} n^l = \phi^{kl,0} n^l \tag{5.66}$$

and also

$$\phi^{00,0} = -\phi^{0l}{}_{,l} = \phi^{0l}{}_{,0} n^l = \phi^{kl,0} n^k n^l \tag{5.67}$$

Substituting Eqs. (5.64), (5.66), and (5.67) into Eq. (5.62), we then obtain

$$t_{(0)}{}^{0s} n^s = \frac{n^s}{4} \left(2\phi^{kl,0} \phi^{kl,0} n^s - 4\phi^{kl,0} n^l \phi^{km,0} n^m n^s + \phi^{kl,0} n^k n^l \phi^{mr,0} n^m n^r n^s \right) \tag{5.68}$$

With $n^s n^s = 1$ and with Eq. (5.60), this becomes

$$\begin{aligned} t_{(0)}{}^{0s} n^s &= \frac{1}{4} \left(2\phi^{kl,0} \phi^{kl,0} - 4\phi^{kl,0} \phi^{km,0} n^l n^m + \phi^{kl,0} \phi^{mr,0} n^k n^l n^m n^r \right) \\ &= \frac{1}{9} \left(\frac{\kappa}{8\pi r} \right)^2 \left(\frac{1}{2} \overset{\cdots kl}{Q} \overset{\cdots kl}{Q} - \overset{\cdots kl}{Q} \overset{\cdots km}{Q} n^l n^m + \frac{1}{4} \overset{\cdots kl}{Q} \overset{\cdots mr}{Q} n^k n^l n^m n^r \right) \end{aligned} \tag{5.69}$$

The energy radiated by the system per unit solid angle and unit time in the direction n^s is

$$-\frac{d^2 E}{dt\, d\Omega} = r^2 t_{(0)}{}^{0s} n^s \tag{5.70}$$

Obviously, the angular distribution is very complicated; however, the total power radiated can be obtained by integrating Eq. (5.70) over all solid angles, and the final expression turns out to be quite simple.

We find

$$-\frac{dE}{dt} = \int r^2 t_{(0)}{}^{0s} n^s d\Omega$$
$$= \frac{1}{9}\left(\frac{\kappa}{8\pi}\right)^2 \left(\frac{1}{2}\overset{\ldots kl}{Q}\overset{\ldots kl}{Q}\int d\Omega - \overset{\ldots kl}{Q}\overset{\ldots km}{Q}\int n^l n^m d\Omega\right. \tag{5.71}$$
$$\left. + \frac{1}{4}\overset{\ldots kl}{Q}\overset{\ldots mr}{Q}\int n^k n^l n^m n^r d\Omega\right)$$

The average values of $n^l n^m$ and of $n^k n^l n^m n^r$ over the surface of a sphere are, respectively,

$$\frac{1}{4\pi}\int n^l n^m d\Omega = \frac{1}{3}\delta_l^m \tag{5.72}$$

$$\frac{1}{4\pi}\int n^k n^l n^m n^r d\Omega = \frac{1}{15}\left(\delta_k^l\delta_m^r + \delta_k^m\delta_l^r + \delta_k^r\delta_l^m\right) \tag{5.73}$$

Substituting Eqs. (5.72) and (5.73) into Eq. (5.71), we obtain

$$-\frac{dE}{dt} = \frac{4\pi}{45}\left(\frac{\kappa}{8\pi}\right)^2 \overset{\ldots kl}{Q}\overset{\ldots kl}{Q} \tag{5.74}$$

If we insert the value $\kappa^2 = 16\pi G$ and return to cgs units, we find a radiated power

$$-\frac{dE}{dt} = \frac{G}{45c^5}\overset{\ldots kl}{Q}\overset{\ldots kl}{Q} \tag{5.75}$$

In Section 1.4 we saw that a shift of the origin of coordinates changes the quadrupole moment by an additive constant [see Eq.(1.30)].

Since the emission of radiation involves only the time-dependent part of Q^{kl}, it follows that changes in the origin will not affect our results; we can choose any convenient point as origin.

The radiation we have calculated earlier is often called *quadrupole radiation*.

Since our results all involve the quadrupole-moment tensor, this is as good a name as any.

However, keep in mind that if we were to make the same kind of approximations in the electromagnetic case that we have made in the gravitational case, we would obtain electromagnetic dipole radiation.

The analog of gravitational "quadrupole" radiation is electromagnetic dipole radiation – these kinds of radiation correspond to the approximation of lowest order we can make in the retarded solution to the field equation [see Eq. (5.49) and the corresponding equation in electromagnetism].

5.4 Emission by a vibrating quadrupole

According to Eq. (5.75), any system of masses with a time-dependent quadrupole moment and with $\ddot{Q}^{\mathrm{k}\,l} \neq 0$ will radiate gravitational waves.

We can classify the sources of gravitational waves into periodic and bursting sources.

The former have a quadrupole moment that varies periodically; they emit harmonic waves.

The latter have a quadrupole moment that varies for a short while in a nonperiodic manner; they emit a burst, or a pulse, of gravitational waves.

For example, vibrating masses and rotating masses are periodic sources, whereas a mass suffering a brief and sharp acceleration while colliding with another mass is a bursting source.

The simplest periodic source of gravitational waves consists of two equal masses connected by a spring (see Fig. 5.8).

This ideal system is called a linear quadrupole.

We treat the masses as particles, but they could equally well be spherical (for a system of spherical masses, the quadrupole moment is the same as for a system of particles).

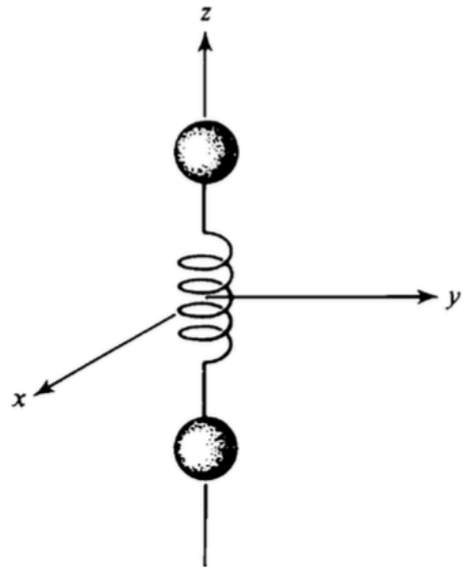


Fig. 5.8 An oscillating quadrupole.

The ideal linear quadrupole is mainly of theoretical interest.

Although such a linear quadrupole would be easy to construct in a laboratory, the amount of gravitational radiation is insignificant for masses of laboratory size.

There are no astrophysical sources of gravitational radiation in the shape of the linear quadrupole.

However, a vibrating star or supernova remnant, in an elongated mode of vibration, exhibits some of the general features of the vibrating linear quadrupole, and we can extract some crude estimates for the radiation emitted by such a star if we use the simple formulas that are valid for the linear quadrupole.

As illustrated in Figure 5.8, the two masses move along the z-axis; their positions are, respectively,

$$z = \pm(b + a \sin \omega t) \tag{5.76}$$

For the sake of simplicity, we assume that $a \ll b$. Hence we can approximate

$$z^2 \cong b^2 + 2ab \sin \omega t \tag{5.77}$$

Since our system consists of two discrete point masses, the integral (5.58) for the quadrupole moment reduces to a sum of two equal terms.

The retarded value of the quadrupole moment of the system can be expressed as

$$Q^{kl} \cong \left[1 + \frac{2a}{b} \sin \omega(t - r)\right] Q^{kl}(0) \tag{5.78}$$

where $Q^{kl}(0)$ is the static quadrupole moment:

$$Q^{kl}(0) = \begin{pmatrix} -2mb^2 & 0 & 0 \\ 0 & -2mb^2 & 0 \\ 0 & 0 & 4mb^2 \end{pmatrix} \tag{5.79}$$

It can then be shown that the radiated gravitational field is

$$\phi^{kl} = \frac{\kappa}{8\pi r} \frac{2a}{3b} \omega^2 \sin \omega(t - r) Q^{kl}(0) \tag{5.80}$$

The angular distribution of the radiated energy given by Eq. (5.70) is

$$\begin{aligned} -\frac{d^2E}{dtd\Omega} = \frac{1}{9}\left(\frac{\kappa}{8\pi}\right)^2 \Bigg[& \frac{1}{2}\left(\dddot{Q}^{11}\right)^2 + \frac{1}{2}\left(\dddot{Q}^{22}\right)^2 + \frac{1}{2}\left(\dddot{Q}^{33}\right)^2 - \left(\dddot{Q}^{11} n^1\right)^2 - \left(\dddot{Q}^{22} n^2\right)^2 \\ & - \left(\dddot{Q}^{33} n^3\right)^2 + \frac{1}{4}\left(\dddot{Q}^{11} n^1 n^1 + \dddot{Q}^{22} n^2 n^2 + \dddot{Q}^{33} n^3 n^3\right)^2 \Bigg] \end{aligned} \tag{5.81}$$

In terms of the usual polar angles, defined in Figure 5.9,

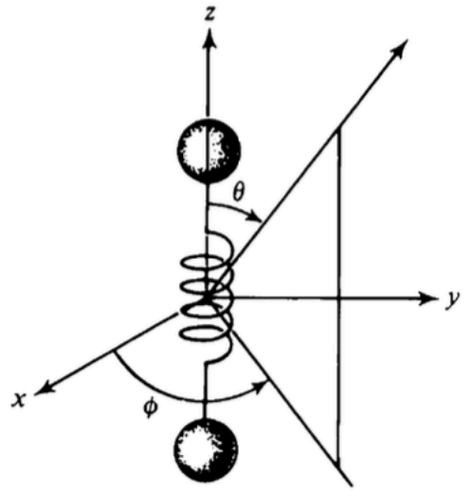


Fig. 5.9 Angular direction in the radiated gravitational wave.

we have

$$n^1 = n_x = \sin\theta\cos\phi$$
$$n^2 = n_y = \sin\theta\sin\phi \tag{5.82}$$
$$n^3 = n_z = \cos\theta$$

and Eq. (5.81) reduces to

$$-\frac{d^2E}{dtd\Omega} = \left(\frac{\kappa}{8\pi}\right)^2 [2mab\omega^3\cos\omega(t-r)]^2\sin^4\theta \tag{5.83}$$

The radiation pattern is shown in Figure 5.10.

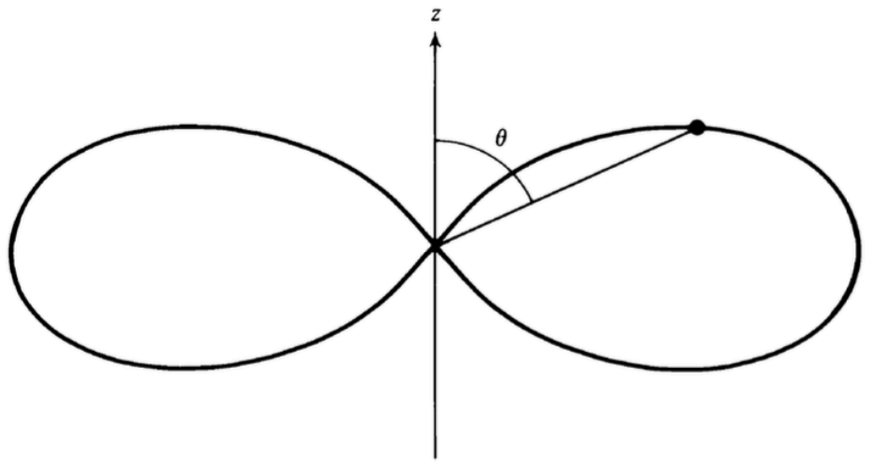


Fig. 5.10 The radiation pattern for the emission of gravitational radiation by a quadrupole oscillating along the *z*-axis. This is a polar plot of the function $\sin^4\theta$. The distance from the origin to a point on the curve is proportional to the flux emitted in that direction. The pattern has cylindrical symmetry about the *z*-axis.

Note that this radiation pattern in no way resembles the electromagnetic quadrupole pattern; in the latter case the flux is zero in at least three perpendicular directions (and their opposite directions).

If anything, Figure 5.10 is vaguely reminiscent of an electromagnetic dipole radiation pattern.

The total emitted power is

$$-\frac{dE}{dt} = \frac{G}{45c^5}\dddot{Q}^{kl}\dddot{Q}^{kl} = \frac{32G}{15c^5}[mab\omega^3\cos\omega(t-r)]^2 \tag{5.84}$$

with a time average

$$-\left\langle \frac{dE}{dt} \right\rangle = \frac{16G}{15c^5}(mab)^2\omega^6 \tag{5.85}$$

Although we have derived Eq. (5.85) for a system of two point masses or two spherical masses connected by a spring, it gives an order of magnitude estimate for the gravitational-wave energy emitted in the vibration of any elastic body, provided that the vibration is, roughly, in one direction.

For example, we can use Eq. (5.85) to estimate the power radiated by a steel rod in longitudinal vibration; in this case, a is the amplitude of oscillation, m the mass of the rod, b its length, and ω the frequency of its oscillation.

Radiation emitted by quadrupole vibrations might be important in the case of novas.

Nova outbursts occur on white-dwarf stars in binary systems, where accretion from a companion star gradually accumulates nuclear fuel on the dwarf's surface until the fuel reaches a critical mass and explodes.

Among other consequences, the explosion initiates vibrations in the body of the white dwarf, at characteristic frequencies of 0.01 to 1 Hz.

The energy released in a nova explosion is typically 10^{45} erg, of which perhaps as much as 10% is deposited in the vibrational motion of the star and is subsequently radiated in the form of gravitational waves.

Higher frequencies and larger amounts of energy are emitted by the vibrations of neutron stars formed in supernova implosions.

Since the damping time for the vibrations of a neutron star is quite short (a fraction of a second), this kind of radiation has the form of a burst, and we will deal with it in a later section.

5.5 Emission by a rotating quadrupole

Another simple periodic source of gravitational waves is the rotating quadrupole, consisting of two particles or two spherical masses moving in circular orbits about their common center of mass (see Fig. 5.11).

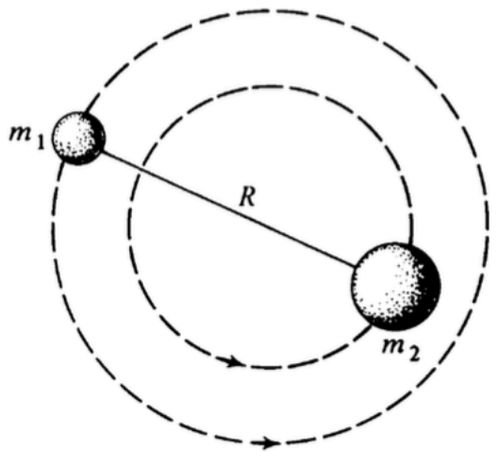


Fig. 5.11 Two spherical masses in circular orbits about their center of mass.

Since the quadrupole moment repeats when the masses move through one-half of their orbit, the frequency of the emitted waves is twice the orbital frequency.

The power radiated by this system is

$$-\frac{dE}{dt} = \frac{32G}{5c^5}\left(\frac{m_1 m_2}{m_1 + m_2}\right)^2 R^4 \omega^6 \tag{5.86}$$

where m_1 and m_2 are the masses, R is the distance between them, and ω is the orbital frequency.

We can apply Eq. (5.86) to a binary star system, consisting of two stars in circular orbits around each other.

In this case, the force holding the two masses in their orbits is gravitational, and R and ω are related by Kepler's third law,

$$\omega^2 = \frac{G(m_1 + m_2)}{R^3} \tag{5.87}$$

so Eq. (5.86) becomes

$$-\frac{dE}{dt} = \frac{32G^4}{5c^5R^5}(m_1m_2)^2(m_1+m_2) \tag{5.88}$$

When the gravitating system loses energy by radiation, the distance between the masses decreases at a rate

$$\frac{dR}{dt} = -\frac{64G^4}{5c^5R^3}m_1m_2(m_1+m_2) \tag{5.89}$$

and the orbital frequency increases at a rate

$$\frac{d\omega}{dt} = -\frac{3\omega}{2R}\frac{dR}{dt} = \frac{96}{5}\left[\frac{G(m_1+m_2)}{c^2R}\right]^{3/2}\frac{G^2m_1m_2}{c^2R^4} \tag{5.90}$$

Appreciable amounts of gravitational radiation are emitted by a binary star system provided the orbital radius is small.

Table 5.1 gives some examples of known binary systems with small orbital radii (and short orbital periods).

Table 5.1 Binary systems as sources of gravitational radiation*

System	Masses ($M_\odot$)	Distance (pc)**	Wave Frequency (10^{-6} Hz)	Radiated Power (10^{30} erg/s)	Flux at Earth (10^{-12} erg/cm^{2}s)	Amplitude at Earth, κA (10^{-22})
Eclipsing binaries						
ι Boo	1.0, 0.5	11.7	86	1.1	68.0	51
μ Sco	12, 12	109	16	51	38.0	210
V Pup	16.5, 9.7	520	16	59	1.9	46
Cataclysmic binaries (novas)						
Am CVn	1.0, 0.041	100	1900	300	240	5
SS Cyg	0.97, 0.83	30	84	2	20	30
Binary X-ray sources (black holes or neutron stars)						
Cyg X-1	19, 15	1800	4.1	2.6	0.006	9
PSR 1913+16	1.4, 1.4	5000	70***	0.6	0.0002	0.12
			140	2.9	0.0011	0.14
			210	5.8	0.0021	0.12

* From Douglass and Braginsky (1979), except Cyg X-1, which is recalculated according to new mass and distance determinations made by Orosz et al. (2011).

** 1 parsec (pc), equals 3.26 light-years.

*** This binary emits radiation at harmonic frequencies, because its orbit has large eccentricity.

The system that emits the most power and produces the largest flux at the Earth is Am CVn, with 3×10^{32} erg/s.

This system consists of a blue white dwarf and a low-mass white dwarf (see Fig. 5.12)

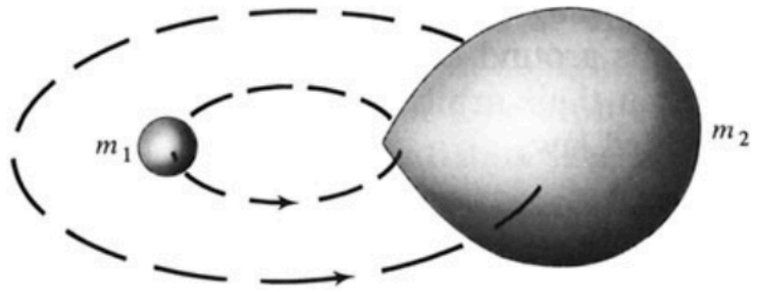


Fig. 5.12 The binary system Am CVn. The compact star has the larger mass (m_1) and is a blue white dwarf; the other star (m_2) is a low-mass white dwarf. There is mass transfer from the latter to the former. (After Faulkner, Flannery, and Warner, 1972)

in an exceptionally small orbit around each other, with a period of only 17.5 min!

However, the binary system that produces the largest gravitational-wave amplitude at the Earth is μ Sco, with $\kappa A \cong 2 \times 10^{-20}$ [although the amplitude of the waves from Am CVn is smaller, their much higher frequency permits them to carry more energy; see Eqs. (5.29) and (5.30)].

Detectors sensitive to gravitational waves of amplitudes as small as $\kappa A \cong 10^{-22}$ are now in operation, but they respond only to waves of much higher frequency than the frequencies radiated by these binary systems, and there are no immediate prospects for the direct detection of gravitational waves of low frequency by resonant or interferometric detectors.

However, gravitational radiation can be detected indirectly, by the effect that the loss of energy has on the orbital motion of the binary system.

In this context, the most interesting system is PSR 1913+16, the binary pulsar discovered in 1974 by Hulse and Taylor during a pulsar search with the Arecibo radio telescope.

This system consists of a pulsar (a rotating neutron star that emits radio pulses) in a tight orbit around a companion, either a white dwarf or another neutron star.

Hulse and Taylor found that the pulsar frequency exhibits a periodic variation, attributed to the Doppler shift generated by the orbital motion of the pulsar around its (undetected) companion.

They carefully analyzed the motion by examining the arrival times of individual pulses over a long time.

Such an analysis of arrival times of individual pulses yields higher accuracy than measurement of the "instantaneous" frequency, because a frequency measurement in a finite time $\Delta\tau$ is subject to the usual uncertainty $\Delta\nu \cong 1/\Delta\tau$.

For instance, if we attempt to measure the "instantaneous" frequency associated with a 1-min segment of data, then $\Delta\nu \cong 1/60$ Hz, which is a rather large error.

Precise values of the parameters of the binary system have been extracted from a least-squares fit of the arrival times.

This analysis not only yielded the orbital period and the projected semi-major axis (a sin i, where *i* is the angle of inclination of the orbit relative to the line of sight) but also the eccentricity and the rate of precession of the periastron (analogous to the precession of the perihelion of Mercury in the Solar System).

Table 5.2 lists the most important parameters of PSR 1913+16.

Table 5.2 Some observed parameters of PSR 1913+16*

Pulsar period	0.05903000321797(±5) s
Projected semi-major axis	2.341782(±3) light-seconds
Eccentricity	0.6171334(±5)
Orbital period	0.322997448911(±4) d
Rate of precession of periastron	4.226598(±5) degree/yr
Amplitude of time-dilation factor	0.0042992(±8)
Rate of change of orbital period	$-2.423(\pm1) \times 10^{-12}$ s per s

* From Weisberg, Nice, and Taylor (2010).

Note that the pulse period and the pulse frequency are known to about 13 significant figures.

The analysis also revealed the presence of significant time-dilation effects, which vary in magnitude as the pulsar moves in its elliptical orbit that brings it closer or farther from its companion star.

The time-dilation effects arise in part from the gravitational time dilation produced by the gravitational field of the companion and in part from the "transverse," or special-relativistic, time dilation produced by the orbital speed.

The nominal pulsar period listed in Table 5.2 has been corrected for these time-dilation effects (the pulsar period is also subject to a time dilation contributed by the pulsar's own gravitational field, but this is a constant factor, which cannot be determined directly from the data).

As expected for a binary system that loses energy by gravitational radiation, the orbital period of PSR 1913+16 is decreasing, by 2×10^{-12} second per second.

For a quantitative test of the theoretical prediction for the rate of change of the orbital period [see Eq. (5.90)], we need to know the masses of the two components.

The values of the masses can be extracted from the measured values of the two relativistic effects in the system: the precession of the periastron and the time dilation.

The periastron precession depends only on the sum $m_1 + m_2$ of the masses, whereas the time dilation depends on a different combination of m_1 and m_2.

Thus, the measured values of the periastron precession and the time dilation (see Table 5.2) determine both m_1 and m_2:

$$\begin{aligned} m_1 &= (1.4398 \pm 0.0002) M_\odot \quad \text{for pulsar} \\ m_2 &= (1.3886 \pm 0.0002) M_\odot \quad \text{for companion} \end{aligned} \tag{5.91}$$

With these values of the masses and with a minor correction for a relative acceleration between the binary pulsar system and the Solar System, the theoretical prediction for the rate of change of the orbital period is 0.999 ± 0.002 times the observed rate of change, so these rates of change are consistent to within the observational error.

shows a plot of the observed data vs. the predicted shift of the time of periastron passage over a span of 30 years; this plot illustrates the excellent agreement between the theoretical prediction for the decrease of the period and the observed decrease.

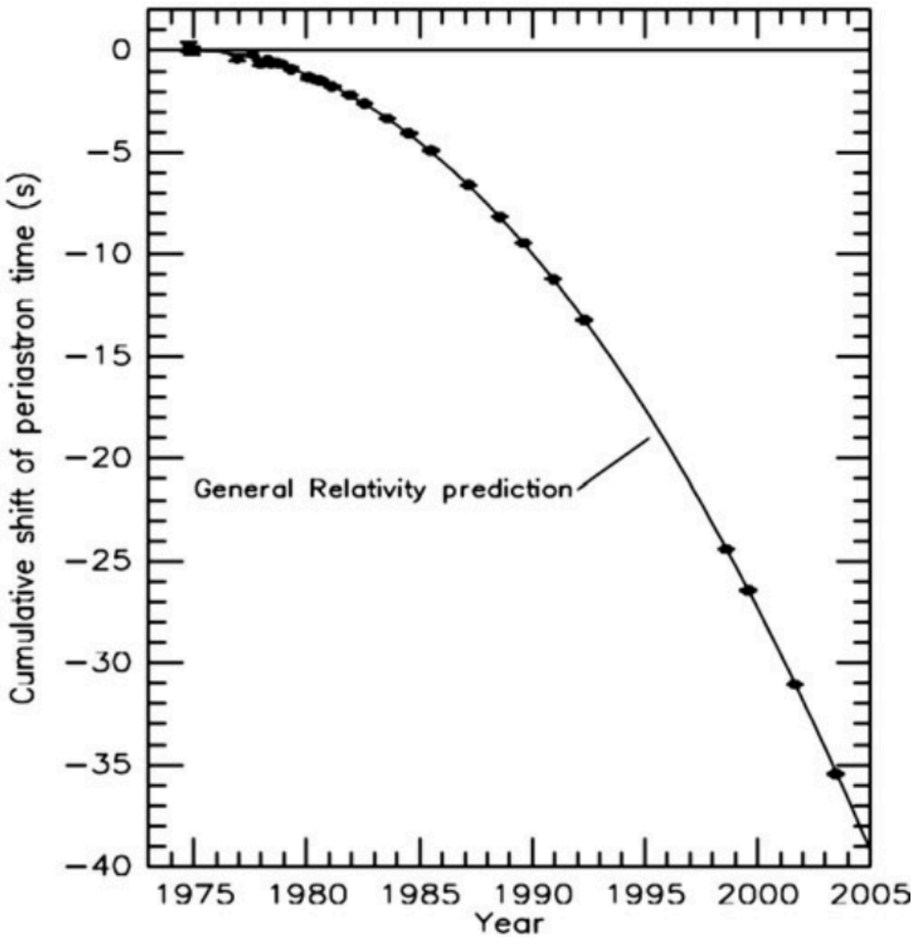


Fig. 5.13 Observed shift of the time of periastron passage (dots) and predicted shift (solid line). Since the period of the orbital motion is decreasing, the system reaches periastron early, which corresponds to a negative time shift. The magnitude of this shift increases from year to year. (From Weisberg and Taylor, 2004)

Note that for this comparison of theory and observation we cannot use the simple formula (5.90), because this formula applies only to circular orbits, whereas the orbit of PSR 1913+16 is elliptical, with a substantial eccentricity ($e = 0.617$).

For an elliptical orbit, gravitational radiation is emitted not only at the fundamental frequency, or the orbital frequency,[5] but also at multiples of this fundamental frequency; this occurs because the x and y coordinates of a mass in an elliptical orbit are not simple harmonic functions, but are harmonic series (Fourier series), which include all multiples of the orbital frequency.

In the case of PSR 1913+16, the most power is emitted at eight times the orbital frequency, but the largest wave amplitude occurs at four times the orbital frequency.
An additional complication is that the gravitational fields in the system are fairly strong. With $M = 1.44M_{\odot}$ and $r = 10$ km, which is a typical radius for a neutron star, we find $GM/rc^2 \cong 0.2$ at the surface of the star.

It therefore becomes necessary to contemplate corrections to the linear approximation on which the quadrupole formula for gravitational radiation was based. The required theoretical corrections have been explored in great detail by Damour with the conclusion that the quadrupole formula is actually more accurate than expected – it correctly gives the power radiated in the form of gravitational waves even when the background gravitational fields are fairly strong.

The excellent agreement between the observed and the predicted rates of decrease of the orbital period of PSR 1913+16 provides us with good – although circumstantial – evidence for the existence of gravitational radiation.

More evidence of the same kind is available from observations on three other binary pulsar systems, although with somewhat lower precision.

For instance, in the two-pulsar binary system PSR J0737−3039 the theoretical prediction for the rate of change of the orbital period attributed to gravitational radiation is 1.00 ±0.01 times the observed rate of change.

At present, such observations of energy losses from binary pulsar systems are the only empirical evidence for the existence of gravitational radiation.

The extreme precision with which the parameters of PSR1913+16 are known permits us to use this pulsar to test for some other relativistic and cosmological effects that might affect the rotational motion or the orbital motion.

For instance, as already mentioned in Section 4.7, the relativistic gravitational theory predicts gravitomagnetic effects, and the data from the binary pulsar provide us with observational evidence for these gravitomagnetic effects and confirm gravitational spin-orbit coupling.

Furthermore, according to some cosmological speculations, the gravitational constant G might gradually decrease, with a characteristic time scale equal to the age of the universe.

Such a decrease in the strength of gravitation would result in a gradual expansion of the orbits of all gravitationally bound systems, and a consequent gradual increase of the orbital periods.

Since the binary pulsar system displays no residual, unexplained changes in its period, we can set a limit on the rate of change of G.

The limit consistent with the observational uncertainties is $|dG/dt|/G < 2\times10^{-12}/y$; similar limits have been set by lunar laser ranging.

The orbital motion of a binary system provides us with the simplest example of emission of gravitational waves by a rotating quadrupole.

But the rotational motion of almost any system consisting of particles orbiting about a center or an asymmetric body spinning about an axis also will produce a time-dependent quadrupole moment and emit gravitational waves.

In this connection, the spinning motion of neutron stars is of considerable astrophysical interest.

The spin period of neutron stars (pulse period of pulsars) is typically in the range from 0.03 to 3 s.

If such a spinning neutron star is endowed with some deviation from exact cylindrical symmetry about the axis of rotation, then it constitutes a rotating quadruple, which will emit gravitational radiation.

For instance, the Crab pulsar rotates with a period of 0.033 s, and if the dimensions in the equatorial plane differ by 1 part in 1,200, then gravitational radiation would be emitted at a rate of 10^{38} erg/s.

This energy loss would account for the observed slowing down of the rate of spin observed for the Crab pulsar.

However, since it is known that the Crab emits large amounts of electromagnetic energy, most of the loss of rotational kinetic energy is already accounted for, and the emission rate of gravitational energy (and the deformation in the equatorial plane) must be smaller.

A search for gravitational waves from the Crab pulsar was attempted with a detector tuned to the expected wave frequency, $2 \times (1/0.033)$ Hz = 60 Hz. The results were negative (see Section 5.7).

5.6 Emission of bursts of gravitational radiation

We saw in the preceding section that, because of the loss of energy by gravitational radiation, the orbits of the two stars in a binary system gradually shrink – the stars gradually spiral toward each other.

This inward motion is slow at first, but as the stars fall into smaller and smaller orbits, of higher and higher frequency, the inward motion accelerates drastically.

The stars spiral toward each other faster and faster, emitting a gravitational wave of increasing amplitude and increasing frequency (an upward screech, crescendo and glissando, sometimes called a “chirp”).

This swan song of the binary system comes to a sudden end when the stars collide.

The collision releases a final blast of gravitational radiation.

The radiation emitted during this coalescence of a binary system is initially periodic or quasiperiodic, but the radiation emitted during the last few revolutions is burstlike.

For example, consider the coalescence of a system of two neutron stars, initially at a distance of, say, $R_0 = 100$ km.

If the mass of each star is 1 $M_\odot$, the radiated power is initially 3.3×10^{51} erg/s, and the rate of decrease of the distance is $-dR/dt|_0 = 2.5 \times 10^6$ cm/s.

Integrating Eq. (5.89), we obtain

$$\frac{1}{4}(R^4 - {R_0}^4) = -\frac{64G^3}{5c^5} m_1 m_2 (m_1 + m_2)(t - t_0) \tag{5.92}$$

which we can also write as

$$t - t_0 = \left(1 - \frac{R^4}{{R_0}^4}\right) \frac{R_0}{4} \frac{1}{(-dR/dt|_0)} \tag{5.93}$$

where the subscripts 0 indicate quantities evaluated at the initial time t_0.

The time interval needed for a decrease in the distance from R_0 to $R \ll R_0$ is therefore

$$\Delta t \cong \frac{R_0}{4} \frac{1}{(-dR/dt|_0)} \tag{5.94}$$

This result is (approximately) independent of the final distance R, because the system spends most of its time at the larger distances and passes through the smaller distances very quickly.

If we insert the values of R_0 and $dR/dt|_0$ given earlier, we obtain $\Delta t \cong 1.0$ s.

A simple calculation shows that in this time interval the binary system completes ≈ 130 revolutions; the shape of the orbit is therefore that of a tightly wound spiral.

The frequency of the radiation increases with time; initially, the frequency is ≈ 160 Hz, but toward the end it is ≈ 10 times as large.

The radiated power increases even more drastically; toward the end it is ≈ 3000 times as large as initially.

The total amount of energy emitted during the coalescence is equal to the change in the total energy of the system,

$$\Delta E = \frac{G}{2} m_1 m_2 \left(\frac{1}{R_0} - \frac{1}{R} \right) \tag{5.95}$$

With $R_0 = 100$ km and $R = 20$ km, the emitted energy amounts to $-\ \Delta \mathrm{E} = 5.3 \times 10^{52}$ erg, which is 1% of the rest-mass energy of the system.

Most of this energy is emitted during the last stages of the coalescence of the binary system.

An extra pulse will be emitted when the neutron stars collide; furthermore, the system that forms after the collision (a larger neutron star or a black hole) may emit more gravitational radiation while it settles into a stationary configuration.

The pulse of radiation emitted by a coalescing binary neutron-star system somewhere in our galaxy or in a neighboring galaxy would be of sufficient intensity to be detected with presently available detectors; however, such events are expected to be rare.

Note that the preceding calculation is only approximate; it ignores orbital eccentricity and relativistic corrections (the eccentricity plays only a small role, because the emission of gravitational radiation tends to decrease the eccentricity, so the final stages of the binary system involve nearly circular orbits).

At $R_0 = 100$ km, the orbital speed is $\mathrm{v/c} \cong 1/10$, and by the time the distance reaches $R = 20$ km, the speed is even larger, $\mathrm{v/c} \cong 2/10$.

It is also necessary to take into account corrections arising from the nonlinear behavior of the gravitational fields.

Our simple calculation must be regarded merely as an order of magnitude estimate.

The sources of gravitational radiation we have mentioned so far are periodic or quasiperiodic, and they are based on vibrating or rotating quadrupoles.

Since any quadrupole with a nonzero third time derivative will radiate, there are many other possible sources of gravitational radiation.

For example, consider a particle that is accelerated by falling directly toward a star along a radial line (an orbit of zero angular momentum).

For simplicity, assume that the mass M of the star is large compared with the mass m of the particle.

This implies that the star remains at rest and only the particle contributes a time-dependent quadrupole moment.

If we take the z-axis along the direction of incidence and the origin at the center of the star (see Fig. 5.14),

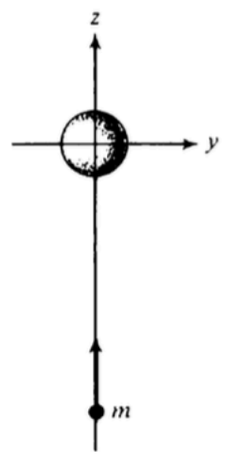


Fig. 5.14 Particle falling radially toward a mass.

then Eq. (5.75) gives

$$-\frac{dE}{dt} = \frac{2Gm^2}{15c^5}(6\dot{z}\ddot{z} + 2z\dddot{z})^2 \tag{5.96}$$

If the particle starts at z = −∞ with zero velocity (orbit of zero energy), then

$$\frac{1}{2}m\dot{z}^2 = \frac{GmM}{|z|} \tag{5.97}$$

and

$$\dot{z} = \frac{1}{|z|^{1/2}}(2GM)^{1/2}, \quad \ddot{z} = \frac{GM}{z^2}, \quad \dddot{z} = \frac{2GM}{|z|^{7/2}}(2GM)^{1/2} \tag{5.98}$$

Equations (5.96) and (5.98) give

$$-dE = \frac{1}{|z|^{9/2}}\frac{2Gm^2}{15c^5}(2GM)^{5/2}dz \tag{5.99}$$

The energy radiated during the plunge from z = −∞ to z = −R is obtained by integrating Eq. (5.99):

$$-\Delta E = \frac{1}{|R|^{7/2}}\frac{4Gm^2}{105c^5}(2GM)^{5/2} \tag{5.100}$$

Obviously, the radiated energy is large when the lower limit *R* is small.

It is therefore advantageous to consider the case of a very compact star or other object.

The most compact object that we can think of is a black hole.

We will discuss these objects later; for now, we need to know only that a black hole of mass *M* has a radius

$$R = \frac{2GM}{c^2} \tag{5.101}$$

The gravitational fields at this radius are so intense that any particle (or light or whatever) that approaches the object nearer than $2GM/c^2$ is unavoidably pulled into the black hole and can nevermore escape

For the case of a particle falling directly into such a black hole, the lower limit on R is that given by Eq. (5.101), because any gravitational radiation emitted afterward will remain within the black hole. With this value of R, Eq. (5.100) yields

$$-\Delta E = \frac{2}{105} mc^2 \times \frac{m}{M} = 0.019 mc^2 \times \frac{m}{M} \tag{5.102}$$

Our simple calculation is defective in that we have failed to take into account that the motion becomes relativistic as the particle approaches $R = 2GM/c^2$.

Also, nonlinear gravitational effects must be considered.

A more precise calculation shows that

$$-\Delta E = 0.0104 mc^2 \times \frac{m}{M} \tag{5.103}$$

which is about half as large as our simple approximate result (5.102).

Note that the amount of gravitational-wave energy radiated *decreases* with M.

To obtain a large pulse of radiation we must choose m as large and M as small as we dare (remember that we assumed $M \gg m$ in our calculation).

If we consider a "particle" of mass $m = M_\odot$ and a black hole of mass $M = 10M_\odot$ and radius

$R = 2GM/c^2 \cong 30\text{km}$, we obtain $-\Delta E = 2 \times 10^{51}$ erg.

Most of this radiation is emitted during the final stages of the plunge, say, when the particle falls from $2R$ to R.

The radiation therefore appears as a pulse of length

$$\Delta t \cong R/(\text{particle velocity}) \cong R/c \cong 30\ \text{km}/c \cong 10^{-4}\ \text{s} \tag{5.104}$$

Such a pulse will contain frequencies of up to 10^4 Hz.

The results of a calculation of the frequency spectrum are shown in Figure 5.15.

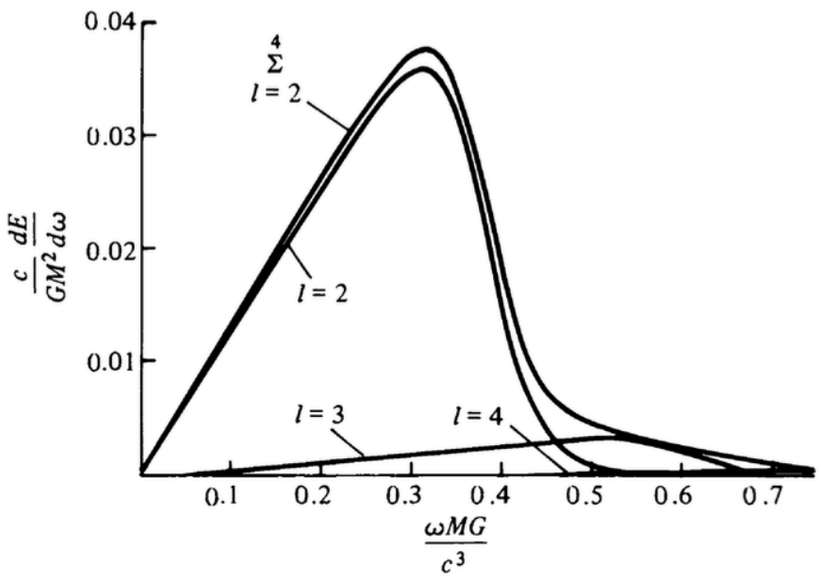


Fig. 5.15 Spectrum of the gravitational radiation emitted by a particle of mass m falling radially into a black hole of mass M. The quantity $dE/d\omega$ gives the amount of energy radiated per unit frequency interval. The curve marked $l = 2$ corresponds to quadrupole radiation; the other curves ($l = 3, l = 4$) correspond to multipole radiation of higher order. Note that most of the radiation is emitted with frequencies below $\omega \cong 0.5c^3/GM$. (From Davis et al., 1971)

As a final example of a source of gravitational radiation we mention gravitational collapse of supernovas (type II supernovas).

When a star of several solar masses evolves into a state of high density, the gravitational field can become so intense that the interior pressure is unable to support the weight of the outer layers of the star.

The core of the star crushes itself in its own gravitational field and collapses on itself, until it reaches nuclear densities, when the collapse suddenly slows or stops.

This sudden stop generates a shock wave that travels outward and rips off the outer layers of the star in a tremendous explosion, which we see as a supernova.

Meanwhile, either the core attains a stable configuration as a neutron star, or if the core is too massive, it continues to collapse and ultimately forms a black hole.

Whether gravitational radiation is emitted during the gravitational collapse depends on the presence of an asymmetry.

If the collapsing star remains spherically symmetric, it has no quadrupole moment and will not emit gravitational radiation.

However, most stars rotate, and their symmetry is cylindrical, not spherical.

During the gravitational collapse, the pole and the equator of such a rotating star collapse at different speeds, and they may even oscillate in and out when the core reaches nuclear densities.

This produces a time-dependent, possibly even an oscillating, quadrupole moment.

Furthermore, the neutron star formed at the end of the collapse is likely to be in some excited vibrational state.

The amount of gravitational energy radiated by the gravitational collapse of a supernova has been estimated at about $10^{-6}M_{\odot}c^2$, although it might be considerably larger in the case of a fast-rotating star, which might break up into several, highly asymmetric fragments during the collapse.

The pulse of gravitational radiation contains a range of frequencies, from 10^3 to 10^4 Hz.

Another kind of gravitational collapse might occur in a dense galactic nucleus, where a cluster of stars might merge to form a large black hole, of a mass between 10^3 and $10^7M_{\odot}$.

Such a collapse could generate low-frequency gravitational waves, between 10 and 10^{-4} Hz.

Table 5.3 is a summary of typical astrophysical sources of gravitational radiation and of the corresponding frequency ranges and amplitudes of the gravitational waves reaching the Earth.

For each kind of bursting source, the distance listed in this table has been selected large enough to yield an estimated three events per year.

Table 5.3 Typical astrophysical sources of gravitational radiation*

Source	Frequency	Distance**	Amplitude at Earth, κA
Periodic sources			
Binaries	10^{-4} Hz	10 pc	10^{-20}
Nova	10^{-2} to 1	500 pc	10^{-22}
Spinning neutron star (Crab)	60	2 kpc	$<10^{-24}$
Bursting sources			
Coalescence of binary	10 to 10^3	100 Mpc	10^{-21}
Infall of star into 10 $M_\odot$ b.h.	10^{-4}	10 Mpc	10^{-21}
Supernova	10^3	10 kpc	10^{-18}
Gravitational collapse of 10^4 $M_\odot$ b.h.	10^{-1}	3 Gpc	10^{-19}

* Based on Thorne (1987).
** Distances have been selected large enough to yield approximately three events per year.

5.7 Detectors of gravitational radiation

Detectors of gravitational radiation were first built by Weber at the University of Maryland, starting in 1966.

These detectors consisted of massive cylinders of aluminum, typically about 1,000 or 2,000 kg, in which any incident gravitational waves would excite quadrupole oscillations.

The resonant frequency for the longitudinal mode of oscillation of such a cylinder was typically somewhat above 1,000 Hz.

The cylinder was suspended inside a vacuum tank by a wire wrapped around its middle.

The beam supporting the wire rested on stacks of rubber and metal blocks, which provided good isolation against external mechanical vibrations.

The oscillations of the cylinder in its fundamental longitudinal mode were measured by piezoelectric strain transducers bonded to the surface of the cylinder around its middle.

Weber's early detectors operated at room temperature, $T \cong 300$ K, and they were severely limited by their internal thermal noise.

Calculations of their sensitivity indicate that their threshold for detection of a gravitational-wave energy flux and amplitude was $t_{(0)}{}^{03} \geq 10^4$ erg/cm$^2 \cdot$ s and $\kappa A \geq 10^{-20}$, respectively.

These estimates of the sensitivity apply only to a wave whose frequency matches the resonant frequency of the detector.

There are no known astrophysical sources of periodic waves in the kilohertz frequency range, and therefore these estimates are of doubtful relevance.

However, as mentioned in Section 5.4, spinning neutron stars might emit some gravitational radiation with a frequency of 10 to 100 Hz (twice the rotation frequency), and it is feasible to build Weber detectors deliberately tuned to that wave frequency.

One such detector was built to match the frequency of (supposed) gravitational waves emitted by the Crab pulsar.

This detector attained a sensitivity of $\kappa A \cong 7 \times 10^{-23}$, and at this level, it did not detect any emissions from the Crab.

Apart from this exceptional instance of a detector tuned to an astrophysical source of known frequency, we cannot rely on a match of the wave frequency and the resonant frequency of the detector.

The best we can hope for is that an astrophysical source produces a short pulse of radiation that contains a spread of frequencies that overlaps the resonant frequency.

The sensitivity of a Weber detector for such a wave is then much worse.

To reach the detection threshold, the gravitational wave pulse must carry an adequate amount of power per unit area per unit frequency interval at the resonant frequency.

Since the power carried by a gravitational wave is proportional to the square of its amplitude κA, the detection threshold can also be expressed as an amplitude per square root of unit frequency interval.

Thus, the sensitivity of detectors is usually quoted in units of $1/(\mathrm{Hz})^{1/2}$.

For instance, the sensitivity of a Weber detector is about $10^{-16}/(\mathrm{Hz})^{1/2}$, which means that Fourier transform of the strain amplitude squared must be $10^{-32}/\mathrm{Hz}$ or larger.

Weber operated several detectors in coincidence, to discriminate between true gravitational signals and instrumental noise. In 1968 he reported signals, with an energy in excess of noise, in his detectors.

He claimed that these signals were arriving at the rate of a few per day, and he claimed that there was some indication that their direction of arrival correlates with the direction of the galactic center.

These sensational claims encouraged other experimenters to seek confirmation of Weber's discoveries.

Several resonant detectors of the Weber type, but of somewhat higher sensitivity, were built in quick succession, but none detected any pulses of gravitational radiation.

The weight of the evidence suggests that Weber's apparent coincidences were the result of faulty data analysis, which permitted chance coincidences ("accidentals") to masquerade as real coincidences.

By now, several groups of experimenters have built larger, much more sensitive resonant detectors (see Table 5.4 and Fig. 5.16).

These detectors attain sensitivities of up to $10^{-21}/(\text{Hz})^{1/2}$, resulting from the use of a new aluminum alloy in most cases, better electronic sensors, low-noise amplifiers, and low-temperature operation.

All the new detectors listed in Table 5.4 operate at a temperature near that of liquid helium; however, their effective temperatures are even lower, as low as a few mK.

None of the detectors has found any gravitational wave signals so far.

Table 5.4 Resonant detectors of gravitational radiation*

Detector	Mass (kg)	Resonant Frequency (Hz)	Temperature (K)	Amplitude Sensitivity $(1/(\text{Hz})^{1/2})$
Weber	1,400	1,660	290	10^{-16}
ALLEGRO (Louisiana, USA)	2,296	916	4.2	10^{-20}–10^{-21}
AURIGA (Legnaro, Italy)	2,230	875	4.5	same
EXPLORER (CERN)	2,270	907	2.6	same
NAUTILUS (Frascati, Italy)	2,260	936	2.5	same

* Based on Ronga (2006). As of 2012, only AURIGA and NAUTILUS are still in operation.



Fig. 5.16 The Auriga detector at Legnaro, near Padova. (Courtesy of Auriga Collaboration, INFN – Laboratori Nazionali di Legnaro)

Other groups of experimenters have built interferometric detectors, which measure the displacement that an incident gravitational wave produces between a pair of free masses, rather than masses joined by springs, as in the quadrupole detectors.

For a pair of free masses, the gravitational wave produces the displacement given by Eqs. (5.40) and (5.41).

In practice, free masses can be simulated by suspending the masses from wires, as pendulums.

If the frequency of the pendulum is much lower than the frequency of the gravitational wave, then the suspended mass will behave like a free mass (as is well known, a harmonic oscillator behaves like a free particle if the resonant frequency is much lower than the frequency of the driving force).

The pendulum suspension typically has a frequency of 1 Hz, and for free behavior of the mass the gravitational-wave frequency must therefore be substantially larger than 1 Hz.

The displacement of the suspended masses is detected by interferometric methods, based on the principle of the Michelson interferometer.

Figure 5.17 shows the basic features of the experimental arrangement in a simplified form.

Mirrors are mounted on the free masses at the ends of the arms, and a beam of light, supplied by a laser, is reflected along the arms and then allowed to interfere with itself.

In the LIGO detectors (see final part of this section) each arm of the interferometer actually consists of a resonant Fabry-Perot optical cavity, with a highly reflecting mirror mounted on the free mass at the far end and a partially transmitting input mirror at the near end; this arrangement makes the light beam bounce back and forth along the arm many times, thereby multiplying the effective path length and sensitivity of the interferometer.

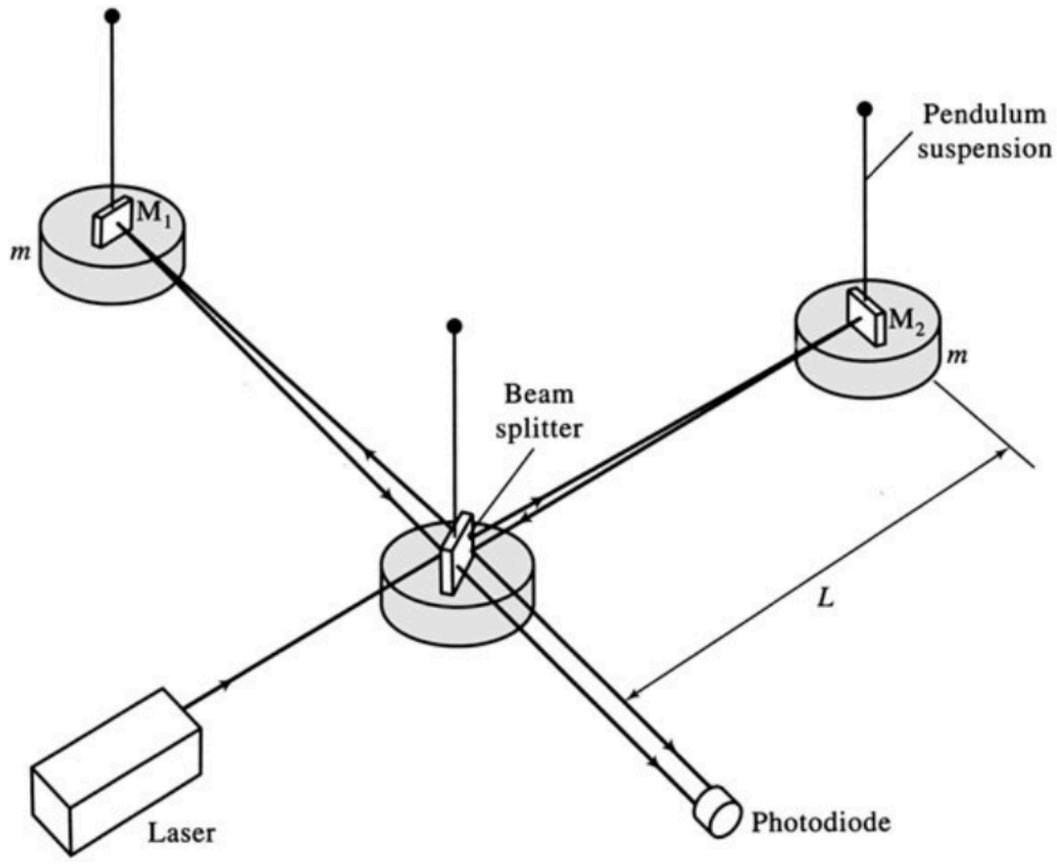


Fig. 5.17 A simple Michelson interferometer for the detection of gravitational waves. The mirrors M_1 and M_2 at the ends of the arms are suspended as pendulums.

An incident gravitational wave displaces the end mirrors in the two arms by different amounts and thereby generates a phase shift between the emerging light beams.

The resulting change of intensity of the combined beams is measured by a photodetector.

The measurement must be completed within a half-cycle of the gravitational wave, because after a half-cycle the wave will displace the end mirrors in opposite directions, and the phase shift between the interfering light beams will return to zero.

For high sensitivity, the arms of the interferometer must be made long, either in actual length or in effective length, by multiple back-and-forth reflections in each arm.

Furthermore, an intense laser beam is needed, so that even a quite small phase shift (and a faint change of intensity) becomes detectable.

The main limitation on the sensitivity of the interferometer arises from fluctuations in the interference fringes due to random fluctuations in the number of photons.

If the number of photons participating in the length measurement is N_p, then the uncertainty in this number is $N_p^{1/2}$. This implies an uncertainty in the phase of the light wave according to the uncertainty principle for phase and number: $\Delta\phi = 1/\Delta N_p = 1/N_p^{1/2}$. Along the light wave, the length increment per unit phase increment is $\lambda_p/2\pi$, where λ_p is the wavelength of light. Hence the uncertainty in the interferometric length measurement resulting from the phase uncertainty $\Delta\phi$ is

$$\Delta l = \frac{\lambda_p}{2\pi} \times \Delta\phi = \frac{\lambda_p}{2\pi N_p^{1/2}} = \frac{c}{2\pi \nu_p N_p^{1/2}} \tag{5.105}$$

If the measurement lasts a time τ and the laser has an effective power P_p, the number of available photons is

$$N_p \cong \frac{\tau P_p}{h\nu_p} \tag{5.106}$$

With this, Eq. (5.105) becomes

$$\Delta l = \sqrt{\frac{c^2\hbar}{2\pi \nu_p \tau P_p}} \tag{5.107}$$

If the effective length of the interferometer is L, including back and forth reflections within the arms, then the fractional uncertainty in length is $\Delta l/L$, and this equals the minimal detectable gravitational wave amplitude κA:

$$\kappa A = \frac{\Delta l}{L} = \sqrt{\frac{c^2\hbar}{2\pi \nu_p \tau P_p L^2}} \tag{5.108}$$

Evidently, to achieve the best sensitivity it is desirable to make the length of the arms as large as possible.

However, we have to impose the constraint that the light in the interferometer must complete its travel within a time shorter than half the period of the gravitational wave, because after half a period the stretch of the arms reverses and the interference effects in the interferometer change sign.

Thus the maximum acceptable arm length in Eq. (5.108) is $L = \lambda/2$.

Furthermore, for a gravitational-wave pulse, we have to impose the constraint that the measurement time τ be shorter than the pulse lifetime, so the measurement can be completed before the pulse fades away.

The pulse lifetime is of the order of magnitude of the inverse of the frequency spread of the pulse; for a typical burst, with a broad frequency spectrum (for instance, see Fig. 5.15), the frequency spread is of the order of magnitude of the average frequency.

Hence we can insert $\tau \cong 1/\nu$ in Eq. (5.108), and we obtain the simple result,

$$\kappa A = \sqrt{\frac{2\hbar\nu^3}{\pi\nu_p P_p}} \tag{5.109}$$

Current interferometric detectors use laser light with a frequency $\nu_p = 3 \times 10^{14}$ Hz and an effective power (delivered into the interferometer at the beam splitter) of $P_p = 200$ watt.

For a gravitational wave of frequency $\nu = 100$ Hz, Eq. (5.109) then gives $\kappa A \cong 10^{-22}$.

To express this in the conventional units $1/(\text{Hz})^{1/2}$, we must divide by the square root of the frequency spread, $\sqrt{\Delta\nu} \cong \sqrt{\nu} = 10(\text{Hz})^{1/2}$, so the sensitivity would be reported as $\kappa A \cong 10^{-23}/(\text{Hz})^{1/2}$.

The first interferometric detector, of modest sensitivity, was built by Forward and Moss, followed by several interferometric detectors of gradually improved sensitivity.

Larger interferometers – optimistically called gravitational-wave observatories – came into operation in 2000.

These large interferometers are listed in Table 5.5.

Table 5.5 Interferometric detectors of gravitational radiation*

Detector	Size	Frequency Range (Hz)	Amplitude Sensitivity $(1/(\text{Hz})^{1/2})$
LIGO (Hanford, WA)	4 km and 2 km	60–1,000	10^{-22}
LIGO (Livingston, LU)	4 km	60–1,000	10^{-22}
TAMA (Tokyo)	300 km	700–1500	10^{-20}
GEO (Hannover, Germany)	600 m	100–4,000	10^{-20}
VIRGO (Pisa, Italy)	3 km	40–1,000	10^{-22}
KAGRA (Kamioka, Japan)**	3 km, underground	20–1,000	10^{-24}
LISA***	5×10^6 km, in orbit	10^{-4}–0.1	10^{-23}

* Based on GWIC Roadmap Report (2010).
** Under development.
*** Under development; listed sensitivity assumes a periodic source.

The LIGO detectors (Laser Interferometric Gravitational-Wave Observatories) at Hanford and Livingston have arm lengths of 4 km (see Fig. 5.18), and they attain sensitivities of the order of $10^{-22}/(\text{Hz})^{1/2}$ for gravitational waves of a frequency of about 50 Hz to 1,000 Hz.

Thus, these interferometric detectors are much more sensitive than the available resonant detectors.

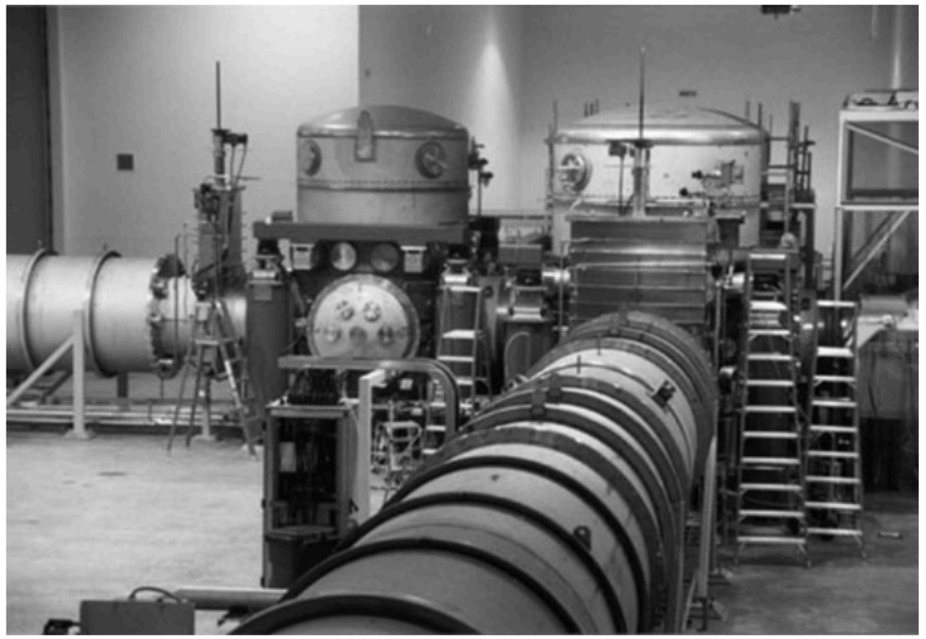


Fig. 5.18 Vacuum chambers enclosing the interferometer arms of the LIGO detector at Livingston, Louisiana. Each of the horizontal vacuum chambers (left and bottom) is 4 km long. (Credit: Jonathan Kern)

Improvements of the LIGO detectors are in progress.

Instead of the simple pendulum suspension of the original LIGO system, the improved, "advanced" LIGO system now being assembled uses compound pendulums, with four masses, each suspended from the mass above it (see Fig. 5.19).

This quadruple pendulum arrangement provides good isolation against seismic noise.

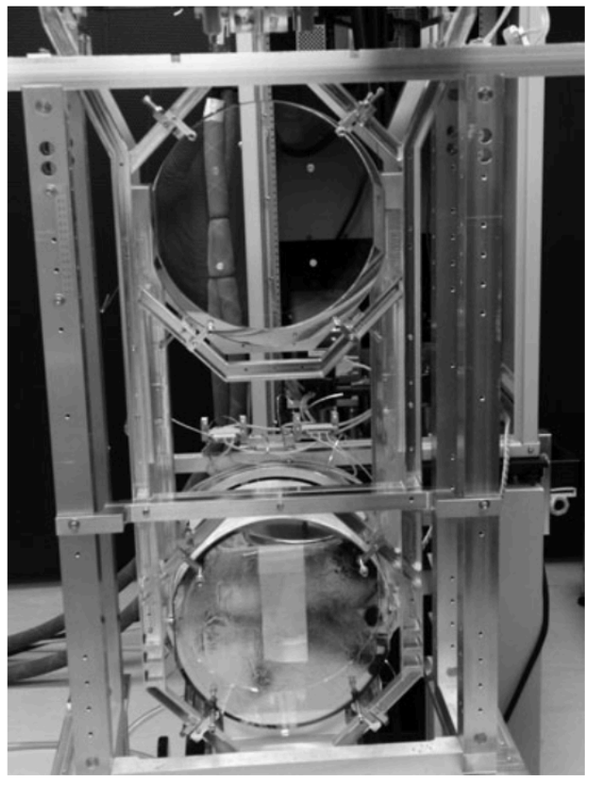


Fig. 5.19 Pendulum mirror suspension for the advanced LIGO detector. The 40-kg mirror at the bottom is the test mass of the detector. It is suspended from the "penultimate" mass (a circular glass plate) by vertical carbon fibers, 60 cm long (hidden in the vertical casings). Two more masses will be attached above, to form a quadruple pendulum. The diameter of the mirror is about 30 cm. (Credit: University of Glasgow)

Enhancements of sensitivity by a factor of 10 are planned, and it is expected that this will result in the detection of bursts of gravitational radiation from the coalescence of neutron-star and black-hole binaries.

The estimated rate of such events, within the threshold of sensitivity of the improved LIGOs, is several dozen per year.

An absence of detected events would be hard to explain, for experimenters and for astrophysicists.

An even larger interferometric detector, for deployment in an orbit around the Sun is under development.

This LISA detector (Laser Interferometer Space Antenna) will consist of three spacecraft placed in orbits arranged so as to keep these spacecraft at the corners of an equilateral triangle, 5×10^6 km on a side.

This detector will be sensitive to gravitational waves of low frequency (down to 10^{-4} Hz), such as might be generated by pairs of black holes of large mass, 10^5 to 10^7 $M_\odot$, orbiting around each other.

Efforts are also underway to detect gravitational waves by fluctuations in pulsar-timing measurements.

This involves monitoring the pulsing times of a large sample of pulsars by means of several arrays of radio telescopes.

Any coherent, joint phase fluctuations in the observed pulse timings can be interpreted as caused by a gravitational wave passing through the space surrounding the Earth.

This method is sensitive only to gravitational waves of very low frequency, around 10^{-9} Hz.

Available pulsar data have already placed a limit of about $\kappa\ A < 10^{-13}$ on the amplitude of such low-frequency gravitational waves.

By comparison with other detection techniques, this is a rather poor constraint on the amplitude.

But the energy density of these waves is proportional to the square of the amplitude multiplied by the square of the frequency, and the low frequency permits us to place a tight limit on the energy density of these waves, despite the poor amplitude constraint – the energy density of these waves is only about 10^{-8} times the "critical" energy density for the universe (see Chapter 9).

6 - Riemannian Geometry

Mathematics? Einstein did not know any mathematics. — Werner Heisenberg to Remo Ruffini

The linear tensor theory of gravitation that we developed by analogy with electrodynamics started out as the theory of a tensor field in a flat spacetime background.

The geometric interpretation of this tensor field emerged only as an afterthought.

However, the analysis of spacetime measurements (with clocks for time measurements and also for distance, by means of the radar-ranging procedure) has shown us that the flat spacetime background is purely fictitious – in a gravitational field, the real geometry measured by our instruments is the geometry of a curved spacetime, that is, the geometry of a Riemannian spacetime.

Mathematically, a Riemannian space is a differentiable manifold endowed with a topological structure and a geometric structure.

In the discussion of the geometric structure of a curved space we must make a distinction between the *affine* geometry and the *metric* geometry.

These two kinds of geometries correspond to two different ways in which we can ascertain the curvature of a space.

One way is by examination of the behavior of parallel line segments, or parallel vectors.

For example, on the surface of a sphere, we can readily detect the curvature by transporting a vector around a closed path, always keeping the vector as parallel to itself as possible.

Figure 6.1 shows what happens if we parallel-transport a vector around a "triangular" path on the sphere.

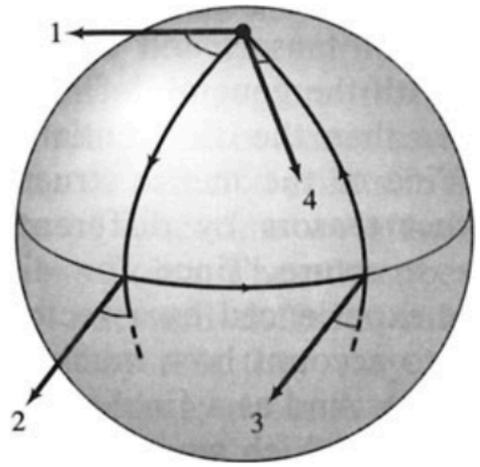


Fig. 6.1 Parallel transport of a vector. We start at the North Pole and parallel-transport the vector along a meridian to the equator, then some distance along the equator, and then return to the North Pole along another meridian. (To visualize the orientation of the vectors 1, 2, 3, 4 keep in mind that the tails of these vectors are tangent to the sphere.)

The final vector differs in direction from the initial vector, whereas on a flat surface the final vector would not differ.

Such changes in a vector produced by parallel transport characterize the affine geometry (the word *affine* means *connected* and refers to how parallels at different places are connected, or related).

Another way in which we can detect curvature is by measurements of lengths and areas.

For example, we can draw a circle on the sphere and check the radius vs. the circumference or the radius vs. the area.

Both the circumference and the area of such a circle are smaller than for a circle on a flat surface (see Fig. 6.2).

Such deviations from what we expect on a flat surface characterize the metric geometry.

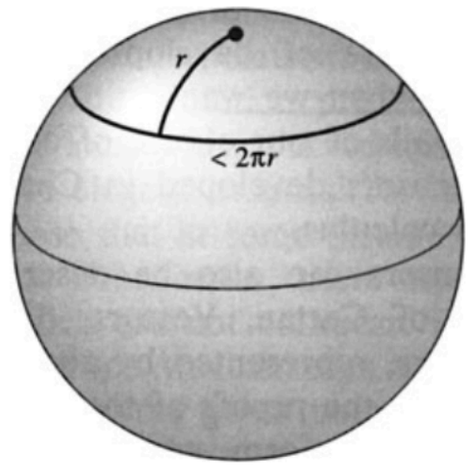


Fig. 6.2 A circle on the sphere has a circumference smaller than 2π times its radius. If the radius exceeds the distance from the pole to the equator, the circumference decreases with increasing radius.

Note that the lateral surface of a cylinder is a flat space, according to both of our criteria for curvature: Parallel transport of a vector along this surface does not alter its direction, and measurements of lengths and areas reveal no deviations from what we find on a flat surface (this can be easily seen by slicing the lateral surface of the cylinder open and spreading it out on a flat surface).

The lateral surface of a cylinder is said to have zero *intrinsic* curvature, but nonzero *extrinsic* curvature.

By the latter is meant how the surface curves into dimensions that lie outside the surface.

When we perform parallel transport or when we measure lengths and areas within some given curved surface, we can detect only the intrinsic curvature, not the extrinsic curvature.

Since we are confined to the four dimensions of spacetime, our measurements permit us to detect only the intrinsic curvature of spacetime, and so this is the only curvature we will deal with in our study of Riemannian geometry.

Before we can proceed with the study of "gravitation as geometry," we need to develop some mathematical concepts, in particular the concepts of tensors and tensor fields in a Riemannian spacetime.

We begin with the general definitions of tensors; these definitions require no more than the differential structure of spacetime – we do not need the affine or the metric structure.

As a next step, we will attempt to construct tensors by differentiation; we then need to introduce the affine structure, because for differentiation we have to calculate the change experienced by a vector during a displacement, and we need to take into account how much of this change is attributed to parallel transport.

And as a final step, we will want to calculate the lengths of vectors, for which we need to introduce the metric structure.

Throughout most of this chapter, we will describe tensors by the formalism of the absolute differential calculus of Ricci and Levi-Civita.

In this formalism, tensors are represented by their components relative to a general set of coordinates, such as spherical, cylindrical, or any other kind of curvilinear coordinates.

The tensor calculus is said to be *absolute* because it is independent of the details of the choice of coordinates; that is, the equations have the same form in all coordinate systems.

Thus, the details of the coordinate system need not be specified in advance for the purpose of developing the mathematics of tensor calculus (of course, at a later stage, when we want to do numerical computations, we need to specify the details of our choice of coordinates).

The tensor calculus for Lorentz tensors developed in Chapter 2 is a special case of the general tensor calculus.

Alternatively, tensors can also be described by the formalism of differential forms of Cartan.

Vectors, differential forms, and tensors are then operators, represented by abstract symbols, as in Section 2.6.

In many instances, the proofs of theorems of tensor calculus take a very elegant and concise form when expressed in the language of differential forms.

But in other instances this language results in excessive "orgies of formalism," and then component language is our salvation.

Furthermore, we must not forget that the physicist who wishes to measure a tensor has no choice but to set up a coordinate system and then measure the numerical values of the components.

Thus, to carry out the comparison of theory and experiment, the physicist cannot ultimately avoid the language of components – only a pure mathematician can adhere exclusively to the abstract, coordinate-free language of differential forms.

We will develop Riemannian geometry in the language of differential forms in Section 6.7, and in later chapters we will occasionally use this language, when convenient.

6.1 General coordinates and tensors

Unless the spacetime is flat, there is no good reason to use rectangular coordinates.

In flat spacetime, rectangular coordinates are preferred, because both the metric tensor ($g_{\mu\nu} = \eta_{\mu\nu}$) and the Lorentz transformations take their simplest form when expressed in these coordinates.

In a curved spacetime, the use of rectangular coordinates or some modified version of rectangular coordinates does not necessarily result in any simplification of the metric tensor, except in those regions that are far away from all gravitational fields, where spacetime becomes asymptotically flat.

We will find it expedient to use general coordinates to describe points in spacetime.

By this is meant that to each point of spacetime we attach a set of four numbers that serve to identify the point.

These numbers can be assigned according to any arbitrary, but well-defined, rule.

The coordinates should be assigned in such a way that they vary more or less continuously as we move through spacetime; a finite number of discontinuities, such as found in polar coordinates at $\phi = 2\pi$, will of course be acceptable.

On a sheet of paper it is easy to label a point by coordinates: We simply mark the point with an X and write the coordinate values next to the X.

In empty spacetime this becomes impossible, and we need a different labeling procedure.

One way to label spacetime points is to imagine all of space to be filled with an array of small clocks, each of which is labeled with the values of its spatial coordinates.

There is no need that these clocks be at "rest;" a continuous streaming will keep the coordinate values smooth and acceptable.

The coordinates of spacetime points in the empty space between adjacent clocks can be obtained by interpolation, using meter sticks or a radar-ranging procedure (the interpolation procedure can be performed as in flat spacetime, since locally any curved spacetime is approximately flat).

A somewhat more practical procedure for labeling spacetime points is the following: Suppose that at some distance from the region where we want to introduce coordinates, there are three "fixed" stars (red, white, and blue), which are not collinear.

For the spatial coordinates of a given point in space we can then adopt the three angles between the stars as seen from the given point (see Fig. 6.3).

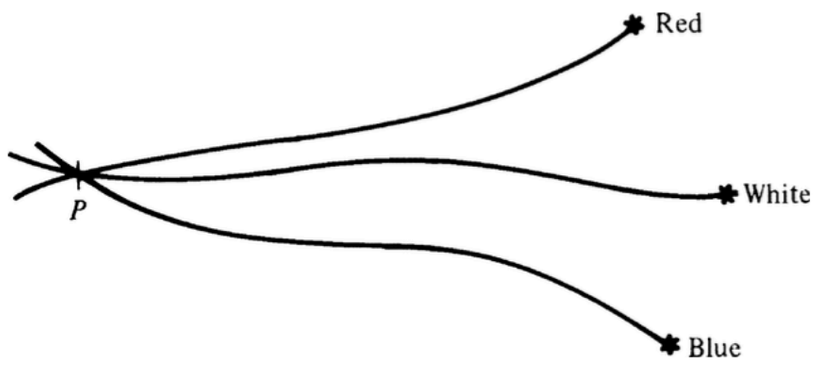


Fig. 6.3 Astrogator coordinates of a point *P* are obtained by measuring the angles between rays of light reaching the point from three stars.

For the time coordinate we can adopt the angle between, say, the red star and a fourth star (yellow), which is in motion with respect to red, white, and blue.

Note that it is not really necessary that the red, white, and blue stars be fixed; even if they are in motion, the coordinates will be well defined.

Also note that the bending of light by gravitational fields does no harm to these coordinates, since we are merely trying to find some well-defined procedure for associating four numbers with each given point of spacetime.

We might call the numbers obtained by this construction astrogator coordinates, since an astro-navigator would probably use coordinates of this kind.

These astrogator coordinates are intended merely as a demonstration that an operational definition of spacetime coordinates is possible.

In practice, our choice of coordinates will be governed by our desire to keep the solution of the field equations as simple as possible.

In the theory of the (electrostatic) potential, the advantages of a clever choice of coordinates are well known.

Unfortunately, in a curved spacetime the coordinates that lead to the simplest mathematics often lack any obvious physical interpretation, and it will often be possible to give an operational definition of the coordinate points only after we have solved the field equation for the metric tensor, and after we have calculated just how much the light signals used to locate coordinate points are deflected and retarded.

We already discussed one instance of this difficulty at the end of Section 4.4, in connection with Fig. 4.10.

We will use the notation $x^\mu = (x^0, x^1, x^2, x^3)$ for the general coordinates.

It is important to keep in mind that the index on the general coordinates is always an upper (contravariant) index.

Although there exists a general procedure for raising and lowering indices (it involves the metric tensor of the spacetime; see Section 6.4), this procedure gives nothing useful when applied to x^μ.

Thus, we will avoid using x_μ, except when dealing with simple rectangular coordinates in flat spacetime, as in earlier chapters.

Let us see what happens if we transform from one set of general coordinates to another.

Suppose that the old coordinates are x^μ and the new x'^μ.

The new coordinates are then functions of the old,

$$x'^\mu = x'^\mu(x) \tag{6.1}$$

and the old are functions of the new,

$$x^\mu = x^\mu(x') \tag{6.2}$$

Accordingly, the differentials of the coordinates obey the transformation law:

$$dx'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} dx^\nu \tag{6.3}$$

$$dx^\mu = \frac{\partial x^\mu}{\partial x'^\nu} dx'^\nu \tag{6.4}$$

In the case of the Lorentz transformations, the partial derivatives $\partial x^\mu/\partial x'^\nu$ were constants.

In the present case of general coordinate transformations, these partial derivatives are functions of x^α or x'^α.

Note that the transformations (6.3) and (6.4) are the inverses of one another, and therefore

$$\frac{\partial x'^\mu}{\partial x^\beta} \frac{\partial x^\beta}{\partial x'^\nu} = \delta^\mu_\nu \tag{6.5}$$

This also is obvious from the chain rule for derivatives.

The transformation laws for the differential operator ∂_μ are

$$\frac{\partial}{\partial x'^{\mu}} = \frac{\partial x^{\nu}}{\partial x'^{\mu}} \frac{\partial}{\partial x^{\nu}} \tag{6.6}$$

$$\frac{\partial}{\partial x^{\mu}} = \frac{\partial x'^{\nu}}{\partial x^{\mu}} \frac{\partial}{\partial x'^{\nu}} \tag{6.7}$$

In Chapter 2 we defined vectors and tensors by their behavior under Lorentz transformations.

Now we will define them by their behavior under general coordinate transformations.

These new definitions are natural generalizations of our earlier definitions.

The transformation laws for differentials and for differential operators serve as prototypes for the transformation equations of vectors.

Thus, contravariant vectors have the transformation law of dx^{μ}, covariant vectors that of ∂_{μ}:

A four-component object A^{μ} *is a contravariant vector under general coordinate transformations if it transforms according to*

$$A'^{\mu} = \frac{\partial x'^{\mu}}{\partial x^{\nu}} A^{\nu} \tag{6.8}$$

A four-component object B_{μ} *is a covariant vector if*

$$B'_{\mu} = \frac{\partial x^{\nu}}{\partial x'^{\mu}} B_{\nu} \tag{6.9}$$

Note that x^μ is *not a vector* with respect to general coordinate transformations: The transformation law for x^μ is given by Eq. (6.1), and this does not have the form of Eq. (6.8).

In the exceptional case of a linear and homogeneous transformation, Eq. (6.1) becomes

$$x'^\mu = b^\mu_\nu x^\nu \tag{6.10}$$

where $b_\nu{}^\mu$ is a constant matrix.

In this case, x^μ does have the transformation law (6.8), since Eq. (6.10) can then be written as

$$x'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} x^\nu \tag{6.11}$$

Thus, x^μ is a vector with respect to such linear transformations.

This is why, in Chapter 2, we were able to treat the rectangular coordinate x^μ as a vector with respect to Lorentz transformations.

Tensors of higher rank have the transformation laws

$$A'^{\alpha\beta\ldots\lambda} = \frac{\partial x'^\alpha}{\partial x^\mu}\frac{\partial x'^\beta}{\partial x^\nu}\cdots\frac{\partial x'^\lambda}{\partial x^\kappa} A^{\mu\nu\ldots\kappa} \tag{6.12}$$

and

$$B'_{\alpha\beta\ldots\lambda} = \frac{\partial x'^\mu}{\partial x^\alpha}\frac{\partial x'^\nu}{\partial x^\beta}\cdots\frac{\partial x'^\kappa}{\partial x^\lambda} B_{\mu\nu\ldots\kappa} \tag{6.13}$$

if they are contravariant and covariant, respectively.

We note that $\delta_\mu{}^\nu$ transforms as a tensor covariant in μ, contravariant in ν.

We must keep in mind that a multicomponent object is (or is not) a tensor with respect to a given class of transformations.

Thus, it is quite possible for a quantity to be a tensor with respect to Lorentz transformations, but not a tensor with respect to general coordinate transformations; the converse is of course not possible.

By the word *tensor*, without qualifiers, we will hereafter mean tensor with respect to general coordinate transformations.

Using Eq. (6.5), we can now easily show that

$$A^{\mu} B_{\mu} \tag{6.14}$$

is a scalar if A^{μ} and B_{μ} are vectors, and that

$$A^{\mu\nu\alpha}{}_{\alpha} \tag{6.15}$$

is a second-rank contravariant tensor if $A^{\mu\nu\alpha}{}_{\beta}$ is a tensor contravariant in $\mu\nu\alpha$, covariant in β, and so on.

We can also easily show that the derivative $\partial_{\mu}\phi$ of a scalar function is a vector.

However, it does not follow from the transformation laws that the derivative $\partial_{\mu}A_{\nu}$ of a vector is a tensor.

As we will see in the next section, to construct tensors by differentiation requires a somewhat more complicated procedure.

6.2 Parallel transport; covariant derivative

In Chapter 2 we saw that in flat spacetime, with rectangular coordinates x^{μ}, it is easy to construct higher-rank tensors by differentiation.

Thus, if A_{μ} is a covariant tensor of first rank, then $\partial_{\nu}A_{\mu}$ is a covariant tensor of second rank.

However, in a curved spacetime, and even in flat spacetime with curvilinear coordinates, simple differentiation does not yield a tensor quantity.

For example, suppose that $A\mu$ is a covariant tensor of first rank in flat spacetime in, say, spherical coordinates, with $x^0 = t$, $x^1 = r$, $x^2 = \theta$, $x^3 = \phi$.

Then simple differentiation yields the 16-component object $\partial_\nu A_\mu$ comprising $\partial A_0/\partial t$, $\partial A_0/\partial r$, $\partial A_0/\partial \theta$, $\partial A_0/\partial \phi$, $\partial A_1/\partial t$, $\partial A_1/\partial r$, $\partial A_1/\partial \theta$, $\partial A_1/\partial \phi$, and so on, which does not have the correct transformation properties for a second-rank tensor.

We can see this by examining what happens if we attempt to transform to some other coordinates, say, from the spherical coordinates x^μ to rectangular coordinates x'^μ.

The transformation law for A_μ is

$$A'_\alpha = \frac{\partial x^\mu}{\partial x'^\alpha} A_\mu \tag{6.16}$$

and hence

$$\frac{\partial A'_\alpha}{\partial x'^\beta} = \frac{\partial x^\mu}{\partial x'^\alpha}\frac{\partial A_\mu}{\partial x'^\beta} + \frac{\partial^2 x^\mu}{\partial x'^\alpha \partial x'^\beta} A_\mu = \frac{\partial x^\mu}{\partial x'^\alpha}\frac{\partial x^\nu}{\partial x'^\beta}\frac{\partial A_\mu}{\partial x^\nu} + \frac{\partial^2 x^\mu}{\partial x'^\alpha \partial x'^\beta} A_\mu \tag{6.17}$$

The right side of Eq. (6.17) differs from the transformation law for a second-rank tensor by the extra term $\partial^2 x^\mu/\partial x'^\alpha \partial x'^\beta A_\mu$, which contains second derivatives of the spherical coordinates x^μ relative to the rectangular coordinates x'^μ.

Since the relation between the spherical and the rectangular coordinates is nonlinear, these derivatives are nonzero.

Their presence "spoils" the transformation law for $\partial A_\mu/\partial x^\nu$, and we see that this quantity is not a second-rank tensor with respect to general coordinate transformations.

We can gain some further insight into the reason for the failure of the tensor transformation law for the derivative $\partial A_\mu/\partial x^\nu$ by examining the definition of this derivative,

$$\frac{\partial A_\mu}{\partial x^\nu} = \lim_{dx\to 0} \frac{A_\mu(x+dx) - A_\mu(x)}{dx^\nu} \tag{6.18}$$

The trouble with this expression is that the numerator is not a vector.

This numerator is the difference between two vectors located at two separate points.

Although the difference between two vectors located at the same point is a vector, the difference between two vectors located at separate points is not a vector, because the transformation law (6.9) depends on position.

This suggests that to obtain a numerator that is a vector, we should parallel-transport A_μ from x to $x + dx$ before performing the subtraction, so that the two vectors in the numerator are located at the same point.

Thus, the construction of tensors by differentiation requires that we first construct a definition of parallel transport of a vector.

We will denote by δA_μ the change produced in the components of the vector A_μ by its parallel transport during a small displacement dx^β.

This change δA_μ must be linear in dx^β, and it must also be linear in A_μ, because the change produced by parallel transport of a vector sum $A_\mu + B_\mu$ must equal the sum of the changes of the individual vectors.

Therefore δA_μ must be of the general form

$$\delta A_\mu = \Gamma^\nu{}_{\mu\beta} A_\nu dx^\beta \tag{6.19}$$

where the $4 \times 4 \times 4$-component object $\Gamma^\nu{}_{\mu\beta}$ is some function of position.

As we will see later, $\Gamma^{\nu}{}_{\mu\beta}$ is related to the derivatives of the metric tensor.

However, since we have not yet introduced a metric tensor in our spacetime, we will for now regard $\Gamma^{\nu}{}_{\mu\beta}$ as some arbitrarily specifiable function that characterizes parallel transport in our spacetime; in other words, it characterizes the affine geometry of spacetime.

The objects $\Gamma^{\nu}{}_{\mu\beta}$ are called the *Christoffel symbols*, or the connection coefficients.

Equation (6.19) defines the parallel transport of a covariant vector.

The parallel transport of a contravariant vector B^{μ} is determined by the requirement that $A_{\mu}B^{\mu}$ must be a scalar, which cannot change under parallel transport.

Thus,

$$0 = \delta(A_\mu B^\mu) = \delta A_\mu B^\mu + A_\mu \delta B^\mu$$
$$= \left(\Gamma^{\nu}{}_{\mu\beta} A_\nu dx^\beta B^\mu + A_\mu \delta B^\mu\right) = A_\mu \left(\Gamma^{\mu}{}_{\nu\beta} dx^\beta B^\nu + \delta B^\mu\right)$$

Since this must be true for any arbitraryA_{μ}, the coefficient multiplying A_{μ} in this equation must vanish, which tells us that

$$\delta B^\mu = -\Gamma^{\mu}{}_{\nu\beta} B^\nu dx^\beta \tag{6.20}$$

We will assume that $\Gamma^{\nu}{}_{\mu\beta}$ is symmetric in its lower indices, so

$$\Gamma^{\nu}{}_{\mu\beta} = \Gamma^{\nu}{}_{\beta\mu} \tag{6.21}$$

To understand the significance of this symmetry condition, let us explore what would happen if it were not satisfied.

Consider two small displacements $d\xi^{\beta}$ and $d\zeta^{\beta}$ that start at the same spacetime point and attempt to construct the parallelogram that has these vectors for two of its sides.

To perform this construction, we can parallel-transport $d\xi^{\beta}$ along $d\zeta^{\beta}$ (see Fig. 6.4);

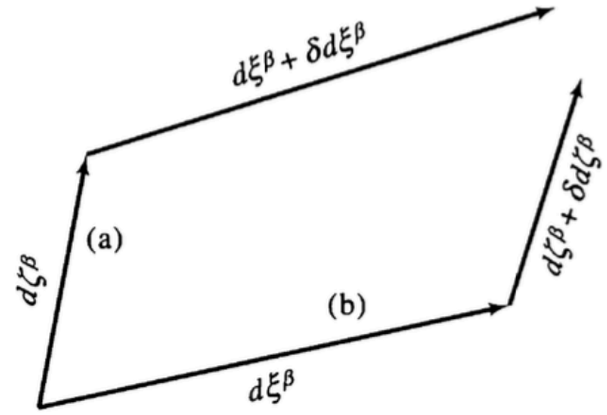


Fig. 6.4 Construction of a parallelogram. (a) Parallel transport of $d\xi^\beta$ along $d\zeta^\beta$. (b) Parallel transport of $d\zeta^\beta$ along $d\xi^\beta$.

this gives us the displacement vector for the "opposite" corner of the parallelogram:

$$d\zeta^\beta + (d\xi^\beta + \delta d\xi^\beta) = d\zeta^\beta + d\xi^\beta - \Gamma^\beta{}_{\mu\nu} d\xi^\mu d\zeta^\nu$$

Alternatively, we can parallel-transport $d\zeta^\beta$ along $d\xi^\beta$; this gives us a different expression for the opposite corner:

$$d\xi^\beta + (d\zeta^\beta + \delta d\zeta^\beta) = d\xi^\beta + d\zeta^\beta - \Gamma^\beta{}_{\mu\nu} d\zeta^\mu d\xi^\nu$$

Comparing these expressions, we see that the two alternative constructions agree if and only if $\Gamma^\beta{}_{\mu\nu}$ is symmetric in μ and ν.

If $\Gamma^\beta{}_{\mu\nu}$ were not symmetric, the parallelogram would fail to close, which indicates a special deviation from flat geometry involving a twist in the angles between the sides.

We can think of the open (or "cracked") configuration in Fig. 6.4 as a loop made of a bent piece of wire whose ends are initially in contact at the upper right corner, but then the angles were bent some more, so the ends fail to match.

A geometry with asymmetric Christoffel symbols is said to have *torsion*.

Although attempts have been made to incorporate torsion into gravitational theory ("Einstein-Cartan theory"), the physical foundations of theses attempts are shaky.

We will ignore these attempts and assume that the Christoffel symbols are symmetric.

That parallel transport produces changes in the components of a vector should not be surprising.

It happens even in flat spacetime, if we use spherical coordinates or other curvilinear coordinates.

When we parallel-transport a constant vector along some curve, its components with respect to rectangular coordinates do not change, but its components with respect to spherical coordinates obviously do change (see Fig. 6.5).

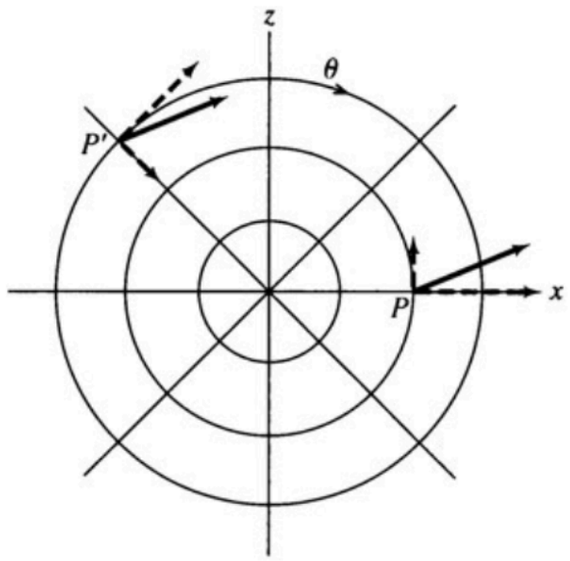


Fig. 6.5 Parallel transport of a constant vector in flat spacetime. The vector is in the meridional plane, $\phi = 0$. The magnitude and the direction of the vector do not change during parallel transport from P to P', but its radial and tangential components (indicated by broken lines) do change.

Note that in flat spacetime, we can exploit the known behavior of a vector under parallel transport to calculate the Christoffel symbols in any given curvilinear coordinate system.

For instance, suppose that a contravariant vector in spherical coordinates is initially in the radial direction, $A^\mu = (0, A^1, 0, 0)$, and we parallel-displace it from (t, r, θ, ϕ) to $(t, r, \theta + d\theta, \phi)$.

Then, as we can see from Fig. 6.6, the change in this vector is $\delta A^\mu = (0, 0, -A^1 d\theta/r, 0)$, and hence Eq. (6.20) tells us that $\Gamma^2{}_{12} = 1/r$.

All the other Christoffel symbols in spherical coordinates can be calculated in a similar way, by simple geometrical considerations.

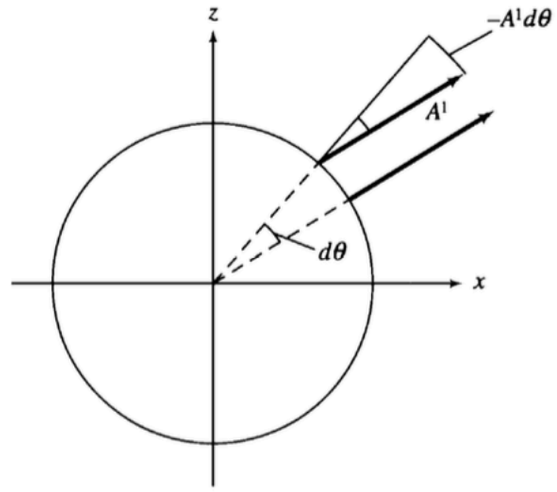


Fig. 6.6 Parallel displacement of a constant vector in flat spacetime. During a small displacement $d\theta$, the vector acquires a tangential component of length $-A^1 d\theta$. However, in spherical coordinates, the corresponding θ-component is this length divided by r, that is, $-A^1 d\theta / r$ (just as the θ-component of an arbitrary displacement dx^μ is the length $r d\theta$ divided by r, that is, $r d\theta / r = d\theta$).

By similar arguments, we can show that the nonzero Christoffel symbols for flat spacetime in spherical coordinates are

$$\begin{aligned} \Gamma^1{}_{22} &= -r, \quad \Gamma^1{}_{33} = -r \sin^2\theta \\ \Gamma^2{}_{12} &= \Gamma^2{}_{21} = 1/r, \quad \Gamma^2{}_{33} = -\sin\theta\cos\theta \\ \Gamma^3{}_{13} &= \Gamma^3{}_{31} = 1/r, \quad \Gamma^3{}_{23} = \Gamma^3{}_{32} = \cot\theta \end{aligned} \tag{6.22}$$

A more interesting example is parallel transport on the surface of a sphere.

This surface is a curved 2-D space, usually called a two-sphere.

Although we are interested mainly in four-dimensional spacetime, the definition (6.19) for parallel transport applies to a Riemannian space of any number of dimensions (but in an n-dimensional space, the indices μ and ν take the values $1, 2, 3, \ldots, n$).

Now consider the surface of a sphere, and use coordinates $x^1 = \theta$ and $x^2 = \phi$, where θ and ϕ are the usual polar and azimuthal angles on this surface.

It is straightforward to show that the Christoffel symbols are

$$\begin{aligned} \Gamma^1{}_{11} &= 0, \quad \Gamma^1{}_{12} = \Gamma^1{}_{21} = 0, \quad \Gamma^1{}_{22} = -\sin\theta\cos\theta \\ \Gamma^2{}_{11} &= 0, \quad \Gamma^2{}_{12} = \Gamma^2{}_{21} = \cot\theta, \quad \Gamma^2{}_{22} = 0 \end{aligned} \tag{6.23}$$

It is easy to convince yourself that the results for the parallel transport of a vector along a meridian (ϕ = constant; see Fig. 6.7) and along the equator ($\theta = \pi/2$) are intuitively reasonable.

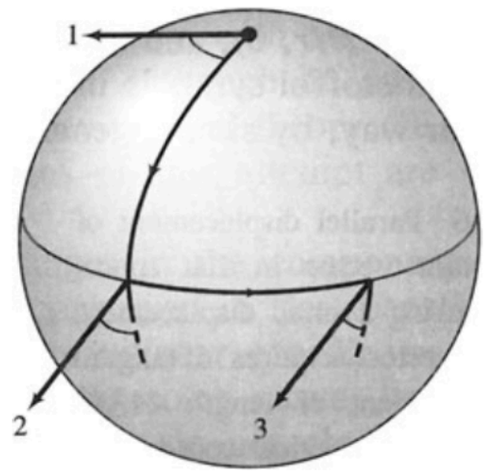


Fig. 6.7 Parallel transport of a vector along a meridian and along the equator of a sphere. During transport, the vector maintains a fixed angle relative to the path.

Parallel transport along other paths that are not great circles is not intuitively obvious and must be calculated by integration of the equations (6.19) or (6.20) for infinitesimal displacements.

Note that in the calculation of the parallel transport of a vector along a finite path, the integration of equations (6.19) or (6.20) for infinitesimal displacements must be performed jointly for all components, because in general, every component of the vector can affect the parallel transport of every other component.

This can make calculations of parallel transport rather messy.

Now that we have available a definition of parallel transport, we can construct a genuine tensor by differentiation.

To make the numerator in the derivative (6.18) into a vector, we parallel-transport $A_\mu(x)$ from x to $x + dx$ before performing the subtraction.

We then obtain

$$\lim_{dx\to 0}\frac{A_\mu(x+dx)-A_\mu(x)-\delta A_\mu}{dx^\nu}=\lim_{dx\to 0}\frac{A_\mu(x+dx)-A_\mu(x)}{dx^\nu}-\Gamma^\alpha{}_{\mu\beta}A_\alpha\frac{dx^\beta}{dx^\nu}$$
$$=\frac{\partial A_\mu}{\partial x^\nu}-\Gamma^\alpha{}_{\mu\nu}A_\alpha \qquad (6.24)$$

This is called the covariant derivative of the vector $A_\mu(x)$.

It is customary to indicate a covariant derivative by a semicolon, in distinction to the ordinary derivative, indicated by a comma:

$$A_{\mu;\nu}=A_{\mu,\nu}-\Gamma^\alpha{}_{\mu\nu}A_\alpha \qquad (6.25)$$

By construction, the covariant derivative is a second-rank tensor.

The word *covariant* as used in this context merely serves to distinguish $A_{\mu;\nu}$ from the ordinary derivative $A_{\mu,\nu}$.

Since we already have been using *covariant* to describe lower indices on a tensor, it might perhaps be better to call (6.25) the *tensor derivative*; but *covariant derivative* is common usage, and what meaning is intended must be decided from the context.

For example, (6.25) is sometimes called the covariant derivative of a covariant vector.

A similar construction based on the parallel-transport formula for a contravariant vector B^μ tells us that the covariant derivative of such a vector is

$$B^\mu{}_{;\nu}=\frac{\partial B^\mu}{\partial x^\nu}+\Gamma^\mu{}_{\alpha\nu}B^\alpha=B^\mu{}_{,\nu}+\Gamma^\mu{}_{\alpha\nu}B^\alpha \qquad (6.26)$$

Furthermore, we can proceed to the construction of covariant derivatives for tensors of rank higher than one.

The construction of these covariant derivatives is based on the formulas for the parallel transport of higher rank tensors, which can be obtained from the requirement that scalars, such as $A_\mu C_\nu B^{\mu\nu}$, must not change under parallel transport.

For such higher-rank tensors, the covariant derivative differs from the ordinary derivative by as many extra terms as there are indices.

For example, the covariant derivatives of the tensors $B^{\mu\nu}$ and $A^{\mu\nu}{}_\sigma$ are, respectively,

$$B^{\mu\nu}{}_{;\beta} = \frac{\partial B^{\mu\nu}}{\partial x^\beta} + \Gamma^\mu{}_{\alpha\beta} B^{\alpha\nu} + \Gamma^\nu{}_{\alpha\beta} B^{\mu\alpha} \tag{6.27}$$

$$A^{\mu\nu}{}_{\sigma;\beta} = \frac{\partial A^{\mu\nu}{}_\sigma}{\partial x^\beta} + \Gamma^\mu{}_{\alpha\beta} A^{\alpha\nu}{}_\sigma + \Gamma^\nu{}_{\alpha\beta} A^{\mu\alpha}{}_\sigma - \Gamma^\alpha{}_{\sigma\beta} A^{\mu\nu}{}_\alpha \tag{6.28}$$

Note that in these covariant derivatives, each tensor index receives the same treatment as the vector index in Eqs. (6.25) and (6.26).

By convention, the covariant derivative of a scalar field is simply taken to be the ordinary derivative, $\phi_{;\alpha} \equiv \phi_{,\alpha}$.

As a consequence of Eq. (6.27), the covariant derivative of the second-rank tensor $A^\mu B^\nu$ can be written

$$(A^\mu B^\nu)_{;\beta} = A^\mu{}_{;\beta} B^\nu + A^\mu B^\nu{}_{;\beta} \tag{6.29}$$

This shows that covariant differentiation obeys the familiar rule for the differentiation of a product.

This rule is also valid for products of tensors of higher rank than those in Eq. (6.29).

Although we have written the Christoffel symbol in a notation usually reserved for third-rank tensors, the Christoffel symbol is not a tensor – it does not obey the transformation law for a third-rank tensor.

The transformation law for the Christoffel symbol must be such that the sum $\partial A_\mu/\partial x^\nu - \Gamma^\alpha{}_{\mu\nu}A_\alpha$ is a tensor.

We know that in this sum the transformation of the first term deviates from a tensor [see Eq. (6.17)], and therefore the transformation of the second term must deviate by an opposite amount.

This requirement leads to the following transformation law for $\Gamma^\mu{}_{\alpha\beta}$:

$$\Gamma'^{\mu}{}_{\alpha\beta} = \frac{\partial x'^{\mu}}{\partial x^{\nu}} \frac{\partial x^{\sigma}}{\partial x'^{\alpha}} \frac{\partial x^{\tau}}{\partial x'^{\beta}} \Gamma^{\nu}{}_{\sigma\tau} + \frac{\partial x'^{\mu}}{\partial x^{\nu}} \frac{\partial^2 x^{\nu}}{\partial x'^{\alpha} \partial x'^{\beta}} \tag{6.30}$$

This equation can also be written as

$$\Gamma'^{\mu}{}_{\alpha\beta} = \frac{\partial x'^{\mu}}{\partial x^{\nu}} \frac{\partial x^{\sigma}}{\partial x'^{\alpha}} \frac{\partial x^{\tau}}{\partial x'^{\beta}} \Gamma^{\nu}{}_{\sigma\tau} - \frac{\partial x^{\nu}}{\partial x'^{\alpha}} \frac{\partial x^{\sigma}}{\partial x'^{\beta}} \frac{\partial^2 x'^{\mu}}{\partial x^{\sigma} \partial x^{\nu}} \tag{6.31}$$

Note that the difference between Eq. (6.30) and the transformation law of a third-rank tensor is due to the presence of the second term on the right side.

In the special case of a linear transformation, say, $x'^\mu = b_\nu{}^\mu x^\nu$, where $b_\nu{}^\mu$ is a constant, this second term is absent, and then $\Gamma^\mu{}_{\alpha\beta}$ does behave like a tensor.

The Christoffel symbol $\Gamma^\mu{}_{\alpha\beta}$ is an example of an object that is a tensor with respect to linear transformations (such as Lorentz transformations), but is not a tensor with respect to general nonlinear coordinate transformations.

To emphasize the nontensor character of the Christoffel symbol, the notation $\left\{ {\mu \atop \alpha\beta} \right\}$ is sometimes used instead of $\Gamma^{\mu}{}_{\alpha\beta}$.

6.3 Geodesic equation

In a curved space, we can define geodesic curves in two different ways: as straightest curves or as curves of extremum length.

To construct a straightest possible curve, we take a short segment of the curve and move it forward, step by step, always parallel to itself (see Fig. 6.8).

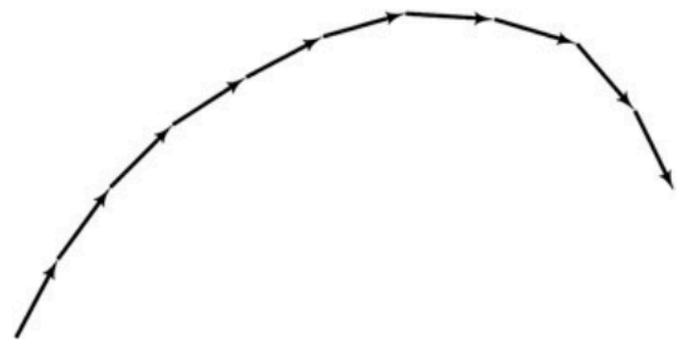


Fig. 6.8 Step-by-step construction of a geodesic curve by parallel displacement of a small segment of the curve.

On a 2-D curved surface – for instance, a sphere – this construction of a straightest curve can be implemented by running a strip of adhesive tape along the surface, without bending it to the left or the right.

Alternatively, to construct a curve of extremal length, we stretch a string (or, in spacetime, we measure proper time with a clock), and we find the path that yields the shortest length (or the extremal proper time).

The first method for determining a geodesic relies on parallel transport; that is, it relies on the affine geometry of the curved space.

The second method relies on measurements of length (or proper time); that is, it relies on the metric geometry.

In this section we will write the geodesic equation according to the first method, exploiting our definition of parallel transport.

In the next section we will introduce the metric tensor, and we will then reexamine the geodesic equation.

We will see that the geodesics determined by the metric geometry coincide with those determined by parallel transport.

Consider a small segment dx^α of a geodesic curve (see Fig. 6.8).

If we displace this segment forward, parallel to itself, through a displacement dx^α equal to itself, the segment suffers a change

$$\delta dx^\alpha = -\Gamma^\alpha{}_{\mu\beta} dx^\mu dx^\beta \tag{6.32}$$

To convert this into a differential equation, we introduce a parametrization of the curve, with a parameter *s* that increases monotonically along the curve.

If *ds* is the increment that corresponds to the displacement dx^α, we can divide Eq. (6.32) by $(ds)^2$, and rewrite it as

$$\frac{d}{ds}\frac{dx^\alpha}{ds} = -\Gamma^\alpha{}_{\mu\beta}\frac{dx^\mu}{ds}\frac{dx^\beta}{ds}$$

or as

$$\frac{d^2x^\alpha}{ds^2} + \Gamma^\alpha{}_{\mu\beta}\frac{dx^\mu}{ds}\frac{dx^\beta}{ds} = 0 \tag{6.33}$$

This is the differential equation for the geodesic.

The parameter *s* in this equation has, as of yet, no physical interpretation; it is merely a mathematical parameter.

However, after we introduce the metric tensor in the next section, we will be able to identify this parameter with the proper time.

Note that if we change the parametrization of the curve, by defining a new parameter dq = f(q)ds, the geodesic equation becomes

$$\frac{d}{dq}\left(f(q)\frac{dx^\alpha}{dq}\right) + f(q)\Gamma^\alpha{}_{\mu\beta}\frac{dx^\mu}{dq}\frac{dx^\beta}{dq} = 0 \tag{6.34}$$

This is an alternative form of the geodesic equation.

It looks more complicated than Eq. (6.33), but of course, it yields exactly the same curve in spacetime – only the parametrization has changed, but the curve has not changed.

If we multiply Eq. (6.34) by $(dq)^2$, we recognize that it expresses the parallel transport of $f(q)dx^\alpha$ along the curve.

Because the direction of the vector $f(q)dx^\alpha$ is the same as the direction of dx^α, we could have obtained Eq. (6.34) by starting our derivation of the geodesic equation with the parallel displacement of $f(q)dx^\alpha$.

Equation (6.33) is called the *normal form* of the geodesic equation, and we will give it preference over the more complicated form displayed in Eq. (6.34).

At any one point *P* of spacetime we can simplify the geodesic equation by introducing a new set of coordinates x'^μ such that the Christoffel symbols are zero at this one point.

To accomplish this trick, we use the quadratic coordinate transformation

$$x'^\mu = x^\mu - x^\mu(P) + \tfrac{1}{2}\Gamma^\mu{}_{\alpha\beta}(P)[x^\alpha - x^\alpha(P)][x^\beta - x^\beta(P)] \tag{6.35}$$

where $x^\mu(P)$ are the old coordinates of the point *P*.

Note that Eq. (6.35) takes *P* as the origin of the new coordinates, with $x'^\mu(P) = 0$.

From Eq. (6.35) we obtain

$$\left.\frac{\partial x'^{\mu}}{\partial x^{\nu}}\right|_P = \delta^{\mu}_{\nu}, \qquad \left.\frac{\partial x^{\mu}}{\partial x'^{\nu}}\right|_P = \delta^{\mu}_{\nu}, \quad \text{and} \quad \left.\frac{\partial^2 x'^{\mu}}{\partial x^{\sigma}\,\partial x^{\nu}}\right|_P = \Gamma^{\mu}{}_{\sigma\nu}(P) \tag{6.36}$$

which, when inserted into Eq. (6.31), gives us $\Gamma'^{\mu}{}_{\alpha\beta}(\mathrm{P}) = 0$.

Note that this result hinges on the symmetry of $\Gamma'^{\mu}{}_{\alpha\beta}$ in α and β.

If $\Gamma'^{\mu}{}_{\alpha\beta}$ had an antisymmetric part, this part would be left unchanged by the transformation (6.35), because only the symmetric part of $\Gamma'^{\mu}{}_{\alpha\beta}$ contributes to Eqs. (6.35) and (6.36).

With $\Gamma'^{\mu}{}_{\alpha\beta}(\mathrm{P}) = 0$, the geodesic equation at this point takes the simple form

$$\frac{d^2 x'^{\mu}}{ds^2} = 0 \tag{6.37}$$

which says that locally, in a small neighborhood of the point *P*, the geodesic is a straight line.

The coordinates x'^{μ} that, at some point *P*, make the Christoffel symbols zero and give the geodesic equation its simple form (6.37) are called *local geodesic coordinates* (at that point).

It must be emphasized that these coordinates are geodesic only at one spacetime point – that is, at one place and at one time – because the Christoffel symbols are zero only at one spacetime point *P*.

If we want geodesic coordinates at a different place or a different time, we must carry out the transformation (6.35) for this different place or time.

These coordinates are also called *local inertial coordinates*, because the geodesic equation (6.37) in these coordinates implies motion at constant velocity, like the inertial motion of a free particle in flat spacetime.

Geometrically, the local geodesic coordinates correspond to replacing the curved space by a small flat patch tangent to the former at the point P.

For instance, in the curved 2-D space consisting of the surface of a sphere, we can visualize the construction of local geodesic coordinates by placing a small flat rectangular patch of paper tangent to the sphere at the given point P (see Fig. 6.9).

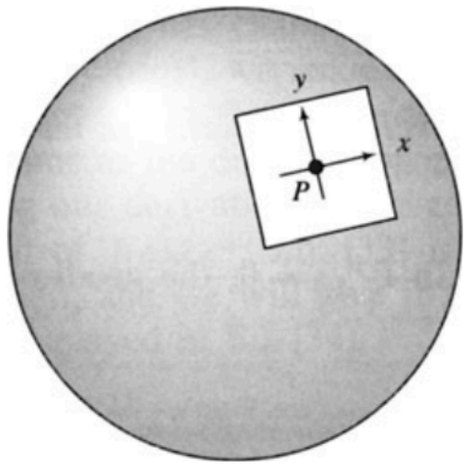


Fig. 6.9 Construction of local geodesic coordinates near a point P on the surface of a sphere. A small flat patch of paper is placed tangent to the sphere at this point. The rectangular coordinates painted on this flat patch give us local geodesic coordinates.

The rectangular coordinates painted on this flat patch constitute local geodesic coordinates.

This construction is, of course, valid only over an infinitesimal region.

When we move away from the point P, the Christoffel symbols begin to deviate from zero.

In fact, as we will see later, the first derivatives of $\Gamma^{\mu}{}_{\alpha\beta}$ cannot all be made zero by choice of coordinates.

In local geodesic coordinates, at a given point, parallel transport of a vector reduces to $\delta A_{\mu} = 0$; that is, parallel transport leaves the components of the vector unchanged, just as in rectangular coordinates in flat spacetime.

Furthermore, in local geodesic coordinates, the covariant derivative of a vector or of a tensor coincides with the ordinary derivative.

This can be exploited to simplify some calculations with covariant derivatives.

For instance, to verify that some expression containing covariant derivatives is equal to some other covariant derivative or some other tensor, we can introduce local geodesic coordinates and eliminate all the terms involving Christoffel symbols.

It is also possible to introduce new coordinates, called *Fermi coordinates*, such that the Christoffel symbols are zero at every point on a given geodesic curve.

These coordinates are constructed by erecting unit vectors and coordinate axes at the initial point P.

The x'^0-axis and the corresponding unit vector are assumed to lie along the geodesic; the x'^1, x'^2, and x'^3-axes and their corresponding unit vectors lie in the hypersurface x'^0 = constant, for the given initial value of x'^0.

The unit vectors are then carried forward along the geodesic by parallel transport, and the new unit vectors obtained in this way are used to define the coordinate axes at any later point on the geodesic (see Fig. 6.10).

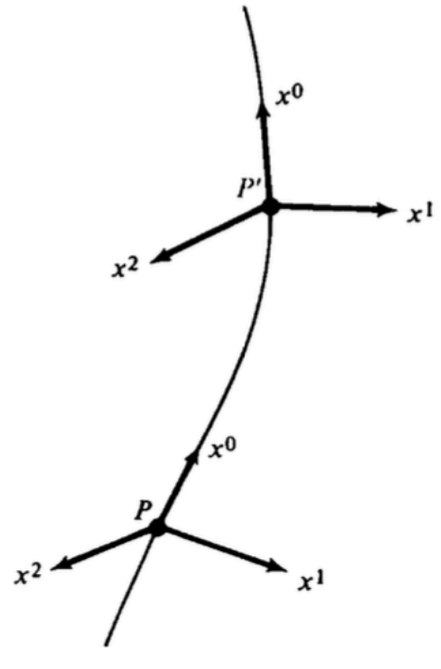


Fig. 6.10 Unit vectors for Fermi coordinates along a geodesic. The unit vectors at different times x^0 are related by parallel transport.

The four unit vectors needed to define the directions of the local coordinate axes are called a *tetrad*.

Since these local coordinate axes are constructed by parallel transport, it is obvious that whenever we parallel-transport some vector along the geodesic, its components cannot change; this immediately implies that all the Christoffel symbols involved in parallel transport along the geodesic must be zero everywhere on the geodesic ($\Gamma'^{\mu}_{\alpha\beta} = 0$).

It can be shown that all the other Christoffel symbols are also zero.

Physically, the Fermi coordinates along a geodesic represent a freely falling reference frame, whose spatial orientation is defined by parallel transport (hence these coordinates are also called “free-fall coordinates”).

In the curved 2-D space consisting of the surface of a sphere, we can visualize the construction of Fermi coordinates by taking a narrow strip or tape of paper and placing it along the geodesic (a great circle; see Fig. 6.11).

The rectangular coordinates painted on this strip of paper constitute Fermi coordinates along the geodesic.

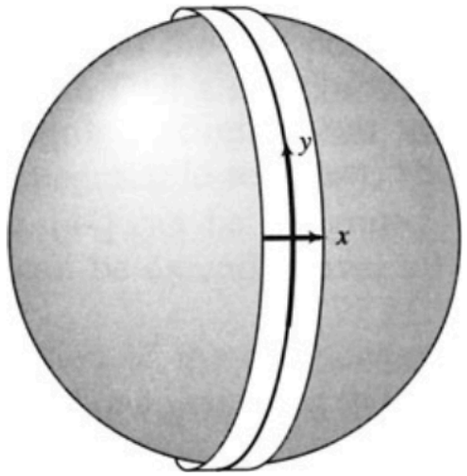


Fig. 6.11 Construction of Fermi coordinates along and near a geodesic (great circle) on the surface of a sphere. A strip of paper is placed along the geodesic. The rectangular coordinates painted on this strip of paper give us Fermi coordinates. Note that the strip of paper is a flat 2-D space; it has no intrinsic curvature, even though it is bent along the great circle (compare the example of the surface of a cylinder discussed in the introduction to this chapter).

Fermi coordinates can also be introduced along an arbitrary, nongeodesic curve, that is, the worldline of a particle subjected to a nongravitational force.

However, along a nongeodesic worldline, the basis vectors cannot be constructed by carrying an initial set of basis vectors forward along the curve by parallel transport; instead, a different kind of transport must be used, called Fermi-Walker transport.

This kind of vector transport will be discussed in Section 6.6.

6.4 Metric tensor

In Chapter 3, we defined the metric tensor $g_{\mu\nu}(x)$ in terms of the spacetime interval ds^2 between two points separated by a coordinate displacement dx^μ [see Eq. (3.110)]:

$$ds^2 = g_{\mu\nu}(x)dx^\mu dx^\nu \tag{6.38}$$

According to our sign convention for $g_{\mu\nu}$, the time component of the metric tensor is positive, $g_{00} > 0$, and the diagonal space components are negative, $g_{kk} < 0$.

Depending on the displacement dx^μ, the spacetime interval ds^2 can be positive, negative, or zero.

If $ds^2 > 0$ (a timelike interval), then ds^2 represents the proper time accumulated by a clock during the displacement dx^μ.

In this case, we will usually adopt the notation $d\tau^2$ for ds^2.

Without any loss of generality, we may assume that the metric tensor is symmetric, $g_{\mu\nu} = g_{\nu\mu}$.

Any antisymmetric part in $g_{\mu\nu}$ would simply drop out of Eq. (6.38), because $dx^\mu dx^\nu$ is symmetric in μ, ν.

To establish that $g_{\mu\nu}$ is indeed a tensor, we rely on the invariance of the spacetime interval ds^2,

$$ds'^2 = ds^2 = g_{\mu\nu}dx^\mu dx^\nu \tag{6.39}$$

By the chain rule for differentials, $dx^\mu = (\partial x^\mu/\partial x'^\alpha)dx'^\alpha$, so

$$ds'^2 = g_{\mu\nu}\frac{\partial x^\mu}{\partial x'^\alpha}\frac{\partial x^\nu}{\partial x'^\beta}dx'^\alpha dx'^\beta \tag{6.40}$$

From this expression we see that the metric tensor in the x'^α coordinates is

$$g'_{\alpha\beta} = \frac{\partial x^\mu}{\partial x'^\alpha}\frac{\partial x^\nu}{\partial x'^\beta}g_{\mu\nu} \tag{6.41}$$

which is exactly what is required for a second-rank, covariant tensor.

Although our main concern is with curved spacetime, the transformation equation (6.41) is perfectly general and can be used to obtain the metric tensor $g'_{\mu\nu}$ that describes flat spacetime in curvilinear coordinates.

For example, if $x'^0 = t$, $x'^1 = r$, $x'^2 = \theta$, $x'^3 = \phi$ are the usual spherical coordinates, then Eq. (6.41) with $x^0 = t$, $x^1 = x$, $x^2 = y$, $x^3 = z$, and $g_{\mu\nu} = \eta_{\mu\nu}$ gives

$$g'_{\alpha\beta} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -r^2 & 0 \\ 0 & 0 & 0 & -r^2\sin^2\theta \end{pmatrix} \tag{6.42}$$

The corresponding expression for the spacetime interval is

$$ds^2 = g'_{\alpha\beta}dx'^\alpha dx'^\beta = dt^2 - dr^2 - r^2d\theta^2 - r^2\sin^2\theta d\phi^2 \tag{6.43}$$

Note that the expression (6.43) can also be obtained directly by expressing the differentials $dt^2 - dx^2 - dy^2 - dz^2$ in terms of dt, dr, dθ, dϕ; the metric tensor $g'_{\alpha\beta}$ can then be found by inspection of (6.43).

This procedure is mathematically equivalent to using the transformation law (6.41), but perhaps somewhat quicker.

By means of a coordinate transformation it is always possible to make the tensor $g'_{\alpha\beta}$ equal to $\eta_{\alpha\beta}$ at any given point *P*.

According to Eq. (6.41), to accomplish this we need a local reference frame with coordinates x'^{α} such that

$$\eta_{\alpha\beta} = \frac{\partial x^{\mu}}{\partial x'^{\alpha}} \frac{\partial x^{\nu}}{\partial x'^{\beta}} g_{\mu\nu} \tag{6.44}$$

This can be regarded as a set of 10 equations for the coefficients $\partial x^{\mu}/\partial x'^{\alpha}$; because there are 16 such coefficients, it is always possible to find a solution of these 10 equations.

This solution gives us a local coordinate system with $g'_{\alpha\beta} = \eta_{\alpha\beta}$ at the one point *P*.

However, in curved spacetime it is not possible to find a solution that simultaneously makes $g'_{\alpha\beta} = \eta_{\alpha\beta}$ at all other points of spacetime; in other words, the solution of Eq. (6.44) is not integrable for the functions $x^{\mu}(x')$.

By contrast, in flat spacetime, the solution of Eq. (6.44) can be integrated, and the local (rectangular) coordinate system can be extended over all spacetime.

We already discussed the physical interpretation of the components of the metric tensor in Section 3.5.

The components g_{00} and g_{kl} determine how measured times and lengths are related to coordinate displacements.

A coordinate displacement dx^0 (with $dx^k = 0$) implies a time $\sqrt{g_{00}}\, dx^0$ as measured by a clock at rest, and a coordinate displacement dx^k (with $dx^0 = 0$) implies a distance $dl^2 = -g_{kl}dx^k dx^l$ as measured with a meter stick at one instant of time.

For the description of the 3-D space geometry at one instant of time, it is convenient to define a 3×3 matrix ${}^{(3)}g_{kl} = -g_{kl}$.

This matrix has positive diagonal components, and it gives us the geometry of 3-D space at one instant of time,

$$dl^2 = {}^{(3)}g_{kl}dx^k dx^l \tag{6.45}$$

Conversely, the measured times and lengths determine the metric tensor; that is, we can measure the metric tensor with clocks and meter sticks.

However, instead of segregating time and distance measurements in this way, we can directly measure the value of ds^2 for any arbitrary spacetime displacement dx^μ by the generalized radar-ranging procedure described in Section 2.1.

We can apply this procedure to an infinitesimal spacetime displacement PP', in which case the emission and reception times for the radar signal are also infinitesimal (see Fig. 6.12), and the measured spacetime interval between P and P' is [compare Eq. (2.7)]

$$ds^2 = (dT_1)(dT_2) \tag{6.46}$$

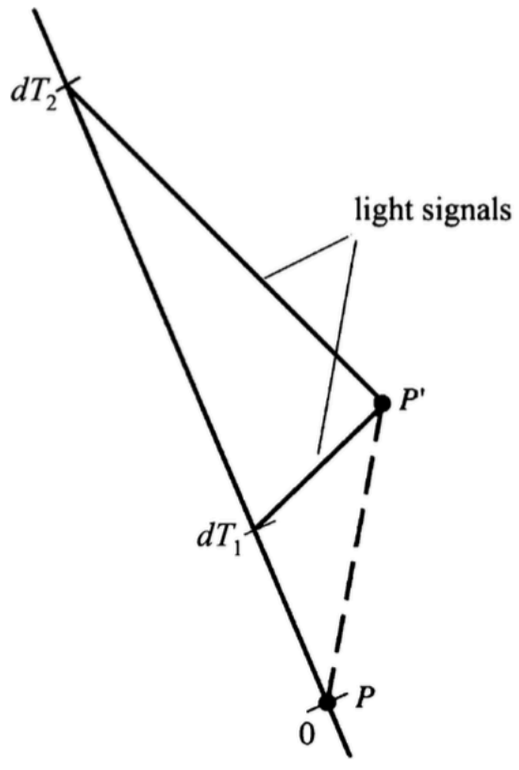


Fig. 6.12. Radar-ranging procedure for the direct measurement of ds^2 for a spacetime displacement PP'. The clock used for radar ranging reads a proper clock time $T = 0$ at the point P; it reads dT_1 when it emits the radar pulse; and it reads dT_2 when it receives the echo.

Ten separate measurements of ds^2 for displacements in 10 different directions (say, the four directions along the four coordinate axes and six directions between pairs of coordinate axes) then give us the 10 components of $g_{\mu\nu}$.

Note that radar-ranging measurements based on Eq. (6.46) automatically assign a value $ds^2 = 0$ to a displacement $P P'$ that coincides with the worldline of a light signal, because then $dT_1 = 0$ (see Fig. 6.12).

Thus the familiar condition $ds^2 = 0$ for light signals in flat spacetime also holds in curved spacetime. This is consistent with the principle of equivalence, because in a freely falling reference frame in this curved spacetime, light propagates at its standard speed, $c = 1$; hence we have $ds^2 = 0$ in this reference frame, and by invariance of the spacetime interval, we then also have $ds^2 = 0$ in any other reference frame.

The inverse $g^{\mu\nu}$ of the metric tensor $g_{\mu\nu}$ is defined by

$$g_{\mu\alpha} g^{\alpha\nu} = \delta^{\nu}_{\mu} \tag{6.47}$$

We will use the tensors $g^{\mu\nu}$ and $g_{\mu\nu}$ to raise and to lower indices, respectively, as in the following examples:

$$dx_\mu \equiv g_{\mu\nu} dx^\nu \tag{6.48}$$

$$\partial^\mu \equiv g^{\mu\nu} \partial_\nu \tag{6.49}$$

$$B^{\mu\alpha} \equiv g^{\alpha\beta} B^\mu{}_\beta \tag{6.50}$$

$$A^{\nu\alpha\beta\ldots} \equiv g^{\mu\nu} A_\mu{}^{\alpha\beta\ldots} \tag{6.51}$$

$$A^{\mu;\beta} \equiv g^{\beta\nu} A^\mu{}_{;\nu} \tag{6.52}$$

Next, we establish the relation between the affine geometry and the metric geometry, that is, the relation between the Christoffel symbol $\Gamma^\mu{}_{\alpha\beta}$ and the metric tensor $g_{\mu\nu}$.

The relation emerges from the requirement that if we parallel-transport a vector A^μ, its "length" $A^\mu A^\nu g_{\mu\nu}$ must remain unchanged. More generally, if we parallel-transport two vectors A^μ and B^μ, their scalar product $A^\mu B^\nu g_{\mu\nu}$ must remain unchanged,

$$(A^\mu B^\nu g_{\mu\nu})_{,\beta} = 0 \tag{6.53}$$

Because $A^\mu B^\nu g_{\mu\nu}$ is a scalar, we can replace the ordinary derivative by a covariant derivative, and if we then differentiate term by term, we find

$$(A^\mu B^\nu g_{\mu\nu})_{;\beta} = A^\mu{}_{;\beta} B^\nu g_{\mu\nu} + A^\mu B^\nu{}_{;\beta} g_{\mu\nu} + A^\mu B^\nu g_{\mu\nu;\beta} = 0 \tag{6.54}$$

Because the only changes in the vectors A^μ and B^ν are those contributed by parallel transport, the covariant derivatives of these vectors are zero, and Eq. (6.54) reduces to

$$A^\mu B^\nu g_{\mu\nu;\beta} = 0 \tag{6.55}$$

In this equation, the vectors A^μ and B^ν are arbitrary; hence the coefficient of $A^\mu B^\nu$ must vanish:

$$g_{\mu\nu;\beta} = 0 \tag{6.56}$$

Thus, the consistency between the affine geometry and the metric geometry demands that the covariant derivative of the metric tensor is zero.

This means that the metric tensor at any place can be obtained from the metric tensor at some other place by parallel transport.

However, as we will see later, it is best to view Eq. (6.56) as a restriction on the Christoffel symbols, rather than a restriction on the metric tensor.

If we use the expression for the covariant derivative of a second-rank tensor, Eq. (6.56) becomes

$$\frac{\partial g_{\mu\nu}}{\partial x^\beta} - g_{\alpha\nu}\Gamma^\alpha{}_{\beta\mu} - g_{\mu\alpha}\Gamma^\alpha{}_{\beta\nu} = 0 \tag{6.57}$$

This equation permits us to express the derivatives of the metric tensor in terms of the Christoffel symbols:

$$\frac{\partial g_{\mu\nu}}{\partial x^\beta} = g_{\alpha\nu}\Gamma^\alpha{}_{\beta\mu} + g_{\mu\alpha}\Gamma^\alpha{}_{\beta\nu} \tag{6.58}$$

Conversely, we can express the Christoffel symbols in terms of the derivatives of the metric tensor.

For this, we permute the indices μ, ν in Eq. (6.58) cyclically, to obtain

$$\frac{\partial g_{\beta\mu}}{\partial x^\nu} = g_{\alpha\mu}\Gamma^\alpha{}_{\nu\beta} + g_{\beta\alpha}\Gamma^\alpha{}_{\nu\mu} \tag{6.59}$$

$$\frac{\partial g_{\nu\beta}}{\partial x^\mu} = g_{\alpha\beta}\Gamma^\alpha{}_{\mu\nu} + g_{\nu\alpha}\Gamma^\alpha{}_{\mu\beta} \tag{6.60}$$

Now, add Eqs. (6.59) and (6.60), and subtract Eq. (6.58), and take into account the symmetries of the components of the metric tensor and the Christoffel symbols; the result is

$$\frac{\partial g_{\beta\mu}}{\partial x^{\nu}} + \frac{\partial g_{\nu\beta}}{\partial x^{\mu}} - \frac{\partial g_{\mu\nu}}{\partial x^{\beta}} = 2g_{\alpha\beta}\Gamma^{\alpha}{}_{\mu\nu} \tag{6.61}$$

which yields

$$\Gamma^{\alpha}{}_{\mu\nu} = \frac{1}{2}g^{\alpha\beta}\left(\frac{\partial g_{\beta\mu}}{\partial x^{\nu}} + \frac{\partial g_{\nu\beta}}{\partial x^{\mu}} - \frac{\partial g_{\mu\nu}}{\partial x^{\beta}}\right) = \frac{1}{2}g^{\alpha\beta}(g_{\beta\mu,\nu} + g_{\nu\beta,\mu} - g_{\mu\nu,\beta}) \tag{6.62}$$

According to this equation, the metric tensor completely determines the Christoffel symbols, and therefore the metric geometry completely determines the affine geometry.

Note that the converse is not true – if we arbitrarily specify the Christoffel symbols, we will not always be able to find a metric tensor consistent with these Christoffel symbols [Eq. (6.58) is not necessarily integrable].

Hereafter, we will always regard the metric tensor as primary, and we will assume that the Christoffel symbols are calculated from the metric tensor according to Eq. (6.62).

It is easy to verify that the covariant derivative of the inverse of the metric tensor is also identically zero:

$$g^{\mu\nu}{}_{;\alpha} = 0 \tag{6.63}$$

It follows that the operation of raising or lowering of indices commutes with covariant differentiation.

For example,

$$(g^{\mu\nu}A_{\nu})_{;\alpha} = g^{\mu\nu}A_{\nu;\alpha} \tag{6.64}$$

The surface of a sphere is a 2-D curved space.

Suppose that the radius of the sphere is b, and use the coordinates $x^1 = \theta$, $x^2 = \phi$.

Then distances along the sphere are given by

$$dl^2 = b^2 d\theta^2 + b^2 \sin^2\theta d\phi^2 \tag{6.65}$$

and therefore the metric tensor ${}^{(2)}g_{\mu\nu}$ for this 2-D space is

$$ {}^{(2)}g_{\mu\nu} = \begin{pmatrix} b^2 & 0 \\ 0 & b^2 \sin^2\theta \end{pmatrix} \tag{6.66} $$

Here we deal only with distances rather than spacetime intervals; we have inserted a superscript (2) to prevent any confusion between ${}^{(2)}g_{\mu\nu}$ and the metric tensor of spacetime).

For the metric tensor ${}^{(2)}g_{\mu\nu}$, we can calculate the Christoffel symbols $\Gamma^k{}_{mn}$ according to Eq. (6.62), with $k, m, n = 1, 2$, and we verify that they are as stated in Eq. (6.23).

$$ \Gamma^1{}_{11} = 0, \quad \Gamma^1{}_{12} = \Gamma^1{}_{21} = 0, \quad \Gamma^1{}_{22} = -\sin\theta\cos\theta $$
$$ \Gamma^2{}_{11} = 0, \quad \Gamma^2{}_{12} = \Gamma^2{}_{21} = \cot\theta, \quad \Gamma^2{}_{22} = 0 $$

In general relativity (see Section 7.4), the following spacetime interval for spherically symmetric curved spacetime will play an important role:

$$ ds^2 = e^{N(r,t)}dt^2 - e^{L(r,t)}dr^2 - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \tag{6.67} $$

where N(r, t) and L(r, t) are functions of r and t, but not functions of θ or ϕ (we write these functions as exponentials e^N and e^L because this simplifies some expressions in the calculations).

With $x^0 = t$, $x^1 = r$, $x^2 = \theta$, $x^3 = \phi$, the metric tensor is

$$ g_{\mu\nu} = \begin{pmatrix} e^N & 0 & 0 & 0 \\ 0 & -e^L & 0 & 0 \\ 0 & 0 & -r^2 & 0 \\ 0 & 0 & 0 & -r^2 \sin^2\theta \end{pmatrix} \tag{6.68} $$

We can then derive the following expressions for the Christoffel symbols (the primes indicate radial derivatives, $L' \equiv \partial L/\partial r$, $N' \equiv \partial N/\partial r$; the dots indicate time derivatives, $\dot{N} \equiv \partial N/\partial t$, $\dot{L} \equiv \partial L/\partial t$):

$$
\begin{aligned}
&\Gamma^0{}_{00} = \tfrac{1}{2}\dot{N}, \quad \Gamma^0{}_{01} = \Gamma^0{}_{10} = \tfrac{1}{2}N', \quad \Gamma^0{}_{11} = \tfrac{1}{2}\dot{L}e^{L-N} \\
&\Gamma^1{}_{00} = \tfrac{1}{2}N'e^{N-L}, \quad \Gamma^1{}_{10} = \Gamma^1{}_{01} = \tfrac{1}{2}\dot{L}, \quad \Gamma^1{}_{11} = \tfrac{1}{2}L', \quad \Gamma^1{}_{22} = -re^{-L}, \\
&\Gamma^1{}_{33} = -r\sin^2\theta e^{-L} \\
&\Gamma^2{}_{12} = \Gamma^2{}_{21} = 1/r, \quad \Gamma^2{}_{33} = -\sin\theta\cos\theta \\
&\Gamma^3{}_{13} = \Gamma^3{}_{31} = 1/r, \quad \Gamma^3{}_{23} = \Gamma^3{}_{32} = \cot\theta
\end{aligned}
\tag{6.69}
$$

All the other Christoffel symbols are zero

Let us now consider geodesics in the metric geometry.

Given the metric tensor $g_{\mu\nu}$, we define the geodesic curves by the requirement of extremal proper time,

$$\delta \int_{P_1}^{P_2} ds = 0 \tag{6.70}$$

or

$$\delta \int_{P_1}^{P_2} \sqrt{g_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau}}\,d\tau = 0 \tag{6.71}$$

As we already saw in Section 3.5, the Euler-Lagrange equation that results from this variational principle is

$$\frac{d}{d\tau}\left(g_{\mu\nu}\frac{dx^\nu}{d\tau}\right) - \frac{1}{2}g_{\alpha\beta,\mu}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0 \tag{6.72}$$

Since

$$\frac{dg_{\mu\nu}}{d\tau} = g_{\mu\nu,\alpha}\frac{dx^{\alpha}}{d\tau} \tag{6.73}$$

we can write Eq. (6.72) as

$$g_{\mu\nu}\frac{d^2x^{\nu}}{d\tau^2} + g_{\mu\nu,\alpha}\frac{dx^{\alpha}}{d\tau}\frac{dx^{\nu}}{d\tau} - \frac{1}{2}g_{\alpha\beta,\mu}\frac{dx^{\alpha}}{d\tau}\frac{dx^{\beta}}{d\tau} = 0 \tag{6.74}$$

and, multiplying this by $g^{\sigma\mu}$, we obtain

$$\frac{d^2x^{\sigma}}{d\tau^2} + \frac{1}{2}g^{\sigma\mu}(2g_{\mu\beta,\alpha} - g_{\alpha\beta,\mu})\frac{dx^{\alpha}}{d\tau}\frac{dx^{\beta}}{d\tau} = 0 \tag{6.75}$$

In view of the symmetry of $dx^{\alpha}dx^{\beta}$ in α and β, the term $2g_{\mu\beta,\alpha}(dx^{\alpha}/d\tau)(dx^{\beta}/d\tau)$ equals $(g_{\mu\beta,\alpha} + g_{\mu\alpha,\beta})(dx^{\alpha}/d\tau)(dx^{\beta}/d\tau)$, and we then recognize that the combination of derivatives of the metric tensor appearing in Eq. (6.75) is just the Christoffel symbol $\Gamma^{\sigma}{}_{\alpha\beta}$.

Thus, Eq. (6.75) becomes

$$\frac{d^2x^{\sigma}}{d\tau^2} + \Gamma^{\sigma}{}_{\alpha\beta}\frac{dx^{\alpha}}{d\tau}\frac{dx^{\beta}}{d\tau} = 0 \tag{6.76}$$

This geodesic equation obtained from the metric tensor coincides with the geodesic equation Eq. (6.33) obtained in the preceding section from parallel transport.

The purely mathematical parameter *s* we introduced in Eq. (6.33) is now seen to play the role of proper time.

We know from the preceding section that it is always possible to find coordinates (local geodesic coordinates) such that at some given point *P* of spacetime the Christoffel symbols are zero.

According to Eq. (6.58), the first derivatives of the metric tensor are then also zero at that point.

Furthermore, by an extra coordinate transformation, it is possible to make the metric tensor at that point equal to the metric tensor of flat spacetime, $g''_{\mu\nu} = \eta_{\mu\nu}$.

To see this, suppose that in the local geodesic coordinates $x'^{\mu}_{\mu\nu}$ defined by Eq. (6.35), the metric tensor at the point P is $g'_{\mu\nu}$.

A transformation from the coordinates x'^{μ} to new coordinates x''^{μ} gives $g''_{\mu\nu} = \eta_{\mu\nu}$ (and $g''^{\mu\nu} = \eta^{\mu\nu}$) at the point P provided

$$\eta^{\mu\nu} = \frac{\partial x''^{\mu}}{\partial x'^{\alpha}} \frac{\partial x''^{\nu}}{\partial x'^{\beta}} g'^{\alpha\beta} \tag{6.77}$$

As we already know from Eq. (6.44), this can be regarded as a set of 10 equations for the unknown coefficients $\partial x''^{\mu}/\partial x'^{\nu}$, evaluated at the point P, and this set of equations always has a solution.

But besides making $g''_{\mu\nu} = \eta_{\mu\nu}$, we also want to keep the Christoffel symbols zero.

We can achieve this by demanding a linear relationship between the coordinates x'^{μ} and x''^{μ} in the vicinity of the point P:

$$x''^{\mu} = \left.\frac{\partial x''^{\mu}}{\partial x'^{\nu}}\right|_{P} x'^{\nu} \tag{6.78}$$

Here the coefficients have been labeled with the subscript P to emphasize that they are evaluated at the point P, that is, they are constants.

Such a linear transformation does not change the Christoffel symbols – for a linear transformation, the second term on the right side of Eq. (6.31) is absent, and the transformation leaves the Christoffel symbols unchanged, at zero.

The local geodesic coordinates x'^{μ}, with the extra local condition that the metric tensor be equal to $\eta_{\mu\nu}$ at the origin (the point *P*), can be regarded as the coordinates of a local inertial reference frame, that is, a freely falling small framework of meter sticks and synchronized clocks.

Such a small reference frame, of negligible size and mass, behaves like a test particle.

According to the principle of equivalence, other test particles in free fall in the local neighborhood will then have zero acceleration relative to this reference frame, so the equation of motion of such test particles is

$$\frac{d^2x'^{\sigma}}{d\tau^2} = 0 \tag{6.79}$$

When we transform this equation of motion from the local geodesic coordinates x'^{μ} of our freely falling inertial reference frame to general spacetime coordinates x^{μ}, where $\Gamma^{\alpha}{}_{\mu\nu} \neq 0$, it becomes the general geodesic equation, Eq. (6.76).

Thus, the physical interpretation of geodesics is that they are the worldlines of freely falling test particles.

Note that we can rewrite the general geodesic equation of motion for a particle in the alternative forms

$$\frac{d}{d\tau}\left(\frac{dx^{\sigma}}{d\tau}\right) + \Gamma^{\sigma}{}_{\alpha\beta}\frac{dx^{\alpha}}{d\tau}\frac{dx^{\beta}}{d\tau} = 0 \tag{6.80}$$

and

$$\frac{dp^{\sigma}}{d\tau} + \Gamma^{\sigma}{}_{\alpha\beta}\,p^{\alpha}\frac{dx^{\beta}}{d\tau} = 0 \tag{6.81}$$

where $p^{\sigma} = m dx^{\sigma}/d\tau$ is the momentum of the particle and *m* the rest mass.

These equations say that the four-velocity and the four-momentum of the particle are parallel-transported along the geodesic.

6.5 Riemann curvature tensor

The simplest way to detect curvature in a geometry is to perform parallel transport of a vector around a closed path, as in Fig. 6.1.

Alternatively, we can transport the vector from the initial point P to a final point P' along two alternative paths 1 and 2 connecting P and P'.

The difference between the two resulting vectors at P' is then the same as the difference for transport around the closed path consisting of path 1 followed by path 2 in reverse.

Figure 6.13 shows such a closed path consisting of two alternative paths 1 and 2, both of them constructed with infinitesimal displacements.

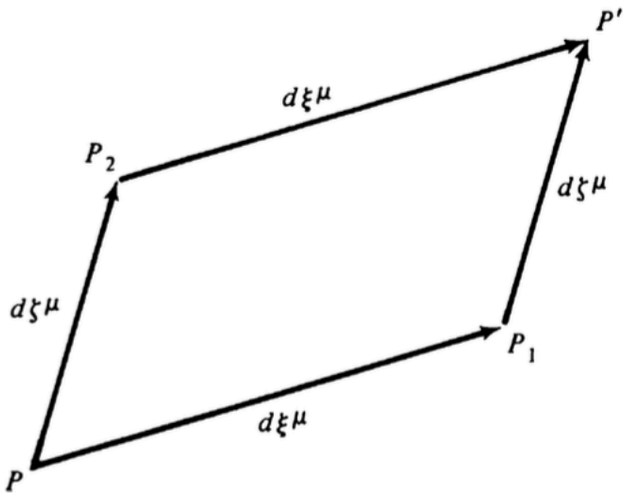


Fig. 6.13 Spacetime points P and P' connected by the two alternative paths PP_1P' and PP_2P'. The displacements $d\xi^\mu$ and $d\zeta^\mu$ are infinitesimal.

The first path PP_1P' consists of a displacement $d\xi^\mu$ followed by a displacement $d\zeta^\mu$.

The second path PP_2P' involves the same displacements, but performed in the opposite order.

The parallel transport of an initial vector A^α from P to P_1 gives the result $A^\alpha - \Gamma^\alpha{}_{\beta\mu}(P)A^\beta d\xi^\mu$, and the further transport of this vector from P_1 to P' gives

$$A^\alpha - \Gamma^\alpha{}_{\beta\mu}(P)A^\beta d\xi^\mu - \Gamma^\alpha{}_{\sigma\nu}(P_1)[A^\sigma - \Gamma^\sigma{}_{\beta\mu}(P)A^\beta d\xi^\mu]d\zeta^\nu \tag{6.82}$$

With $\Gamma^{\alpha}{}_{\sigma\nu}(P_1) \cong \Gamma^{\alpha}{}_{\sigma\nu}(P) + \Gamma^{\alpha}{}_{\sigma\nu,\mu}(P)d\xi^\mu$, we can write (6.82) as

$$A^\alpha - \Gamma^\alpha{}_{\beta\mu}A^\beta d\xi^\mu - \Gamma^\alpha{}_{\sigma\nu}A^\sigma d\zeta^\nu + \Gamma^\alpha{}_{\sigma\nu}\Gamma^\sigma{}_{\beta\mu}A^\beta d\xi^\mu d\zeta^\nu - \Gamma^\alpha{}_{\sigma\nu,\mu}A^\sigma d\xi^\mu d\zeta^\nu \tag{6.83}$$

where a term of third order in the (small) displacements has been omitted and where all Γ symbols are evaluated at the point P.

The parallel transport of A^α along the path $P\,P_2P'$ gives a result similar to (6.83), but with $d\xi$ and $d\zeta$ interchanged:

$$A^\alpha - \Gamma^\alpha{}_{\beta\mu}A^\beta d\zeta^\mu - \Gamma^\alpha{}_{\sigma\nu}A^\sigma d\xi^\nu + \Gamma^\alpha{}_{\sigma\nu}\Gamma^\sigma{}_{\beta\mu}A^\beta d\zeta^\mu d\xi^\nu - \Gamma^\alpha{}_{\sigma\nu,\mu}A^\sigma d\zeta^\mu d\xi^\nu \tag{6.84}$$

If we exchange the dummy indices μ and ν in all the terms in (6.83), we find that the difference between the expressions (6.83) and (6.84) is

$$\Delta A^\alpha = \left(-\Gamma^\alpha{}_{\beta\mu,\nu} + \Gamma^\alpha{}_{\beta\nu,\mu} + \Gamma^\alpha{}_{\sigma\mu}\Gamma^\sigma{}_{\beta\nu} - \Gamma^\alpha{}_{\sigma\nu}\Gamma^\sigma{}_{\beta\mu}\right) A^\beta d\xi^\nu d\zeta^\mu \tag{6.85}$$

This shows that the result of parallel transport is path dependent.

Since the change produced by parallel transport along the path $P\,P_2P'$ is opposite to that produced along the reversed path $P'P_2P$, it is obvious that (6.85) can also be regarded as the change in the vector A^β due to parallel transport around the closed path $PP_1P'P_2P$.

We can rewrite Eq. (6.85) as

$$\Delta A^\alpha = R^\alpha{}_{\beta\mu\nu}A^\beta d\xi^\nu d\zeta^\mu \tag{6.86}$$

where

$$R^\alpha{}_{\beta\mu\nu} \equiv -\Gamma^\alpha{}_{\beta\mu,\nu} + \Gamma^\alpha{}_{\beta\nu,\mu} + \Gamma^\alpha{}_{\sigma\mu}\Gamma^\sigma{}_{\beta\nu} - \Gamma^\alpha{}_{\sigma\nu}\Gamma^\sigma{}_{\beta\mu} \tag{6.87}$$

The expression (6.86) is a vector, because it is the difference between two vectors at the same point P'.

Since a^{β}, $d\xi^{\nu}$, and $d\zeta^{\mu}$ are arbitrary vectors, the factor $R^{\alpha}{}_{\beta\mu\nu}$ multiplying these vectors must be a tensor of rank four, that is, a tensor with $4 \times 4 \times 4 \times 4$ components.

It is called the *Riemann curvature tensor*.

Similarly, we can show that the result for parallel transport of a covariant vector is

$$\Delta B_{\alpha} = -R^{\beta}{}_{\alpha\mu\nu} B_{\beta} d\xi^{\nu} d\zeta^{\mu} \tag{6.88}$$

Rather than repeating the long calculation, we can use the condition $\Delta(B_{\alpha}A^{\alpha}) = 0$ for arbitrary A^{α}.

In a flat spacetime, the result of parallel transport does not depend on the path taken; hence, in flat spacetime we necessarily have $R^{\alpha}{}_{\beta\mu\nu} = 0$. In a curved spacetime, the result of parallel transport depends on the path taken.

The tensor $R^{\alpha}{}_{\beta\mu\nu}$ characterizes the path dependence in parallel transport and serves as a quantitative measure of the curvature of spacetime.

If $R^{\alpha}{}_{\beta\mu\nu} \neq 0$, the spacetime is curved.

We can also show the converse: If $R^{\alpha}{}_{\beta\mu\nu} = 0$ throughout spacetime, then the space-time is flat.

In other words, if the curvature tensor is zero, then the metric tensor can differ from $\eta_{\mu\nu}$ by only a coordinate transformation.

A simple argument is the following: Choose a point *P* and introduce local geodesic coordinates at that point.

Then parallel-transport the coordinate unit vectors from the point P to all other points, and construct coordinate axes at all other points.

If $R^{\alpha}{}_{\beta\mu\nu} = 0$, the result is path independent and therefore uniquely defined (an intuitive, if not quite rigorous, proof of this path independence is sketched in Fig. 6.14).

If the unit vectors and coordinate axes are constructed by parallel transport, any vector transported with respect to these axes will suffer no change.

Hence, $\Gamma^{\mu}{}_{\alpha\beta} = 0$ everywhere, and therefore $g_{\mu\nu}$ is constant.

If $g_{\mu\nu} = \eta_{\mu\nu}$ at the initial point P, the metric tensor must then be $\eta_{\mu\nu}$ everywhere; that is, the spacetime is flat, with a Minkowski metric tensor.

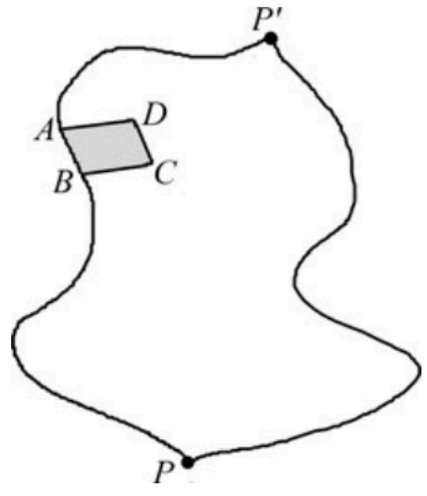


Fig. 6.14 A vector is parallel-transported from an initial point P to an arbitrary point P' along two alternative paths, one path on the left and one path on the right. If $R^{\alpha}{}_{\beta\mu\nu} = 0$, the parallel transport of the vector around the small loop $ABCD$ does not change the vector, and we can therefore replace the original path $PBAP'$ by the shifted path $PBCDAP'$. We can repeat such shifts of path segments in succession, until the entire path on the left has been shifted to coincide with the path on the right. This establishes that the two alternative paths from P to P' give the same result for the parallel transport of the vector.

Given the metric tensor $g_{\mu\nu}$ of some arbitrary spacetime, we can calculate the Riemann curvature tensor and decide unambiguously whether this metric tensor describes flat spacetime expressed in some complicated curvilinear coordinates or whether it describes curved spacetime.

We have seen that, in gravitational theory, $g_{\mu\nu}$ (or $h_{\mu\nu}$) plays a role analogous to that of the vector potential of electrodynamics.

The curvature tensor $R^{\alpha}{}_{\beta\mu\nu}$ plays a role analogous to that of the **E** and **B** fields of electrodynamics.

A gravitational field is said to be present if, and only if, $R^{\alpha}{}_{\beta\mu\nu}$ has a nonzero component.

As we will see in the next section, the tensor $R^{\alpha}{}_{\beta\mu\nu}$ is the relativistic generalization of the tidal-force tensor introduced in Eq. (1.53).

We can lower the first index in $R^{\alpha}{}_{\beta\mu\nu}$ to obtain a purely covariant curvature tensor,

$$R_{\alpha\beta\mu\nu} = g_{\alpha\sigma} R^{\sigma}{}_{\beta\mu\nu} \tag{6.89}$$

This tensor satisfies several important identities:

$$R_{\alpha\beta\mu\nu} = -R_{\beta\alpha\mu\nu} \tag{6.90}$$

$$R_{\alpha\beta\mu\nu} = -R_{\alpha\beta\nu\mu} \tag{6.91}$$

$$R_{\alpha\beta\mu\nu} = R_{\mu\nu\alpha\beta} \tag{6.92}$$

$$R_{\alpha\beta\mu\nu} + R_{\alpha\mu\nu\beta} + R_{\alpha\nu\beta\mu} = 0 \tag{6.93}$$

Although the tensor $R_{\alpha\beta\mu\nu}$ has $4 \times 4 \times 4 \times 4 = 256$ components, only 20 of these components are independent.

Taking into account the identities (6.90) and (6.91), we see that $R_{\alpha\beta\mu\nu}$ can be regarded as an antisymmetric second-rank tensor in α, β and as an antisymmetric second-rank tensor in μ, ν.

Since a second-rank antisymmetric tensor has six independent components, we are then left with no more than 6×6 components for $R_{\alpha\beta\mu\nu}$.

We can write these 6×6 components in the form of a 6×6 matrix R_{AB}, where A stands for the pair $\alpha\beta$ and B stands for the pair $\mu\nu$; the values of A and B may be taken as $01, 02, 03, 12, 13, 23$.

The identity (6.92) says that R_{AB} is a symmetric matrix, and such a 6×6 symmetric matrix has 21 independent components.

Finally, the identity (6.93) imposes one constraint $R_{0123} + R_{0231} + R_{0312} = 0$ among these 21 components [it is easy to see that unless all the indices in Eq. (6.93) are different, this equation coincides with one of Eqs. (6.90)–(6.92)].

Hence we are left with only 20 independent components for the Riemann tensor.
The Riemann tensor is the only fourth-rank tensor that can be constructed by taking linear combinations of the second derivatives of the metric.

To prove this, we can use a simple counting argument.

We begin with geodesic coordinates x'^{μ} with origin at some point P, and we perform a coordinate transformation to a new set of coordinates

$$x''^{\mu} = x'^{\mu} + \tfrac{1}{6} a^{\mu}{}_{\alpha\beta\gamma} x'^{\alpha} x'^{\beta} x'^{\gamma} \tag{6.94}$$

where $a^{\mu}{}_{\alpha\beta\gamma}$ is a constant matrix.

At the point $x'^{\mu} = 0$, this gives

$$\frac{\partial x''^{\mu}}{\partial x'^{\nu}} = \delta^{\mu}_{\nu}, \qquad \frac{\partial^2 x''^{\mu}}{\partial x'^{\alpha} \partial x'^{\beta}} = 0, \qquad \frac{\partial^3 x''^{\mu}}{\partial x'^{\alpha} \partial x'^{\beta} \partial x'^{\gamma}} = a^{\mu}{}_{\alpha\beta\gamma}$$

Since the transformation laws for $g_{\mu\nu}$ and $g_{\mu\nu,\alpha}$ involve only the first and second derivatives of x''^{μ}, it follows that $g''_{\mu\nu} = g'_{\mu\nu}$ and $g''_{\mu\nu,\alpha} = g'_{\mu\nu,\alpha}$.

The transformation law for $g''_{\mu\nu,\alpha\beta}$, obtained from Eq. (6.41) by differentiating twice, involves the third $\mu\nu,\alpha,\beta$ derivative of x''^{μ}, and hence $g''_{\mu\nu,\alpha\beta}$ is not equal to $g'_{\mu\nu,\alpha\beta}$.

In fact, the values of $a^{\mu}{}_{\alpha\beta\gamma}$ are at our disposal, and we can adjust the values of the second derivatives $g''_{\mu\nu,\alpha\beta}$ to some extent.

There are 100 different derivatives $g''_{\mu\nu,\alpha\beta}$.

Since there are 80 independent coefficients $a^{\mu}{}_{\alpha\beta\gamma}$, we can impose 80 conditions on these 100 derivatives. We can therefore conclude that there exist only 20 independent linear combinations of second derivatives that are unaffected by the choice of $a^{\mu}{}_{\alpha\beta\gamma}$.

If we now want to construct tensors out of the second derivatives, we must use these 20 linear combinations exclusively; the reason for this is that the tensor transformation law involves only $\partial x''^{\mu}/\partial x'^{\nu}$, and therefore $a^{\mu}{}_{\alpha\beta\gamma}$ cannot affect it.

What tensors can be constructed with the available 20 linear combinations?

There is one and only one: the tensor $R''_{\alpha\beta\mu\nu}$.

This tensor is constructed of second derivatives, and it has exactly 20 independent components; it therefore exhausts all the viable linear combinations of second derivatives.

The fact that the Riemann tensor is the only tensor that contains the second derivatives of the metric linearly plays an important role in the formulation of the differential equations for the gravitational field (see Sections 7.2 and 7.3).

Besides the purely algebraic identities (6.90)–(6.93), the Riemann tensor also satisfies several differential identities.

The covariant derivative $R^{\alpha}{}_{\beta\mu\nu;\sigma}$ is a tensor of fifth rank.

The following identity holds:

$$R^{\alpha}{}_{\beta\mu\nu;\sigma} + R^{\alpha}{}_{\beta\nu\sigma;\mu} + R^{\alpha}{}_{\beta\sigma\mu;\nu} = 0 \tag{6.95}$$

To derive this equation, we can use an instructive trick.

If we introduce geodesic coordinates, then $\Gamma^{\mu}{}_{\alpha\beta} = 0$ and

$$R^{\alpha}{}_{\beta\mu\nu} = -\Gamma^{\alpha}{}_{\beta\mu,\nu} + \Gamma^{\alpha}{}_{\beta\nu,\mu} \tag{6.96}$$

Therefore Eq. (6.95) reduces to

$$-\Gamma^{\alpha}{}_{\beta\mu,\nu,\sigma} + \Gamma^{\alpha}{}_{\beta\nu,\mu,\sigma} - \Gamma^{\alpha}{}_{\beta\nu,\sigma,\mu} + \Gamma^{\alpha}{}_{\beta\sigma,\nu,\mu} - \Gamma^{\alpha}{}_{\beta\sigma,\mu,\nu} + \Gamma^{\alpha}{}_{\beta\mu,\sigma,\nu} = 0$$

which is a trivial identity.

We have therefore established that Eq. (6.95) is valid in geodesic coordinates.

Since this equation is a tensor equation, its validity in one coordinate system implies its validity in all.

The identity (6.95) has a nice geometric interpretation in terms of parallel transport of a vector along the edges of a small parallelepiped.

If we parallel-transport a vector A^{α} from an initial point *P* (see Fig. 6.15) sequentially along the edges of each of the six faces of the parallelepiped, with opposite directions of transport for adjacent faces, then the net change in A^{α} resulting from this transport is zero, because each edge is traversed twice, in opposite directions.

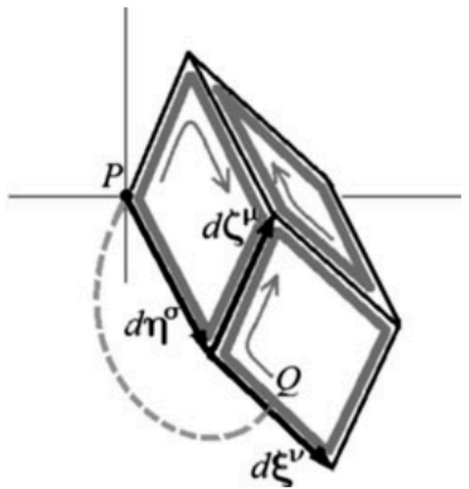


Fig. 6.15 A small parallelepiped $d\xi^{\nu} \times d\zeta^{\mu} \times d\eta^{\sigma}$ shown in local geodesic coordinates. A vector A^{α}, initially at *P*, is parallel-transported sequentially around each face. To reach the faces that do not have a vertex at *P*, such as the front face, we transport the vector along the dashed path *PQ*, then proceed around the face, and then return along the dashed path *QP*. The contributions from *PQ* and *QP* cancel, and only the contribution from transport around the front face remains. The difference between the Riemann tensors at the front and the rear is $R^{\alpha}{}_{\beta\mu\nu}|_{\text{front}} - R^{\alpha}{}_{\beta\mu\nu}|_{\text{rear}} = R^{\alpha}{}_{\beta\mu\nu,\sigma} d\eta^{\sigma}$.

To extract the identity (6.95) from this zero net change condition, consider the contributions from each face and the opposite face, say, the front face in Fig. 6.15 and the rear face.

These faces are traversed in opposite directions, so the sum of the contributions from these two faces is, in local geodesic coordinates,

$$\begin{aligned}\Delta A^{\alpha}|_{\text{front}} + \Delta A^{\alpha}|_{\text{rear}} &= R^{\alpha}{}_{\beta\mu\nu}\big|_{\text{front}} A^{\beta} d\xi^{\nu} d\zeta^{\mu} + R^{\alpha}{}_{\beta\mu\nu}\big|_{\text{rear}} A^{\beta} d\xi^{\mu} d\zeta^{\nu} \\ &= R^{\alpha}{}_{\beta\mu\nu}\big|_{\text{front}} A^{\beta} d\xi^{\nu} d\zeta^{\mu} - R^{\alpha}{}_{\beta\mu\nu}\big|_{\text{rear}} A^{\beta} d\xi^{\nu} d\zeta^{\mu} \\ &= R^{\alpha}{}_{\beta\mu\nu,\sigma} A^{\beta} d\eta^{\sigma} d\xi^{\nu} d\zeta^{\mu}\end{aligned}$$

Similar contributions arise from the left-right faces and the top-bottom faces.

The sum of these contributions is then $\Delta A^{\alpha} = R^{\alpha}{}_{\beta\mu\nu,\sigma} A^{\beta} d\eta^{\sigma} d\xi^{\nu} d\zeta^{\mu} + R^{\alpha}{}_{\beta\nu\sigma,\mu} A^{\beta} d\eta^{\sigma} d\xi^{\nu} d\zeta^{\mu} + R^{\alpha}{}_{\beta\sigma\mu,\nu} A^{\beta} d\eta^{\sigma} d\xi^{\nu} d\zeta^{\mu}$, and because this sum is already known to be identically zero for every choice of the vectors $d\eta^{\sigma}$, $d\xi^{\nu}$, and $d\zeta^{\mu}$, Eq. (6.95) follows.

Thus, this equation represents the geometric condition of zero net change for transport of a vector along the edges of a parallelepiped.

By contraction on the first and last index of $R^{\alpha}{}_{\beta\mu\nu}$ we obtain a second-rank tensor $R_{\beta\mu}$

$$R_{\beta\mu} \equiv R^{\alpha}{}_{\beta\mu\alpha} \tag{6.97}$$

This tensor is called the *Ricci tensor*; it is symmetric in β, μ:

$$R_{\beta\mu} = R_{\mu\beta} \tag{6.98}$$

By a further contraction, we obtain a scalar,

$$R \equiv R^{\beta}{}_{\beta} = R^{\alpha\beta}{}_{\beta\alpha} \tag{6.99}$$

This one-component object is called the *curvature scalar*.

The following differential equation holds identically:

$$\left(R^{\mu}{}_{\nu} - \tfrac{1}{2}\delta^{\mu}_{\nu} R\right)_{;\mu} = 0 \tag{6.100}$$

The identities (6.95) and (6.100) are known as *Bianchi identities*.

As we saw in Section 6.2, covariant derivatives of vectors can be defined in terms of parallel displacements. Equations (6.86) and (6.88) show that successive parallel displacements do not commute, and it is therefore not surprising that successive covariant derivatives also fail to commute.

Thus, when we take consecutive covariant derivatives of a vector or a tensor field, we must pay careful attention to the order in which we carry out the differentiations.

The failure of commutativity of successive parallel displacements involves the Riemann tensor, and the failure of commutativity of successive covariant derivatives likewise involves the Riemann tensor.

For example, suppose we take two consecutive covariant derivatives of a vector field A_β, say, with respect to x^μ and x^ν.

Depending on which derivative we take first, we obtain $A_{\beta;\mu;\nu}$ or $A_{\beta;\nu;\mu}$.

In general, these covariant derivatives are not equal; that is, these derivatives do not commute.
Let us calculate the difference between $A_{\beta;\mu;\nu}$ and $A_{\beta;\nu;\mu}$.

Since $A_{\beta;\mu}$ is a second-rank tensor, we have

$$A_{\beta;\mu;\nu} = A_{\beta;\mu,\nu} - \Gamma^{\sigma}{}_{\beta\nu} A_{\sigma;\mu} - \Gamma^{\sigma}{}_{\mu\nu} A_{\beta;\sigma} \tag{6.104}$$

On the right side of this equation, we must insert the expressions for the covariant derivative of a first-rank tensor:

$$
\begin{aligned}
A_{\beta;\mu;\nu} &= \left(A_{\beta,\mu} - \Gamma^{\alpha}{}_{\beta\mu} A_{\alpha}\right)_{,\nu} - \Gamma^{\sigma}{}_{\beta\nu}\left(A_{\sigma,\mu} - \Gamma^{\alpha}{}_{\sigma\mu} A_{\alpha}\right) - \Gamma^{\sigma}{}_{\mu\nu}\left(A_{\beta,\sigma} - \Gamma^{\alpha}{}_{\beta\sigma} A_{\alpha}\right) \\
&= \left[A_{\beta,\mu,\nu} - \Gamma^{\alpha}{}_{\beta\mu} A_{\alpha,\nu} - \Gamma^{\sigma}{}_{\beta\nu} A_{\sigma,\mu} - \Gamma^{\sigma}{}_{\mu\nu}\left(A_{\beta,\sigma} - \Gamma^{\alpha}{}_{\beta\sigma} A_{\alpha}\right)\right] \\
&\quad - \Gamma^{\alpha}{}_{\beta\mu,\nu} A_{\alpha} + \Gamma^{\sigma}{}_{\beta\nu}\Gamma^{\alpha}{}_{\sigma\mu} A_{\alpha}
\end{aligned}
\tag{6.105}
$$

The corresponding expression for $A_{\beta;\nu;\mu}$ can be obtained from (6.105) by simply exchanging μ and ν.

Obviously the quantity contained in the brackets in Eq. (6.105) is symmetric in μ and ν and hence will not contribute to $A_{\beta;\mu;\nu} - A_{\beta;\nu;\mu}$.

The remaining terms give

$$
A_{\beta;\mu;\nu} - A_{\beta;\nu;\mu} = \left(-\Gamma^{\alpha}{}_{\beta\mu,\nu} + \Gamma^{\sigma}{}_{\beta\nu}\Gamma^{\alpha}{}_{\sigma\mu}\right) A_{\alpha} + \left(\Gamma^{\alpha}{}_{\beta\nu,\mu} - \Gamma^{\sigma}{}_{\beta\mu}\Gamma^{\alpha}{}_{\sigma\nu}\right) A_{\alpha} \tag{6.106}
$$

We can rewrite this as

$$
A_{\beta;\mu;\nu} - A_{\beta;\nu;\mu} = R^{\alpha}{}_{\beta\mu\nu} A_{\alpha} \tag{6.107}
$$

where $R^{\alpha}{}_{\beta\mu\nu}$ is, again, the Riemann curvature tensor introduced in Eq. (6.87).

This reappearance of the curvature tensor is not surprising, because it arises from the parallel transport implicitly involved in the definitions of the covariant derivates; that is, successive covariant derivatives do not commute because the successive parallel transports do not commute.

Table 6.1 summarizes the most important formulas of tensor analysis in Riemannian geometry.

The formulas are given in terms of a first-rank tensor (a vector); the generalization to higher-rank tensors is obvious.

The concepts of derivative along a curve and of geodesic deviation will be discussed in the next section; the corresponding formulas have been included in Table 6.1 for the sake of completeness.

Table 6.1 Tensor-analysis formulas

Transformation of contravariant vector	$A'^{\mu} = \dfrac{\partial x'^{\mu}}{\partial x^{\nu}} A^{\nu}$
Transformation of covariant vector	$A'_{\mu} = \dfrac{\partial x^{\nu}}{\partial x'^{\mu}} A_{\nu}$
Parallel transport	$\delta A^{\mu} = -\Gamma^{\mu}{}_{\nu\beta} A^{\nu} dx^{\beta}$ and $\delta A_{\mu} = \Gamma^{\nu}{}_{\mu\beta} A_{\nu} dx^{\beta}$
Covariant derivative of contravariant vector	$A^{\mu}{}_{;\nu} = A^{\mu}{}_{,\nu} + \Gamma^{\mu}{}_{\alpha\nu} A^{\alpha}$
Covariant derivative of covariant vector	$A_{\mu;\nu} = A_{\mu,\nu} - \Gamma^{\alpha}{}_{\mu\nu} A_{\alpha}$
Spacetime interval and metric tensor	$ds^2 = g_{\mu\nu} dx^{\mu} dx^{\nu}$
Inverse metric tensor	$g_{\mu\alpha} g^{\alpha\nu} = \delta^{\nu}_{\mu}$
Christoffel symbol	$\Gamma^{\alpha}{}_{\mu\nu} = \frac{1}{2} g^{\alpha\beta} (g_{\beta\mu,\nu} + g_{\nu\beta,\mu} - g_{\mu\nu,\beta})$
Geodesic coordinates (at one spacetime point)	$g'_{\mu\nu,\alpha} = 0, \quad g'_{\mu\nu} = \eta_{\mu\nu}$
Geodesic equation	$\dfrac{d^2 x^{\mu}}{d\tau^2} + \Gamma^{\mu}{}_{\alpha\beta} \dfrac{dx^{\alpha}}{d\tau} \dfrac{dx^{\beta}}{d\tau} = 0$
Riemann tensor	$R^{\alpha}{}_{\beta\mu\nu} = -\Gamma^{\alpha}{}_{\beta\mu,\nu} + \Gamma^{\alpha}{}_{\beta\nu,\mu} + \Gamma^{\alpha}{}_{\sigma\mu} \Gamma^{\sigma}{}_{\beta\nu} - \Gamma^{\alpha}{}_{\sigma\nu} \Gamma^{\sigma}{}_{\beta\mu}$
Ricci tensor	$R_{\mu\nu} = R^{\alpha}{}_{\mu\nu\alpha}$
Curvature scalar	$R = R^{\mu}{}_{\mu}$
Change in vector produced by parallel transport around closed path	$\Delta A^{\alpha} = R^{\alpha}{}_{\beta\mu\nu} A^{\beta} d\xi^{\nu} d\zeta^{\mu}$
Commutation relation for covariant derivatives	$A_{\beta;\mu;\nu} - A_{\beta;\nu;\mu} = R^{\alpha}{}_{\beta\mu\nu} A_{\alpha}$
Metric tensor in normal coordinates	$g_{\mu\nu} \cong \eta_{\mu\nu} - (1/3) R(0)_{\mu\alpha\nu\beta} x^{\alpha} x^{\beta}$
Derivative along a curve	$\dfrac{DA^{\mu}}{D\tau} = A^{\mu}{}_{;\nu} \dfrac{dx^{\nu}}{d\tau} = \dfrac{dA^{\mu}}{d\tau} + \Gamma^{\mu}{}_{\alpha\beta} A^{\alpha} \dfrac{dx^{\beta}}{d\tau}$
Geodesic deviation	$\dfrac{D^2 s^{\mu}}{D\tau^2} = -R^{\mu}{}_{\alpha\sigma\beta} s^{\sigma} \dfrac{dx^{\alpha}}{d\tau} \dfrac{dx^{\beta}}{d\tau}$

In this section, we have emphasized the role of the Riemann tensor in parallel transport and the differentiation of vectors.

But the Riemann tensor also plays a direct role in the determination of quintessentially geometric quantities, such as magnitudes of volumes, areas, and circumferences.

For instance, the volume of a small sphere and the area of a small circle deviate from the familiar flat-space values by amounts proportional to appropriate components of the Riemann tensor.

For the calculation of such "metric" deviations from flat-space geometry we can take the local geodesic coordinates x^α as a starting point, and we can regard the metric tensor $g_{\mu\nu} = \eta_{\mu\nu}$ as the zero-order term in a Taylor-series expansion in powers of x^α.

The first-order term is zero, because $g_{\mu\nu,\alpha} = 0$ in local geodesic coordinates.

But the second-order term is not zero, so

$$g_{\mu\nu} \cong \eta_{\mu\nu} + \frac{1}{2}\frac{\partial g(0)_{\mu\nu}}{\partial x^\alpha \partial x^\beta} x^\alpha x^\beta \tag{6.108}$$

Here the argument (0) indicates that the derivatives are evaluated at the origin.

For a general metric tensor, there are 100 distinct second derivatives.

However, we know from the discussion following Eq. (6.94) that, by a clever choice of coordinates, 80 of these derivatives can be eliminated (either because they are zero, or because they can be expressed in terms of other derivatives), so only 20 independent combinations of second derivatives remain, and these correspond to the 20 independent components of the Riemann tensor $R(0)_{\mu\alpha\nu\beta}$.

Accordingly, instead of writing Eq. (6.108) in terms of derivatives of the metric tensor, we can write this equation in terms of components of the Riemann tensor.

When we do this, we must include all possible combinations of these components,

$$g_{\mu\nu} \cong \eta_{\mu\nu} + aR(0)_{\mu\alpha\nu\beta}x^{\alpha}x^{\beta} + [bR(0)_{\mu\nu}\eta_{\alpha\beta} + b'\eta_{\mu\nu}R(0)_{\alpha\beta} + b''R(0)\eta_{\mu\nu}\eta_{\alpha\beta}]x^{\alpha}x^{\beta} \tag{6.109}$$

The values of the *a* and *b* coefficients can be determined by exploiting a consistency requirement: The Riemann tensor calculated by differentiation of the right side of Eq. (6.109) must be equal to $R(0)_{\mu\alpha\nu\beta}$; that is, the differentiation must give us back the Riemann tensor we used for the construction of the right side.

In geodesic coordinates, the general formula (6.87) for the Riemann tensor reduces to

$$R_{\alpha\beta\mu\nu} = \left(-\frac{1}{2}\right)(g_{\mu\alpha,\beta,\nu} - g_{\beta\mu,\alpha,\nu} - g_{\nu\alpha,\beta,\mu} + g_{\beta\nu,\alpha,\mu}) \tag{6.110}$$

It is then easy to check that when we insert Eq. (6.109) into the right side of Eq. (6.110), we obtain the desired identity $R_{\alpha\beta\mu\nu} = R(0)_{\alpha\beta\mu\nu}$ if and only if $a = -1/3$ and $b = b' = b'' = 0$.

Hence the correct expression for the Taylor-series expansion of the metric tensor takes the simple form

$$g_{\mu\nu} \cong \eta_{\mu\nu} - \frac{1}{3}R(0)_{\mu\alpha\nu\beta}x^{\alpha}x^{\beta} \tag{6.111}$$

Of course, this expression depends on a special choice of coordinates; that is, a special extrapolation of the geodesic coordinates.

These special coordinates are called *normal coordinates*.

A characteristic feature of these coordinates is that, along any radial line from the origin to some given point X^{μ}, the invariant spacetime interval is simply $\int ds = (X^{\mu}X_{\mu})^{1/2}$, which is the same result as for flat spacetime [this assumes that X^{μ} is at a finite distance, but within the range of validity of the approximation (6.111)].

To confirm this result, parametrize the line from the origin to the point X^μ by writing the coordinate along the line as $x^\mu = \lambda X^\mu$, where λ ranges from 0 (at the origin) to 1 (at the final point X^μ).

Then $dx^\mu = X^\mu d\lambda$, and the spacetime interval for the metric tensor (6.111) becomes $ds^2 = g_{\mu\nu}dx^\mu dx^\nu = \eta_{\mu\nu}X^\mu X^\nu (d\lambda)^2 - (1/3)R(0)_{\mu\alpha\nu\beta}X^\alpha X^\beta X^\mu X^\nu \lambda^2 (d\lambda)^2$.

Here, the second term is zero, because of the antisymmetry of $R(0)_{\mu\alpha\nu\beta}$ in μ, α, and the first term is exactly the same as for flat spacetime.

However, note that this simple result is valid only for radial displacements; the result for tangential displacements differs from that in flat spacetime and is affected by the presence of the curvature tensor in Eq. (6.111).

From the expression for $g_{\mu\nu}$, we can readily calculate volumes, areas, and lengths for geometric figures of small size.

For instance, the volume of a sphere of radius L (at one instant of time, $x^0 = 0$) can be shown to be

$$\text{volume} \cong \frac{4\pi L^3}{3}\left(1 + \frac{L^2}{30}R(0)_{knkn}\right) \tag{6.112}$$

Similar results can be obtained for the area of a circle or the circumference of a circle.

6.6 Geodesic deviation and tidal forces; Fermi-Walker transport

We defined the Riemann curvature tensor in terms of the entirely geometric concept of parallel transport.

But this tensor also has a simple physical interpretation: $R^{\alpha}{}_{\beta\mu\nu}$ is the tidal-force field, that is, it gives the relative acceleration between two particles in free fall.

Before we can calculate this relative acceleration, we need to introduce the concept of differentiation along a curve.

For this purpose, consider a curve in spacetime, possibly, but not necessarily, a geodesic.

We can specify displacements along the curve by giving the corresponding increment in the "length" or the proper time τ measured along this curve.

Suppose that the curve is immersed in a vector field A^{μ}.

If we want to know the rate of change of A^{μ} along the curve, we must calculate $dA^{\mu}/d\tau$.

However, this is not a good measure of how much A^{μ} is "really" changing, since part, or maybe all, of the contribution to $dA^{\mu}/d\tau$ could be due to the curvilinear coordinates used to define the components A^{μ}.

Also, $dA^{\mu}/d\tau$ is not a vector (for the same reason that $\partial A^{\mu}/\partial x^{\nu}$ is not a tensor).

A better measure of the rate of change of A^{μ} along the curve is the vector quantity

$$\frac{DA^{\mu}}{D\tau} \equiv A^{\mu}{}_{;\nu}\frac{dx^{\nu}}{d\tau} = \frac{dA^{\mu}}{d\tau} + \Gamma^{\mu}{}_{\alpha\beta}A^{\alpha}\frac{dx^{\beta}}{d\tau} \tag{6.113}$$

This is a vector, because it is the product of the tensor $A^{\mu}{}_{;\beta}$ and the vector $dx^{\beta}/d\tau$.

The quantity $DA^{\mu}/D\tau$ is called the *derivative along the curve*, or the *absolute derivative*.

In local geodesic coordinates (in which $\Gamma^{\mu}{}_{\alpha\beta} = 0$), the derivative $DA^{\mu}/D\tau$ along a curve reduces to the ordinary derivative $dA^{\mu}/d\tau$.

In terms of parallel transport, we can understand Eq. (6.113) as follows: $DA^{\mu}/D\tau$ is obtained from $dA^{\mu}/d\tau$ by subtracting a correction that represents the parallel transport of the vector A^{μ} along the curve [compare Eq. (6.20)].

Note that if the vector field $A^{\mu}(x)$ is a "constant" vector field, obtained by parallel transporting some initial vector from an initial point to all regions of spacetime, then the derivative $DA^{\mu}/D\tau$ is zero.

The expression (6.113) is valid for any curve. Let us now concentrate on geodesics and find the second derivative along a geodesic curve:

$$
\begin{aligned}
\frac{D^2A^{\mu}}{D\tau^2} &= \frac{D}{D\tau}\left(\frac{DA^{\mu}}{D\tau}\right) = \frac{d}{d\tau}\left(\frac{DA^{\mu}}{D\tau}\right) + \Gamma^{\mu}{}_{\alpha\beta}\frac{DA^{\alpha}}{D\tau}\frac{dx^{\beta}}{d\tau} \\
&= \frac{d}{d\tau}\left(\frac{dA^{\mu}}{d\tau} + \Gamma^{\mu}{}_{\alpha\beta}A^{\alpha}\frac{dx^{\beta}}{d\tau}\right) + \Gamma^{\mu}{}_{\alpha\beta}\left(\frac{dA^{\alpha}}{d\tau} + \Gamma^{\alpha}{}_{\kappa\lambda}A^{\kappa}\frac{dx^{\lambda}}{d\tau}\right)\frac{dx^{\beta}}{d\tau} \\
&= \frac{d^2A^{\mu}}{d\tau^2} + \Gamma^{\mu}{}_{\alpha\beta,\nu}\frac{dx^{\nu}}{d\tau}A^{\alpha}\frac{dx^{\beta}}{d\tau} + 2\Gamma^{\mu}{}_{\alpha\beta}\frac{dA^{\alpha}}{d\tau}\frac{dx^{\beta}}{d\tau} \\
&\quad - \Gamma^{\mu}{}_{\alpha\beta}A^{\alpha}\Gamma^{\beta}{}_{\kappa\lambda}\frac{dx^{\kappa}}{d\tau}\frac{dx^{\lambda}}{d\tau} + \Gamma^{\mu}{}_{\alpha\beta}\Gamma^{\alpha}{}_{\kappa\lambda}A^{\kappa}\frac{dx^{\lambda}}{d\tau}\frac{dx^{\beta}}{d\tau}
\end{aligned}
\tag{6.114}
$$

The next to last term in this expression arises from evaluating $(d/d\tau)(dx^\beta/d\tau)$ with the geodesic equation.

We can now exploit this expression to calculate the relative acceleration of two particles moving on neighboring geodesics.

First, we write down the equation of motion of each particle.

Suppose that for a given value of τ along each geodesic, the coordinates of the particles are $x^\mu(\tau)$ and $x^\mu(\tau) + s^\mu(\tau)$ (see Fig. 6.16).

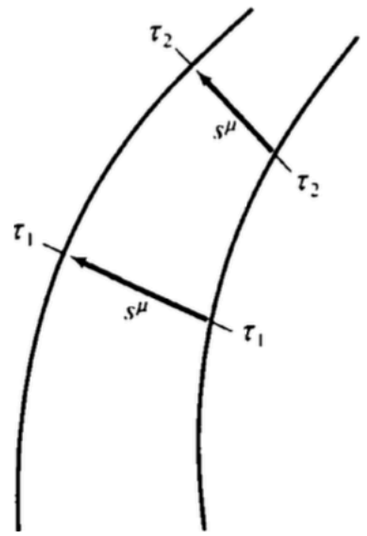


Fig. 6.16 Each of two adjacent geodesics is parametrized by its proper time τ. The displacement vector s^μ connects pairs of points with the same value of τ on the two geodesics.

Then the geodesic equations for the two particles are, respectively,

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu{}_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0 \tag{6.115}$$

and

$$\frac{d^2(x^\mu + s^\mu)}{d\tau^2} + \Gamma^\mu{}_{\alpha\beta}(x+s)\left(\frac{dx^\alpha}{d\tau} + \frac{ds^\alpha}{d\tau}\right)\left(\frac{dx^\beta}{d\tau} + \frac{ds^\beta}{d\tau}\right) = 0 \tag{6.116}$$

We will assume that s^μ and $ds^\mu/d\tau$ are infinitesimal; this means that the particles are near to each other and remain near for a fairly long time.

If we approximate $\Gamma^{\mu}{}_{\alpha\beta}(x+s) \cong \Gamma^{\mu}{}_{\alpha\beta}(x) + \Gamma^{\mu}{}_{\alpha\beta,\sigma}\, s^{\sigma}$ and keep only first-order terms in s^{σ}, the difference between Eqs. (6.115) and (6.116) yields

$$\frac{d^2 s^{\mu}}{d\tau^2} = -\Gamma^{\mu}{}_{\alpha\beta,\sigma} s^{\sigma} \frac{dx^{\alpha}}{d\tau}\frac{dx^{\beta}}{d\tau} - 2\Gamma^{\mu}{}_{\alpha\beta}\frac{ds^{\alpha}}{d\tau}\frac{dx^{\beta}}{d\tau} \tag{6.117}$$

The absolute second derivative $D^2 s^{\mu}/D\tau^2$ of s^{μ} along the geodesic can be calculated by substituting Eq. (6.117) into(6.114), with $A^{\mu} = s^{\mu}$.

The several Christoffel symbols and their derivatives then combine to give the simple final result

$$\frac{D^2 s^{\mu}}{D\tau^2} = -R^{\mu}{}_{\alpha\sigma\beta} s^{\sigma} \frac{dx^{\alpha}}{d\tau}\frac{dx^{\beta}}{d\tau} \tag{6.118}$$

Equation (6.118) is called the equation of *geodesic deviation*.

The deviation between neighboring geodesics is a tidal effect, and we can attribute it to a tidal force.

Suppose that the particles moving on the two geodesics have equal masses *m* and that in local freely falling coordinates they are moving slowly, so

$$\frac{dx'^{\alpha}}{d\tau} \cong (1, 0, 0, 0) \tag{6.119}$$

Furthermore, suppose that $s'^{0} = 0$; this simply means that the particle accelerations are compared at equal times. Then Eq. (6.118) reduces to

$$\frac{d^2 s'^{k}}{d\tau^2} \cong -R'^{k}{}_{0l0} s'^{l} \tag{6.120}$$

The tidal force is therefore

$$f^{k} \cong -m R'^{k}{}_{0l0} s'^{l} \tag{6.121}$$

This equation is valid only if the displacement s^{l} and the velocity $ds^{l}/d\tau$ are small.

To establish that this equation for the tidal force is in agreement with the Newtonian expression [Eq. (1.51)], we must check that the old [Eq. (1.50)] and new [Eq. (6.87)] definitions of $R^k{}_{0l0}$ coincide in the case of weak static gravitational fields.

In the linear approximation, the two terms quadratic in $\Gamma^\alpha{}_{\beta\mu}$ can be omitted from Eq. (6.87).

Furthermore, in this approximation,

$$\Gamma^\alpha{}_{\beta\mu} \cong \frac{\kappa}{2}\eta^{\alpha\sigma}(h_{\sigma\beta,\mu} + h_{\mu\sigma,\beta} - h_{\beta\mu,\sigma}) \tag{6.122}$$

This results in

$$R^\alpha{}_{\beta\mu\nu} \cong -\frac{\kappa}{2}\eta^{\alpha\sigma}(h_{\mu\sigma,\beta,\nu} - h_{\beta\mu,\sigma,\nu} - h_{\nu\sigma,\beta,\mu} + h_{\beta\nu,\sigma,\mu}) \tag{6.123}$$

Therefore

$$R^k{}_{0l0} \cong \frac{\kappa}{2}(h_{lk,0,0} - h_{0l,k,0} - h_{0k,0,l} + h_{00,k,l}) \tag{6.124}$$

In the static case, all the terms containing time derivatives can be omitted.

This leaves only a single term involving h_{00}, which, by Eq. (4.10), is proportional to the Newtonian potential,

$$\frac{1}{2}\kappa h_{00} = \Phi \tag{6.125}$$

We then find

$$R^k{}_{0l0} = \frac{\partial^2\Phi}{\partial x^k \partial x^l} \tag{6.126}$$

which is in agreement with Eq. (1.50).

The tidal-force equation (6.121) can be used for measuring the components $R^k{}_{0l0}$ of the Riemann tensor.

We already discussed in Section 1.8 how such measurements of the tidal force can be carried out; we emphasize once more that these measurements can be performed *locally*.

What about the other components of the Riemann tensor?

These components can be obtained by measuring the components $R^{k}{}_{0l0}$ in a sufficiently large number of other reference frames that have some velocity relative to the original reference frame.

Lorentz transformations between these reference frames relate the components $R^{k}{}_{0l0}$ in one frame to the general components $R^{\alpha}{}_{\beta\mu\nu}$ in another and therefore can be used to find all the components.

In principle, the Riemann tensor is therefore entirely determined by tidal-force measurements.

Finally, we can show that the tidal force (6.121) satisfies the condition

$$\frac{\partial f^k}{\partial x^k} = 0 \tag{6.127}$$

This means that the tidal-force field can be represented graphically by field lines.

Of course, we already came across the condition (6.127) in Newton's theory [Eq. (1.54)] and in the linear tensor theory [Eq. (5.45)], but for a general proof we need to use the exact field equations.

According to Eq. (6.121),

$$\frac{\partial f'^k}{\partial x'^k} = -mR'^{k}{}_{0l0}\frac{\partial s'^l}{\partial s'^k} = -mR'^{k}{}_{0k0} \tag{6.128}$$

Since $R^{\alpha}{}_{\beta\mu\nu}$ is antisymmetric in μ and ν, we have $R'^{0}{}_{000} = 0$, and Eq. (6.128) is therefore equivalent to

$$\frac{\partial f'^k}{\partial x'^k} = -mR'^{\mu}{}_{0\mu0} = mR'_{00} \tag{6.129}$$

$R'_{00} \equiv R'^{\mu}{}_{00\mu} = -R'^{\mu}{}_{0\mu0}$ is a component of the Ricci tensor.

As we will see in the next chapter, Einstein's field equations in vacuo tell us that $R_{\mu\nu} = 0$, and hence Eq. (6.127) is valid as long as we remain outside of the mass distribution that generates the gravitational field.

The derivative along a curve also has an important application to the problem of vector transport along a nongeodesic worldline.

In Section 6.3, we saw that to construct Fermi coordinates along a geodesic we need to parallel-transport a set of unit vectors (a tetrad) along the geodesic.

Ordinary parallel transport keeps the time axis and its unit vector aligned with the geodesic – this time unit vector is simply the four-velocity, which is evidently always aligned with the worldline and is parallel-transported along the worldline, if the worldline is a geodesic.

But if the worldline is not a geodesic, then u^α is aligned along the worldline, but it is not parallel-transported.

Instead of parallel transport, we need to invent a different kind of transport, so that u^α becomes a transported vector.

The appropriate kind of transport is *Fermi-Walker transport*, according to the differential equation

$$\frac{DA^\alpha}{D\tau} = \left(u^\beta \frac{Du^\alpha}{D\tau} - u^\alpha \frac{Du^\beta}{D\tau} \right) A_\beta \tag{6.130}$$

If the vector A^α is specified at some initial point of a nongeodesic worldline, then this differential equation determines the vector along the entire worldline.

It is easy to see that the four-velocity u^α is always a solution of this equation and therefore satisfies our new transport criterion.

Furthermore, for any pair of vectors transported according to Eq. (6.130), the scalar product $A^\alpha B_\alpha$ remains constant during this transport.

Thus, Fermi-Walker transport of a set of orthogonal unit vectors keeps these unit vectors normalized and orthogonal.

Note that if the worldline is a geodesic, with $Du^\alpha/D\tau = 0$, then Eq. (6.130) reduces to $DA^\alpha/D\tau = 0$, which means that, along a geodesic, Fermi-Walker transport coincides with parallel transport.

We can gain a clearer understanding of the significance of Eq. (6.130) by introducing local geodesic coordinates at some given point of the nongeodesic worldline, so the four-velocity at this point becomes $u^\alpha = (1,0,0,0)$.

Furthermore, for convenience, let us place the x^1-axis in the direction of the instantaneous acceleration [the acceleration must have a spacelike direction, because $0 = d(u^\alpha u_\alpha) = 2u^\alpha du_\alpha$, which requires $du_0 = 0$ in our geodesic coordinates with $u^\alpha = (1, 0, 0, 0)$]. In our geodesic coordinates we can then write $Du^1/D\tau = dv^1/dt$, and Eq. (6.130) becomes

$$\frac{dA^0}{dt} = -\frac{dv^1}{dt}A_1 \qquad \frac{dA^1}{dt} = \frac{dv^1}{dt}A_0 \qquad \frac{dA^2}{dt} = 0 \qquad \frac{dA^3}{dt} = 0 \tag{6.131}$$

These equations indicate that the changes in the components of the vector are $A^0 \rightarrow A^0 + (dv^1)A^1$ and $A^1 \rightarrow A^1 + (dv^1)A^0$, which are exactly the changes that occur when we give this vector a boost by performing an infinitesimal Lorentz transformation.

Hence in Fermi-Walker transport, the transported vector is merely accelerated along the nongeodesic worldline, in the direction of the instantaneous acceleration, and no other change is made to the vector.

In particular, the vector is not rotated during each of the infinitesimal Lorentz transformations, although, of course, by compounding such accelerations in succession, the spatial direction of the vector will rotate, as in the well-known case of the Thomas precession of special relativity (the Thomas precession can actually be deduced from the Fermi-Walker transport equation in flat spacetime.

To summarize: In Fermi-Walker transport, the timelike unit vector simply maintains its alignment with the worldline, while the spacelike unit vectors perform successive infinitesimal Lorentz transformations in the direction of the instantaneous acceleration.

The final direction of the timelike unit vector is that of the worldline, but the directions of the spacelike unit vectors depend on the detailed acceleration history during the transport.

Fermi-Walker transport permits us to construct convenient Fermi coordinates along a nongeodesic worldline (although not completely convenient, because, in contrast to Fermi coordinates along a geodesic, the Christoffel symbols are not zero).

Fermi-Walker transport can be implemented physically by the transport of a torque-free gyroscope, that is, a gyroscope free of tidal torques and propelled by a force acting on its center of mass.

The spin axis of such a gyroscope tracks the Fermi-Walker transport of a spatial vector.

In the case of a geodesic worldline, this means that the spin axis is parallel-transported.

We will deal with this case in detail in Section 7.7.

6.7 Differential forms in curved spacetime

In the preceding sections we have developed the mathematics of Riemannian geometry in component language.

In this section we will redevelop the mathematics in the language of tangent vectors and differential forms.

This gives us more concise equations and also permits us to derive some identities in a more elegant manner.

The notions of tangent vector and of differential forms are equally valid in curved spacetime and in flat spacetime.

As in Section 2.6, the tangent vectors are linear differential operators defined on test functions, and 1-forms are duals of tangent vectors.

Given an arbitrary system of coordinates, we can introduce the basis vectors $\partial/\partial x^{\mu}$ and the basis 1-forms $\mathbf{d}x^{\mu}$, so any tangent vector $\mathbf{u}$ and any 1-form $\boldsymbol{\alpha}$ can be expressed as a superposition of basis vectors or basis 1-forms, respectively:

$$\mathbf{u} = u^{\mu}\frac{\partial}{\partial x^{\mu}} \tag{6.132}$$

$$\boldsymbol{\alpha} = \alpha_{\mu}\mathbf{d}x^{\mu} \tag{6.133}$$

When the basis 1-forms act on the basis vectors, the result is the same as in flat space- time,

$$\mathbf{d}x^{\mu}\left(\frac{\partial}{\partial x^{\nu}}\right) = \frac{\partial x^{\mu}}{\partial x^{\nu}} = \delta^{\mu}_{\nu} \tag{6.134}$$

Under a transformation of coordinates, the components u^{μ} and α_{μ} and Eqs. (6.132) and (6.133) obey the usual transformation laws for contravariant and covariant vector components, respectively.

As in Section 2.6, these transformation laws follow directly from the transformation laws for the basis vectors and the basis 1-forms,

$$\frac{\partial}{\partial x'^{\nu}} = \frac{\partial x^{\mu}}{\partial x'^{\nu}}\frac{\partial}{\partial x^{\mu}} \tag{6.135}$$

$$\mathbf{d}x'^{\nu} = \frac{\partial x'^{\nu}}{\partial x^{\mu}}\mathbf{d}x^{\mu} \tag{6.136}$$

The notion of tensor can also be taken over unchanged from Section 2.6.

For instance, a tensor of type $\binom{1}{2}$ is a linear map that yields a real number when acting on a 1-form and on two tangent vectors.

Such a tensor can be written as a superposition of tensor products of the basis vectors and basis 1-forms,

$$\mathbf{T} = T^{\sigma}{}_{\mu\nu} \frac{\partial}{\partial x^{\sigma}} \otimes \mathbf{d}x^{\mu} \otimes \mathbf{d}x^{\nu} \tag{6.137}$$

It is easy to show that, under a coordinate transformation, the components $T^{\sigma}{}_{\mu\nu}$ of such a tensor obey the usual transformation law, contravariant in the first index, and covariant in the other two indices.

In the language of differential forms, the definition of the covariant derivative requires parallel transport, in much the same way as in component language.

However, the definition of the covariant derivative per se is a bit awkward, and it is easier to begin with the covariant derivative along a curve [compare Eq. (6.113)].

Suppose that the curve is parametrized by λ, and that, at a given point, the curve has the tangent vector $\mathbf{u} = d/d\lambda$.

Along this curve, an arbitrary tensor field $\mathbf{T}$ can be regarded as a function of λ.

The covariant derivative of such a tensor field $\mathbf{T}$ along the curve is then defined by

$$\nabla_{\mathbf{u}}\mathbf{T} = \lim_{d\lambda \to 0} \frac{\mathbf{T}(\lambda + d\lambda) - \mathbf{T}(\lambda) - \delta\mathbf{T}}{d\lambda} \tag{6.138}$$

where, as in Section 6.2, $\delta\mathbf{T}$ is the change in $\mathbf{T}$ produced by parallel transport.

Note that this covariant derivative along a curve is a tensor of the same rank as the original tensor.

The covariant derivative per se can next be constructed in terms of the covariant derivative along a curve.

To accomplish this construction, we demand that the covariant derivative $\nabla\mathbf{T}$ be a tensor of rank higher by one than $\mathbf{T}$, such that when $\nabla\mathbf{T}$ acts on an arbitrary tangent vector $\mathbf{u}$ the result is $\nabla_{\mathbf{u}}\mathbf{T}$.

However, in most of the calculations hereafter, we will find it sufficient to deal with the covariant derivative along a curve, and we will not need the covariant derivative per se.

The definition (6.138) is valid for tensors of arbitrary rank. In the special case of a function f (that is, a tensor of rank zero), the covariant derivative reduces to an ordinary derivative,

$$\nabla_{\mathbf{u}} f = \frac{df}{d\lambda} \tag{6.139}$$

For a tensor product, the derivative obeys the usual rules,

$$\nabla_{\mathbf{u}}(\mathbf{T} \otimes \mathbf{R}) = \nabla_{\mathbf{u}}\mathbf{T} \otimes \mathbf{R} + \mathbf{T} \otimes \nabla_{\mathbf{u}}\mathbf{R} \tag{6.140}$$

To discover the components of the covariant derivative, take a curve that coincides with one of the lines of the coordinate grid, so the tangent vector $\mathbf{u}$ coincides with one of the basis vectors, say, $\mathbf{u} = \partial/\partial x^{\mu}$.

For convenience, we will adopt the notation $\nabla_{\partial/\partial x^{\mu}} \equiv \nabla_{\mu}$, because it is awkward to write $\partial/\partial x^{\mu}$ as a subscript.

To evaluate the derivative ∇_{μ} of some vector $\mathbf{w}$, we write the vector in terms of basis vectors,

$$\mathbf{w} = w^{\nu} \frac{\partial}{\partial x^{\nu}} \tag{6.141}$$

If we differentiate this term by term, we find

$$\nabla_{\mu}\mathbf{w} = \frac{\partial w^{\nu}}{\partial x^{\mu}} \frac{\partial}{\partial x^{\nu}} + w^{\nu} \nabla_{\mu} \frac{\partial}{\partial x^{\nu}} \tag{6.142}$$

where the first term arises from differentiating w^{ν}, which for present purposes can be treated as a function multiplying the tangent vector $\partial/\partial x^{\nu}$.

Accordingly, the components of this derivative are $\partial w^{\nu}/\partial x^{\mu} + w^{\nu} \nabla_{\mu}$.

Now suppose that the vector field **w** has been constructed by parallel transport.

For such a vector field, the derivative is zero,

$$\frac{\partial w^{\nu}}{\partial x^{\mu}} \frac{\partial}{\partial x^{\nu}} + w^{\nu} \nabla_{\boldsymbol{\mu}} \frac{\partial}{\partial x^{\nu}} = 0 \tag{6.143}$$

To extract the components from this equation, we act with the dual vector $\mathbf{d}x^{\alpha}$ on it:

$$\frac{\partial w^{\nu}}{\partial x^{\mu}} \mathbf{d}x^{\alpha} \left(\frac{\partial}{\partial x^{\nu}} \right) + w^{\nu} \mathbf{d}x^{\alpha} \left(\nabla_{\boldsymbol{\mu}} \frac{\partial}{\partial x^{\nu}} \right) = 0 \tag{6.144}$$

so

$$\frac{\partial w^{\alpha}}{\partial x^{\mu}} + w^{\nu} \mathbf{d}x^{\alpha} \left(\nabla_{\boldsymbol{\mu}} \frac{\partial}{\partial x^{\nu}} \right) = 0 \tag{6.145}$$

This tells us how the vector components w^{ν} change under parallel transport.

Comparing this with the standard expression for parallel transport in component language, $\delta w^{\alpha} + \Gamma^{\alpha}{}_{\mu\nu} w^{\nu} dx^{\mu} = 0$, or

$$\frac{\partial w^{\alpha}}{\partial x^{\mu}} + \Gamma^{\alpha}{}_{\mu\nu} w^{\nu} = 0 \tag{6.146}$$

we see that

$$\mathbf{d}x^{\alpha} \left(\nabla_{\boldsymbol{\mu}} \frac{\partial}{\partial x^{\nu}} \right) = \Gamma^{\alpha}{}_{\mu\nu} \tag{6.147}$$

Thus, the covariant derivatives of basis vectors are directly related to the Christoffel symbols.
It is straightforward to show that

$$(\nabla_{\boldsymbol{\mu}} \mathbf{d}x^{\alpha}) \left(\frac{\partial}{\partial x^{\nu}} \right) = -\Gamma^{\alpha}{}_{\mu\nu} \tag{6.148}$$

In the language of differential forms, the geodesic equation asserts that the tangent vector $\mathbf{u} = d/d\lambda$ is parallel-transported along the curve,

$$\nabla_{\mathbf{u}}\mathbf{u} = 0 \tag{6.149}$$

If we write this out in components, it yields the standard form (6.33) of the geodesic equation.

The metric tensor $\mathbf{g}$ of the curved spacetime is a second-rank tensor of the type $\binom{0}{2}$.

Acting on a pair of vectors, this tensor produces a real number $\mathbf{g}(\mathbf{u}, \mathbf{w})$.

In terms of basis 1-forms, the metric tensor can be expressed as

$$\mathbf{g} = g_{\mu\nu}\mathbf{d}x^{\mu} \otimes \mathbf{d}x^{\nu} \tag{6.150}$$

The inverse metric tensor $\mathbf{g}^{-1}$ is a second-rank tensor of the type $\binom{2}{0}$, such that (6.150)

$$\mathbf{g}\mathbf{g}^{-1} = \mathbf{g}^{-1}\mathbf{g} = \mathbf{1} \tag{6.151}$$

In terms of basis vectors,

$$\mathbf{g}^{-1} = g^{\mu\nu}\frac{\partial}{\partial x^{\mu}} \otimes \frac{\partial}{\partial x^{\nu}} \tag{6.152}$$

It is then easy to verify that Eq. (6.151) is satisfied, provided that

$$g_{\alpha\mu}g^{\alpha\nu} = \delta^{\nu}_{\mu} \tag{6.153}$$

The covariant derivative of the metric tensor must be identically zero,

$$\nabla_{\mathbf{u}}\mathbf{g} \equiv 0 \tag{6.154}$$

so that $\mathbf{g}(\mathbf{v}, \mathbf{w})$= constant whenever $\mathbf{v}$ and $\mathbf{w}$ are parallel-transported vectors.

In terms of components, this requirement implies the usual relation between $g_{\mu\nu,\alpha}$ and the Christoffel symbols.

The definition of the Riemann curvature tensor can be introduced by examining the difference between successive covariant derivatives taken in opposite order.

Thus, consider a 1-form α, and compare the covariant derivatives $\nabla_\mu \nabla_\nu \alpha$ and $\nabla_\nu \nabla_\mu \alpha$, evaluated at some given point.

The difference between these derivatives is

$$(\nabla_\mu \nabla_\nu - \nabla_\nu \nabla_\mu)\boldsymbol{\alpha} \tag{6.155}$$

We can show that this difference depends only on the value of the 1-form at the given point, but does not depend on the value of the 1-form at nearby points; that is, the derivative terms in the definition of the second covariant derivatives cancel against each other, leaving only the terms arising from parallel transport, which are linear functions of the 1-form.

The proof can be most easily given by using components, as we did in Eqs. (6.104)–(6.107), but for purists there is a rather more tedious proof that does not use components.

In any case, the operator ($\nabla_\mu \nabla_\nu - \nabla_\nu \nabla_\mu$) generates a linear map from the1-form α to the 1-form ($\nabla_\mu \nabla_\nu - \nabla_\nu \nabla_\mu$) α . If we let this1-form act on a vector **s**, the result is a real number.

Since this real number also depends on the basis vectors $\partial/\partial x^\mu$ and $\partial/\partial x^\nu$, as indicated by the abbreviated subscript notation in Eq. (6.155), we write it as $\mathbf{R}(\alpha, \partial/\partial x^\mu, \partial/\partial x^\nu, \mathbf{s})$; that is,

$$(\nabla_\mu \nabla_\nu - \nabla_\nu \nabla_\mu)\boldsymbol{\alpha}(\mathbf{s}) = \mathbf{R}\left(\boldsymbol{\alpha}, \frac{\partial}{\partial x^\mu}, \frac{\partial}{\partial x^\nu}, \mathbf{s}\right) \tag{6.156}$$

Here we see that **R** is a map from one 1-form α and three tangent vectors (two of which are basis vectors, $\partial/\partial x^\mu$ and $\partial/\partial x^\nu$) to a real number. Thus, **R** is a tensor of type $\binom{1}{3}$.

Actually, for the general definition of the tensor **R**, it is not sufficient to examine its action on the basis vectors $\partial/\partial x^\mu$ and $\partial/\partial x^\nu$, but on general tangent vectors, such as $\mathbf{u} = u^\mu \partial/\partial x^\mu$ and $\mathbf{w} = w^\nu \partial/\partial x^\nu$. Such vectors do not commute, since the derivative in each acts on the coefficients appearing in the other,

$$\mathbf{uw} - \mathbf{wu} = u^\mu \frac{\partial}{\partial x^\mu}\left(w^\nu \frac{\partial}{\partial x^\nu}\right) - w^\nu \frac{\partial}{\partial x^\nu}\left(u^\mu \frac{\partial}{\partial x^\mu}\right)$$
$$= u^\mu \frac{\partial w^\nu}{\partial x^\mu}\frac{\partial}{\partial x^\nu} - w^\nu \frac{\partial u^\mu}{\partial x^\nu}\frac{\partial}{\partial x^\mu}$$

Thus, the commutator of two tangent vectors is, again, a tangent vector, different from zero.

The extra term resulting from the noncommutativity of **u** and **w** spoils the cancellation of derivatives in the operator on the left side of Eq. (6.156) and therefore destroys the tensorial character of the operator.

But we can restore the cancellation by subtracting **uw − wu**, which gives us a corrected operator:

$$(\nabla_\mathbf{u}\nabla_\mathbf{w} - \nabla_\mathbf{w}\nabla_\mathbf{u} - \mathbf{uw} + \mathbf{wu})\boldsymbol{\alpha}(\mathbf{s}) = \mathbf{R}(\boldsymbol{\alpha}, \mathbf{u}, \mathbf{w}, \mathbf{s}) \tag{6.157}$$

This is the general definition of the Riemann tensor **R** in the language of differential forms.

The components of this tensor are $R^\alpha{}_{\beta\mu\nu}$,

$$\mathbf{R} = R^\alpha{}_{\beta\mu\nu}\frac{\partial}{\partial x^\alpha} \otimes \mathbf{d}x^\beta \otimes \mathbf{d}x^\mu \otimes \mathbf{d}x^\nu \tag{6.158}$$

The symmetry properties of **R** are somewhat awkward to state in the language of differential forms, and we will not trouble to re-derive them.

The second-rank Ricci tensor $\overline{\mathbf{R}}$ is a tensor of type $\binom{0}{2}$, obtained in the usual way by contraction on the first and last slots in Eq. (6.157).

Acting on two vectors, it produces a real number:

$$\overline{\mathbf{R}}(\mathbf{u}, \mathbf{w}) = \mathbf{R}\left(\frac{\partial}{\partial x^\sigma}, \mathbf{u}, \mathbf{w}, \mathbf{d}x^\sigma\right) \tag{6.159}$$

From the Ricci tensor we can construct the curvature scalar by first multiplying with **g**, which gives us a tensor of type $\binom{1}{1}$, and then contracting,

$$\mathbf{g} \otimes \overline{\mathbf{R}}\left(\frac{\partial}{\partial x^\sigma}, \mathbf{d}x^\lambda, \frac{\partial}{\partial x^\lambda}, \mathbf{d}x^\sigma\right) \tag{6.160}$$

Finally, let us rewrite the equation of geodesic deviation in the language of differential forms.

Instead of re-deriving the equation of geodesic deviation from first principles, we can simply translate Eq. (6.118) from component language into the new language.

We designate the tangent vectors along the geodesic by **u**, and we designate the tangent vector from one geodesic to the adjacent geodesic by **s**.

The components of **u** are $dx^\alpha/d\tau$, and the components of **s** are s^μ,

$$\mathbf{u} = \frac{dx^\alpha}{d\tau}\frac{\partial}{\partial x^\alpha} \quad \text{and} \quad \mathbf{s} = s^\mu \frac{\partial}{\partial x^\mu} \tag{6.161}$$

The derivative $D/D\tau$ along the curve corresponds to the covariant derivative $\nabla_{\mathbf{u}}$.

Hence our old equation of geodesic deviation takes the new form

$$\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\mathbf{s} = -\mathbf{R}(\quad, \mathbf{u}, \mathbf{u}, \mathbf{s}) \tag{6.162}$$

Here the empty first slot in **R** means that instead of mapping one 1-form and three vectors into a real number, **R**(, **u**, **u**, **s**) maps three vectors into a vector [this can be best seen from Eq. (6.158); when this expression for **R** is allowed to act on the three vectors **u**, **u**, and **s**, the three 1-forms in Eq. (6.158) disappear, and only the vector $\partial/\partial x^\alpha$ remains].

Thus, despite the strange notation, each side of Eq. (6.162) is a vector, as it should be.

6.8 Isometries of spacetime; Killing vectors

If the metric tensor of a given spacetime geometry is invariant under a coordinate trans- formation $x'^{\mu} = x'^{\mu}(x)$, then this coordinate transformation is said to be a symmetry, or an *isometry*, of the spacetime.

For example, translations, 3-D rotations, and Lorentz transformations (with arbitrary direction of the velocity) are symmetries, or isometries, of flat spacetime – the metric tensor $\eta_{\mu\nu}$ remains invariant under each of these transformations.

Instead of regarding such coordinate transformations as changes from one coordinate system to another, we can adopt an alternative point of view and regard them as changes of position within the same, fixed coordinate system.

This makes it more obvious why invariance under the transformation is a symmetry – it corresponds to the metric tensor being the same at different, related, positions.

A general coordinate transformation changes the metric tensor $g_{\mu\nu}$ into $g'_{\mu\nu}$, according to the familiar tensor transformation formula (6.41),

$$g'_{\alpha\beta}(x'(x)) = \frac{\partial x^{\mu}}{\partial x'^{\alpha}} \frac{\partial x^{\nu}}{\partial x'^{\beta}} g_{\mu\nu}(x) \tag{6.163}$$

On the left side of this equation, the coordinate x' has been expressed as a function of x, to emphasize that the two sides of the equation are evaluated at the same point of spacetime.

If we invert this formula, it gives us $g_{\mu\nu}$ in terms of $g'_{\mu\nu}$:

$$g_{\mu\nu}(x) = \frac{\partial x'^{\alpha}}{\partial x^{\mu}} \frac{\partial x'^{\beta}}{\partial x^{\nu}} g'_{\alpha\beta}(x'(x)) \tag{6.164}$$

If the coordinate transformation is an isometry, the metric tensor is invariant under this transformation, which means that the new metric tensor must be the same function of its argument as the old metric tensor, that is,

$$g'_{\alpha\beta}(x') = g_{\alpha\beta}(x') \quad \text{and} \quad g'_{\alpha\beta}(x'(x)) = g_{\alpha\beta}(x'(x)) \tag{6.165}$$

If so, Eq. (6.164) becomes

$$g_{\mu\nu}(x) = \frac{\partial x'^{\alpha}}{\partial x^{\mu}} \frac{\partial x'^{\beta}}{\partial x^{\nu}} g_{\alpha\beta}(x'(x)) \tag{6.166}$$

The isometries of a given spacetime will often be evident by inspection of the metric tensor.

If not, we must try to discover them by "solving" Eq. (6.166) for the function $x'(x)$ that characterizes the isometry.

This equation is hard or impossible to solve if $x'(x)$ is a finite coordinate transformation.

But the equation becomes much simpler if we consider the special case of an infinitesimal coordinate transformation,

$$x'^{\alpha} = x^{\alpha} + \varepsilon\xi^{\alpha}(x) \quad \text{with} \quad \varepsilon \to 0 \tag{6.167}$$

where $\xi^{\alpha}(x)$ is a function of position, that is, $\xi^{\alpha}(x)$ is a vector field.

For this transformation,

$$\frac{\partial x'^{\alpha}}{\partial x^{\mu}} = \delta^{\alpha}_{\mu} + \varepsilon \frac{\partial \xi^{\alpha}}{\partial x^{\mu}} \tag{6.168}$$

and the transformation of the metric tensor, Eq. (6.164), takes the form

$$\begin{aligned} g_{\mu\nu}(x) &= \left(\delta^{\alpha}_{\mu} + \varepsilon \frac{\partial \xi^{\alpha}}{\partial x^{\mu}}\right)\left(\delta^{\beta}_{\nu} + \varepsilon \frac{\partial \xi^{\beta}}{\partial x^{\nu}}\right) g'_{\alpha\beta}(x'(x)) \\ &\cong g'_{\mu\nu}(x'(x)) + \varepsilon \frac{\partial \xi^{\alpha}}{\partial x^{\mu}} g_{\alpha\nu}(x'(x)) + \varepsilon \frac{\partial \xi^{\beta}}{\partial x^{\nu}} g_{\mu\beta}(x'(x)) \end{aligned} \tag{6.169}$$

Here, we neglected terms of order ε^2.

Note that in the last two terms on the right side of Eq. (6.169) we replaced $g'_{\mu\nu}$ by $g_{\mu\nu}$.

This is permissible, since the difference between these two tensors is of order ε; hence the difference can be neglected in these two terms, because they already contain another factor of ε.

To express everything as an explicit function of x, we use the Taylor-series expansion

$$g'_{\mu\nu}(x'(x)) \cong g'_{\mu\nu}(x) + \varepsilon\xi^{\alpha}\frac{\partial g'_{\mu\nu}}{\partial x^{\alpha}} \cong g'_{\mu\nu}(x) + \varepsilon\xi^{\alpha}\frac{\partial g_{\mu\nu}}{\partial x^{\alpha}} \tag{6.170}$$

With this, Eq. (6.169) becomes

$$g_{\mu\nu}(x) \cong g'_{\mu\nu}(x) + \varepsilon\xi^{\alpha}\frac{\partial g_{\mu\nu}(x)}{\partial x^{\alpha}} + \varepsilon\frac{\partial \xi^{\alpha}}{\partial x^{\mu}}g_{\alpha\nu}(x) + \varepsilon\frac{\partial \xi^{\beta}}{\partial x^{\nu}}g_{\mu\beta}(x) \tag{6.171}$$

The condition $g'_{\mu\nu}(x) = g_{\mu\nu}(x)$ for the invariance of the metric tensor under the infinitesimal transformation is then

$$\xi^{\alpha}\frac{\partial g_{\mu\nu}(x)}{\partial x^{\alpha}} + \frac{\partial \xi^{\alpha}}{\partial x^{\mu}}g_{\alpha\nu}(x) + \frac{\partial \xi^{\beta}}{\partial x^{\nu}}g_{\mu\beta}(x) = 0 \tag{6.172}$$

This equation is the infinitesimal version of Eq. (6.166).

It is convenient to write the derivatives of the contravariant components ξ^{α} in terms of covariant components, by differentiating $\xi_{\nu} = g_{\alpha\nu}\xi^{\alpha}$:

$$\xi^{\alpha}\frac{\partial g_{\mu\nu}}{\partial x^{\alpha}} + \left(\frac{\partial \xi_{\nu}}{\partial x^{\mu}} - \xi^{\alpha}\frac{\partial g_{\alpha\nu}}{\partial x^{\mu}}\right) + \left(\frac{\partial \xi_{\mu}}{\partial x^{\nu}} - \xi^{\beta}\frac{\partial g_{\mu\beta}}{\partial x^{\nu}}\right) = 0 \tag{6.173}$$

Here the three derivatives of the metric tensor combine into a Christoffel symbol, so

$$\frac{\partial \xi_{\nu}}{\partial x^{\mu}} + \frac{\partial \xi_{\mu}}{\partial x^{\nu}} - 2\xi_{\alpha}\Gamma^{\alpha}{}_{\mu\nu} = 0 \tag{6.174}$$

which is the same as

$$\xi_{\nu;\mu} + \xi_{\mu;\nu} = 0 \tag{6.175}$$

This differential equation is called *Killing's equation*.

The condition for an infinitesimal symmetry transformation of the metric can also be expressed in terms of the *Lie derivative*.

If we rewrite Eq. (6.171) as

$$g_{\mu\nu}(x) - g'_{\mu\nu}(x) = \varepsilon\xi^{\alpha}\frac{\partial g_{\mu\nu}(x)}{\partial x^{\alpha}} + \varepsilon\frac{\partial \xi^{\alpha}}{\partial x^{\mu}}g_{\alpha\nu}(x) + \varepsilon\frac{\partial \xi^{\beta}}{\partial x^{\nu}}g_{\mu\beta}(x) \tag{6.176}$$

we see that the three terms on the right side represent the change in the tensor $g_{\mu\nu}(x)$ under the infinitesimal transformation. This change, divided by ε, is called the *Lie derivative* of $g_{\mu\nu}(x)$ with respect to ξ^{μ}, and it is designated by $\mathcal{L}_{\xi} g_{\mu\nu}(x)$:

$$\mathcal{L}_{\xi} g_{\mu\nu}(x) = \lim_{\varepsilon\to 0}\frac{g_{\mu\nu}(x) - g'_{\mu\nu}(x)}{\varepsilon} = \xi^{\alpha}\frac{\partial g_{\mu\nu}(x)}{\partial x^{\alpha}} + \frac{\partial \xi^{\alpha}}{\partial x^{\mu}}g_{\alpha\nu}(x) + \frac{\partial \xi^{\beta}}{\partial x^{\nu}}g_{\mu\beta}(x) \tag{6.177}$$

The Lie derivative of an arbitrary second-rank tensor can be defined similarly, and Lie derivatives for tensors of other ranks can also be defined, by a simple calculation of the change of the tensor under an infinitesimal transformation.

In terms of the Lie derivative, the condition for the invariance of the metric under an infinitesimal transformation then merely says that the Lie derivative is zero, $\mathcal{L}_{\xi} g_{\mu\nu}(x) = 0.$

After this digression, let us return to the Killing equation.

The solutions of this equation, called *Killing vectors*, give us the symmetry transformations of the metric, both the infinitesimal transformations and the finite transformations.

It is obvious that for any finite transformation there exists an infinitesimal transformation, since the Taylor-series expansion of the finite transformation $x' = x(x)$ yields the infinitesimal transformation $x'^{\alpha} = x^{\alpha} + \varepsilon dx^{\alpha}/d\varepsilon$, where ε is a parameter that characterizes the transformation

(for instance, an infinitesimal angle in the case of a rotation, or an infinitesimal velocity in the case of a Lorentz transformation).

We can then identify the vector field ξ^α in Eq. (6.167) with $dx^\alpha/d\varepsilon$.

But the converse is also true: Given an infinitesimal transformation we can always reconstruct the finite transformation.

For this purpose, we introduce a new set of coordinates $\tilde{x}^\mu$ such that the Killing vector takes the simple form

$$\tilde{\xi}^1 = \text{constant} = b \tag{6.178}$$

$$\tilde{\xi}^\mu = 0 \text{ for } \mu \neq 1 \tag{6.179}$$

This means that the $x^{\sim 1}$-coordinate at each point has the direction of the vector ξ^α.

The conditions (6.178) and (6.179) can always be satisfied.

To see this, we express $\tilde{\xi}^\mu$ in terms of ξ^μ, by means of the usual transformation equation $\tilde{\xi}^\mu = (\partial \tilde{x}^\mu/\partial x^\alpha)\xi^\alpha$, and we obtain

$$\frac{\partial \tilde{x}^1}{\partial x^\alpha}\xi^\alpha = b \tag{6.180}$$

$$\frac{\partial \tilde{x}^\mu}{\partial x^\alpha}\xi^\alpha = 0 \text{ for } \mu \neq 1 \tag{6.181}$$

This is a set of linear differential equations, with coefficients $\xi^\alpha(x)$, for the functions $\tilde{x}^\mu(x)$; such equations always have a solution.

If we now write Eq. (6.172) in terms of the new coordinates $\tilde{x}^\mu(x)$ and we insert the expressions (6.178) and (6.179), we find that

$$\frac{\partial \tilde{g}_{\mu\nu}}{\partial \tilde{x}^1} = 0 \tag{6.182}$$

Thus, $\tilde{g}_{\mu\nu}(x)$ does not depend on the coordinate $\tilde{x}^1$.

Consequently, any finite translation of this coordinate is a symmetry of the metric tensor $\tilde{g}_{\mu\nu}(x)$.

This establishes that the infinitesimal symmetry transformation generates a finite symmetry transformation.

Note that the argument also establishes that, with a suitable choice of coordinates, any symmetry transformation of the metric tensor can be regarded as a translation of one of the coordinates, and the metric tensor is then independent of this one coordinate.

Checking for the independence of the metric tensor of some coordinate is, of course, a familiar method for discovering symmetries "by inspection."

For example, in this way we immediately recognize the translation symmetry of the flat spacetime metric tensor $\eta_{\mu\nu}$ from its independence of t, x, y, and z.

And we recognize the rotation symmetry of this metric tensor about the, say, z-axis by transforming it to spherical or cylindrical coordinates, where the metric tensor is seen to be independent of the polar angle ϕ.

(However, if we want to recognize the symmetry of the metric tensor $\eta_{\mu\nu}$ under Lorentz transformations, it is not so obvious how we ought to pick new coordinates that lead to independence from one of them.)

For an infinitesimal rotation about the z-axis, with an angle $\phi \ll 1$ and with $(x^0, x^1, x^2, x^3) = (t, x, y, z)$,

$$x'^1 = x^1 + \phi x^2 \quad \text{and} \quad x'^2 = -\phi x^1 + x^2$$

Hence $\xi^\alpha = (0, x^2, -x^1, 0)$.

If we integrate the equations (6.180) and (6.181) for $\tilde{x}^1$ and $\tilde{x}^2$, we get the results

$$\tilde{x}^1 = -b \tan^{-1} x^2/x^1 = -b\phi$$

$$\tilde{x}^2 = \sqrt{(x^1)^2 + (x^2)^2} = r$$

For an infinitesimal Lorentz transformation, with a speed v ≪ 1,

$$x'^0 = x^0 - vx^1 \quad \text{and} \quad x'^1 = x^1 - vx^0 \tag{6.183}$$

Hence $\xi^\alpha = (-x^1, -x^0, 0, 0)$.

Figure 6.17a displays this Killing vector field.

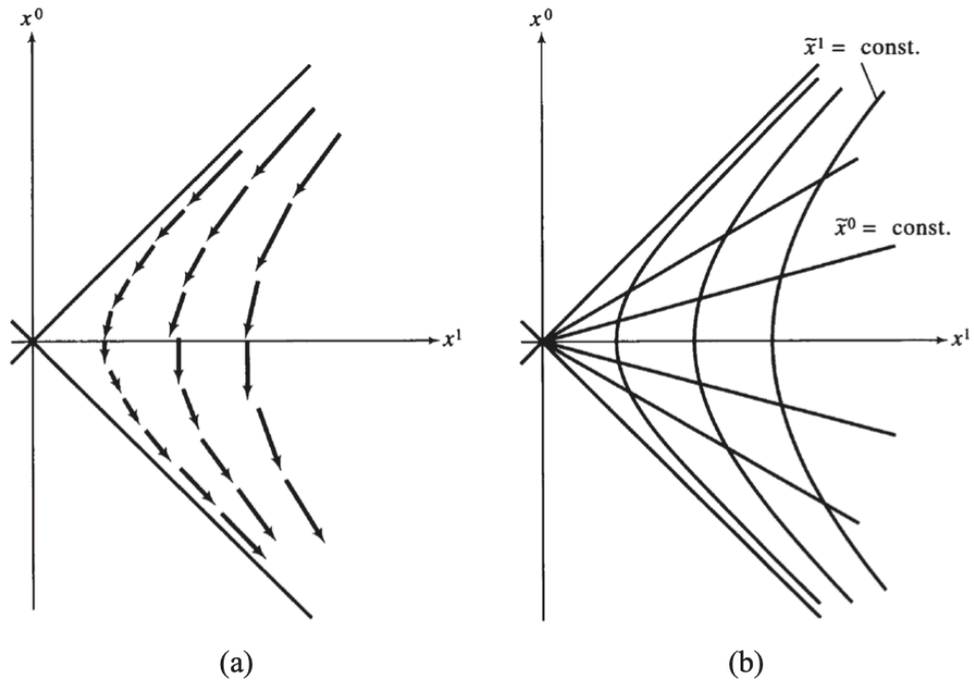


Fig. 6.17 (a) The Killing vector field $\xi^\alpha = (-x^1, -x^0, 0, 0)$. (b) Rindler coordinates in the right wedge.

If we integrate the equations (6.180) and (6.181) for $\tilde{x}^0$ and $\tilde{x}^1$ the solutions are

$$\tilde{x}^0 = b \tanh^{-1} x^0/x^1 \tag{6.184}$$

$$\tilde{x}^1 = \sqrt{(x^1)^2 - (x^0)^2} \tag{6.185}$$

Since the curve $\sqrt{(x^1)^2 - (x^0)^2}$ = constant is timelike and the curve $\tanh^{-1} x^0/x^1$ = constant is spacelike, convention requires that we redefine the coordinates $\tilde{x}^0$ and $\tilde{x}^1$:

$$\tilde{x}^0 = b \tanh^{-1} x^0/x^1 \tag{6.186}$$

$$\tilde{x}^1 = \sqrt{(x^1)^2 - (x^0)^2} \tag{6.187}$$

Correspondingly, the Killing vector $\tilde{\xi}^\alpha$ becomes $(-b, 0, 0, 0)$. [We could have avoided this awkward redefinition by selecting $\tilde{\xi}^0$ as the nonzero, constant component in Eq. (6.178) instead of $\tilde{\xi}^1$; this would have led directly to Eqs (6.186), (6.187).]

The solutions (6.184)–(6.187) apply in the wedge, or quadrant, of spacetime characterized by $x^1 > 0$ and $(x^1)^2 - (x^0)^2 > 0$ (see Fig. 6.17b).

It is straightforward to find the solutions that apply in the other wedges.

In the coordinates $\tilde{x}^0$, $\tilde{x}^1$ it can be shown that the spacetime interval is

$$ds^2 = \frac{(\tilde{x}^1)^2}{b^2}(d\tilde{x}^0)^2 - (d\tilde{x}^1)^2 \tag{6.188}$$

As expected, the metric tensor is independent of $\tilde{x}^0$.

It is easy to verify that a finite Lorentz transformation amounts to a translation of $\tilde{x}^0$.

The coordinates $\tilde{x}^0, \tilde{x}^1$ given by Eqs. (6.186) and (6.187) are called *Rindler coordinates*.

The curves $\tilde{x}^1$ = constant in these coordinates are the worldlines of particles moving with a constant proper acceleration ("hyperbolic motion").

The constant *b* is inversely proportional to the proper acceleration, b = 1/a.

Although in principle all the symmetries of the metric tensor can be recognized by its independence from coordinates upon making suitable choices of coordinates, this is not a convenient method for finding symmetries, since the "suitable" choices of coordinates are not always obvious.

The Killing equation gives us a straightforward, systematic method for finding the symmetries.

By constructing all possible solutions of this differential equation, we can find all the symmetries.

Two or more Killing vectors $\xi^{\alpha}_{(1)}, \xi^{\alpha}_{(2)}, \ldots$ are said to be linearly dependent if there exists a set of constant coefficients c_n such that

$$\sum_n c_n \xi^{\alpha}_{(n)} = 0 \tag{6.189}$$

If there exist no such coefficients, the Killing vectors are said to be linearly independent.

Each linearly independent Killing vector corresponds to a separate symmetry of the spacetime.

The maximum number of independent solutions, and the maximum number of symmetries, is 10.

This agrees with the number of symmetries of flat spacetime, which has exactly 10 symmetries: four independent directions for translations, three independent directions for the axes of spatial rotations, and three independent directions for the velocities of Lorentz transformations (these symmetries constitute the 10-parameter group of general Lorentz transformations).

We expect that the number of symmetries of a curved spacetime is no larger than that of flat spacetime, and that therefore curved spacetime has no more than 10 symmetries.

A spacetime with 10 symmetries is called *maximally symmetric*.

Apart from flat spacetime, the only maximally symmetric spacetime is the de Sitter spacetime, an isotropic and homogeneous model of the universe we will discuss in Chapter 9.

It is well known that symmetries lead to conservation laws.

According to Noether's theorem, to every continuous symmetry of a physical system there corresponds a conservation law.

For example, in flat spacetime, symmetry under spatial translations implies conservation of momentum, symmetry under time translations implies conservation of energy, and symmetry under rotations implies conservation of angular momentum.

In a general curved spacetime, we can formulate the conservation laws for the motion of a particle on the basis of Killing vectors.

We can prove that if ξ^μ is a Killing vector, then, for a particle moving along a geodesic, the scalar product of this Killing vector and the (canonical) momentum $P_\mu = mg_{\mu\nu}dx^\nu/d\tau$ of the particle is a constant:

$$\xi^\mu P_\mu = \text{constant} \tag{6.190}$$

To prove this conservation theorem, we evaluate the derivative of $\xi^\mu P_\mu$ along the geodesic curve,

$$\frac{d}{d\tau}\xi^\mu P_\mu = \frac{d}{d\tau}\xi_\mu P^\mu = \xi_\mu \frac{dP^\mu}{d\tau} + P^\mu \frac{d\xi_\mu}{d\tau} \tag{6.191}$$

But, according to the geodesic equation and the definition of the covariant derivative,

$$\frac{dP^\mu}{d\tau} = -\Gamma^\mu{}_{\alpha\nu} P^\alpha \frac{dx^\nu}{d\tau}$$

and

$$\frac{d\xi_\mu}{d\tau} = \frac{\partial \xi_\mu}{\partial x^\nu}\frac{dx^\nu}{d\tau} = \left(\xi_{\mu;\nu} + \Gamma^\alpha{}_{\mu\nu}\xi_\alpha\right)\frac{dx^\nu}{d\tau} \tag{6.192}$$

When we substitute these expressions into Eq. (6.191), the two terms involving the Christoffel symbols cancel, leaving us with

$$\frac{d}{d\tau}(\xi_\mu P^\mu) = \xi_{\mu;\nu} P^\mu \frac{dx^\nu}{d\tau} = \frac{1}{2m}(\xi_{\mu;\nu} + \xi_{\nu;\mu})P^\mu P^\nu \tag{6.193}$$

This is zero in consequence of the Killing equation.

Note that if we adopt the coordinates $\tilde{x}^{\mu}$ that make $\tilde{\xi}^{1}$ = b and the other components of $\tilde{\xi}^{\mu}$ zero, then $\xi^{\mu}P_{\mu} = \tilde{\xi}^{\mu}\,\tilde{P}_{\mu}$ = constant reduces to $\tilde{P}_{1}$ = constant.

This says that the component of the momentum conjugate to the coordinate $\tilde{x}^{1}$ is constant.

In Lagrangian language, the coordinate $\tilde{x}^{1}$ is an *ignorable coordinate*, because it does not appear in the Lagrangian $L = m\sqrt{\tilde{g}_{\mu\nu}(d\tilde{x}^{\mu}/d\tau)(d\tilde{x}^{\nu}/d\tau)}$ $\tilde{g}_{\mu\nu}(d\,\tilde{x}^{\mu}/d\tau)(d\,\tilde{x}^{\nu}/d\tau)$.

The conservation of the canonical momentum $\tilde{P}_{1} = m\,\tilde{g}_{1\mu}(d\,\tilde{x}^{\mu}/d\tau)$ conjugate to this ignorable coordinate is an immediate corollary of the Lagrangian equations.

If the spacetime geometry is time independent, so $g_{\mu\nu}$ does not depend on x^{0}, then ξ^{μ} =(b,0,0,0) is the corresponding Killing vector, and the conservation theorem (6.190) tells us that P_0 is constant.

This is the law of conservation of energy.

This conservation law also applies to photons moving in curved spacetime.

Thus, if the spacetime geometry is time independent, the photon "energy" P_0 is constant.

As we will see in Section 7.6, the gravitational redshift for photons can be extracted from this conservation law.

7 - Einstein's Gravitational Theory

You boil it in sawdust: you salt it in glue
You condense it with locusts and tape
Still keeping one principal object in view –
To preserve its symmetrical shape.

Lewis Carrroll,
The Hunting of the Snark

The preceding chapter dealt with the kinematics of the curved spacetime geometry, that is, the description of the geometry and its curvature.

We now come to the dynamics of the geometry, that is, the interaction of the geometry and matter.

This interaction is the essence of Einstein's equations for the gravitational field.

There are several routes that lead to Einstein's equations; they differ in their starting points.

One route begins with the equations of the linear approximation of Chapter 3 and adds the assumption that the exact, nonlinear equations are of second differential order and are endowed with general invariance.

These assumptions suffice to completely determine the exact, nonlinear equations for the gravitational field.

That the linear equations imply the full nonlinear equations is a quite remarkable feature of Einstein's theory of gravitation.

Given some complicated set of nonlinear equations, it is always easy to derive the corresponding linear approximation; but in general, if we know only the linear approximation, we cannot reconstruct the nonlinear equations.

What permits us to perform this feat in gravitational theory is the requirement of general invariance.

As we will see later, this requirement states that the form and the content of the equations must remain unchanged under all coordinate transformations.

An alternative route to Einstein's equations begins by assuming that gravitation should be a geometric theory, with a curved spacetime in which test particles move along geodesics.

This assumption can either be treated as an outright stipulation, or it can be justified by some plausibility argument.

In addition we need to assume that all gravitational effects arise exclusively from the metric tensor (no extra gravitational fields!) and that, again, the equations are endowed with general invariance.

The known properties of Riemannian geometry then lead us to the equations for the gravitational field.

In both of these approaches to the Einstein field equations, the requirement of general invariance plays a crucial role.

Regarded naively, this seems to be a simple generalization of the requirement of Lorentz invariance, which states that the equations must remain unchanged under all Lorentz transformations.

In his original work, Einstein sought to justify general invariance by arguing that the laws of physics should not depend on our choice of coordinates and that therefore all reference frames – even accelerated reference frames – should obey a principle of relativity of motion.

Thus, Einstein's theory was intended to be a theory of "general relativity," which extends the relativity of uniform translational motion to accelerated motion.

Einstein sought to justify the relativity of accelerated motion by claiming that, locally, an accelerated reference frame is equivalent to a stationary reference frame placed in a gravitational field.

But, as we already saw in Chapter 1, tidal forces distinguish between such reference frames and spoil their equivalence.

Hence the principle of general invariance is not a relativity principle; rather, it is an independent symmetry principle that we impose on the laws of physics to restrict the possible interactions of gravitation with itself and with matter.

In contrast to the principle of special relativity, which we are compelled to accept by abundant empirical evidence, the principle of general invariance is an elective, optional, symmetry for which the empirical evidence is much less compelling.

In fact, it is possible to construct various gravitational theories without the principle of general invariance.

Since Einstein's theory of "general relativity" is no more relativistic than special relativity, it would be preferable to adopt the name *geometrodynamics* for this theory.

This name, coined by Wheeler, puts the emphasis where it belongs – on the dynamical geometry that acts on and reacts to matter.

7.1 General covariance and invariance; gauge transformations

While laying the groundwork for general relativity, Einstein proposed a principle of covariance for all equations of physics.

The new meaning that he gave to *covariance* in this context was different from that used in earlier chapters:

> *An equation is said to be covariant under general coordinate transformations if the form of the equation is left unchanged by the transformations*.

He thought that a principle of covariance had to be imposed on all physical laws to make them independent of the coordinates in which they are expressed.

We have covariant vectors, covariant derivatives, and now covariant equations.

The word “covariant” is unfortunately much overworked, and which of the three meanings is intended must be guessed from context.

But in a critical examination of Einstein’s theory, Kretschmann argued that the content of all reasonable physical laws is automatically independent of the coordinates, so any explicit requirement of coordinate independence is redundant.

Physical laws are ultimately about events involving physical bodies and causal relationships between such events.

Coordinates are merely a convenient bookkeeping tool for describing events and their relationships and do not constrain the content of the laws that govern these relationships.

All laws are either automatically covariant (for instance, the conservation law of baryon number) or else they can be made covariant by a simple mathematical reformulation (for instance, the various fundamental laws of physics expressed as differential equations).

As an example of a covariant reformulation of a differential equation, consider the special-relativistic equation of motion for a particle in an electromagnetic field.

In an inertial reference frame this equation of motion is

$$m\frac{du^\mu}{d\tau} = qu^\nu F^\mu{}_\nu \tag{7.1}$$

This equation preserves its form under Lorentz transformations but not under general coordinate transformations.

Hence, the equation is not covariant in the sense of Einstein’s new definition.

However, it is easy to rewrite the equation in a mathematical form that is covariant by performing a transformation from the original inertial coordinates x^{μ} to general coordinates x'^{μ}.

We dealt with several such transformations in the preceding chapter, but those mostly involved transformations from general coordinates to inertial coordinates (local geodesic coordinates).

It is instructive to perform the reverse transformation, so that we can see what steps and what assumptions are involved.

As a first step, we multiply both sides of Eq. (7.1) by $\partial x'^{\beta}/\partial x^{\mu}$ and sum over μ, which transforms the contravariant vector on the right side of Eq. (7.1) into $qu'^{\nu}F'^{\beta}{}_{\nu}$

Then we substitute $u^{\mu} = (\partial x^{\mu}/\partial x'^{\alpha})u'^{\alpha}$ on the left side and perform the differentiation indicated by $d/d\tau$.

After some simple manipulations, the net result is

$$m\frac{du'^{\beta}}{d\tau} + m\frac{\partial x'^{\beta}}{\partial x^{\mu}}\frac{\partial^2 x^{\mu}}{\partial x'^{\alpha}\partial x'^{\nu}}u'^{\alpha}u'^{\nu} = qu'^{\nu}F'^{\beta}{}_{\nu} \qquad (7.2)$$

As an optional further step, we can express the second term on the left side more concisely in terms of the Christoffel symbol $\Gamma'^{\beta}{}_{\alpha\nu} = (\partial x'^{\beta}/\partial x^{\mu})(\partial^2 x^{\mu}/\partial x'^{\alpha}\partial x'^{\nu})$ obtained by transforming the Christoffel symbol $\Gamma^{\mu}{}_{\alpha\beta} = 0$ associated with the original inertial coordinates to general coordinates [see Eq. (6.30)]. Equation (7.2) then becomes

$$m\frac{du'^{\beta}}{d\tau} + m\Gamma'^{\beta}{}_{\alpha\nu}u'^{\alpha}u'^{\nu} = qu'^{\nu}F'^{\beta}{}_{\nu} \qquad (7.3)$$

Rewritten in the form (7.2) or (7.3) the equation of motion is covariant – any further transformation from the general coordinates x'^{μ} to other general coordinates x''^{μ} leaves the *form* of the equation unchanged.

Evidently, the steps that took us from Eq. (7.1) to Eqs. (7.2) or (7.3) introduced no new physics into these equations.

These steps are merely a change of variables in a differential equation; that is, a purely mathematical reformulation without any change of the physical content.

The final covariant form of the equation is no more and no less relativistic or geometric or symmetric than the original form.

Covariance imposes no restrictions on the physical content of the equation, but only on the mathematical form in which it is written.

Covariance is totally devoid of any physical implications.

This absence of physical implications of covariance is further illustrated by the circumstance that even Newton's equation of motion can be rewritten in a covariant form.

Consider a particle subjected to the action of a force f^k.

The Newtonian equation of motion in inertial coordinates is

$$m\frac{d}{dt}v^k = f^k$$

where v^k is the ordinary velocity, $v^k = dx^k/dt$.

Since this equation of motion is the nonrelativistic, slow-motion limit of Eq. (7.1), it is obvious that the covariant form of this equation must be the slow-motion limit of the covariant equation (7.2), that is,

$$m\frac{dv'^\beta}{dt} + m\frac{\partial x'^\beta}{\partial x^\mu}\frac{\partial^2 x^\mu}{\partial x'^\alpha \partial x'^\nu}v'^\alpha v'^\nu = f'^\beta \tag{7.4}$$

Here, again, v'^k is the ordinary velocity, $v'^k = dx'^k/dt$.

In general coordinates, the absolute Newtonian time *t* in Eq. (7.4) is to be regarded as the time registered by a clock moving with the particle, and this time is not necessarily equal to the general time coordinate x'^0, so $v'^0 \equiv dx'^0/dt/= 1$.

The absolute Newtonian time *t* is equal to the inertial time coordinate x^0, so $v^0 \equiv dx^0/dt = 1$.

Although the original Newtonian equation in inertial coordinates is not covariant, Eq. (7.4) is covariant.

When we perform a coordinate transformation from the general coordinates x'^μ to different general coordinates x''^μ, the only change in Eq. (7.4) is that all primes are replaced by double primes; that is, the equation retains its form.
These examples show that noncovariant equations can be rewritten as equivalent covariant equations.

Since covariance is a statement about the form of an equation, it is not surprising that it depends on the form in which the equation is written.

In essence, we achieved covariance by introducing into our equations of motion extra terms constructed from the transformation functions $\partial x'^\mu/\partial x^k$ and $\partial x^\nu/\partial x'^\mu$.

These functions play the role of pseudo-force fields, analogous to the centrifugal and Coriolis pseudo-force fields in a rotating reference frame.

The presence of these extra functions reveals that inertial coordinates retain a special significance for our equations, but the presence of these extra functions does not contradict covariance.

Because equations that are not covariant can be changed into equations that are covariant by the insertion of extra transformation functions, the covariant reformulation of equations is merely a mathematical exercise, without physical significance.

In response to Kretschmann's criticism about the trivial and redundant character of the covariance principle, Einstein wrote

> Even though it is true that one must be able to bring every empirical law into generally covariant form, yet the Principle has considerable heuristic force, which has proved itself in the problem of gravitation and which relies on the following. Of two theoretical systems which agree with experience, that one is to be preferred, which from the point of view of the absolute differential calculus is the simplest and most transparent. One need only try to bring Newtonian mechanics and gravitation into the form of . . . covariant equations (four-dimensional) and one will surely be convinced that the Principle rules out this theory practically if not theoretically! (Einstein, 1918)

Thus, Einstein proposed to reject some covariant equations as too complicated.

But such an attempt to discriminate between "simple equations" and "complicated" equations relies on vague subjective selection criteria.

Furthermore, Einstein supposition that the covariant formulation of the Newtonian equations would be excessively complicated was merely an unverified conjecture – as Eq. (7.4) shows, the complications found in covariant Newtonian equations are actually not very different from those in covariant relativistic equations.

Instead of Einstein's vague criterion for "simplest and most transparent" equations, we will rely on the criterion of *invariance* of equations.

Invariance goes beyond covariance in that it demands that not only the form but also the content of the equations be left unchanged by the coordinate transformation.

Roughly, invariance is achieved by imposing the extra condition that all constants that appear in the equation remain exactly the same.

In this context, by "constants" are meant not only those things that are true numerical constants (the speed of light c, or the gravitational constant G, or Planck's constant $\hbar$) but also any quantity independent of the state of matter.

A better terminology is to call such quantities *absolute objects*.

Thus, an absolute object is a quantity, with one or more components, that is independent of the state of matter.

The value of such an absolute object is completely unaffected by any changes in the condition of the matter in the universe.

For example, c, G, and $\hbar$ are absolute objects; and so are $\partial x'^{\mu}/\partial x^{\nu}$ and $\partial x^{\nu}/\partial x'^{\mu}$ (where x' is some given function of x), and so on.

In contrast, dynamical objects are physical variables such as particle positions and momenta, field strengths, energy densities, and the like that do depend on the state of matter.

Given any law of physics, we can then classify the quantities that appear in it into absolute objects and dynamical objects.

With these preliminaries we are now ready to give a more precise definition of invariance under coordinate transformations:

> *An equation is said to be invariant under coordinate transformations if* (*i*) *it is covariant and* (*ii*) *all absolute objects contained in it are left unchanged*.

In brief, the form and content of the equation are unchanged.

Note that this definition tells us that Eq. (7.4) is not invariant.

This equation is covariant, but it contains the absolute objects $\partial x'^{\mu}/\partial x^{k}$ and $\partial x^{\nu}/\partial x'^{\mu}$, and a coordinate transformation changes these objects into $\partial x''^{\mu}/\partial x^{k}$ and $\partial x^{\nu}/\partial x''^{\mu}$, in violation of item (*ii*).

Einstein's gravitational theory is based on the following principle of general invariance, which we elect to impose on all the equations of physics:

> *Principle of general invariance: All laws of physics must be invariant under general coordinate transformations*.

This principle demands that the laws have a higher symmetry, a symmetry that goes beyond Lorentz symmetry.

As we will see later, this extra symmetry is a gauge symmetry, and it is motivated by the gauge symmetry of the linear gravitational field we encountered in Chapter 3.

One immediate consequence of this general-invariance symmetry is that we must regard the metric tensor of spacetime as dependent on the state of matter – the metric tensor must be a dynamical field, satisfying some field equation of its own.

The reason is obvious: The general coordinate transformations do not leave $g_{\mu\nu}$ unchanged; hence the presence of $g_{\mu\nu}$ as an absolute object in an equation contradicts the requirement that all such objects keep their values unchanged under coordinate transformations.

The only alternative is to suppose that $g_{\mu\nu}$ is not an absolute object, but rather a dynamical variable.

Thus, the principle of general invariance excludes "prior geometry;" that is, it excludes a fixed, immutable, pre-ordained geometry.

Equations (7.3) and (7.4) serve as a clear illustration of the distinction between an equation that is invariant and an equation that is not invariant.

In equation (7.3) we can endow the metric tensor and the Christoffel symbols $\Gamma'^{\beta}_{\alpha\nu}$ calculated from this metric tensor with the status of dynamical variables, and we thereby make Eq. (7.3) into an invariant equation.

The Christoffel symbols in this equation are then not absolute objects and their presence does not conflict with the invariance of this equation.

In contrast, in Eq. (7.4) we cannot express the derivatives in second term on the left side as dynamical variables, because the geometry underlying Newtonian physics is fixed and immutable, so these derivatives must be regarded as absolute objects that cannot be re-expressed as Christoffel symbols related to a metric tensor that acts as a dynamical variable.

Furthermore, Eq. (7.4) explicitly contains the absolute Newtonian time t, which is a pre-eminent absolute object.

Hence, although Eq. (7.4) is covariant, it is not invariant.

Since we have taken some pains to make it clear that covariance is trivial (and therefore unavoidable, if the equations are suitably rewritten), but that invariance is nontrivial (and therefore avoidable and elective), it will be necessary to give some argument in favor of invariance.

The strongest argument emerges from our discussion of the linear tensor field theory in Chapters 3–5.

There we saw that the field equation of the linear theory is invariant under the gauge transformation [see Eq. (3.39)],

$$h^{\mu\nu}(x) \to h'^{\mu\nu}(x) = h^{\mu\nu}(x) + \partial^{\nu}\Lambda^{\mu} + \partial^{\mu}\Lambda^{\nu}. \tag{7.5}$$

Let us compare this gauge transformation with a coordinate transformation.

Consider the infinitesimal coordinate transformation

$$x'^{\mu} = x^{\mu} + \varepsilon\xi^{\mu}(x) \quad \text{with } \varepsilon \to 0 \tag{7.6}$$

where $\xi^{\mu}(x)$ is a function of position. In this equation indices are raised with the Minkowski tensor $\eta^{\mu\nu}$.

From this we obtain

$$\frac{\partial x'^{\mu}}{\partial x^{\alpha}} = \delta^{\mu}_{\alpha} + \varepsilon\xi^{\mu}{}_{,\alpha} \tag{7.7}$$

and

$$\begin{aligned} g'^{\mu\nu}(x') &= \frac{\partial x'^{\mu}}{\partial x^{\alpha}} \frac{\partial x'^{\nu}}{\partial x^{\beta}} g^{\alpha\beta}(x) \\ &= \left(\delta^{\mu}_{\alpha} + \varepsilon\xi^{\mu}{}_{,\alpha}\right)\left(\delta^{\nu}_{\beta} + \varepsilon\xi^{\nu}{}_{,\beta}\right) g^{\alpha\beta}(x) \\ &\cong g^{\mu\nu}(x) + g^{\alpha\nu}(x)\varepsilon\xi^{\mu}{}_{,\alpha} + g^{\mu\beta}(x)\varepsilon\xi^{\nu}{}_{,\beta} \end{aligned} \tag{7.8}$$

In this equation, as in all tensor transformation equations, it is assumed that the x appearing in $g^{\mu\nu}(x)$ and in $\xi^{\mu}(x)$ on the right side is expressed as a function of x', so the right side yields a function of x', in accord with the functional dependence indicated on the left side.

Alternatively, we can express the x' appearing in $g'^{\mu\nu}(x')$ on the left side as a function of x, and leave the right side as a function of x.

This is advantageous if we want to see just how much the new function $g'^{\mu\nu}$ differs from the old function $g^{\mu\nu}$.

For this purpose, we use the Taylor-series expansion

$$g'^{\mu\nu}(x'(x)) \cong g'^{\mu\nu}(x) + g^{\mu\nu}(x)_{,\alpha}\varepsilon\xi^{\alpha} \tag{7.9}$$

which leads to

$$g'^{\mu\nu}(x) = g^{\mu\nu}(x) - g^{\mu\nu}(x)_{,\alpha}\varepsilon\xi^{\alpha} + g^{\alpha\nu}(x)\varepsilon\xi^{\mu}{}_{,\alpha} + g^{\mu\beta}(x)\varepsilon\xi^{\nu}{}_{,\beta} \tag{7.10}$$

If $\kappa h^{\mu\nu}$ is small, the relation between the inverse metric tensor and $h_{\mu\nu}$ is

$$g^{\mu\nu}(x) = \eta^{\mu\nu} - \kappa h^{\mu\nu}(x) \tag{7.11}$$

where the indices on the right side are raised with the Minkowski tensor $\eta^{\mu\nu}$ (this relation can be verified by multiplying each side into the usual relation $g_{\nu\alpha} = \eta_{\nu\alpha} + \kappa h_{\nu\alpha}$; except for a negligible term of order κ^2, the result gives the identity $\delta_\alpha{}^\mu = \delta_\alpha{}^\mu$).

Substituting such expressions for $g^{\mu\nu}$ and $g'^{\mu\nu}$ into Eq. (7.10), we then obtain, except for a negligible term of order $\varepsilon\kappa h^{\mu\nu}$,

$$-\kappa h'^{\mu\nu}(x) = -\kappa h^{\mu\nu}(x) + \varepsilon\xi^{\mu,\nu} + \varepsilon\xi^{\nu,\mu} \tag{7.12}$$

Comparing this equation with Eq. (7.5), we see that the equations coincide, provided we make $\varepsilon\xi^\mu$ proportional to Λ^μ:

$$\varepsilon\xi^\mu(x) = -\kappa\Lambda^\mu(x) \tag{7.13}$$

This shows that the gauge transformation (7.5) is nothing but an infinitesimal coordinate transformation, and the invariance of the linear theory under such gauge transformations is really invariance under infinitesimal coordinate transformations.

By integration of infinitesimal coordinate transformations we can construct finite coordinate transformations.

Thus, we can regard such finite coordinate transformations as the appropriate gauge transformations of the nonlinear theory.

From this point of view, the symmetry expressed by the principle of general invariance imposes a gauge symmetry on the gravitational theory, and the motivation for this gauge symmetry arises from the physics of gravitational wave fields discussed in Chapter 5:

We want the gravitational wave fields to correspond to a specific representation of the Lorentz group; that is, we want the quanta of the field to correspond to spin 2, mass zero.

This requires gauge symmetry to eliminate all wave solutions except the two polarizations $\oplus$ and $\otimes$ we discussed in Chapter 5.

Of course, we could also contemplate a gravitational theory that involves several other types of gravitational wave fields (for instance, spin 2 *and* spin 0, or spin 2 *and* spin 1), which might require additional gauge symmetries.

But such mixed fields are an undesirable complication, best avoided, unless compelled by empirical data.

At present, to the best of our knowledge, a single tensor field suffices to accommodate the empirical data, and there is no compelling need to introduce extra, mixed fields.

Another relevant argument in favor of general invariance emerges from theoretical calculations of the inertial and gravitational masses of an arbitrary system of particles and fields.

In this context, as in Section 1.5, the gravitational mass is defined by the gravitational field that the system produces at large distances (where the gravitational field asymptotically becomes Newtonian); and the inertial mass is defined by the total energy of the system, including the energy in the gravitational field, calculated from the 00 component of the total energy-momentum tensor.

The general-invariance gauge symmetry plays a crucial role in the general proof of the equality of the inertial and gravitational masses.

We will not present the details here, but it seems that, without this gauge symmetry, there is no known general proof of the equality $M_G = M_I$.

The physical interpretation of the principle of special relativity is that velocity is relative – no experiment can detect any intrinsic difference between reference frames in uniform motion with different velocities.

It is tempting to give the principle of general invariance the physical interpretation that acceleration is also relative.

Einstein named his theory of gravitation "general relativity" because he thought that (locally) the phenomena observed in a gravitational field are indistinguishable from those observed in an accelerated system of reference and that "according to this conception one cannot speak of the absolute acceleration of a system of reference, just as in the ordinary theory of relativity one cannot speak of the absolute velocity of a system".

However, we saw in Section 1.8 that the tidal effects allow us to make an absolute distinction between the gravitational forces and the pseudo-forces found in accelerated reference frames.

It is therefore false to speak of a general relativity of motion.

The principle of special relativity is a relativity principle, whereas the principle of general invariance is not.

The principles of special relativity and general invariance are different in kind because the symmetries involved in these principles are different in kind.

Lorentz invariance restricts the spacetime geometry to be flat, with a Minkowski metric tensor $\eta_{\mu\nu}$.

It tells us what combinations of components of (Lorentz) tensors can appear in the equations of physics, and it thereby restricts all the kinds of interactions that can occur in physics.

In contrast, general invariance does not restrict the spacetime geometry to be flat.

It tells us how the nonflat metric tensor $g_{\mu\nu}$ can appear in the equations of physics, and it thereby restricts gravitational interactions whenever gravity couples to itself or to other forms of matter, but it does not restrict nongravitational interactions.

The role of general invariance is analogous to that of the gauge invariance of electrodynamics, which tells us how the electromagnetic field A^{μ} can appear in the equations, and thereby restricts electromagnetic interactions.

As an example of the help offered by the principle of general invariance in the search for the couplings between gravitation and matter, consider the problem of finding the equations that describe the effect of gravitation on electrodynamics.

In flat spacetime, the Maxwell equations are [see Eqs. (2.111) and (2.112)]:

$$F^{\mu\nu}{}_{,\mu} = 4\pi j^{\mu} \tag{7.14}$$

$$F^{\mu\nu,\alpha} + F^{\nu\alpha,\mu} + F^{\alpha\mu,\nu} = 0 \tag{7.15}$$

Here, as always, the commas indicate ordinary derivatives.

Obviously, these equations are not invariant under general coordinate transformations.

The left sides are not tensors – we know that ordinary derivatives of tensors do not give tensors.

To make the left sides into tensors, we must replace the ordinary derivatives by covariant derivatives:

$$F^{\mu\nu}{}_{;\mu} = 4\pi j^{\mu} \tag{7.16}$$

$$F^{\mu\nu;\alpha} + F^{\nu\alpha;\mu} + F^{\alpha\mu;\nu} = 0 \tag{7.17}$$

These equations satisfy the principle of general invariance.

The gravitational field variables appear in these equations through the Christoffel symbols that are contained in the covariant derivatives.

The gravitational field therefore affects the electromagnetic field – for instance, the light deflection given in Chapter 4 can be obtained by a study of the wave solutions of Eqs. (7.16) and (7.17).

Unfortunately, although the principle of general invariance places considerable restrictions on the possible interactions of gravitation and matter, it does not determine these interactions uniquely.

For example, Eq. (7.16) remains entirely consistent with general invariance if we add an extra term constant $\times (F^{\mu\nu} R)_{;\mu}$ to the left side.

To rule out such extra terms, it is customary to appeal to the strong equivalence principle (SEP), according to which, in a local freely falling reference frame, the laws of nongravitational physics are supposed to be those of special relativity.

The local freely falling reference frame corresponds to local geodesic coordinates x'^{μ}, and therefore SEP fixes the laws of physics in these geodesic coordinates.

The laws of physics in general coordinates x^{μ} can then be obtained by reversing the transformation $x^{\mu} \rightarrow x'^{\mu}$ given in Eq. (6.35).

This inverse transformation changes the metric tensor $\eta_{\mu\nu}$ of geodesic coordinates into the metric tensor $g_{\mu\nu}$ of general coordinates, it changes the zero Christoffel symbols into nonzero Christoffel symbols, and it changes ordinary derivatives into covariant derivatives (the "comma to semicolon rule").

In the case of electromagnetic fields, this yields Eqs. (7.16) and (7.17).

In essence, SEP plays the role of a minimal-coupling principle.

It prescribes that the equations of motion of matter in the presence of gravitation are to be obtained from those that hold in the absence of gravitation (that is, the equations of special relativity) by replacing $\eta_{\mu\nu}$ by $g_{\mu\nu}$, and replacing ordinary derivatives by covariant derivatives, *and no other changes are to be made*.

However, SEP is not a symmetry principle, and there is no fundamental justification for it, except that it keeps the equations as simple as possible.

In fact, this principle is not really a principle ("general law"), because it cannot be applied to all the equations of physics.

Already in the early years of general relativity, Eddington expressed skepticism about what we today call SEP:

> It is essentially an hypothesis to be tested by experiment as opportunity offers. Moreover it is to be regarded as a suggestion, rather than a dogma admitting of no exceptions. It is likely that some of the phenomena will be determined by comparatively simple equations in which the components of the curvature of the world do not appear. . . . But there are more complex phenomena governed by equations in which the curvatures of the world are involved. . . . For these the Principle of Equivalence breaks down. (Eddington, 1923)

For example, if we rashly apply SEP to the equation of motion of the spin of a rigid body, it would tell us that the equation of motion in local geodesic coordinates is $d\mathbf{S}/d\tau = 0$, since this is the equation that the spin satisfies in the absence of gravitation.

But this is the wrong equation of motion for spin.

We know from Section 1.8 that tidal torques act on spin and that the equation of motion of the spin depends on the Riemann tensor [see Eq. (1.68); the connection between the tidal torque and the Riemann tensor will be further discussed in Section 7.7].

Thus SEP is an unreliable and dangerous thing.

It may be used to determine the form that laws take in the presence of gravitation only as a last resort, when we are sure that these laws cannot be determined by other means.

Furthermore, it is advisable to keep in mind the precedent set by electrodynamics.

It is well known that a minimal-coupling principle can be formulated for electrodynamic interactions, but that this principle is violated in nature (for instance, by the existence of the anomalous magnetic moments of protons and neutrons).

Hence, we had better be cautious with the analogous principle in geometrodynamics.

7.2 Einstein's field equation

In our discussion of the linear theory in Chapter 3, we obtained an approximate equation for the gravitational field.

Expressed in terms of the metric tensor $g_{\mu\nu} = \eta_{\mu\nu} + \kappa h_{\mu\nu}$, this equation takes the following form [see Eq. (3.38)]:

$$\partial^\lambda \partial_\lambda g_{\mu\nu} - \partial^\lambda \partial_\nu g_{\mu\lambda} - \partial^\lambda \partial_\mu g_{\nu\lambda} + \partial_\mu \partial_\nu g_\sigma{}^\sigma - \eta_{\mu\nu} \partial_\lambda \partial^\lambda g_\sigma{}^\sigma + \eta_{\mu\nu} \partial^\lambda \partial^\sigma g_{\lambda\sigma} = -\kappa^2 T_{\mu\nu} \qquad (7.18)$$

In this equation all indices are raised with the Minkowski metric tensor $\eta^{\mu\nu}$; for example, $g_\sigma{}^\sigma \equiv \eta^{\sigma\tau} g_{\sigma\tau}$.

As we already remarked in Chapter 3, Eq. (7.18) suffers from several defects because it fails to take into account that the gravitational field gravitates.

We will now derive the exact, nonlinear field equation.

One possible derivation relies on a method of successive approximations to discover the exact energy-momentum tensor of the gravitational field.

However, we will use a different and more direct method that relies heavily on general invariance.

To be precise, the exact, nonlinear field equation satisfied by the metric tensor $g_{\mu\nu}$ can be derived from the following assumptions:

(i) The equation is invariant under general coordinate transformations.
(ii) The equation reduces to Eq. (7.18) in the linear approximation.
(iii) The equation is of second differential order and is linear in second derivatives.

The first two assumptions require no further comment.

The third assumption expresses a prejudice of physicists in favor of field equations of the general form

$$F(\psi, \partial\psi)\partial^2\psi + G(\psi, \partial\psi) = 0 \tag{7.19}$$

where F and G are functions of the field ψ and of the first derivative $\partial\psi$, but not of the second derivative $\partial^2\psi$.

This form is suggested by our experience with the dynamical equations of Lagrangian mechanics.

Note that the second derivative enters linearly in Eq. (7.19).

Nonlinear equations of this special type are called *quasilinear*.

Mathematical theorems on the uniqueness and existence of solutions, wave propagation, initial-value problem, and so on, can be proved for such quasilinear equations; not much is known about the solution of more general nonlinear equations.

To derive the gravitational field equation, we begin by asking what form the equation takes in local geodesic coordinates at some point.

Since in these coordinates $g'_{\mu\nu} = \eta_{\mu\nu}$ and $g'_{\mu\nu,\alpha}=0$, it is clear that the differential equation must be some expression involving only the second derivatives $g'_{\mu\nu,\alpha,\beta}$.

By assumption (*iii*), this implies that the equation is necessarily linear.

But assumption (*ii*) tells us that if the equation is linear it must have the form

$$\partial'^{\lambda}\partial'_{\lambda}g'_{\mu\nu} - \partial'^{\lambda}\partial'_{\nu}g'_{\mu\lambda} - \partial'^{\lambda}\partial'_{\mu}g'_{\nu\lambda} + \partial'_{\mu}\partial'_{\nu}g'^{\sigma}_{\sigma} - \eta_{\mu\nu}\partial'_{\lambda}\partial'^{\lambda}g'^{\sigma}_{\sigma} + \eta_{\mu\nu}\partial'^{\lambda}\partial'^{\sigma}g'_{\lambda\sigma} = -\kappa^2 T'_{\mu\nu} \tag{7.20}$$

Because Eq. (7.20) is the *exact* field equation in geodesic coordinates x'^{μ}, we can find the *exact* field equation in the original coordinates x^{μ} by a coordinate transformation from x'^{μ} to x^{μ}; it is of course at this stage that assumption (*i*) enters.

Because the transformation $x'^{\mu} \rightarrow x^{\mu}$ from local geodesic coordinates to general coordinates is somewhat messy, it is more convenient to try to express the left side of Eq. (7.20) in terms of a tensor quantity whose transformation law is manifest.

Because the Riemann tensor is the only tensor that can be constructed from linear combinations of the second derivatives of $g'_{\mu\nu}$, the left side of Eq. (7.20) must involve some contracted form of the Riemann tensor.

To discover just what combination is required, we begin by noting that in local geodesic coordinates the Ricci tensor is

$$R'_{\mu\nu} = \frac{1}{2}\left(\partial'^{\lambda}\partial'_{\lambda}g'_{\mu\nu} - \partial'^{\lambda}\partial'_{\nu}g'_{\mu\lambda} - \partial'^{\lambda}\partial'_{\mu}g'_{\nu\lambda} + \partial'_{\mu}\partial'_{\nu}g'^{\sigma}_{\sigma}\right) \tag{7.21}$$

On the right side of this equation, the indices are raised with the Minkowski metric $\eta^{\mu\nu}$.

This expression for $R'_{\mu\nu}$ lacks the terms proportional to $\eta\mu\nu$ that are present in Eq. (7.20). To obtain these terms, we add $-\frac{1}{2}\eta_{\mu\nu}R'$ to both sides of Eq. (7.21), which gives us

$$\begin{aligned} R'_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}R' = \frac{1}{2}\big(&\partial'^{\lambda}\partial'_{\lambda}g'_{\mu\nu} - \partial'^{\lambda}\partial'_{\nu}g'_{\mu\lambda} - \partial'^{\lambda}\partial'_{\mu}g'_{\nu\lambda} + \partial'_{\mu}\partial'_{\nu}g'^{\sigma}_{\sigma} \\ &- \eta_{\mu\nu}\partial'_{\lambda}\partial'^{\lambda}g'^{\sigma}_{\sigma} + \eta_{\mu\nu}\partial'^{\lambda}\partial'^{\sigma}g'_{\lambda\sigma}\big) \end{aligned} \tag{7.22}$$

By comparing Eqs. (7.20) and (7.22), we see that our field equation in local geodesic coordinates is simply

$$R'_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}R' = -\frac{1}{2}\kappa^2 T'_{\mu\nu}$$

or

$$R'_{\mu\nu} - \frac{1}{2}g'_{\mu\nu}R' = -\frac{1}{2}\kappa^2 T'_{\mu\nu} \tag{7.23}$$

Because the transformation properties of $R'_{\mu\nu}$ are those of a tensor, it is immediately obvious that the coordinate transformation from geodesic coordinates x'^{μ} to general coordinates x^{μ} gives,

with $\kappa^2 = 16\pi G$,

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -8\pi G T_{\mu\nu} \tag{7.24}$$

This is *Einstein's field equation*.

Although this field equation was discovered by Einstein, it was also discovered independently by Hilbert, who therefore deserves to share the credit.

Hilbert adopted the physical basis that Einstein had gradually laid for the theory of gravitation in work extended over several years, and he actually discovered and published the field equations a few days before Einstein.

However, he then promptly embarked on a misguided quest to explain all of matter in terms of electromagnetic fields, for which purpose he discarded the term $\frac{1}{2}g_{\mu\nu}R$ in the equation.

This has led to much confusion about how much credit for Eq. (7.24) should be awarded to Hilbert.

In cgs units, the field equation becomes

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -\frac{8\pi G}{c^4}T_{\mu\nu} \tag{7.25}$$

That the exact nonlinear equations are implied by the linear equations (the converse is of course trivial) is a remarkable feature of Einstein's theory.

This intimate link between the exact equations and the linear approximation would not exist were it not for the principle of general invariance.

The field equation (7.24) in conjunction with conservation law for the energy- momentum tensor and the geodesic equation of motion implied by this conservation law provide us with the mathematical formulation of the fundamental idea that "gravitation is geometry."

As Wheeler expressed it with his characteristic verve: Mass grips spacetime, telling it how to curve, and spacetime grips mass, telling it how to move.

The second half of this chapter and the following chapters will deal with various solutions of Eq. (7.24).

We will obtain the exact solution for the gravitational field surrounding a static, spherically symmetric mass (Schwarzschild solution) and that surrounding a rotating mass (Kerr solution); and later we will apply Eq. (7.24) to cosmology and obtain solutions that describe the evolution of the large-scale geometry of the universe.

Unfortunately, the differential equation (7.24) is very complicated.

The second derivatives of the metric field appear linearly, but the first derivatives enter quadratically.

The dependence on $g_{\mu\nu}$ is even worse; the left side of Eq. (7.24) contains the inverse $g^{\mu\nu}$ (used to raise indices), and this is a horrible thing when expressed as a function of $g_{\mu\nu}$.

No general procedure exists for solving the equations analytically; one guesses solutions as best one can.

The tensor $T_{\mu\nu}$ appearing on the right side of Eq. (7.24) is the energy-momentum tensor of matter.

The energy-momentum of the gravitational field is already included (implicitly) on the left side of this equation – most of the nonlinear terms on the left side represent the energy and momentum of the gravitational field.

In empty space, the field equation reduces to

$$R_{\mu\nu} = 0 \tag{7.26}$$

This equation will give us the gravitational field in the empty space surrounding a mass distribution; for instance, the gravitational field surrounding the Sun or the gravitational field surrounding a black hole.

Note that, by raising an index, we can also change Eq. (7.24) into the form

$$R_\mu{}^\nu - \frac{1}{2}\delta^\nu_\mu R = -8\pi G T_\mu{}^\nu \tag{7.27}$$

Since the left side of this equation satisfies the Bianchi identity $(R_\mu{}^\nu - \frac{1}{2}\delta^\nu_\mu R)_{;\nu} = 0$, it follows that

$$T_\mu{}^\nu{}_{;\nu} = 0 \tag{7.28}$$

This expresses the exchange of energy-momentum between matter and the gravitational field.

With the definition of the covariant derivative, Eq. (7.28) becomes

$$\partial_\nu T_\mu{}^\nu + \Gamma^\nu{}_{\alpha\nu} T_\mu{}^\alpha - \Gamma^\alpha{}_{\mu\nu} T_\alpha{}^\nu = 0 \tag{7.29}$$

This equation is the generalization of our approximate conservation law $\partial_\nu T_\mu{}^\nu = 0$ that we used in Chapter 3, where we ignored the extra terms in Eq. (7.29).

Later, in Appendix A.4, we will see that the conservation law (7.28) is a direct consequence of the general-invariance symmetry of the dynamical equations of matter and that it holds independently of Einstein's equations.

The equation of motion of a test particle can be obtained easily from Eq. (7.29) by integration of the energy-momentum tensor over the volume of the particle.

Of course we already know the equation of motion of a test particle – it is the geodesic equation.

Nevertheless, the calculation based on Eq. (7.29) is important in principle because we must make sure that the equation of motion is free of contradictions.

With the usual assumption that the mass of the test particle is negligible, the Christoffel symbols in Eq. (7.29) are not affected by the presence of the test particle – they are entirely determined by the background geometry.

To obtain the equation of motion from Eq. (7.29), it is simplest to go to geodesic coordinates.

Then Eq. (7.29) reduces to

$$\partial'_\nu T'^{\nu}_{\mu} = 0 \tag{7.30}$$

Although this equation is strictly valid only at one point, in the case of a sufficiently small (pointlike) particle we can regard it as valid throughout the volume of the particle.

If we integrate this equation over the volume of the particle, we find

$$\int \partial'_0 T'^{0}_{\mu} d^3x' + \int \partial'_k T'^{k}_{\mu} d^3x' = 0 \tag{7.31}$$

Here, the second term is the integral of a spatial divergence, which, by Gauss's theorem, can be rewritten as a surface integral enclosing the volume of the particle.

If we take this surface to lie outside of the particle, the surface integral vanishes, and only the first term in Eq. (7.31) remains,

$$\int \partial'_0 T'^{0}_{\mu} d^3x' = \frac{d}{dt'} \int T'^{0}_{\mu} d^3x' = \frac{d}{dt'} p'_\mu = 0 \tag{7.32}$$

Accordingly, in local geodesic coordinates, the energy and the momentum of the particle do not change.

This agrees with the result (6.79) that we obtained directly from the equivalence principle.

We can therefore conclude that the particle motion implied by Eq. (7.29) is the same as that implied by Eq. (6.79), that is, geodesic motion.

Because in Einstein's theory all particles obey the geodesic equation of motion, it follows that the Galileo principle of equivalence is satisfied.

It can also be shown that the Newton principle of equivalence is satisfied; that is, for any arbitrary system (not necessarily a particle), the gravitational mass equals the inertial mass.

As already mentioned in Section 7.1, the principle of general invariance plays an important role in the proof of the equality of these masses.

A final general remark about the field equations (7.24).

The tensors $R_{\mu\nu}$ and $g_{\mu\nu}$ are symmetric, and therefore Eq. (7.24) comprises 10 separate equations.

However, not all these equations are independent; the Bianchi identity, consisting of a set of four equations, shows that we have not 10, but only 6 independent differential equations.

Since the unknown functions are the 10 components of $g_{\mu\nu}$, the Einstein equations do not determine the unknown functions entirely.

We already encountered this situation in the linear case, where the solutions suffered from an ambiguity due to gauge transformations, and we eliminated this ambiguity by using the Hilbert gauge condition.

In the present case, the ambiguity in the solution for $g_{\mu\nu}(x)$ arises from the ambiguity in the choice of coordinates.

This comes as no surprise, because we know that the gauge transformations of the linear theory are in fact (infinitesimal) coordinate transformations.

That the Einstein field equations determine the field $g_{\mu\nu}(x)$ only up to a general coordinate transformation is actually highly desirable: It would be absurd that the field equations should not only determine the geometry but also prescribe what coordinates we must use to describe the geometry.

To remove the ambiguity in the solution for $g_{\mu\nu}(x)$, we can impose extra "coordinate conditions." For example, we can impose the harmonic coordinate condition

$$g^{\mu\nu}\Gamma^{\alpha}{}_{\mu\nu} = 0 \tag{7.33}$$

This condition replaces the Hilbert condition of the linear theory.

Other coordinate conditions are also possible.

It is often convenient to impose a purely algebraic condition on $g_{\mu\nu}$, rather than a differential equation.

One such possibility is the condition for time-orthogonal (or "Gaussian") coordinates

$$g_{00} = 1, \quad g_{0k} = 0 \tag{7.34}$$

In general, the condition (7.34) is not very useful because the coordinate frame in which this condition holds will usually not be rigid – relative to the rectangular coordinates at infinity, the coordinate frame will have to bend and twist in a convoluted way.

We will make our choice of coordinate conditions opportunistically, whenever and however it is convenient.

7.3 Another approach to Einstein's equation; cosmological term

In our approach to Einstein's equation for the gravitational field, we took the linear tensor theory of Chapter 3 as our starting point, and we found that this theory approximately describes a curved spacetime.

We then searched for an exact theory consistent with this approximate linear tensor, and this led us to the exact nonlinear Einstein equation, that is, Eq. (7.24).

We will now approach these Einstein equations in a different way, without using the linear tensor theory as a starting point.

This different approach is close to Einstein's own original approach, and it is also usually followed by inventors of alternative theories of gravitation.

In view of the great success of Einstein's theory, the quest for alternative theories is a somewhat quixotic endeavor, but such alternative theories serve as a foil to Einstein's theory and permit us to gain a better appreciation of its advantages over its prospective rivals.

It is therefore useful to have some understanding of the restrictions that apply to the construction of alternative theories.

Einstein's first step toward a geometric theory of gravity arose from the insight that the relativistic generalization of the Newtonian equation of motion in a gravitational field leads to a curved spacetime geometry.

This geometric interpretation of gravitation becomes obvious if we start with the nonrelativistic Lagrangian

$$L = \frac{1}{2}mv^2 - m\Phi \tag{7.35}$$

and construct its simplest relativistic generalization:

$$L = -mc^2\sqrt{1 - v^2/c^2} - m\Phi \tag{7.36}$$

But in the realm of Newtonian mechanics, the potential is small, $|\Phi| \ll c^2$, so it is also possible to write Eq. (7.36) as

$$L \cong -mc^2\sqrt{1 - v^2/c^2 + 2\Phi/c^2} \tag{7.37}$$

Both (7.36) and (7.37) are possible relativistic generalizations of (7.35), and it is not immediately evident which is the better choice (from our discussion of the linear tensor theory, we actually know that both of these equations are approximate, but the latter is somewhat better).

The Lagrangian (7.37) carries a surprising implication.

The principle of least action for this Lagrangian is (with c = 1 and the factor *m* omitted)

$$0 = \delta \int \sqrt{1 - v^2 + 2\Phi}dt = \delta \int \sqrt{(1 + 2\Phi)dt^2 - dx^2 - dy^2 - dz^2} \tag{7.38}$$

Several years before he formulated general relativity, Einstein discovered this equation by an indirect route, while attempting to introduce a position-dependent speed into the equations of special relativity.

He immediately recognized it as the variational equation for a geodesic in a curved spacetime with a proper time interval $d\tau = \sqrt{(1 + 2\Phi)dt^2 - dx^2 - dy^2 - dz^2}$, or

$$d\tau^2 = (1 + 2\Phi)dt^2 - dx^2 - dy^2 - dz^2 \tag{7.39}$$

Accordingly, the metric tensor of this curved spacetime is $g_{00} = (1 + 2\Phi)$ (note that this component of the metric tensor agrees with our approximate tensor theory of Chapter 3, but that the other components are wrong).

This insight was the starting point of Einstein's quest for a gravitational theory based on curved spacetime.

The adoption of a curved spacetime and a geodesic equation for the motion of test particles is one of the essential ingredients of Einstein's theory.

The other is the adoption of a field equation with invariance under general coordinate transformations.

This assumption is the same as assumption (*i*) of Section 7.2, but Einstein's motivation was different.

Einstein confused covariance with invariance, and he thought that in imposing invariance on his field equation he was merely demanding that the laws of physics be valid in all systems of coordinates.

As we saw in Section 7.1, invariance goes beyond covariance – it is a symmetry of choice, not compelled by any obvious principle of physics, except by the gauge symmetry of the linear approximation, of which Einstein was unaware.

By adopting this "elective" symmetry, Einstein instinctively made the right choice, although he had no solid reason for doing so (and Hilbert followed Einstein's lead, for no better reason).

General invariance implies that both the right side of his field equation (the energy-momentum tensor that acts as a source for the gravitational field) and the left side (the metric tensor and the differential operators that act on it) are tensors of rank two under general coordinate transformations, so the field equation is of the form

$$H_{\mu\nu} = -8\pi G T_{\mu\nu} \tag{7.40}$$

Here $H_{\mu\nu}$ is a second-rank tensor constructed from the metric tensor $g_{\mu\nu}$ and its derivatives.

We have inserted a constant 8π G on the right side of this equation for convenience, but what constant we insert is unimportant at this stage, because we still have the freedom to adjust the constants on the left side.

To determine $H_{\mu\nu}$ Einstein assumed that this tensor contains derivatives of no more than second order and is linear in these second derivatives.

This is the same assumption as (*iii*) of Section 7.2, but we now do *not* assume (*ii*).

We know from Section 6.5 that the only tensor that contains the second derivatives linearly is $R_{\mu\nu\alpha\beta}$.

Hence this tensor and $g_{\mu\nu}$ are the only tensors available for the construction of $H_{\mu\nu}$.

The most general second-rank tensor that we can construct, with second derivatives entering linearly, is then

$$H_{\mu\nu} = aR_{\mu\nu} + bg_{\mu\nu}R \tag{7.41}$$

and this gives us a field equation

$$aR_{\mu\nu} + bg_{\mu\nu}R = -8\pi GT_{\mu\nu} \tag{7.42}$$

To establish a relation between a and b, Einstein exploited the conservation law

$$T_{\mu\ ;\nu}^{\ \nu} = 0 \tag{7.43}$$

which he regarded as the appropriate generalization of the conservation law for the energy-momentum tensor in the presence of a gravitational field.

As we already saw in Eq. (7.29), this conservation law determines the exchange of energy and momentum between matter and the gravitational field.

Einstein based Eq. (7.43) on the strong equivalence principle (SEP), according to which gravitation is absent in a local geodesic reference frame, so the conservation law is simply that of special relativity, $T'^{\nu}{}_{\mu,\nu} = 0$.

Transformation of this expression from geodesic coordinates to general coordinates then gives the conservation law (7.43).

However, SEP is merely a minimal-coupling principle, which may or may not be valid (see the discussion in Section 7.1).

That is, maybe the conservation law in geodesic coordinates is not the conservation law $T'^{\nu}{}_{\mu,\nu} = 0$ of special relativity, but something more complicated, with nonzero, curvature-dependent terms that represent residual gravitational interactions that are not eliminated by free fall.

For that reason, it is better not to rely on SEP as the basis for Eq. (7.43).

Instead, Eq. (7.43) can be derived from the general-invariance symmetry, by means of which it can be shown that Eq. (7.43) is valid even when SEP is not valid.

If we raise the index ν in Eq. (7.42) and take the covariant divergence of both sides, we obtain

$$aR_{\mu}{}^{\nu}{}_{;\nu} + bR_{,\mu} = 0 \tag{7.44}$$

According to the Bianchi identity [Eq. (6.100)], $R_{\mu}{}^{\nu}{}_{;\nu} - \frac{1}{2}R_{,\mu} = 0.$

So Eq. (7.44) will be identically satisfied if b = –a/2, and Einstein stipulated this value.

The field equation (7.42) then becomes

$$a\left(R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R\right) = -8\pi G T_{\mu\nu} \tag{7.45}$$

To determine the value of a, we must take into account that in the linear, nonrelativistic limit, Eq. (7.45) should agree with the equation for the Newtonian potential.

It is easy to see that this requires $a = 1$, which leads to the final field equation (7.24), as expected.

For cosmological applications, we need to consider a possible generalization of this field equation.

Equation (7.41) relied on the assumption that the tensors involve second derivatives, as expected for typical dynamical field equations.

But it is consistent with general invariance to include an extra term $\Lambda\, g_{\mu\nu}$ in Eq. (7.41), where Λ is a constant, called the *cosmological constant*.

This is a term without any derivatives, but it is, evidently, a second-rank tensor.

Then $H_{\mu\nu} = aR_{\mu\nu} + bg_{\mu\nu}R + \Lambda g_{\mu\nu}$, and the final field equation becomes

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + \Lambda g_{\mu\nu} = -8\pi G T_{\mu\nu} \tag{7.46}$$

The extra term $\Lambda\, g_{\mu\nu}$ does not affect the conservation law $T^{\nu}{}_{\mu;\nu} = 0$ because $g^{\nu}{}_{\mu;\nu}$ is identically zero [see Eq. (6.56)].

If Λ is sufficiently small, then within the Solar System Eq. (7.46) will be in tolerable agreement with the equation for the Newtonian potential.

However, even a small value of Λ will have drastic effects on the behavior of gravitation at large, cosmological, distances and on the evolution of the universe, as we will see in Sections 9.9 and 9.10.

The cosmological term can be interpreted as an energy-momentum of the vacuum.

To recognize this, we rewrite Eq. (7.46) with the cosmological term on the right side of the equation:

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -8\pi G\left(T_{\mu\nu} + \frac{\Lambda}{8\pi G}g_{\mu\nu}\right) \tag{7.47}$$

In this equation, $(\Lambda/8\pi G)g_{\mu\nu}$ plays the role of an extra energy-momentum tensor, that is, an energy-momentum tensor associated not with matter but with empty (matter-free) space.

Note that in local geodesic coordinates, the extra energy-momentum tensor is proportional to $\eta_{\mu\nu}$.

Thus, the energy and momentum distribution is homogeneous and isotropic when viewed in these coordinates (furthermore, it is Lorentz invariant in these coordinates).

To grasp the implications of a nonzero value of Λ more clearly, it is helpful to look at the linear approximation corresponding to Eq. (7.47).

If Λ is small, we can approximate $\Lambda g_{\mu\nu} \cong \Lambda \eta_{\mu\nu}$, and with the usual Hilbert gauge condition [Eq. (3.40)] we obtain

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -8\pi G \left(T_{\mu\nu} + \frac{\Lambda}{8\pi G} g_{\mu\nu} \right) \tag{7.48}$$

We can also write this equation as

$$\partial_\lambda \partial^\lambda \left(h_{\mu\nu} - \frac{1}{2} \eta_{\mu\nu} h \right) = -\kappa T_{\mu\nu} - \frac{2\Lambda}{\kappa} \eta_{\mu\nu} \tag{7.49}$$

In the absence of matter, $T_{\mu\nu} = 0$, and from Eq. (7.49) we then obtain the following differential equation for the static Newtonian potential $\Phi = \frac{1}{2}\kappa h_{00}$:

$$\nabla^2 \Phi = -\Lambda \tag{7.50}$$

If we compare this with the Newtonian equation $\nabla^2 \Phi = 4\pi G \varrho$, we recognize that the Λ term in Einstein's equation corresponds to a uniform effective mass density

$$\rho_{eff} = -\frac{\Lambda}{4\pi G} \tag{7.51}$$

Thus, if Λ is positive, the effective gravitational mass density of the vacuum is negative, and conversely if Λ is negative, then mass density is positive.

Note that these signs for the effective mass in the Newtonian equation are the opposite of what we expect if we examine the 00 component of the vacuum energy-momentum tensor $(\Lambda/8\pi G)\eta_{\mu\nu}$.

The reason is that what appears as source for h_{00} in the linear equation is not T_{00}, but rather $T_{00} - \frac{1}{2}T = \frac{1}{2}(T_{00} + T_{11} + T_{22} + T_{33})$. .

If Λ is positive, the 00 component of the vacuum energy-momentum tensor is positive, but the 11, 22, and 33 components are negative (they play the role of a negative pressure).

These negative components make the combination $\frac{1}{2}(\Lambda/8\pi G)(\eta_{00}+\eta_{11}+\eta_{22}+\eta_{33}) = -\Lambda/8\pi G$ negative, leading to a negative effective mass in the Newtonian equation (7.50).

Expressed in another way, the inertial mass density for the cosmological term is positive, but the gravitational mass density is negative and twice as large in magnitude.

If we arbitrarily set $\Phi = 0$ at the origin, then in spherical coordinates Eq. (7.50) has the solution,

$$\Phi = -\frac{\Lambda}{6}r^2 \tag{7.52}$$

This potential indicates that between any two particles there acts an effective harmonic-oscillator force of Λ r/3 per unit mass, attractive in the case $\Lambda < 0$, and repulsive in the case $\Lambda > 0$.

In the context of Chapter 3, we did not attempt to include a term $\Lambda\eta_{\mu\nu}$ in our field equation, because we wanted to work with weak fields in an asymptotically flat spacetime.

The cosmological term violates this assumption.

Obviously, at large distances the potential (7.52) becomes very large, and this contradicts both the weak-field condition ($\Phi \ll c^2$) and the boundary condition at infinity ($\Phi \to 0$).

However, a nonzero value of Λ does not entirely destroy the validity of the approximate equations of Chapter 3.

If we are dealing with a central mass that is not enormously large, then there will exist a wide range of radii $c/\sqrt{|\Lambda|} \gg r \gg GM/c^2$ in which the spacetime is very nearly flat; in this range the linear theory is a good approximation.

For large distances, the potential (7.52) and the force grow very large.

But for small distances they are negligible.

We can set an upper limit on Λ by noting that the Newtonian inverse-square law seems to be valid for the motion of galaxies in multiple systems of galaxies.

For example, Page has applied the inverse-square law to determine the mass of binary systems of galaxies, that is, pairs of galaxies in a tight orbit about each other.

The masses obtained for binary systems (via the virial theorem) are in reasonable agreement with independent mass determination of the components (via the rotation curve; see Section 9.6).

This implies that the deviations from the inverse-square law are not significant; that is, within regions of a size at least as large as these binary systems, the acceleration Λ r/3 is small compared with the gravitational acceleration GM/r^2.

From the dimension ($r \approx 10^{23}$ cm) and the mass ($M \approx 10^{11} M_\odot$) of typical binary systems it is possible to set the limit

$$|\Lambda| < 10^{-33}/s^2 \tag{7.53}$$

The effects of the Λ term will be most noticeable in the large-scale motion of the universe.

Hence, cosmological data give the best values for Λ .

Our universe is expanding (see Section 9.3).

The attractive Newtonian gravitational forces between the masses in the universe tend to decelerate this expansion.

The extra force associated with the cosmological term tends to decelerate the expansion if Λ is negative, and accelerate it if Λ is positive.

Throughout most of the 20th century, it was assumed that the cosmological term is insignificant and that the expansion of the universe decelerates.

But in the 1990s, measurements of the speed of recession of distant supernovas showed that the expansion is actually accelerating.

As we will see in Chapter 9, the value of Λ calculated from this cosmological acceleration is

$$\Lambda \cong 2 \times 10^{-35}/s^2 \tag{7.54}$$

And this corresponds to a mass density $\Lambda/8\pi G \sim= 1 \times 10^{-29}$ g/cm^3 distributed throughout "empty" space [this is the mass density, or energy density, that corresponds to the 00 component of the cosmological term on the right side of Eq. (7.47); the effective gravitational mass density from Eq. (7.51) is twice as large, and negative].

Although the implications of the Λ term are cosmological, the origin of this term probably is to be found in quantum theory rather than cosmology.

As we saw, this term can be interpreted as an energy density of the vacuum.

The quantum-theoretical vacuum is a very active place.

It is populated with electron-positron pairs, proton-antiproton pairs, photons, gravitons, and so on, which are spontaneously created and destroyed by vacuum fluctuations.

It is then not surprising that the vacuum should have an energy density; the surprise is rather that the energy density is as small as our value (7.54) for the cosmological term indicates: only $\approx 10^{-29}$ g/cm^3.

Quantum-theoretical estimates suggest that the energy density should be of the order of a Planck mass per Planck length cubed – that is, $(\hbar\, c/G)^{1/2} \times (\hbar\, G/c^3)^{-3/2} \approx 10^{92}$ g/cm^3 – which is about 120 orders of magnitude larger than the observed value. This indicates an enormous defect in our theories, which is unlikely to be resolved anytime soon.

According to some recent speculations, the value of Λ may have been much larger (and positive) during the early stages of the evolution of the universe.

Such a large value of Λ would have resulted in a dramatic, fast expansion of the early universe (an "inflation").

We will discuss this inflationary model of the early universe in Chapter 10.

7.4 Schwarzschild solution and Birkhoff theorem

For the discussion of the deflection, time delay, and redshift of light in the gravitational field of the Sun, the linear approximation of Chapter 3 was quite adequate.

The same is true for the Lense-Thirring precession of a gyroscope, which is another experimental test of gravitational theory that has been successfully completed.

However, the perihelion precession of planetary orbits in the field of the Sun cannot be treated by the linear approximation.

We need an exact, or at least more exact, solution for the gravitational field surrounding a spherically symmetric mass distribution.

Einstein at first calculated the precession from an approximate solution that included terms of order GM and $(GM)^2$.

The exact solution was obtained shortly thereafter by Schwarzschild.

The importance of this solution goes much beyond its application in the Solar System.

It is believed that the gravitational collapse of any nonrotating, electrically neutral star will necessarily lead to the Schwarzschild geometry as its final result; even if the initial geometry is asymmetric, any deviations from the symmetric Schwarzschild field will ultimately be radiated away in the form of gravitational waves.

In what follows we will solve the Einstein equations in the exterior of the mass distribution, that is, the vacuum region surrounding the mass distribution.

The solution in the interior of a mass distribution can be obtained analytically only in some exceptional cases, such as the case of an ideal incompressible fluid or the case of a fluid without pressure (cloud of dust); these special cases are of little interest in the real world.

The solution for the case of matter with a realistic equation of state can be obtained only numerically.

If the mass distribution is spherically symmetric, the gravitational field generated by this mass distribution must also have spherical symmetry.

Since it is sometimes of interest to deal with a collapsing spherical mass – that is, a sphere that shrinks in size – we will begin with an *Ansatz* for the spacetime metric that includes a time dependence for the components of the metric tensor.

However, at the end of our calculation we will find that in the exterior region this time dependence disappears, because, as in the case of the Newtonian gravitational potential, the exterior solution for the metric tensor of a spherical mass does not depend on the size of the mass and remains static even when the mass collapses.

The spherical symmetry of the field imposes severe restrictions on the form of the spacetime interval.

The rectangular displacements dx, dy, dz must occur in the combination $dx^2 + dy^2 + dz^2$ characteristic of spherical symmetry; this combination takes the familiar form $dr^2 + r^2d\theta^2 + r^2\sin^2\theta d\phi^2$ when expressed in spherical polar coordinates.

Furthermore, although the product drdt is consistent with spherical symmetry, the products $d\theta dt$ and $d\phi dt$ are not consistent with spherical symmetry, because they would imply that a light signal (with $ds^2 = 0$) has different speeds when moving in the direction of increasing or decreasing θ and ϕ.

Accordingly, the spacetime interval must be of the form

$$ds^2 = A(r,t)dt^2 - B(r,t)(dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2) - 2F(r,t)drdt \tag{7.55}$$

where A(r, t), B (r, t), and F (r, t) are some functions of the radial coordinate *r* and the time coordinate *t*.

The angular coordinates θ and ϕ are unambiguous; the measurement of these coordinates depends only on our ability to divide a circumference concentric with the mass into equal parts, which we can do even when the functions A(r, t), B (r, t), and F (r, t) are not specified.

The radial coordinate *r* is ambiguous because we do not yet know its precise relation to the measurement of distance.

For a start, we will treat *r* simply as a parameter that identifies different spherical surfaces concentric with the mass.

The time coordinate *t* suffers from similar ambiguities.

However, we will insist that when r $\rightarrow \infty$ and the spacetime becomes flat, the increments in *r* and *t* should equal the increments in the true distance and time, respectively.

This requires that for r $\rightarrow \infty$, A(r,t) and B(r,t) $\rightarrow 1$.

The term involving *drdt* in Eq. (7.55) can be eliminated by a change in the time coordinate.

We introduce a new time coordinate $\tilde{t}$ such that

$$d\tilde{t} = (Adt - Fdr)Q(r,t) \tag{7.56}$$

where Q(r, t) is a function of *r* and *t* that is to be chosen so as to make the right side of Eq. (7.56) into a perfect differential.

From Eq. (7.56) we obtain

$$Adt^2 - 2Fdrdt = \frac{1}{Q^2 A}d\tilde{t}^2 - \frac{F^2}{A}dr^2 \tag{7.57}$$

which gives

$$ds^2 = \frac{1}{Q^2 A}d\tilde{t}^2 - \left(B + \frac{F^2}{A}\right)dr^2 - B(r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2) \tag{7.58}$$

We can simplify Eq. (7.58) further by introducing a new radial coordinate $\tilde{r} = \mathrm{r}\sqrt{B(r)}$.

This has the advantage that the terms involving angular displacements reduce to $\tilde{r}^{\,2}\mathrm{d}\theta^2 + \tilde{r}^{\,2}\sin^2\theta\mathrm{d}\phi^2$, so we obtain

$$ds^2 = \frac{1}{Q^2 A}d\tilde{t}^2 - \left(B + \frac{F^2}{A}\right)\left(\frac{\partial r}{\partial \tilde{r}}\right)^2 d\tilde{r}^2 - \tilde{r}^2 d\theta^2 - \tilde{r}^2 \sin^2\theta d\phi^2 \tag{7.59}$$

Note that according to Eq. (7.59), the circumference of a circle of radius $\tilde{r}$ is $2\pi\tilde{r}$.

For instance, along a meridional circle, the element of length is $d\ell = \sqrt{-ds^2} = \tilde{r}d\theta$, and therefore, as in flat space, .

However, the length interval along the radial line is not $d\tilde{r}$, but $d\tilde{r}$ multiplied by an unknown function of $\tilde{r}$, yet to be determined.

The ratio of the length of the circumference to the length of the radius differs from 2π, so the geometry of space is not flat.

For the solution of Einstein's equations, it is convenient to omit the tildes in Eq. (7.59) and to write the unknown functions multiplying dt^2 and dr^2 as exponentials:

$$ds^2 = \mathrm{e}^{N(r,t)}dt^2 - \mathrm{e}^{L(r,t)}dr^2 - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \tag{7.60}$$

With $x^0 = t$, $x^1 = r$, $x^2 = \theta$, $x^3 = \phi$, the metric tensor corresponding to Eq. (7.60) is

$$g_{\mu\nu} = \begin{pmatrix} e^N & 0 & 0 & 0 \\ 0 & -e^L & 0 & 0 \\ 0 & 0 & -r^2 & 0 \\ 0 & 0 & 0 & -r^2 \sin^2\theta \end{pmatrix} \tag{7.61}$$

The unknown functions are now N (r, t) and L(r, t).

We will use Einstein's equations to find them.

The Christoffel symbols for a metric tensor of the form (7.61) have already been calculated in Chapter 6 [see Eq. (6.69)].

We repeat them here for convenience (the primes indicate derivatives, $N' = \partial N/\partial r$, $L' = \partial L/\partial r$; the dots indicate time derivatives, $\dot{N} = \partial N/\partial t$, $\dot{L} = \partial L/\partial t$):

$$\begin{aligned}
&\Gamma^0{}_{00} = \frac{1}{2}\dot{N}, \quad \Gamma^0{}_{01} = \Gamma^0{}_{10} = \frac{1}{2}N', \quad \Gamma^0{}_{11} = \frac{1}{2}\dot{L}e^{L-N} \\
&\Gamma^1{}_{00} = \frac{1}{2}N'e^{N-L}, \quad \Gamma^1{}_{10} = \Gamma^1{}_{01} = \frac{1}{2}\dot{L}, \quad \Gamma^1{}_{11} = \frac{1}{2}L', \quad \Gamma^1{}_{22} = -re^{-L} \\
&\Gamma^1{}_{33} = -r\sin^2\theta e^{-L} \\
&\Gamma^2{}_{12} = \Gamma^2{}_{21} = 1/r, \quad \Gamma^2{}_{33} = -\sin\theta\cos\theta \\
&\Gamma^3{}_{13} = \Gamma^3{}_{31} = 1/r, \quad \Gamma^3{}_{23} = \Gamma^3{}_{32} = \cot\theta
\end{aligned} \tag{7.62}$$

All other Christoffel symbols are zero.

The Ricci tensor can be calculated from the Christoffel symbols [see Eq. (6.102)]; only the 00, 01, 10, 11, 22, 33 components of $R_{\mu\nu}$ are nonzero.

The Einstein equation in vacuo is

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 0 \tag{7.63}$$

By taking the trace of this equation, we find that $R = 0$.

Hence the equation reduces to

$$R_{\mu\nu} = 0 \tag{7.64}$$

The 00, 01, 10, 11, 22, and 33 components of this equation are as follows:

$$R_{00} = \mathrm{e}^{N-L} \left(-\frac{N''}{2} + \frac{L'N'}{4} - \frac{N'^2}{4} - \frac{N'}{r} \right) + \frac{\ddot{L}}{2} + \frac{\dot{L}(\dot{L} - \dot{N})}{4} = 0 \tag{7.65}$$

$$R_{01} = -R_{10} = -\frac{\dot{L}}{r} = 0 \tag{7.66}$$

$$R_{11} = \frac{N''}{2} - \frac{L'N'}{4} + \frac{N'^2}{4} - \frac{L'}{r} + \mathrm{e}^{L-N} \left(\frac{\ddot{L}}{2} + \frac{\dot{L}(\dot{L} - \dot{N})}{4} \right) = 0 \tag{7.67}$$

$$R_{22} = \mathrm{e}^{-L} \left[1 + \frac{1}{2} r(N' - L') \right] - 1 = 0 \tag{7.68}$$

$$R_{33} = \sin^2\theta \mathrm{e}^{-L} \left[1 + \frac{1}{2} r(N' - L') \right] - \sin^2\theta = 0 \tag{7.69}$$

Although these equations look complicated, they can be integrated quite easily.

Equation (7.66) says that L is time independent, and this implies that *all* the terms involving time derivatives in Eqs. (7.65) and (7.67) drop out.

The sum of Eq. (7.67) and e^{L-N} times Eq. (7.65) then gives

$$-\frac{L'}{r}-\frac{N'}{r}=0 \tag{7.70}$$

from which

$$N=-L+h(t) \tag{7.71}$$

where h(t) is an arbitrary function of time [in regard to the space derivatives in Eq. (7.70), this function h(t) behaves like a constant of integration].

When we substitute Eq. (7.71) into Eq. (7.68), we find

$$\mathrm{e}^{-L}(-rL'+1)=1 \tag{7.72}$$

This has the general solution

$$\mathrm{e}^{L}=\frac{1}{1-C/r} \tag{7.73}$$

where C is a constant.

According to Eq. (7.71), the solution for N(r) is then

$$\mathrm{e}^{N(r)}=\mathrm{e}^{-L(r)+h(t)}=(1-C/r)\mathrm{e}^{h(t)} \tag{7.74}$$

As a final step, we must check that the solutions (7.73) and (7.74) also satisfy the differential equation (7.65) or (7.67) (we used the *sum* of these equations in the construction of our solution; now we must check at least one of these equations by itself).

With the solutions (7.73) and (7.74), the spacetime interval for the spherically symmetric field takes the form

$$ds^2=\left(1-\frac{C}{r}\right)\mathrm{e}^{h(t)}dt^2-\frac{1}{1-C/r}dr^2-r^2d\theta^2-r^2\sin^2\theta d\phi^2 \tag{7.75}$$

Here, the time-dependent function h(t) remains undetermined and unknown.

The field equations do not determine this function, but we can eliminate it by a transformation of the time coordinate.

If we adopt a new time coordinate $\tilde{t}$ such that $d\tilde{t} = e^{h(t)/2}dt$, then the function h(t) disappears from view, and the first term in Eq. (7.75) becomes $(1 - C/r)d\tilde{t}^2$.

We can then omit the tilde, so we obtain

$$ds^2 = \left(1 - \frac{C}{r}\right)dt^2 - \frac{1}{1 - C/r}dr^2 - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \tag{7.76}$$

To find the value of C, we want to compare this expression with the result obtained from the linear theory [Eq. (4.12), with $r^2 = x^2 + y^2 + z^2$],

$$ds^2 = \left(1 - \frac{2GM}{r}\right)dt^2 - \left(1 + \frac{2GM}{r}\right)(dx^2 + dy^2 + dz^2) \tag{7.77}$$

where M is the central mass.

But we cannot compare (7.76) and (7.77) directly, because in these equations the radial variables r are defined in somewhat different ways: The variable r in Eq. (7.76) is such that the measured circumference of a circle concentric with the origin is $2\pi r$, whereas the variable r in Eq. (7.77) is such that the circumference is approximately $2\pi r(1 + GM/r)$.

To deal with this problem, let us change Eq. (7.76) by introducing a new coordinate $\tilde{r}$:

$$\tilde{r} = \frac{1}{2}\sqrt{r^2 - Cr} + \frac{1}{2}r - \tfrac{1}{4}C \tag{7.78}$$

or

$$r = \tilde{r}\left(1 + \frac{C}{4\tilde{r}}\right)^2 \tag{7.79}$$

This transformation gives

$$ds^2 = \left(\frac{1 - C/4\tilde{r}}{1 + C/4\tilde{r}}\right)^2 dt^2 - \left(1 + \frac{C}{4\tilde{r}}\right)^4 (d\tilde{r}^2 + \tilde{r}^2 d\theta^2 + \tilde{r}^2 \sin^2\theta d\phi^2) \tag{7.80}$$

The coordinates used in Eq. (7.80) are called *isotropic*.

In the weak field limit ($\tilde{r} \to \infty$), Eq. (7.80) reduces to

$$ds^2 = \left(1 - \frac{C}{\tilde{r}}\right) dt^2 - \left(1 + \frac{C}{\tilde{r}}\right) (d\tilde{r}^2 + \tilde{r}^2 d\theta^2 + \tilde{r}^2 \sin^2\theta d\phi^2) \tag{7.81}$$

Obviously, this equation has the same form as Eq. (7.77). By comparing, we see that

$$C = 2GM \tag{7.82}$$

With this, Eq. (7.76) becomes

$$ds^2 = \left(1 - \frac{2GM}{r}\right) dt^2 - \frac{1}{1 - 2GM/r} dr^2 - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \tag{7.83}$$

This is the final form of the Schwarzschild solution.

Note that the final result is time independent, even though the only original assumption was merely spherical symmetry.

The time dependence that we included in the original spherical symmetric metric tensor [see Eq. (7.61)] disappears in the final result.

This curious circumstance is called *Birkhoff's theorem*.

It is the general-relativistic analog of Newton's theorem; namely, that the external gravitational potential of a spherical mass ($\Phi = -GM/r$) is independent of the size of the mass, so this potential remains time independent, even if the spherical mass shrinks in size or expands.

In general relativity, this time independence of the spherically symmetric solution is important in the study of gravitational collapse, when a mass is crushed by its own weight and shrinks to a smaller and smaller size.

In the exterior region, the gravitational field of such a collapsing mass remains unchanged, but in the interior region, the gravitational field changes with time.

It is a corollary of Birkhoff's theorem that a spherically symmetric mass distribution produces no gravitational field inside an empty spherical cavity centered on the mass distribution, even if the mass distribution shrinks or expands in size.

This result is of course well known in Newton's theory.

In Einstein's theory, the Birkhoff theorem ensures that the solution inside the cavity must be the Schwarzschild solution given by Eq. (7.76).

Since an empty cavity cannot contain any singularities at r = 0, we must take $C = 0$, and hence the spacetime is flat inside the cavity.

Also note that the mass M in the Schwarzschild solution is the total mass of the system; the mass-energy contributed by the gravitational fields is included in M.

It is clear that unless M is the total energy of the system, the principle of equivalence ($M_I = M_G$) cannot be satisfied.

In this chapter, we will not give the proof that the gravitational mass M in Eq. (7.83) is in fact equal to the inertial mass of the system.

Such a proof can be given, but requires some knowledge of the properties of the exact energy-momentum tensors of matter and of gravitation.

An additional remark: In the preceding calculations we have assumed that the cosmological constant is zero.

If we want to drop this assumption, then we must replace Eq. (7.63) by

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -\Lambda g_{\mu\nu} \tag{7.84}$$

It can be shown that this equation leads to a modified Schwarzschild solution

$$ds^2 = \left(1 - \frac{2GM}{r} - \frac{\Lambda r^2}{3}\right) dt^2 - \frac{1}{1 - 2GM/r - \Lambda r^2/3} dr^2 - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \tag{7.85}$$

This spacetime geometry does not become asymptotically flat as $r \to \infty$.

However, because Λ is very small (see Section 7.3), there is a range of radii $GM/c^2 \ll r \ll c/\sqrt{\Lambda}$ in which the spacetime geometry is nearly flat.

At the lower end of this range, the effect of the mass M dominates; at the upper end of this range, the effect of the cosmological term dominates. (At very large values of r we must also take into account the large-scale curvature of the universe; see Chapter 9).

We know from Section 3.5 that in the Newtonian approximation the quantity $\frac{1}{2}\kappa h_{00}$ is to be identified with the Newtonian potential [see Eq. (3.84)],

$$\Phi = \frac{1}{2}\kappa h_{00} = \frac{1}{2}(g_{00} - 1)$$

According to Eq. (7.85), this leads to

$$\Phi = -\frac{GM}{r} - \frac{\Lambda}{6}r^2 \tag{7.86}$$

The second term on the right side of this equation represents a cosmological correction to the Newtonian potential; this term of course agrees with Eq. (7.52).

For the motion of the planets, the cosmological correction is completely insignificant.

We will, therefore, assume that $\Lambda = 0$ until further notice.

7.5 Motion of planets; perihelion precession

The equation of motion of a particle in a gravitational field is the geodesic equation, Eq. (6.76),

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu{}_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0 \tag{7.87}$$

The following quantity is a constant of the motion:

$$g_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = \text{constant} = 1 \tag{7.88}$$

This equation is obviously true by definition of dτ.

But we can also use the equation of motion (7.87) to show explicitly that the left side of Eq. (7.88) is a constant.

Thus, Eq. (7.88) can be regarded as a first integral of the equation of motion.

An alternative and often useful way of writing Eq. (7.87) is

$$\frac{d}{d\tau}\left(g_{\mu\nu}\frac{dx^\nu}{d\tau}\right) - \frac{1}{2}g_{\alpha\beta,\mu}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0 \tag{7.89}$$

For the Schwarzschild geometry, the Christoffel symbols are given in Eq. (7.62), but we can now omit the terms ∂L/∂t and ∂N/∂t, because we know that L and N are time independent.

The components of the geodesic equation are then

$$\ddot{t} + N'\dot{r}\dot{t} = 0 \tag{7.90}$$

$$\ddot{r} + \frac{1}{2}N'e^{N-L}\dot{t}^2 + \frac{1}{2}L'\dot{r}^2 - re^{-L}\dot{\theta}^2 - r\sin^2\theta e^{-L}\dot{\phi}^2 = 0 \tag{7.91}$$

$$\ddot{\theta} + \frac{2}{r}\dot{r}\dot{\theta} - \cos\theta\sin\theta\dot{\phi}^2 = 0 \tag{7.92}$$

$$\ddot{\phi} + \frac{2}{r}\dot{r}\dot{\phi} + 2\cot\theta\dot{\phi}\,\dot{\theta} = 0 \tag{7.93}$$

where dots now indicate derivatives with respect to τ along the worldline, such as $\dot{t} \equiv dt/d\tau$, $\dot{r} = dr/d\tau$, and so on .

Note that the dots in Eq.(7.62) had a different meaning; they indicated derivatives with respect to t.

We will assume that the orbit is in the equatorial plane, $\theta = \pi/2$.

Equation (7.92) shows that if $\theta = \pi/2$ and $\dot{\theta} = 0$ initially, then $\ddot{\theta} = 0$, and therefore the orbit remains in this plane.

Equations (7.90) and (7.93) can then be integrated directly, giving

$$\dot{t} = \mathcal{E}e^{-N} \tag{7.94}$$

and

$$\dot{\phi} = \ell/r^2 \tag{7.95}$$

where $\mathcal{E}$ and ℓ are constants.

These constants are proportional to the energy and the angular momentum, respectively.

We can see this by calculating the energy and the angular momentum from the Lagrangian

$$L = m\sqrt{g_{\mu\nu}u^{\mu}u^{\nu}} \tag{7.96}$$

The energy is

$$E = P_0 = \frac{\partial L}{\partial u^0} = mg_{00}u^0 = m\mathrm{e}^N \dot{t} \tag{7.97}$$

and, with $x^3 = \phi$ and $x^2 = \theta = \pi/2$, the angular momentum is

$$l = -P_3 = -\frac{\partial L}{\partial u^3} = -mg_{33}u^3 = mr^2\dot{\phi} \tag{7.98}$$

Comparison with Eqs. (7.94) and (7.95) shows that the constants $\mathcal{E}$ and ℓ are, respectively, the energy and the angular momentum per unit mass, $\mathcal{E}$ = E/m and $\ell = l$/m.

That these quantities are constants of the motion is immediately obvious in the context of the Lagrangian formalism, since the Lagrangian is independent of t and of ϕ, and hence the momenta conjugate to these ignorable coordinates are constant.

Equation (7.91) is somewhat harder to solve, but we can bypass it if we use the first integral of the motion given by Eq. (7.88):

$$\mathrm{e}^N \dot{t}^2 - \mathrm{e}^L \dot{r}^2 - r^2\dot{\theta}^2 - r^2 \sin^2\theta\, \dot{\phi}^2 = 1 \tag{7.99}$$

If we insert $\dot{t}$ and $\dot{\phi}$ from Eqs. (7.94) and (7.95), and $\theta = \pi/2$, $\mathrm{e}^{-\mathrm{L}} = \mathrm{e}^{\mathrm{N}} = 1 - 2\mathrm{GM/r}$, we obtain

$$\frac{\mathcal{E}^2}{1 - 2GM/r} - \frac{\dot{r}^2}{1 - 2GM/r} - \frac{\ell^2}{r^2} = 1 \tag{7.100}$$

Here it is convenient to introduce the variable

$$u = \frac{1}{r} \tag{7.101}$$

which is also often used in nonrelativistic celestial mechanics. Then

$$\dot{r} = \frac{dr}{d\phi}\dot{\phi} = -\frac{1}{u^2}\frac{du}{d\phi}\frac{\ell}{r^2} = -\frac{du}{d\phi}\ell \tag{7.102}$$

and Eq. (7.100) reduces to

$$\mathcal{E}^2 - \ell^2\left(\frac{du}{d\phi}\right)^2 - \ell^2u^2(1-2GMu) = (1-2GMu) \tag{7.103}$$

By differentiating this with respect to ϕ, we obtain a second-order differential equation for the orbit:

$$\frac{d^2u}{d\phi^2} + u - \frac{GM}{\ell^2} - 3GMu^2 = 0 \tag{7.104}$$

If we compare this with the corresponding orbital equation of Newtonian theory,

$$\frac{d^2u}{d\phi^2} + u - \frac{GM}{\ell^2} = 0 \tag{7.105}$$

we see that these equations differ only by the term $3GMu^2$.

This term represents the relativistic correction to the motion.

For the special case of a planet moving in the Schwarzschild field of the Sun, the relativistic correction to the motion is extremely small.

This can be seen by comparing the second and fourth terms on the left side of Eq. (7.104).

These terms differ by a factor 3GMu, or, in cgs units, $3GM/rc^2$.

Even for the planet Mercury, this is a very small number; with $M = M_\odot = 2.0\times10^{33}$ g, $r = 5.5\times10^{12}$ cm, it is $3GM/rc^2 \cong 10^{-7}$.

Since $3GMu^2$ is so small compared with the other terms, it will be sufficient to use a method of successive approximations in the solution of Eq. (7.104).

We first write down the solution of the Newtonian equation,

$$u = \frac{GM}{\ell^2}[1 + \varepsilon \cos(\phi - \phi_0)] \tag{7.106}$$

where ε and ϕ_0 are constants.

The equation (7.106) is that of an ellipse with an eccentricity ε and a perihelion located at ϕ_0.

If we now replace the small term $3GMu^2$ in Eq. (7.104) by its Newtonian approximation as given by Eq. (7.106), we obtain

$$\frac{d^2u}{d\phi^2} + u - \frac{GM}{\ell^2} - \frac{3(GM)^3}{\ell^4} - \frac{6\varepsilon(GM)^3}{\ell^4}\cos(\phi - \phi_0) - 3\varepsilon^2\frac{(GM)^3}{\ell^4}\cos^2(\phi - \phi_0) = 0 \tag{7.107}$$

This equation is mathematically analogous to that of a harmonic oscillator, with a restoring force –u, two oscillating driving forces proportional to $\cos(\phi - \phi_0)$ and $\cos^2(\phi - \phi_0)$, and two extra constant forces.

The first of the oscillating driving forces is in resonance with the natural oscillations, and therefore it gradually builds up a (relatively) large secular perturbation of the orbit; but the second oscillating driving force is not in resonance and merely produces periodic perturbations of small amplitude.

We will therefore ignore the second oscillating driving force, and we will also ignore the constant term $3(GM)^3/\ell^4$, which is small by comparison with the constant term $GM)/\ell^2$ and produces no interesting observable effects.

We are then left with

$$\frac{d^2u}{d\phi^2} + u - \frac{GM}{\ell^2} - \frac{6\varepsilon(GM)^3}{\ell^4}\cos(\phi - \phi_0) = 0 \tag{7.108}$$

The solution of this differential equation is

$$u = \frac{GM}{\ell^2}[1 + \varepsilon\cos(\phi - \phi_0)] + \frac{3(GM)^3}{\ell^4}\phi\sin(\phi - \phi_0) \tag{7.109}$$

which we can also write in the approximate form

$$u \cong \frac{GM}{\ell^2}\left[1 + \varepsilon\cos\left(\phi - \phi_0 - \frac{3(GM)^2}{\ell^2}\phi\right)\right] \tag{7.110}$$

The orbit described by Eq. (7.110) is a precessing ellipse.

The argument of the cosine changes by 2π when ϕ changes by

$$\Delta\varphi = 2\pi\left[1 - \frac{3(GM)^2}{\ell^2}\right]^{-1} \cong 2\pi\left[1 + \frac{3(GM)^2}{\ell^2}\right] \tag{7.111}$$

This shows that the angular distance between one perihelion and the next is larger than 2π by

$$6\pi\frac{(GM)^2}{\ell^2} \tag{7.112}$$

This quantity gives the angular precession of the perihelion per revolution.

Note that the perihelion precesses in the direction of motion, that is, the perihelion advances ($\Delta\phi > 2\pi$; see Fig. 7.1).

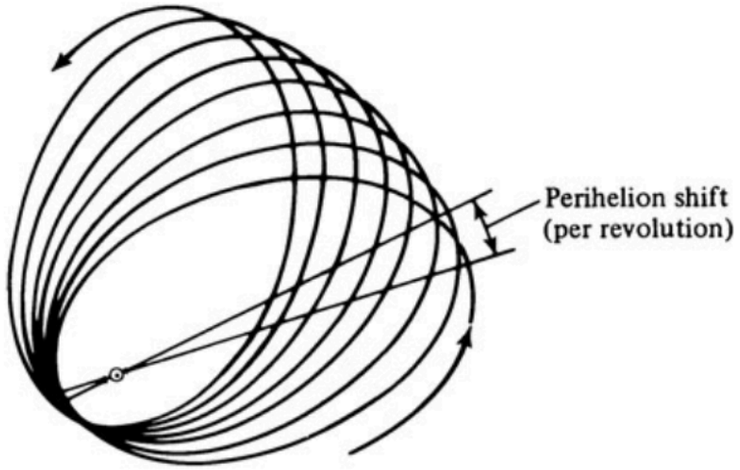


Fig. 7.1 Planetary orbit with perihelion precession. The angular displacement, reckoned around the orbit, between one perihelion and the next is $\Delta\phi > 2\pi$.

For an ellipse of semi-major axis a, the perihelion distance is a(1 − ε).

By evaluating Eq. (7.106) at the perihelion, where $\cos(\phi - \phi_0) = 1$, we find

$$\frac{1}{a(1-\varepsilon)} = \frac{GM}{\ell^2}(1+\varepsilon) \tag{7.113}$$

or

$$\frac{GM}{\ell^2} = \frac{1}{a(1-\varepsilon^2)} \tag{7.114}$$

Accordingly, the angular advance of the perihelion per revolution is, in cgs units,

$$6\pi \frac{GM}{a(1-\varepsilon^2)c^2} \tag{7.115}$$

For the case of Mercury, this amounts to 0.1035 arcsec per revolution, or 42.98 arcsec per century.

Table 7.1 gives the observed and predicted precessions for all the inner planets, according to an investigation by Anderson et al.

Table 7.1 Perihelion precession of planets

Planet	Observed Precession* (arcsec per century)	Predicted Precession (arcsec per century)
Mercury	43.1	42.98
Venus	8.65	8.62
Earth	3.85	3.84
Mars	1.36	1.35

* The values quoted for the observed precessions of the planets are not independent, because in the data analysis a single parameter is used to characterize all these precessions, so if the planets deviate from the predicted precession, they are assumed to deviate in the same way.

In this table, the values for the observed perihelion precession have been corrected for the purely Newtonian perturbations of the orbits.

For instance, the perturbation of Mercury's orbit by other planets contributes a precession of $\cong 500$ arcsec per century.

Furthermore, the equinoctial precession of the celestial coordinates used in planetary observations gives an apparent advance of $\cong 5000$ arcsec per century.

All these known corrections have been subtracted out in Table 7.1.

No correction has been made for a possible quadrupole moment of the Sun, because, from observation of vibrations of the body of the Sun (helioseismology; see Section 1.4), it is known that the quadrupole moment is small and its contribution to the perihelion precession of planets is below the observational uncertainties, about $\pm 0.3\%$, for all the planets listed in the table.

The agreement between observed and predicted values is as good as can be expected.

The results of Table 7.1 are based on several thousand optical observations of planetary positions obtained with Transit Circles at the U.S. Naval Observatory and also on several hundred radar-ranging observations obtained with the Arecibo, Haystack, and Goldstone radio telescopes.

Either the radar pulses were reflected directly on the surfaces of the inner planets, or they were reflected on Mariner or Pioneer spacecraft while in orbit around a planet, or on Viking landers parked on the surface.

By a least-squares fit, the data were used to determine the orbital elements of all the planets; the sizes of Venus, Mercury, and Mars; and the electron density in the solar corona (these parameters affect the travel time of radar signals).

The perturbations of the planetary orbits by each other were evaluated by numerical integration of the equations of motion over the entire 20th century.

Whereas the relativistic perihelion precessions in the Solar System are very small, the precession can be quite large in the case of close binary star systems.

For a system consisting of two white dwarfs or two neutron stars of mass $1M\odot$ separated by a distance of 10^{11} cm, Eq. (7.115) gives a periastron advance of 3×10^{-5} radian per revolution.

In one year, such a system goes through 10^3 revolutions, and hence the periastron precession amounts to several degrees per year.

The binary pulsar PSR 1913+16, discussed in Section 5.5, is an example of such a close binary star system, with a large periastron precession of 4.2° per year.

Because the masses of the stars are not known, the measured periastron precession cannot be used as a direct test of the theoretical formula for the precession.

Instead, as mentioned in Section 5.5, the measured precession and other orbital data are used to determine the masses of the stars, and the rate of energy loss by gravitational radiation is then calculated from these masses and the orbital parameters.

Note, however, that if the theoretical formula for the periastron precession were wrong, then the masses would be wrong, and the excellent agreement between the predicted and the observed rates of change of the orbital period of PSR 1913+16 would become an accidental, and incredible, coincidence.

Several other similar binary pulsar systems are known, with periastron precessions of the same order of magnitude.

Although the perturbative solution (7.110) of the geodesic equation is adequate for the motion of planets around the Sun, a general solution of the geodesic equation is needed for the motion of a particle around, say, a black hole.

Such a solution can be given in terms of elliptic integrals.

However, it is more illuminating to discuss the motion qualitatively, by means of an effective potential.

For this purpose, we rewrite Eq. (7.100) as

$$\dot{r}^2 + \left(1 - \frac{2GM}{r}\right)\left(1 + \frac{\ell^2}{r^2}\right) = \mathcal{E}^2 \tag{7.116}$$

In this equation, the second term on the left plays the role of an effective potential.

It is convenient to define this effective potential as

$$\mathcal{V}(r) = \sqrt{\left(1 - \frac{2GM}{r}\right)\left(1 + \frac{\ell^2}{r^2}\right)} \tag{7.117}$$

With this definition, at large distances, the effective potential approaches $\mathcal{V}(r) \to 1 - GM/r + \ell^2/2r^2$. This coincides with the Newtonian effective potential, except for an additive constant. Equation (7.116) then becomes

$$\dot{r}^2 + \mathcal{V}^2(r) = \mathcal{E}^2 \tag{7.118}$$

Note that the effective potential is zero at r = 2GM.

According to Eq. (7.118), a particle instantaneously at rest ($\dot{r} = 0$) at this point then has zero energy, $\mathcal{E} = 0$ and E = 0.

Since the energy *E* includes the rest-mass energy of the particle, this means that for a particle instantaneously at rest at r = 2GM, the (negative) gravitational binding energy equals the rest-mass energy.

For a qualitative discussion of the orbital motion, we must examine the turning points.

The turning points of the motion are at V(r) = E.

Figure 7.2 presents plots of V(r) for several values of the angular momentum ℓ.

The centrifugal barrier at small values of *r* is of finite height, in contrast to the Newtonian case, where the centrifugal barrier tends to infinity as $r \to 0$.

Hence, for a range of values of ℓ and $\mathcal{E}$, the interior turning point is absent – the particle is then pulled into the singularity at *r* = 0 even if it has angular momentum.

The minima in the curves in Figure 7.2 correspond to stable circular orbits; the maxima correspond to unstable circular orbits.

The smallest stable circular orbit corresponds to the inflection point in the curve for $\ell = 2\sqrt{3}\,GM$; this smallest stable circular orbit has a radius r = 6GM, and it has an energy $\mathcal{E} = \sqrt{8/9}$, or $E = \sqrt{8/9}m$.

Note that for angular momentum $\ell < 2\sqrt{3}\,\mathrm{GM}$, there is no interior turning point (no angular-momentum barrier); if the particle is initially moving inward, it continues to move inward with ever increasing radial speed $|\dot{r}|$.

For the smallest stable circular orbit, the binding energy relative to a particle at rest at infinity is

$\mathrm{m} - \sqrt{8/9}\,\mathrm{m} = 0.057\mathrm{m}$, or about 6% of the rest-mass energy.

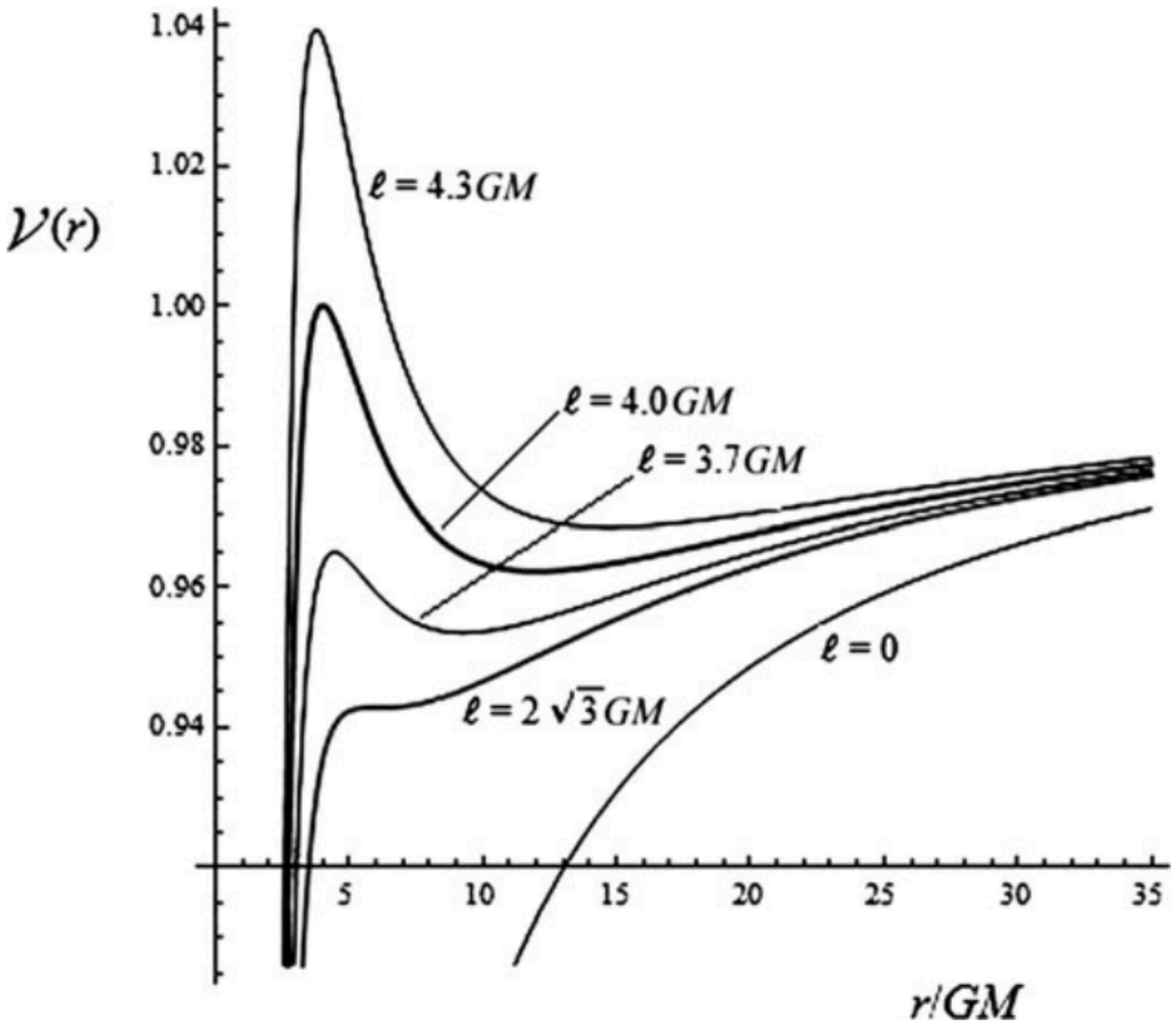


Fig. 7.2 The effective potential $\mathcal{V}(r)$ for several values of the orbital angular momentum ℓ.

7.6 Propagation of light; gravitational redshift

The trajectories of light rays that pass near the Sun are almost straight lines, and the deflection of such light rays can be adequately treated by the linear approximation we discussed in Chapter 4.

However, an exact treatment of the deflection of light in the Schwarzschild geometry is of interest when the central mass is compact and the spacetime in its vicinity is strongly curved.

Such extreme cases of the Schwarzschild geometry, with strong curvature, are found near neutron stars and near black holes.

The differential equation for the trajectory of a light signal in the Schwarzschild geometry can be obtained by taking a suitable limit in the equation for the trajectory of a particle, Eq. (7.103).

We begin by rewriting this equation as

$$\frac{\mathcal{E}^2}{\ell^2} - \left(\frac{du}{d\phi}\right)^2 - u^2(1-2GMu) = \frac{(1-2GMu)}{\ell^2} \tag{7.119}$$

To convert this into an equation for the trajectory of a light signal, we must proceed to the limit of zero particle mass and zero proper time, $m \to 0$ and $d\tau \to 0$.

This implies $\ell = r^2 d\phi/d\tau \to \infty$ and $\mathcal{E} = e^N\, dt/d\tau \to \infty$; however, $\mathcal{E}/\ell$ remains finite (and constant).

Accordingly, Eq. (7.119) becomes

$$\frac{1}{b^2} - \left(\frac{du}{d\phi}\right)^2 - u^2(1-2GMu) = 0 \tag{7.120}$$

where $b = \ell/\mathcal{E}$.

By differentiating this equation with respect to ϕ, we could obtain a second-order differential equation for the trajectory [compare Eq. (7.104)]; but we will deal with the first-order differential equation instead, because this is more convenient.

If the light ray is incident from a large distance, we can identify the constant b as the impact parameter.

This becomes obvious in the limit $u \to 0$ (or $r \to \infty$), where Eq. (7.120) reduces to $1/b^2 - (du/d\phi)^2 = 0$, which has the solution $u = \phi/b + \text{constant}$, as expected for a distant ($u \ll 1/b$) straight line of impact parameter b.

Equation (7.120) gives us the relation between the change in azimuthal angle and the change in u; that is, $d\phi = du/\sqrt{1/b^2 - u^2(1 - 2GMu)}$.

By integration, we can then find the net change $\Delta\phi$ in the azimuthal angle for a ray that arrives from some large distance, bends around the central mass, and proceeds to a large distance in a new direction.

At the point of closest approach, $du/d\phi = 0$.

Designating this point by u_0, we see from Eq. (7.120) that

$$\frac{1}{b^2} = u_0^2(1 - 2GMu_0) \tag{7.121}$$

This is a cubic equation, which determines u_0 for a given value of b.

Since the trajectory is symmetric about the point of closest approach, we can express the net change $\Delta\phi$ as twice the change that occurs between $u = 0$ and $u = u_0$:

$$\Delta\phi = 2\int_0^{u_0} \frac{du}{\sqrt{1/b^2 - u^2(1 - 2GMu)}} \tag{7.122}$$

This integral can be reduced to an elliptic integral.

However, it is usually easiest to obtain $\Delta\phi$ by numerical integration of Eq. (7.122).

Figure 7.3 displays a plot of the bending angle $\Delta\phi - \pi$ of the light ray vs. impact parameter.

This bending angle represents the amount by which the ray is deflected from a straight line (note that $\Delta\phi = \pi$ corresponds to a straight ray).

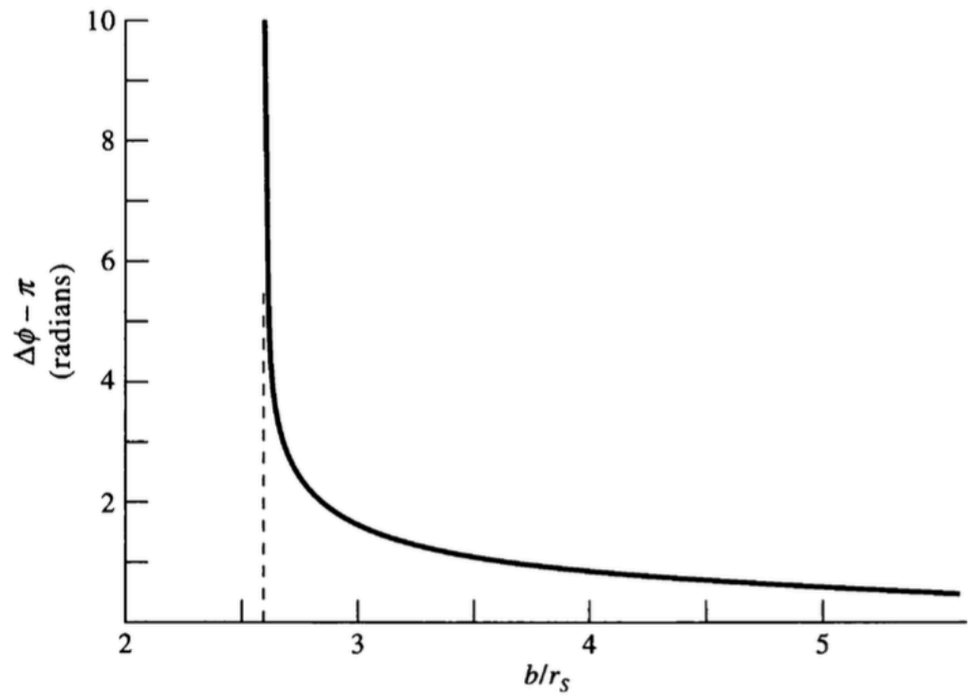


Fig. 7.3 Bending angle vs. impact parameter.

$b \to 3\sqrt{3}\,GM$, which corresponds to $r_0 = 3GM$.

In this limiting case, the ray revolves around the central mass many times, in a quasicircular orbit of radius $\cong 3GM$.

The sphere of radius $r_0 = 3GM$ is called the *photosphere*.

A light ray that starts at this exact radius is in an exact circular orbit.

However, this circular orbit is unstable, and after a few revolutions the light ray is liable to spiral either in or out.

Of course, a large bending angle and a quasicircular orbit for light rays are possible only if the mass is sufficiently compact, so the point of closest approach lies outside of the mass distribution.

This condition is always satisfied for a black hole, in which all the mass has collapsed to a central singularity.

Figure 7.4 shows the trajectories of several light rays that approach a compact mass from a large distance with different impact parameters.

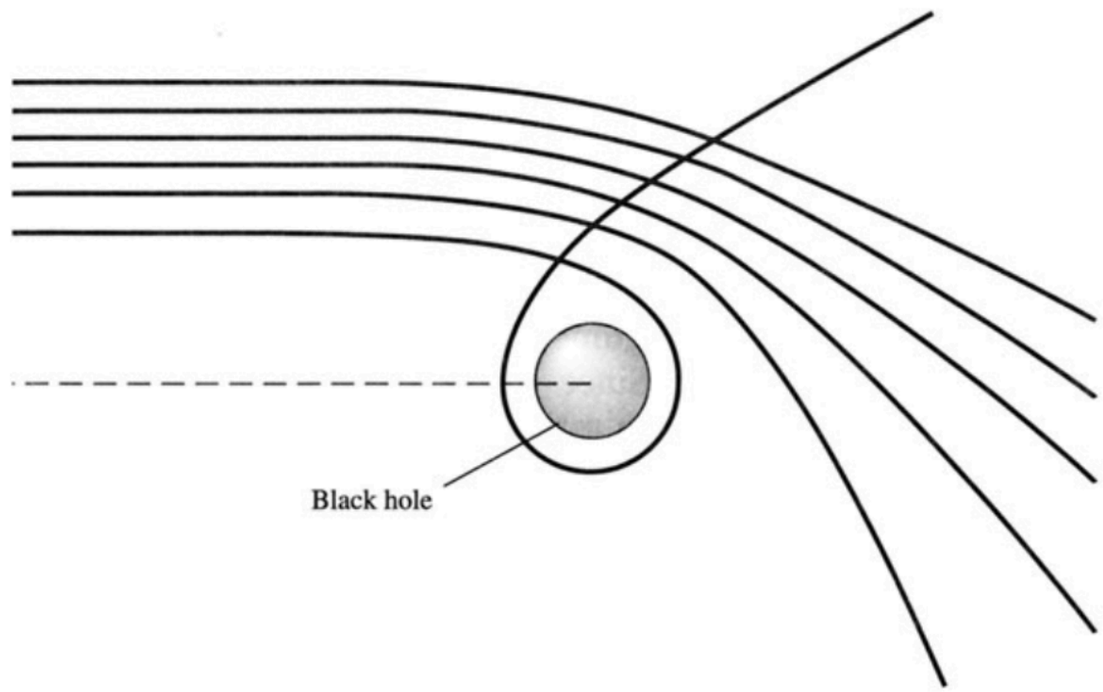


Fig. 7.4 Trajectories of light rays of different impact parameters. (From Ohanian, 1987.)

If the source of light is far to the left, and the observer far to the right, light rays can reach the observer along a path with a deflection angle of less than 2π (perhaps much less than 2π), and also along a multitude of paths with bending angles of more than 2π, which arise from the rays that circle the compact mass one, two, three, or more times.

The paths can circle in either a clockwise or a counterclockwise direction. However, as seen by the observer, all such rays with large bending angles will seem to come from the immediate vicinity of the compact mass (their impact parameters are of the order of $3\sqrt{3}\,GM$).

If the observer is far from the mass, the angular separations between these images will be small, and the observer will not be able to resolve them.

Hence, all these images will appear to merge into a single image seen at (approximately) the position of the compact mass.

The preceding discussion dealt only with the trajectories of light rays.

But it is also of interest to consider the change of frequency of the light during its propagation in the Schwarzschild geometry; that is, the gravitational redshift.

In Section 4.2, we already obtained a gravitational time-dilation formula.

According to this formula [see Eq. (4.14)], the ratio of the proper times that two stationary clocks, at radial positions r_1 and r_2 measure between two successive signals sent from r_1 to r_2 is

$$\frac{d\tau_2}{d\tau_1} = \frac{\sqrt{g_{00}(r_2)}}{\sqrt{g_{00}(r_1)}} \tag{7.123}$$

In the Schwarzschild geometry, this becomes

$$\frac{d\tau_2}{d\tau_1} = \frac{\sqrt{1-2GM/r_2}}{\sqrt{1-2GM/r_1}} \tag{7.124}$$

If the two signals are two successive wavecrests of a periodic light wave, then the measured frequencies are $\nu_1 = 1/d\tau_1$ and $\nu_2 = 1/d\tau_2$, and

$$\frac{\nu_1}{\nu_2} = \frac{\sqrt{1-2GM/r_2}}{\sqrt{1-2GM/r_1}} \tag{7.125}$$

This is the (exact) formula for the gravitational redshift of light in the Schwarzschild geometry.

The result for the gravitational redshift of light can also be obtained by an instructive argument involving the Killing vector of the Schwarzschild geometry.

Because the Schwarzschild metric is independent of time, there must be a timelike Killing vector, of the form $\xi^\mu = (1, 0, 0, 0)$ [it is obvious from Eq. (6.172) that whenever the metric tensor is independent of some coordinate, a constant vector ξ^μ in the direction of that coordinate is a Killing vector].

Note that although ξ^μ is constant, ξ_μ is not constant, that is,

$$\xi_\mu = g_{\mu\nu}\xi^\mu = (g_{00}\xi^0, 0, 0, 0) = ((1 - 2GM/r), 0, 0, 0) \tag{7.126}$$

Figure 7.5 shows the worldlines of the two clocks used for the measurement of the frequency of the light signal.

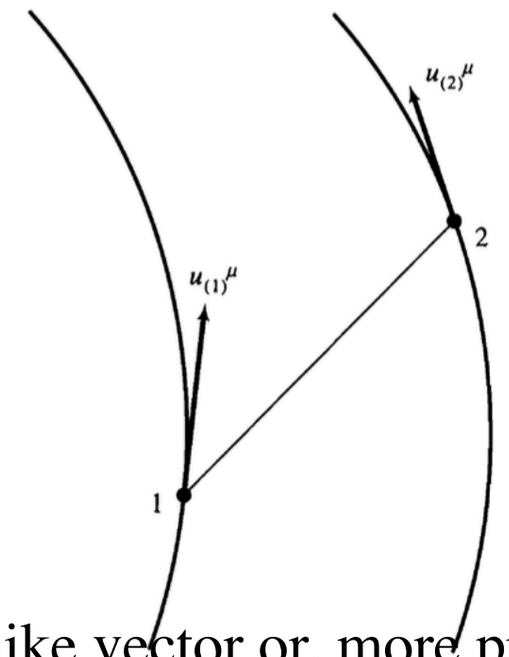


Fig. 7.5 Worldlines of two stationary clocks at different positions in a time-independent gravitational field. The four-velocity u^μ of each clock is parallel to the timelike Killing vector ξ^μ of this gravitational field.

Each clock is stationary, so its four-velocity is a timelike vector or, more precisely, a timelike unit vector ($u^\mu u_\mu = 1$).

This means that the four-velocity u^μ of each clock is parallel to the Killing vector, that is $u^\mu \propto \xi^\mu$.

To make this proportionality into an equality, we must renormalize ξ^μ so it becomes a unit vector; this requires dividing by $(\xi^\alpha \xi_\alpha)^{1/2}$,

$$u^\mu = \frac{\xi^\mu}{(\xi^\alpha \xi_\alpha)^{1/2}} \tag{7.127}$$

The frequency of the light measured by a clock is proportional to the P_0 component of the energy-momentum vector of a photon evaluated in the local rest frame of the clock, according to the usual relation $h\nu = P_0$.

We can also write this as $h\nu = u^\mu P_\mu$, where u^μ is the four-velocity of the clock and P_μ the energy-momentum vector of the photon.

We can then express the ratio of the frequencies measured by the two clocks at the radial positions r_1 and r_2 as

$$\frac{\nu_1}{\nu_2} = \frac{u^\mu(1)P_\mu(1)}{u^\kappa(2)P_\kappa(2)} = \frac{\xi^\mu(1)P_\mu(1)}{\xi^\kappa(2)P_\kappa(2)} \times \frac{[\xi^\beta(2)\xi_\beta(2)]^{1/2}}{[\xi^\alpha(1)\xi_\alpha(1)]^{1/2}} \tag{7.128}$$

But according to the conservation theorem we proved in Section 6.8 [see Eq. (6.190)], $\xi^\mu P_\mu$ is constant during the motion of the photon, and hence the first factor on the right side of Eq. (7.128) equals 1, and

$$\frac{\nu_1}{\nu_2} = \frac{[\xi^\beta(2)\xi_\beta(2)]^{1/2}}{[\xi^\alpha(1)\xi_\alpha(1)]^{1/2}} \tag{7.129}$$

This expresses the gravitational redshift entirely in terms of the Killing vectors evaluated at the two positions.

If we insert the explicit values of the Killing vectors into this equation, we find

$$\frac{\nu_1}{\nu_2} = \frac{\sqrt{1 - 2GM/r_2}}{\sqrt{1 - 2GM/r_1}} \tag{7.130}$$

which is the same result as before.

7.7 Geodetic precession

One of the characteristic features of curved space is that parallel transport of a vector alters its direction.

This suggests that we should try to detect the curvature of spacetime near the Earth by examining parallel transport.

But to perform such an experiment, we must first devise some physical procedure for the parallel transport of a vector.

In nongravitational physics, our experience with gyroscopes tells us that the physical transport of a gyroscope suspended in frictionless gimbals results in the parallel transport of its direction of spin.

However, from this we cannot immediately draw the conclusion that in gravitational physics the transport of such a gyroscope will also result in parallel transport of the direction of its spin.

To determine under what conditions transport of a gyroscope results in parallel transport, we begin with the equation of motion for the spin of a rigid body according to Newton's theory.

From Section 1.8 we know that a rigid body in a gravitational field is subject to a tidal torque [see Eq. (1.66)], which leads to a rate of change of the spin,

$$\frac{dS^n}{dt} = \varepsilon^{nk\ell} R'^{k}{}_{0s0} \left(-I^{ls} + \frac{1}{2}\delta^s_l I^{rr} \right) \tag{7.131}$$

Here $R'^{k}{}_{0s0}$ is the Riemann tensor evaluated in the rest frame of the gyroscope.

The presence of the Riemann tensor on the right side of Eq. (7.131) tells us that the equation of motion for the spin does not obey the SEP, or the principle of minimal coupling, and that the transport of the spin of a nonspherical gyroscope does not imitate parallel transport.

However, if we use a spherical gyroscope, with $I^{ls} \propto \delta_l{}^s$, then the tidal torque in Eq. (7.131) vanishes, and the Newtonian equation of motion becomes $dS^n/dt = 0$.

This Newtonian equation remains valid in curved spacetime, in a reference frame in free fall along a geodesic (Fermi coordinates).

The Newtonian time coordinate *t* must now be interpreted as the proper time measured along the geodesic,

$$\frac{dS^n}{d\tau} = 0 \tag{7.132}$$

Accordingly, in the freely falling reference frame, the spin of our spherical gyroscope remains constant in magnitude and direction.

This means the spin moves by parallel transport.

A warning: The agreement between parallel transport and gyroscope transport holds only along a geodesic, that is, the gyroscope has to be in free fall.

If an extra nongravitational force pushes on the gyroscope and makes it move along some nongeodesic worldline, then the equation of motion for the spin does not reduce to the simple form (7.132).

Along a nongeodesic worldline, the equation of motion of the spin is given by Fermi-Walker transport (see Section 6.6), not parallel transport.

Along a nongeodesic worldline in flat spacetime (that is, a trajectory with acceleration), this gives rise to a Thomas precession.

In general coordinates, the equation for the parallel transport of the spin vector S^n involves the Christoffel symbols:

$$\frac{dS^\mu}{d\tau} = -\Gamma^\mu{}_{\alpha\beta} S^\alpha \frac{dx^\alpha}{d\tau} \frac{dx^\beta}{d\tau} \tag{7.133}$$

From this equation we can calculate how the spin of our gyroscope changes direction while the gyroscope moves along some free-fall trajectory.

To be specific, let us assume that the gyroscope is in a circular orbit of radius r_0 around the Earth.

Experimentally, we measure the change of the direction of spin relative to the fixed stars; this is equivalent to finding the change with respect to a fixed set of coordinates at infinity.

Rectangular coordinates are more convenient for the calculation of the change of direction than polar coordinates, since a change of the rectangular components of the spin vector can be immediately attributed to the curvature of spacetime, whereas a change of the polar components contains contributions from both the curvature of the coordinates and the curvature of spacetime.

We will therefore use the "isotropic" rectangular coordinates x, y, z introduced in Section 7.4.

Since it will suffice to calculate the change of direction to the lowest order in GM, we will not need the exact Schwarzschild solution (7.80), but only the approximate solution (7.77),

$$ds^2 = \left(1 - \frac{2GM}{r}\right) dt^2 - \left(1 + \frac{2GM}{r}\right)(dx^2 + dy^2 + dz^2) \tag{7.134}$$

We assume that the orbit is in the x-y plane (see Fig. 7.6).

For convenience we will evaluate (7.133) at one point of the orbit, say, the point $x = r_0, y = 0, z = 0$.

All points on a circular orbit are equivalent, and from the rate of change of the spin at one point we can deduce the rate of change at any other point.

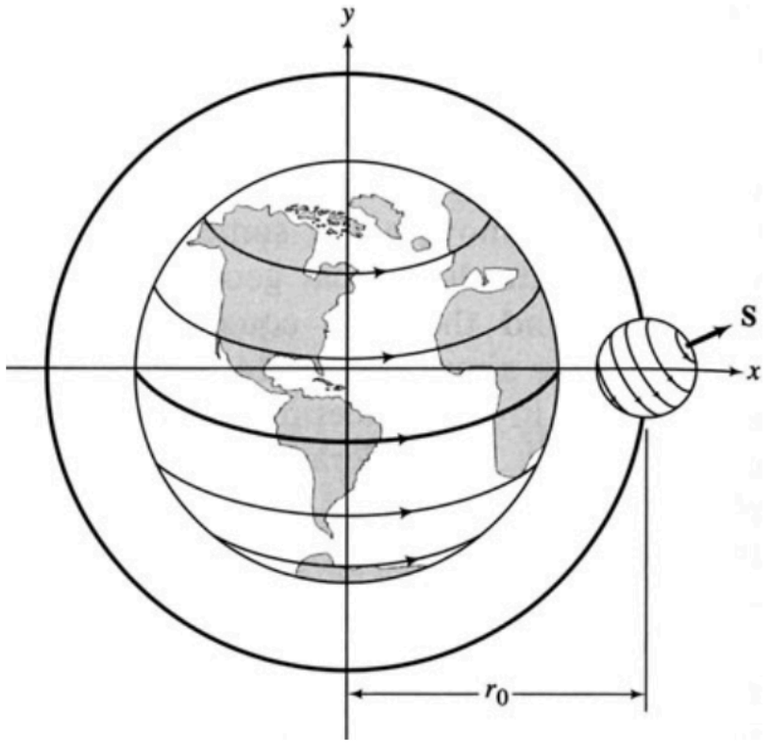


Fig. 7.6 Gyroscope orbiting the Earth in the *x*-*y* plane. At one instant, the gyroscope is at the point $x = r_0, y = 0, z = 0$.

At the chosen point, the Christoffel symbols are, to lowest order in GM,

$$\begin{aligned}
&\Gamma^0{}_{01} = \Gamma^0{}_{10} = GM/{r_0}^2 \\
&\Gamma^1{}_{00} = GM/{r_0}^2, \quad \Gamma^1{}_{11} = -GM/{r_0}^2, \quad \Gamma^1{}_{22} = GM/{r_0}^2, \quad \Gamma^1{}_{33} = GM/{r_0}^2 \\
&\Gamma^2{}_{12} = \Gamma^2{}_{21} = -GM/{r_0}^2 \\
&\Gamma^3{}_{13} = \Gamma^3{}_{31} = -GM/{r_0}^2
\end{aligned} \tag{7.135}$$

All the other Christoffel symbols are zero.

For the spatial components of the spin, Eq. (7.133) reads

$$\frac{dS^k}{d\tau} = -\Gamma^k{}_{\alpha\beta} S^\alpha \frac{dx^\alpha}{d\tau} \frac{dx^\beta}{d\tau} \tag{7.136}$$

To evaluate the right side, we need the four-velocity, which, at the chosen point, is $dx^\beta/d\tau \cong (1, 0, v, 0)$.

We also need the S^0 component of the spin.

To find this component, we note that in the rest frame of the gyroscope, $S'^0 = 0$ and $dx^\beta/d\tau = (1, 0, 0, 0)$, and therefore $g'_{\mu\nu} S'^\mu (dx'^\nu/d\tau) = 0$.

Consequently, this must also be true in our isotropic coordinates, so $g_{\mu\nu}S^\mu(dx^\nu/d\tau) = 0$, from which

$$S^0 = -\frac{1}{g_{00}}\left(S^1 g_{11}\frac{dx}{d\tau} + S^2 g_{22}\frac{dy}{d\tau} + S^3 g_{33}\frac{dz}{d\tau}\right) \cong S^2 v \tag{7.137}$$

We can now evaluate the right side of Eq. (7.136), and we obtain

$$\frac{dS^1}{d\tau} = -2\frac{GM}{r_0{}^2}S^2 v \tag{7.138}$$

$$\frac{dS^2}{d\tau} = \frac{GM}{r_0{}^2}S^1 v \tag{7.139}$$

$$\frac{dS^3}{d\tau} = 0 \tag{7.140}$$

These equations are valid only at the chosen point $x = r_0, y = 0, z = 0$.

But we can rewrite these equations in a form valid at any point around the orbit if we recognize that the three equations (7.138)–(7.140) can be combined into a single 3-D vector equation

$$\frac{d\mathbf{S}}{d\tau} = -2\mathbf{v}\cdot\mathbf{S}\nabla\Phi + \mathbf{v}\,\mathbf{S}\cdot\nabla\Phi \tag{7.141}$$

where $\Phi = -GM/r$ is the Newtonian potential.

Since both the velocity **v** and the gradient $\nabla\,\Phi$ vary with position around the orbit, the behavior of **S** given by Eq. (7.141) is rather complicated, with small periodic oscillations in both the magnitude and the direction of **S**.

However, we are not interested in such periodic wobbles, but only in the long-term, secular change in **S**.

We can calculate this secular change by taking the average of Eq. (7.141) over the orbit.

For this purpose, we express **v** and $\nabla\Phi$ as functions of time:

$$\mathbf{v} = -\hat{\mathbf{x}} v \sin\omega t + \hat{\mathbf{y}} v \cos\omega t \tag{7.142}$$

$$\nabla\Phi = \hat{\mathbf{x}} \frac{GM}{{r_0}^2} \cos\omega t + \hat{\mathbf{y}} \frac{GM}{{r_0}^2} \sin\omega t \tag{7.143}$$

If we insert these expressions into Eq. (7.141) and average over one period of the orbit while treating S^1 and S^2 on the right side as constants, we find

$$\left\langle \frac{d\mathbf{S}}{d\tau} \right\rangle = \frac{3}{2} \frac{GM}{{r_0}^2} v(-\hat{\mathbf{x}} S^2 + \hat{\mathbf{y}} S^1) \tag{7.144}$$

We can rewrite this result more neatly as follows:

$$\left\langle \frac{d\mathbf{S}}{d\tau} \right\rangle = \frac{3}{2} \frac{GM}{{r_0}^3} (\mathbf{r} \times \mathbf{v}) \times \mathbf{S} \tag{7.145}$$

From Eq. (7.145) we see that, on the average, **S** precesses about the axis of the orbit with an angular velocity

$$\Omega_G = \frac{3}{2} \frac{GM}{{r_0}^3} \mathbf{r} \times \mathbf{v} \tag{7.146}$$

This is the *geodetic precession*, or the *de Sitter precession*.

The geodetic precession has sometimes been described as analogous to the Thomas precession of special relativity. But this analogy is somewhat misleading: The Thomas precession in the flat spacetime of special relativity results from nongeodesic motion; that is, it results from accelerated motion brought about by the push of some force (without torque) acting on a gyroscope or on a spinning particle, such as an electron.

To find a true analog of the Thomas precession in curved spacetime, we would have to examine the behavior of a gyroscope when some extra, nongravitational force makes it deviate from geodesic motion.

The Thomas precession would then arise as an extra contribution to the precession, in excess of the contribution from parallel transport.

Although we derived Eq. (7.146) for a circular orbit, it can be shown that the result is valid in general.

Note that the time-average rate of change of the spin can be written in the concise form

$$\left\langle \frac{d\mathbf{S}}{d\tau} \right\rangle = \frac{3}{2} \frac{GM}{m r_0{}^3} \mathbf{L} \times \mathbf{S} \tag{7.147}$$

where $\mathbf{L} = m\mathbf{r} \times \mathbf{v}$ is the orbital angular momentum of the gyroscope.

In Newtonian mechanics, we would interpret the right side of Eq. (7.147) as a torque due to spin-orbit coupling with a potential energy (in cgs units)

$$U = \frac{3}{2} \frac{GM}{m c^2 r_0{}^3} \mathbf{L} \cdot \mathbf{S} \tag{7.148}$$

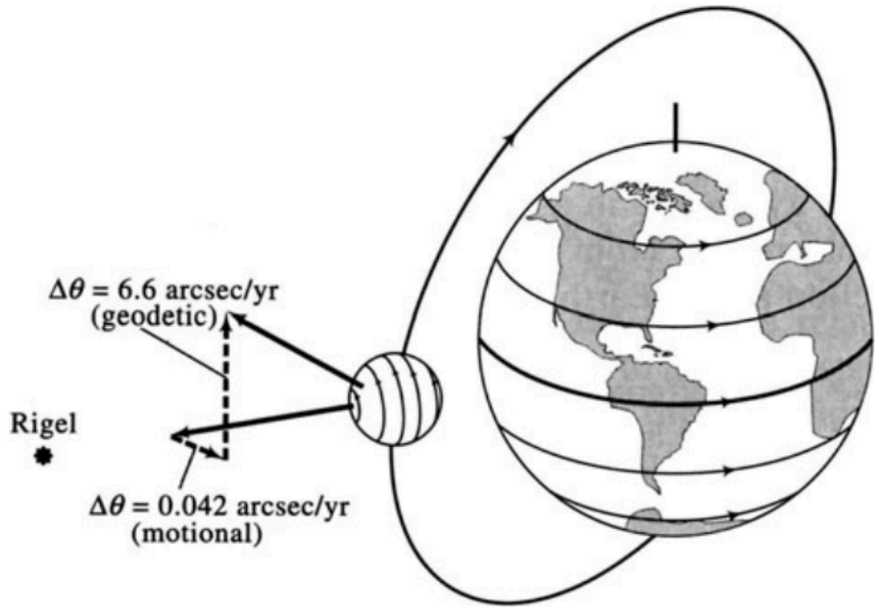
For a gyroscope in orbit 650 km above the Earth, with the spin in the plane of the orbit (see Fig. 7.7), the geodetic precession (7.146) amounts to

$$\frac{3}{2} \frac{GM}{c^2 r_0{}^2} \sqrt{\frac{GM}{r_0}} \cong 6.6 \text{ arcsec per year} \tag{7.149}$$

Note that the calculation of the rate of change of spin in Eqs. (7.135)–(7.145) did not use the exact Schwarzschild solution and could have been carried out in Chapter 4 in the context of the linear theory.

But at that stage we did not have the necessary mathematical tools to deal with parallel transport.

Fig. 7.7 A gyroscope in a 650-km polar orbit. The spin of the gyroscope is (initially) in the plane of the orbit. The geodetic precession causes the spin to swing through 6.6 arcsec per year, in the direction of the orbital motion. In addition, the Lense-Thirring precession causes the spin to swing through 0.042 arcsec per year, in the direction perpendicular to the orbit.



Also note that a spinning particle with the potential (7.148) will suffer not only a spin precession but also a translational acceleration caused by a spin-orbit force.

This suggests that a spinning body in a gravitational field will deviate from geodesic motion.

This conjecture is confirmed by a careful examination of the equation for the translational equation of motion for spinning bodies in curved spacetime.

Although the deviation from geodesic motion is quite small and practically undetectable, it is of some theoretical importance in that it shows that Galileo's principle of equivalence fails for spinning particles.

For example, electrons, protons, neutrons, and so on, do not move exactly on geodesics in a gravitational field.

For a gyroscope in orbit around the Earth, there is an additional precession effect caused by the rotation of the Earth.

The gravitational field of the Earth is not exactly a Schwarzschild field; rather, it is the field of a rotating mass [see Eq. (4.93)].

If we take into account the extra "gravitomagnetic" terms in the metric tensor associated with the rotation of the Earth, we find an extra precession, called the *Lense-Thirring* precession already mentioned in Section 4.7 [see Eq. (4.103)],

$$\Omega_{LT} = \frac{G}{c^2} \frac{3\mathbf{n}\,\mathbf{n}\cdot\mathbf{S}_E - \mathbf{S}_E}{{r_0}^3} \tag{7.150}$$

where $\mathbf{S}_E$ is the spin angular momentum of the Earth, and $\mathbf{n}$ is a unit vector in the radial direction.

In Newtonian terms, the precession (7.150) may be thought of as due to a spin-spin coupling between the spins of the gyroscope and the Earth.

For a gyroscope in a polar orbit at 650 km, the average precession rate that results from Eq. (7.150) is $\cong$ 0.042 arcsec per year, in the same direction as the Earth's rotation.

In Section 4.7 we obtained this precession by analogy with the precession of a magnetic moment in a magnetic field.

But the result can also be obtained by evaluating the change of the spin vector during parallel transport in the field of a rotating mass.

When de Sitter first considered the geodetic precession effect, there were no prospects of a test with gyroscopes.

Instead, de Sitter proposed that the effect could be detected by a precession that the gravitational field of the Sun produces in the orbit of the Moon around the Earth.

We can regard the Earth-Moon system as a gyroscope in free fall in the gravitational field of the Sun.

The axis of spin of this gyroscope – that is, the axis of the geocentric orbit of the Moon – then suffers a geodetic precession according to Eq. (7.150).

The predicted precession rate amounts to 0.0192 arcsec per year.

Although this precession is too small to be detected by conventional astronomical observations, it can be detected by lunar laser ranging, which permits us to determine any change in the direction of the Moon's perigee with extreme precision.

Analysis of the accumulated lunar laser-ranging data confirms the predicted value of the geodetic precession to within the observational error, about 0.7%.

The geodetic precession effect has also been detected in the Hulse-Taylor binary pulsar.

This spinning neutron star is effectively a gyroscope, and its axis of spin is expected to precess as it orbits in the gravitational field of its companion star.

In consequence of the precession, the pulsar radio beam gradually dips in relation to the Earth.

Monitoring of the pulse strength received at the Earth over several years has revealed a gradual change in the intensity profile of the pulses, consistent with the expected precession rate of 1.21° per year.

However, since we do not know the detailed angular shape of the beam nor the possible intrinsic changes of this shape with time, we cannot make any quantitative comparison between theory and observation.

The development of an experiment to measure both the geodetic and the Lense-Thirring precession with gyroscopes was initiated in the 1960s, and the experiment was finally performed in 2004–05, followed by several years of data analysis.

This Stanford Relativity Gyroscope Experiment, also called Gravity Probe B, used gyroscopes installed on a satellite launched into a polar orbit from the Space Shuttle.

The rotors of the gyroscopes were quartz spheres, 3.8 cm in diameter, suspended in their housings by electrostatic forces supplied by capacitor plates (see Fig. 7.8).

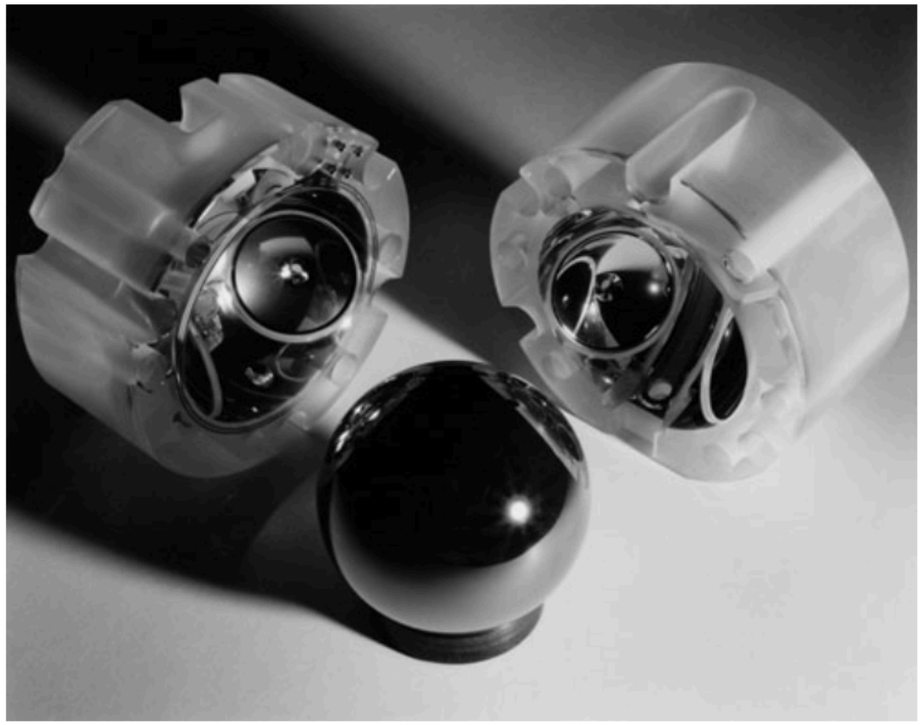


Fig. 7.8 Quartz sphere used as rotor for the gyroscope and its housing. (The Gravity Probe B Image and Media Archive, Stanford University.)

To avoid any extra precession produced by the tidal torque (7.131), the quartz spheres had to be accurately round and homogeneous; deviations from spherical shape larger than 10^{-6} cm would have been unacceptable.

The quartz spheres were coated with a thin layer of superconducting niobium.

When the spheres are spinning, the rotating superconducting niobium generates a magnetic moment (London moment) parallel to the spin.

Changes in the direction of the magnetic moment were detected by the current it induces in a stationary pickup loop surrounding the sphere; this induced current served as an indicator of the change in the direction of spin.

The gyroscopes were surrounded by a large dewar filled with superfluid liquid helium, to keep the temperature of the instrument package at 1.8 K (see Fig. 7.9).

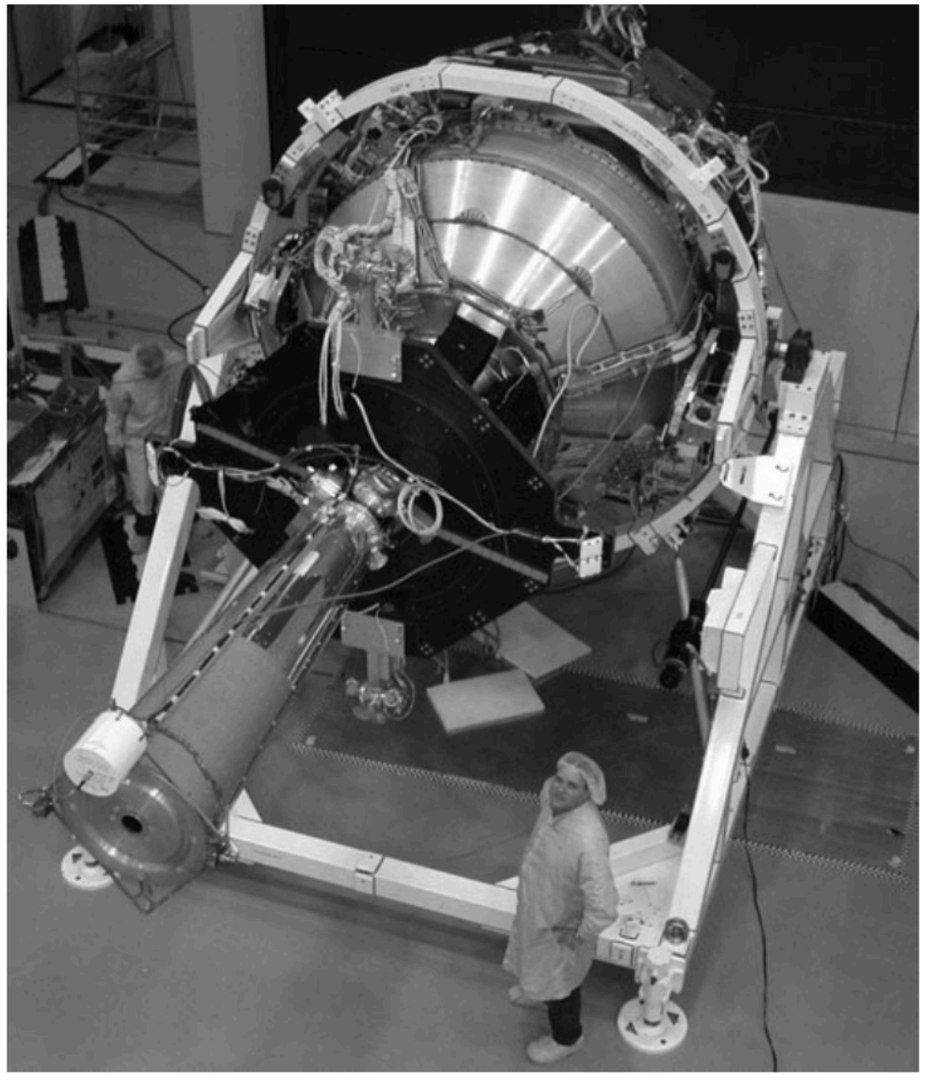


Fig. 7.9 Instrument package and dewar for the Stanford Relativity Gyroscope Experiment. The long protruding tube at the front is a shield for the guidance telescope. (The Gravity Probe B Image and Media Archive, Stanford University.)

The longitudinal axis of the satellite was held in a fixed orientation by an attitude-control system consisting of thrusters governed by a telescope locked on a guide star (Rigel) located in the plane of the orbit.

The precession of the gyroscopes relative to the body of the satellite is therefore equivalent to their precession relative to an inertial reference frame at infinity.

The satellite was also equipped with a drag-free control system consisting of thrusters that compensate for residual atmospheric friction.

The thrusters were governed by a proof mass floating freely in a chamber within the satellite, where it was shielded from atmospheric friction.

Sensors detected any incipient deceleration of the body of the satellite relative to this free-floating mass and sent suitable signals to the thrusters so as to keep the satellite at rest relative to the free-floating mass.

Unfortunately, the experimenters underestimated a difficulty with stray patches of electrostatic charge on the gyroscopes, which contributed unwanted torques and caused drifts of the direction of spin.

Extensive mathematically modeling of the charge distributions and corrections to the observed data were required, and this introduced uncertainties in the final results, in excess of the expected uncertainties.

The uncertainty in the final measured results for the geodetic precession was ±0.3%, and for the Lense-Thirring precession ±20%.

Within these uncertainties, the experiment confirmed the predictions of general relativity.

The Gravity Probe B experiment attained roughly the same precision as previous measurements by laser ranging [±0.7% for the geodetic precession of Moon and ±10% for the Lense-Thirring precession of LAGEOS orbits].

There was, however, an important distinction between the LAGEOS and the GPB results: the former probed the spin-orbit Lense-Thirring interaction, whereas the latter probed the spin-spin interaction, which required much more challenging experimental techniques.

8 - Black Holes and Gravitational Collapse

Lasciate ogni speranza voi ch'entrate.

Dante Alighieri, *The Inferno*

The dimensionless quantity GM/rc^2 may be regarded as a measure of the strength of the gravitational field.

This quantity enters into the formulas for light deflection, time delay, gravitational redshift, perihelion precession, and so on.

The small magnitude of the relativistic gravitational effects in the Solar System is related to the small magnitude of this quantity; even at the solar surface, GM/rc^2 is only 2×10^{-6}.

Large relativistic effects are found in the gravitational field in the neighborhood of an extremely compact mass, where GM/rc^2 can attain values of the order of magnitude of 1.

For example, near such a compact mass, at a radius $r = 3GM/c^2$, the deflection of light in the Schwarzschild field becomes so large that a light signal will move in a closed circular orbit around the central mass.

The relativistic effects become spectacular when $r = 2GM/c^2$.

The gravitational fields at this radius are so strong that nothing can escape from their grip.

Light signals, particles, and even spacecraft with the most powerful engines are inexorably pulled inward.

The existence of a "no-escape" radius for light is suggested by the following naive calculation based on Newton's gravitational theory: At a radius r, the escape velocity for a particle projected upward is $v = \sqrt{2GM}/r$.

Because this expression is independent of the mass of the particle, we may hope that it also applies to the case of light.

For $v = c$, we then obtain a no-escape radius $r = GM/c^2$.

This result was obtained long ago by Michell and by Laplace, who speculated that a sufficiently massive and compact star might appear dark and black.

By dumb luck, this result agrees exactly with the result of a careful relativistic calculation (see the later discussion).

Although the naive calculation gives the correct value for the no-escape radius, it gives the wrong picture of what happens.

Light signals emitted at or inside the no-escape radius do not first rise, then stop, and then fall back.

Rather, they fall immediately and never begin to move in the outward direction.

The same is true of particles, even particles with propulsive devices (spacecraft).

If light cannot escape, no particle can, because the speed of any particle must remain below the speed of light, and the worldline must remain within the light cone.

A bounded region of space into which signals can enter, but from which no signal can ever emerge, is called a black hole.

8.1 Singularities and pseudosingularities

For a spherically symmetric and stationary gravitational field, we found that the space- time geometry is the Schwarzschild geometry with the spacetime interval

$$ds^2 = \left(1 - \frac{2GM}{r}\right) dt^2 - \frac{dr^2}{1 - 2GM/r} - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \qquad (8.1)$$

By inspection, we see that this spacetime interval develops a "singularity" as $r \to 2GM$.

At this radius,

$$g_{00} = 1 - \frac{2GM}{r} \to 0 \tag{8.2}$$

and

$$g_{11} = -\frac{1}{1 - 2GM/r} \to -\infty \tag{8.3}$$

In cgs units, the critical radius is

$$r_S = \frac{2GM}{c^2} \tag{8.4}$$

This is called the *Schwarzschild radius*, or the *gravitational radius* of the mass *M*.

For a mass equal to that of the Sun, $M = M_\odot = 2\times10^{33}$ g, this radius is $r_S = 3.0$ km.

If a body has an actual radius larger than r_S, then the singularities (8.2) and (8.3) are of no concern, because the solution (8.1) is applicable only in the region exterior to the body.

If a body with $M \cong M_\odot$ is to have a radius smaller than its Schwarzschild radius r_S, it must have a density of $\approx 10^{16}$ g/cm^3, which is even larger than the density of a neutron star.

In fact, as we will see, a body that has collapsed to a radius smaller than r_S is unable to come to equilibrium, and it will continue to collapse – the gravitational forces are so strong that nothing can resist them.

The critical density at which a body has a size smaller than its Schwarzschild radius r_S decreases with mass.

For example, suppose that the nucleus of a galaxy consists of 10^{11} stars, each of a mass $M_\odot$, uniformly distributed over a spherical volume.

In this case the critical density is only 10^{-6} g/cm^3; at this density, the typical distance between adjacent stars in the galactic nucleus is about the same as the Earth-Sun distance.

For the discussion of the general features of the Schwarzschild solution, we will assume that the body has collapsed completely, so the mass density is zero at all points (except perhaps at $r = 0$).

This means we treat the Schwarzschild solution as an exact vacuum solution of Einstein's equations.

The surface $r = r_S$ is a surface of infinite redshift.

A clock placed at rest near $r = r_S$ shows a proper time

$$d\tau = \sqrt{1 - 2GM/r}\,dt \tag{8.5}$$

which approaches zero as $r \to r_S$, so the clock runs infinitely slowly compared with a clock at a larger distance.

Thus, if an astronaut at rest near $r = r_S$ sends out light pulses, or signal rockets, with a time interval of 1 s (as shown by his clock) between one pulse and the next, then an observer at large distance will receive these pulses with a time interval much larger than 1 s (as shown by her clock) between successive pulses.

Equation (8.5) gives an infinite time dilation for a clock at rest at $r = r_S$.

Actually, a clock cannot remain at rest on this surface.

The vanishing of $d\tau$ is characteristic of the worldline of a light signal, and hence only a light signal, aimed in the radial direction, can remain at rest at r_S.

The infinite-redshift surface $r = r_S$ is sometimes called the static limit, because material particles cannot remain at rest on it.

If r is in the range $r_S > r > 0$, then the Schwarzschild solution is free of singularities.

However, in this region

$$g_{00} = 1 - r_S/r < 0 \quad \text{and} \quad g_{11} = -1/(1 - r_S/r) > 0 \tag{8.6}$$

Thus, the signs of g_{00} and g_{11} are the opposite of what is normal.

Acccordingly, in the region $r < r_S$, t is a spacelike coordinate, and r is a timelike coordinate.

Because the metric tensor is a function of r, and because r now measures the progress of time, it follows that when $r < r_S$, the metric tensor is actually time dependent.

Stated more precisely, in the exterior region $r > r_S$, the static character of the metric tensor is a consequence of our particular choice of coordinates; if we were to use new coordinates x'^{μ} that deform in time with respect to the t, r, θ, ϕ coordinates, then the metric tensor as seen in these new coordinates would of course be time dependent.

However, in the interior region $r < r_S$, the time dependence of the metric tensor is not a consequence of the choice of coordinates.

This time dependence is unavoidable, because regardless of how the new coordinates x'^{μ} are defined in terms of the old t, r, θ, ϕ coordinates, an advance of time ($ds^2 > 0$) is impossible unless $dr^2 > 0$; hence the advance of time necessarily entails a change in r, and hence a change in the metric tensor.

It is important to recognize that the Schwarzschild "singularity" at $r = r_S$ is not a physical singularity.

The "singularity" in Eqs. (8.2) and (8.3) is spurious – it is a pseudosingularity, or a coordinate singularity.

It arises from an inappropriate choice of coordinates and can be eliminated by a change of coordinates.

An astronaut falling toward and crossing the surface $r = r_S$ will not feel anything unusual.

In his immediate vicinity physics will go on in the way it always does in a freely falling reference frame.

Of course, there will be tidal forces, and these tidal forces will grow stronger and stronger as the astronaut falls deeper and deeper toward the center.

However, the tidal forces remain finite at $r = r_S$; only at $r = 0$ do the tidal forces become infinite and indicate the presence of a real physical singularity.

Mathematically, the absence of any genuine singularity at $r = r_S$ can be seen from an examination of the Riemann curvature tensor.

It turns out that, even though in Schwarzschild coordinates g_{00} and g_{11} misbehave, all components of $R^{\alpha}{}_{\beta\mu\nu}$ are finite at $r = r_S$ when calculated in local geodesic coordinates.

For example, consider the component $R^{0}{}_{101}$.

This component is given by [see Eq. (6.101)), with $\dot{L} = 0, \ddot{L} = 0$]:

$$R^{0}{}_{101} = -\frac{1}{2}N'' + \frac{1}{4}L'N' - \frac{1}{4}(N')^2 = \frac{r_S}{r^3}\frac{1}{1 - r_S/r}$$

At $r = r_S$, this function is singular.

However, this singularity is spurious, like the singularity in Eq. (8.3).

To recognize this, let $x_\mu = (t_0, r_0, \pi/2, \phi_0)$ be a given spacetime point in the equatorial plane, and introduce geodesic coordinates x'^{μ} at this point by means of the transformation

$$x'^0 = (t - t_0)\sqrt{1 - r_S/r_0} + \cdots, \quad x'^1 = \frac{r - r_0}{\sqrt{1 - r_S/r_0}} + \cdots$$
$$x'^2 = r_0(\theta - \pi/2) + \cdots, \qquad x'^3 = r_0(\phi - \phi_0) + \ldots, \tag{8.7}$$

The dots in these equations stand for quadratic terms, such as those appearing in Eq. (6.35).

These quadratic terms are, of course, crucial to achieve the geodesic coordinate condition $\Gamma'^{\alpha}{}_{\mu\nu} = 0$; but we are now interested only in the transformation law for the Riemann tensor, and this law does not depend on the quadratic terms.

The transformation coefficients at $x'^{\mu} = 0$ are

$$\frac{\partial x^0}{\partial x'^0} = \frac{1}{\sqrt{1 - r_S/r_0}}, \quad \frac{\partial x^1}{\partial x'^1} = \sqrt{1 - r_S/r_0}$$
$$\frac{\partial x^2}{\partial x'^2} = \frac{1}{r_0}, \qquad \frac{\partial x^3}{\partial x'^3} = \frac{1}{r_0} \tag{8.8}$$

All other coefficients are zero.

The coordinate transformations given by Eq. (8.7) are singular if r coincides with r_S.

The elimination of the singularity of the Schwarzschild solution hinges on the use of a singular coordinate transformation.

We will of course adopt the view that the coordinates that go bad at $r = r_S$ are the Schwarzschild coordinates x^{μ}, specifically, the coordinates t and r.

The geodesic coordinates x'^{μ} can be given the physical interpretation of coordinates of a reference frame in free fall (see Section 6.3), and they must therefore necessarily be regarded as good coordinates.

For the transformed Riemann tensor we find

$$R'^{0}{}_{101} = \frac{\partial x'^{0}}{\partial x^{\alpha}} \frac{\partial x^{\beta}}{\partial x'^{1}} \frac{\partial x^{\mu}}{\partial x'^{0}} \frac{\partial x^{\nu}}{\partial x'^{1}} R^{\alpha}{}_{\beta\mu\nu} = \frac{\partial x'^{0}}{\partial x^{0}} \frac{\partial x^{1}}{\partial x'^{1}} \frac{\partial x^{0}}{\partial x'^{0}} \frac{\partial x^{1}}{\partial x'^{1}} R^{0}{}_{101}$$
$$= \left(1 - \frac{r_S}{r_0}\right) \frac{r_S}{{r_0}^3} \frac{1}{1 - r_S/r_0} = \frac{r_S}{{r_0}^3} \qquad (8.9)$$

If we now let r approach r_S, the function given by Eq. (8.9) remains finite.

This shows that in the x'^{μ} coordinates, the $R'^{0}{}_{101}$ component of the Riemann tensor is free of singularities.

It is easy to check that all other components of the Riemann tensor are also free of singularities.

The tidal force therefore remains finite at $r = r_S$.

A more convenient way to recognize that in local geodesic coordinates all the components of the Riemann tensor remain finite is by evaluation of the scalar quantity $R^{\alpha}{}_{\beta\mu\nu} R_{\alpha}{}^{\beta\mu\nu}$.

With the explicit values for $R^{\alpha}{}_{\beta\mu\nu}$ in Schwarzschild coordinates, it is easy to verify that this scalar quantity is finite, $R^{\alpha}{}_{\beta\mu\nu} R_{\alpha}{}^{\beta\mu\nu} = 48(GM)^2/{r_0}^6$.

Since in local geodesic coordinates (with $g'_{\mu\nu} = \eta_{\mu\nu}$), this scalar is the sum of the squares of all the components of $R'^{\alpha}{}_{\beta\mu\nu}$, it is then obvious that all these components must be finite.

Note that at $r = r_0 = 0$, the components of $R'^{\alpha}{}_{\beta\mu\nu}$ diverge, as can be seen from the expression (8.9).

Hence, at the "center," the tidal forces diverge and we have a true singularity.

Near this point, r is a timelike coordinate and t a spacelike coordinate, so this singularity happens at a given instant of time ($r = 0$) in all of space (at all values of t).

Thus the time-dependent, dynamical geometry in the interior region evolves into a singularity and comes to an end.

But the geometry on the outside remains static forever.

In some respects the Schwarzschild pseudosingularity at $r = r_S$ is similar to the pseudosingularity found when the metric tensor of flat spacetime is expressed in Rindler coordinates (see Fig. 8.1).

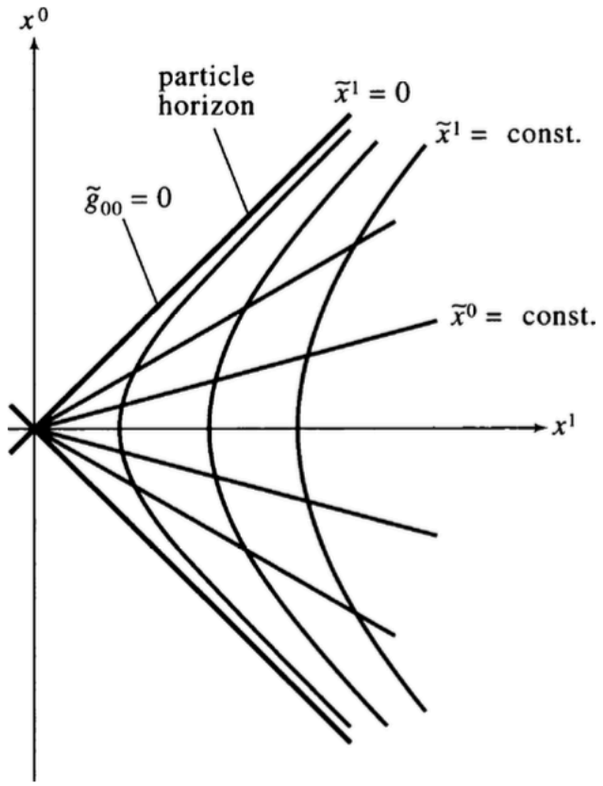


Fig. 8.1 Grid of Rindler coordinates $\tilde{x}^0$ and $\tilde{x}^1$ in flat spacetime. A particle at rest in Rindler coordinates at $\tilde{x}^1 =$ constant has accelerated ("hyperbolic") motion relative to the inertial coordinates x^0, x^1. In the limiting case $\tilde{x}^1 \to 0$, the particle has the worldline of a light signal, with $ds^2 = 0$.

As we saw in Section 6.8, the Rindler coordinates $\tilde{x}^0$ and $\tilde{x}^1$ are related to the inertial rectangular coordinates x^0 and x^1 by

$$\tilde{x}^0 = b \tanh^{-1} x^0/x^1, \quad \tilde{x}^1 = \sqrt{(x^1)^2 - (x^0)^2} \tag{8.10}$$

In terms of these coordinates $\tilde{x}^0$ and $\tilde{x}^1$, the spacetime interval of flat spacetime, $ds^2 = (dx^0)^2 - (dx^1)^2$, takes the form [see Eq. (6.188)]

$$ds^2 = \frac{(\tilde{x}^1)^2}{b^2}(d\tilde{x}^0)^2 - (d\tilde{x}^1)^2 \tag{8.11}$$

For $\tilde{x}^1 \to 0$, the $\tilde{g}_{00}$ component of the Rindler metric tensor develops a singularity similar to the singularity in Eq. (8.2),

$$\tilde{g}_{00} = \frac{(\tilde{x}^1)^2}{b^2} \to 0 \tag{8.12}$$

This means that a clock located at $\tilde{x}^1 = 0$ has an infinite redshift relative to a clock located elsewhere.

We can readily understand this infinite redshift by noting that the worldline $\tilde{x}^1 = 0$ is actually the worldline of a light signal; the infinite redshift is simply the familiar special-relativistic time dilation for a clock moving at a speed approaching the speed of light.

Actually, no clock or particle can remain at rest at $\tilde{x}^1 = 0$, because this would require that it move at the speed of light relative to the inertial coordinates x^0, x^1.

This is analogous to what happens in Schwarzschild spacetime, where no particle can remain at rest at $r = r_S$.

The singularity (8.12) in Rindler coordinates at $\tilde{x}^1 = 0$ arises from an inappropriate choice of coordinates.

It is a pseudosingularity – flat spacetime has no physical singularity at $\tilde{x}^1 = 0$ or anywhere else.

The pseudosingularity can be eliminated by a coordinate transformation back to the inertial coordinates.

The example of Rindler coordinates shows very clearly that the presence of surfaces of infinite redshift depends on the choice of coordinates.

8.2 The black hole and its horizon

Although the region $r < r_S$ has no unusual properties of a local kind (except at $r = 0$, where there is a genuine singularity), it does have some very unusual properties of a global kind.

As we will see from a careful analysis of the spacetime geometry, the region $r < r_S$ is a black hole.

By this is meant that no signal of any kind can emerge from the region $r < r_S$ and reach the region $r > r_S$.

The surface $r = r_S$, regarded as a surface in spacetime (see Figs. 8.2a, b), is the boundary between the region of spacetime that is observable by outside observers, and the region that will never be observable by these observers.

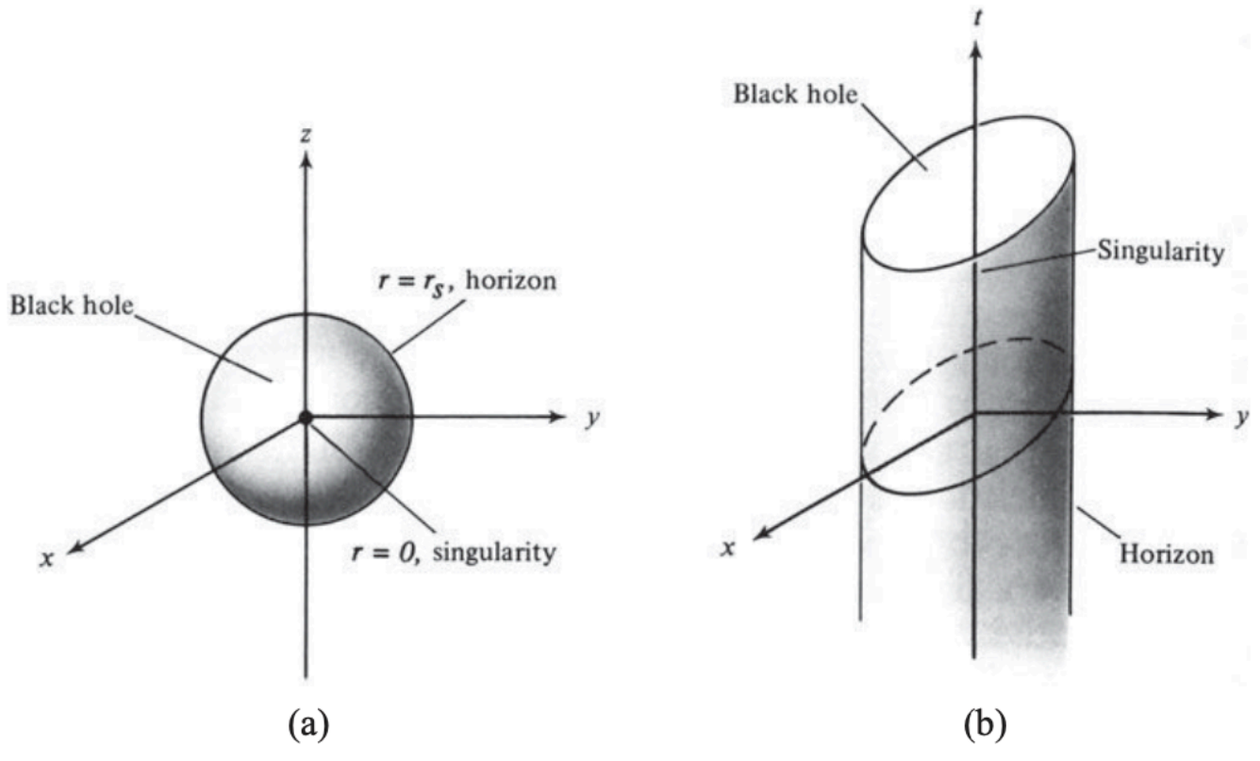


Fig. 8.2 (a) The Schwarzschild black hole and its event horizon displayed in x, y, z coordinates. The event horizon is a sphere of radius r_S. Note that r_S is merely the value of a radial coordinate, not a measured distance. (b) The Schwarzschild black hole and its event horizon displayed in x, y, t coordinates. The event horizon is a cylinder of radius r.

This boundary of the black hole is called the *event horizon*.

It is the place beyond which we cannot see, reminiscent of the horizon on a large lake or on the ocean.

The surface $r = r_S$ acts as a "one-way membrane," through which signals can be sent in, but not out.

This is a global (or nonlocal) property because to test it, we must examine the propagation of light signals and other signals and check what happens to them in the long run.

Note that in flat spacetime in Rindler coordinates, there is also some kind of event horizon.

From the discussion in the preceding section, we know that observers at rest in Rindler coordinates (and permanently accelerated relative to inertial coordinates) perceive a surface of infinite redshift at $\tilde{x}^1 = 0$ in Fig. 8.1.

For these observers, this surface of infinite redshift is also a horizon, since no signal from the region above $\tilde{x}^1 = 0$ can ever reach the region below $\tilde{x}^1 = 0$ (the worldlines of light signals starting anywhere in the region above $\tilde{x}^1 = 0$ are parallel to the boundary line $\tilde{x}^1 = 0$, and they never intersect any of the worldlines $\tilde{x}^1 =$ constant).

But this Rindler horizon is perceived only by a special class of observers, at rest in Rindler coordinates.

The worldline of an observer at rest in inertial coordinates crosses the Rindler horizon, and such an observer can see the region of spacetime beyond.

Thus, the Rindler horizon is an artifact resulting from the accelerated motion of the observers, whereas the Schwarzschild horizon must be thought of as an attribute of the Schwarzschild geometry, since it is perceived by all observers.

This does not mean that at the event horizon the Schwarzschild geometry possesses any kind of distinctive local property, such as a local singularity.

But it does mean that all observers agree on the existence and location of the surfaces across which two-way communication is impossible.

Event horizons refer to global rather than local properties of spacetime, but that does not make event horizons any less real.

Although the surface of infinite redshift and the horizon coincide in the case of the Schwarzschild solution, there exist other solutions of Einstein's equations for which these surfaces are distinct (for example, the Kerr solution of Section 8.4).

If the redshift, for static observers, is infinite on some given surface, this does not by itself necessarily prevent communication across that surface.

It could happen that with an emitter located inside the infinite-redshift surface, the redshift, relative to an observer at large distance, is again reduced to a finite value.

Also, we must keep in mind that the total redshift of a signal depends not only on the location of the emitter in the gravitational field, but also on the velocity of the emitter.

For an emitter in motion, in some suitable direction, the Doppler shift may compensate for part or all of the gravitational redshift.

Thus, an "infinite-redshift surface" is defined only relative to a special class of emitters and receptors.

By contrast, an event horizon is an intrinsic feature of spacetime that does not directly depend on the state of motion of the observers.

All external observers will agree that the surface $r = r_S$ in the Schwarzschild spacetime is an event horizon (although some observers, such as observers with a permanent acceleration, may also perceive additional Rindler horizons).

To understand how signals are cut off at the Schwarzschild horizon $r = r_S$, consider a light signal propagating in the radial direction.

The velocity of this light signal, with respect to t, r coordinates, can be calculated from Eq. (8.1) by taking $ds^2 = 0$:

$$\frac{dr}{dt} = \pm\left(1 - \frac{r_S}{r}\right) \tag{8.13}$$

When the signal approaches $r = r_S$, this coordinate velocity tends to zero.

Figure 8.3 shows the forward light cones obtained from Eq. (8.13) at different values of r.

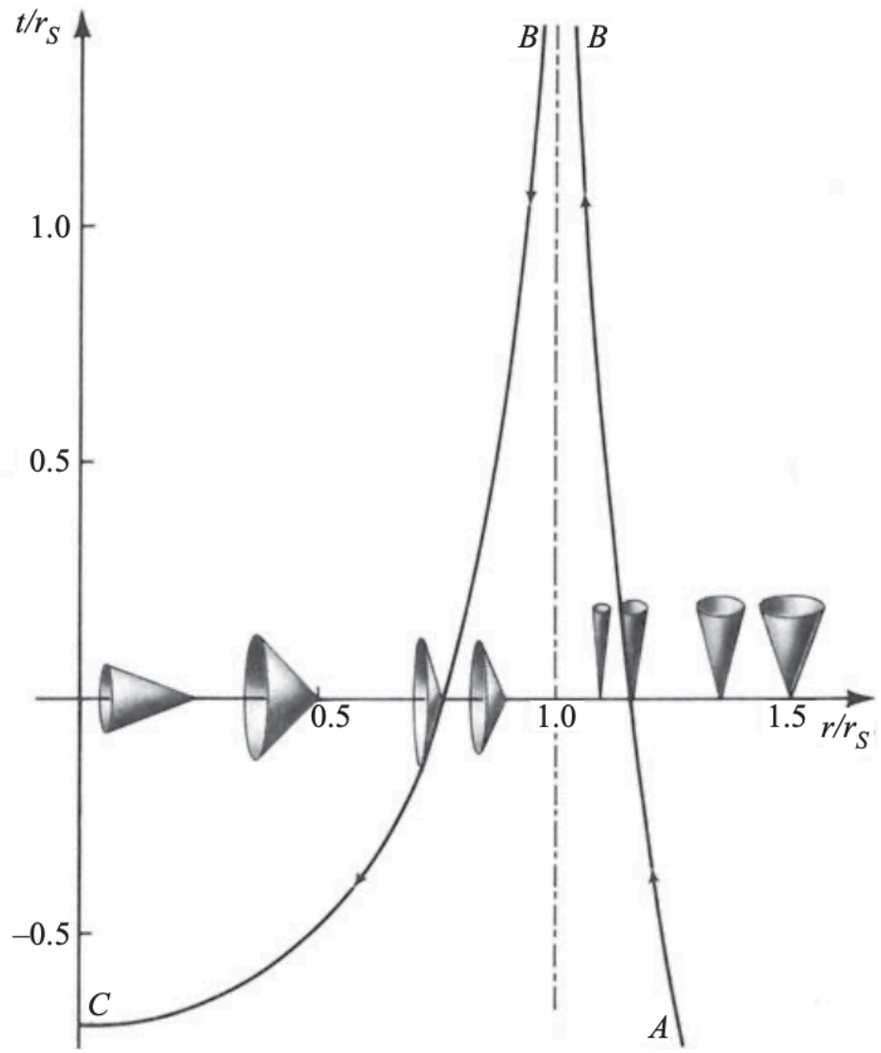


Fig. 8.3 The forward light cones near and inside a black hole. As $r \to \infty$, the light cone assumes its usual shape and direction; that is, $dr/dt = \pm 1$. The curve *ABBC* is the worldline of an ingoing light signal.

In the exterior of the black hole ($r > r_S$), the axis of the light cones is parallel to the *t*-axis.

In the interior of the black hole ($r < r_S$), the axis of the light cones is parallel to the *r*-axis.

The strange orientation of the light cones in the interior region is a simple consequence of reversal of the character of the coordinates in this region: *r* is a timelike coordinate and *t* is a spacelike coordinate.

In the interior region, the quantity dt/dr gives what we would normally call the "velocity," that is, the ratio of spacelike increment to timelike increment.

The existence of an event horizon at $r = r_S$ is obvious from inspection of the light cones in Fig. 8.3.

Any kind of signal must necessarily travel in a spacetime direction that lies within a light cone.

Since the light cones in the black-hole region are oriented toward $r = 0$, any signal in this region is unavoidably pulled toward decreasing values of *r* and can never leave the black hole.

Note that the light cones are tangent to the surface $r = r_S$ (indicated by the dashed line in Fig. 8.3); this means that, viewed in spacetime, the horizon is a null surface.

This is a general property of event horizons, since a light signal that starts exactly on a horizon and is aimed in the outward direction is sandwiched between those signals just outside of the horizon (which escape outward) and those just inside the horizon (which fall inward); the light signal therefore propagates neither out nor in – it hovers in place forever, and thereby indicates that the place at which it is hovering is a null surface.

Such a hovering light signal on the horizon of a black hole should not be confused with a signal in a circular orbit around the black hole; the hovering signal has no circular motion, and it is completely stationary.

Although signals cannot emerge from the black hole, they can enter it freely.

The curve in Fig. 8.3 is the worldline of a light signal that travels inward.

This curve is obtained by integrating Eq. (8.13).

The signal follows the worldline AB to $t = \infty$, and then it follows the worldline BC to $r = 0$.

As measured in t, r coordinates, the signal velocity tends to zero as $r \to r_S$, and the signal takes an infinite t time to reach $r = r_S$.

Hence from the point of view of an observer at infinity, whose clocks indicate t-time, the signal never reaches the horizon.

To make this concrete, suppose that the black hole is surrounded by some dust that scatters a small part of the light signal, so the position of the signal becomes visible to an observer at infinity.

Then this observer will see the lighted spot asymptotically approach the Schwarzschild radius without ever reaching it.

This is, in part, a manifestation of the gravitational time-dilation effect.

From the point of view of an astronaut in free fall in the vicinity of $r = r_S$, no such slowing down occurs; the signal always has the speed of light relative to him; the signal and the astronaut both cross into the black hole in a finite proper time.

If the observer at infinity keeps in touch with the astronaut by audio and video radio links, she finds that the motion and metabolic rate of this astronaut are slowed down in much the same way as the motion and oscillation rate of the light signal.

The astronaut appears to go into “slow motion,” and his movements freeze asymptotically.

The last syllable he sends through his radio before crossing into the black hole is drawn out to infinite length when received by the outside observer; the words he speaks after that remain inside the black hole.

Not only will all signals of an astronaut who has entered the black hole remain trapped, but the astronaut himself is trapped.

His worldline intersects the singularity at $r = 0$, and when he comes near that place, the large tidal forces rip him apart.

Since all worldlines within the astronaut's future light cone terminate on $r = 0$, the collision with the singularity cannot be avoided.

Not even a spacecraft with the most powerful rocket engine can resist the pull of gravity in a black hole. (In fact, the use of rocket engines inside a black hole can never delay the collision with the singularity, but only speed it up, in the sense of shortening the proper time that elapses until collision.)

Incidentally, we can also conclude that if a spherical star is compressed by some astrophysical process to a radius smaller than its Schwarzschild radius, then gravitational collapse necessarily ensues.

Consider a particle on the surface of the star, and remember that, according to the Birkhoff theorem, the exterior geometry must be that of Schwarzschild.

The particle on the surface then has the same equation of motion as a spacecraft with a rocket engine in an empty Schwarzschild spacetime – we may treat the pressure with which the star pushes outward on the particle as mathematically analogous to the thrust of the rocket engine.

The surface of the star must therefore, necessarily fall toward the singularity $r = 0$, just as the spacecraft does.

Gravitational collapse is inevitable.

In the rest frame of a falling astronaut the amount of proper time needed to enter a black hole and crash into the singularity at $r = 0$ is not only finite but also quite short.

For a typical worldline, the proper time between r_S and $r = 0$ is of the order of r_S/c.

This means that after an astronaut crosses the horizon of a black hole with $M \cong M_\odot$, he has only $\approx 10^{-5}$ s to live.

Although the Schwarzschild coordinates suffer from the inconvenience of a singularity at $r = r_S$ that spoils calculations of t-time, this does not affect the calculation of the proper time needed to fall into a black hole and into the singularity.

Equation (7.100) gives us $dr/d\tau$ for an astronaut in free fall.

For purely radial motion, $\ell = 0$ [see Eq. (7.95)], and hence Eq. (7.100) reduces to

$$\left(\frac{dr}{d\tau}\right)^2 = \frac{r_S}{r} - (1 - \mathcal{E}^2)$$

This can be integrated to yield

$$\tau = \text{constant} + \frac{r(r_S/r - 1 + \mathcal{E}^2)^{1/2}}{1 - \mathcal{E}^2} + \frac{r_S}{(1 - \mathcal{E}^2)^{3/2}} \tan^{-1}\left(\frac{r_S/r - 1 + \mathcal{E}^2}{1 - \mathcal{E}^2}\right)^{1/2} \tag{8.14}$$

For a finite change in r, the corresponding change in proper time is finite.

For example, suppose the astronaut is initially at rest at the radius $r_0 = r_S/(1 - \mathcal{E}^2)$; this radius is larger than r_S and hence outside of the black hole.

The proper time to fall freely all the way to $r = 0$ is then

$$\Delta\tau = r_S \frac{\pi/2}{(1 - \mathcal{E}^2)^{3/2}} = r_S \frac{\pi}{2}\left(\frac{r_0}{r_S}\right)^{3/2} \tag{8.15}$$

If the initial value r_0 is near r_S (and $\mathcal{E}$ is near zero), then $r_0/r_S \cong 1$ and $\Delta\tau \cong \pi r_S/2$, which is of the order of r_S, as claimed previously.

The region in the immediate vicinity of the horizon not only has drastic effects on the speed of light, the light cones, and the lapse of proper time but it also has remarkable effects on electric and magnetic fields of particles or currents that might be present in this region.

For instance, Fig. 8.4 shows the electric field lines of electrically charged particles placed at a fixed location, at $r = 2r_S$ or at $r = 1.1r_S$.

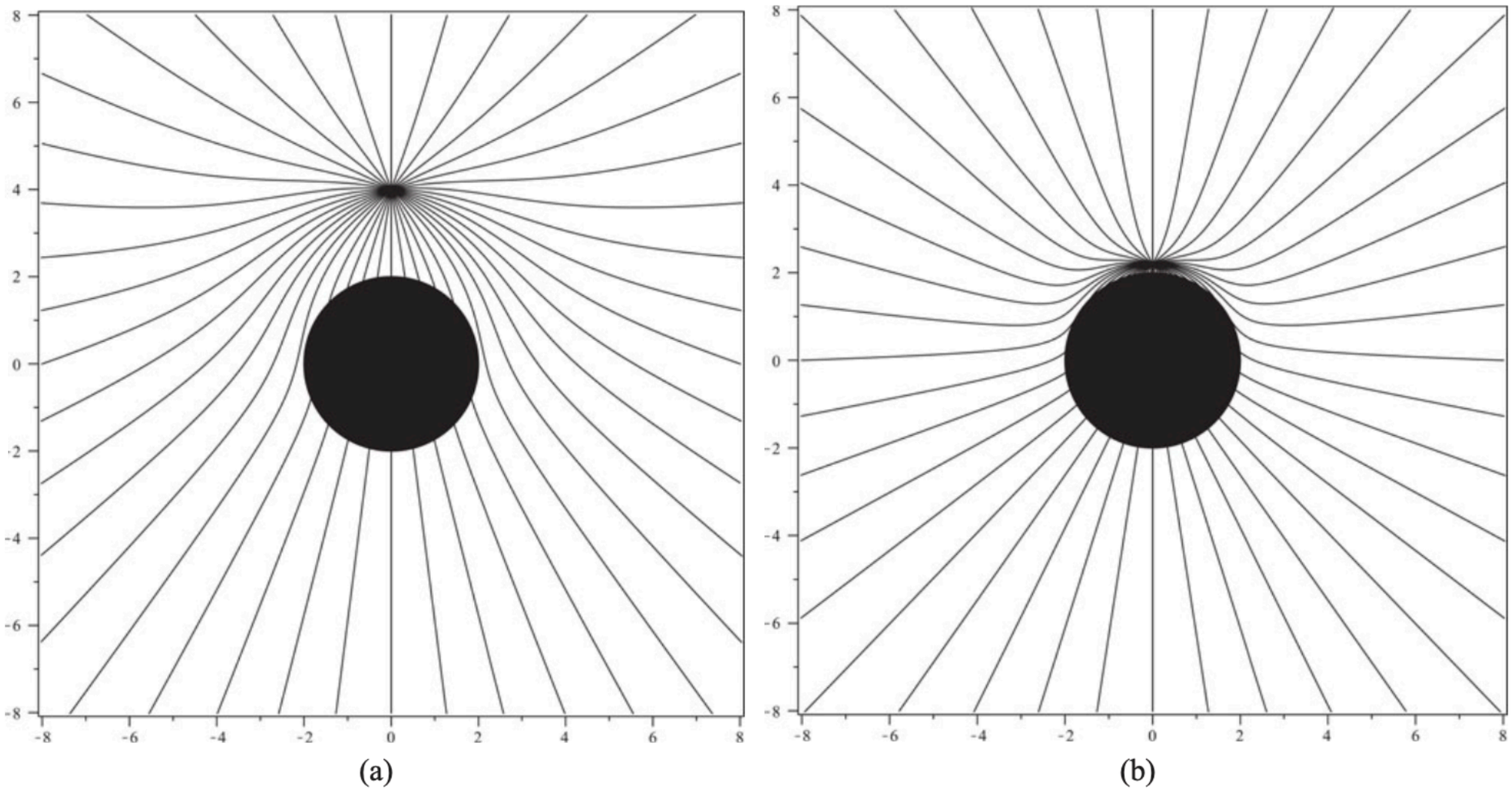


Fig. 8.4 (a) Electric field lines of a point charge located at $r = 2r_S$. (b) Electric field lines of a point charge located at $r = 1.1r_S$. (From Hanni and Ruffini, 1973)

These field lines were obtained by solution of the Maxwell equation in the curved spacetime background (the electric field is assumed to be weak, so the Schwarzschild geometry remains undisturbed).

The strongly curved spacetime geometry near the horizon produces strong distortions of the field lines.

Note that asymptotically the pattern of electric field lines in Fig. 8.4b is centered on the black hole rather than on the point charge above it – the black hole acts as though it had acquired the electric charge, and an observer at a large distance would perceive the charge distribution as located on or in the black hole, whereas the point charge would not be perceptible in the asymptotic electric field.

This is an instance of the "membrane paradigm" for black-hole physics.

The black hole acts as though it were covered by an electrically charged membrane.

If the point charge is given a motion toward or away from the black hole, the apparent membrane charge distribution rearranges itself; that is, currents seem to flow on the membrane, and the resulting magnetic fields can be calculated from these apparent currents.

Other physical properties – such as electric resistance, viscosity, temperature, and entropy – can be associated with the membrane.

In general, the membrane paradigm can be used to describe the physics as seen by observers at rest in the Schwarzschild geometry, at fixed locations in Schwarzschild coordinates.

But for observers who allow themselves to fall into the black hole, the membrane will be revealed as an illusion.

The electric field lines do not start at the black-hole horizon, as it seems in Fig. 8.4.

Instead the electric field lines are everywhere continuous in the empty Schwarzschild geometry, except at physical point charges, as demanded by Maxwell's equations.

8.3 Maximal Schwarzschild geometry

Although event horizons represent physical and coordinate-independent properties of a spacetime, it is often easier to recognize the presence, or absence, of a horizon by particular, cleverly chosen coordinates.

Thus, the absence of an event horizon in flat spacetime is obvious in inertial coordinates, but not so obvious in Rindler coordinates.

For the case of the Schwarzschild geometry, the coordinates that make things easy are the *Kruskal coordinates*, also called *Kruskal-Szekeres* coordinates, which we describe in this section.

In contrast to the Schwarzschild coordinates, these new coordinates are well behaved at the horizon at $r = r_S$.

But the price we must pay for this good behavior is that the metric tensor will not appear static, not even in the exterior region.

In the interior of the black hole, the Kruskal coordinates v, u, are defined as

$$\text{For } r < r_S: \quad v = \sqrt{1 - r/r_S}\, e^{r/2r_S} \cosh t/2r_S, \quad u = \sqrt{1 - r/r_S}\, e^{r/2r_S} \sinh t/2r_S \tag{8.16}$$

In the exterior region, the definitions are

$$\text{For } r > r_S: \quad v = \sqrt{r/r_S - 1}\; e^{r/2r_S} \sinh t/2r_S, \quad u = \sqrt{r/r_S - 1}\; e^{r/2r_S} \cosh t/2r_S \tag{8.17}$$

The inverse transformation is given (implicitly) by

$$\text{For } r < r_S: \quad (r/r_S - 1)e^{r/r_S} = u^2 - v^2, \quad t = 2r_S \tanh^{-1} u/v \tag{8.18}$$

and

$$\text{For } r > r_S: \quad (r/r_S - 1)e^{r/r_S} = u^2 - v^2, \quad t = 2r_S \tanh^{-1} v/u \tag{8.19}$$

The coordinate v is timelike and *u* is spacelike.

Note that the functions that appear on the right sides of Eqs. (8.18) and (8.19) are similar to the functions that appear in Eq. (8.10) in the transformation to Rindler coordinates.

This similarity is no accident.

Since we know that the Schwarzschild pseudosingularity at $r = r_S$ is similar to the Rindler pseudosingularity, we expect that a transformation of roughly the same form as the inverse of the Rindler transformation (from Rindler coordinates to inertial coordinates) will help remove the pseudosingularity.

In Kruskal coordinates, the spacetime interval takes the new form

$$ds^2 = \frac{4{r_S}^3}{r} e^{-r/r_S}(dv^2 - du^2) - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \qquad (8.20)$$

where *r* is to be regarded as a function of v and *u*, as given by Eq. (8.18) or (8.19).

The new metric tensor has no singularity at $r = r_S$.

Of course, there remains a singularity at $r = 0$; this is unavoidable, because the curvature becomes infinite at that point

Let us now disregard displacements in the θ and ϕ directions, and concentrate on the radial and time dependence of the geometry.

Figure 8.5 shows the *u*, v coordinate grid and the corresponding values of the *r*, *t* coordinates.

In this diagram, *r* and *t* are measured in units of r_S; thus, $r = 1.5$ means $r = 1.5r_S$, etc.

As in the case of Rindler coordinates, the curves *r* = constant are hyperbolas, and the curves *t* = constant are straight lines through the origin.

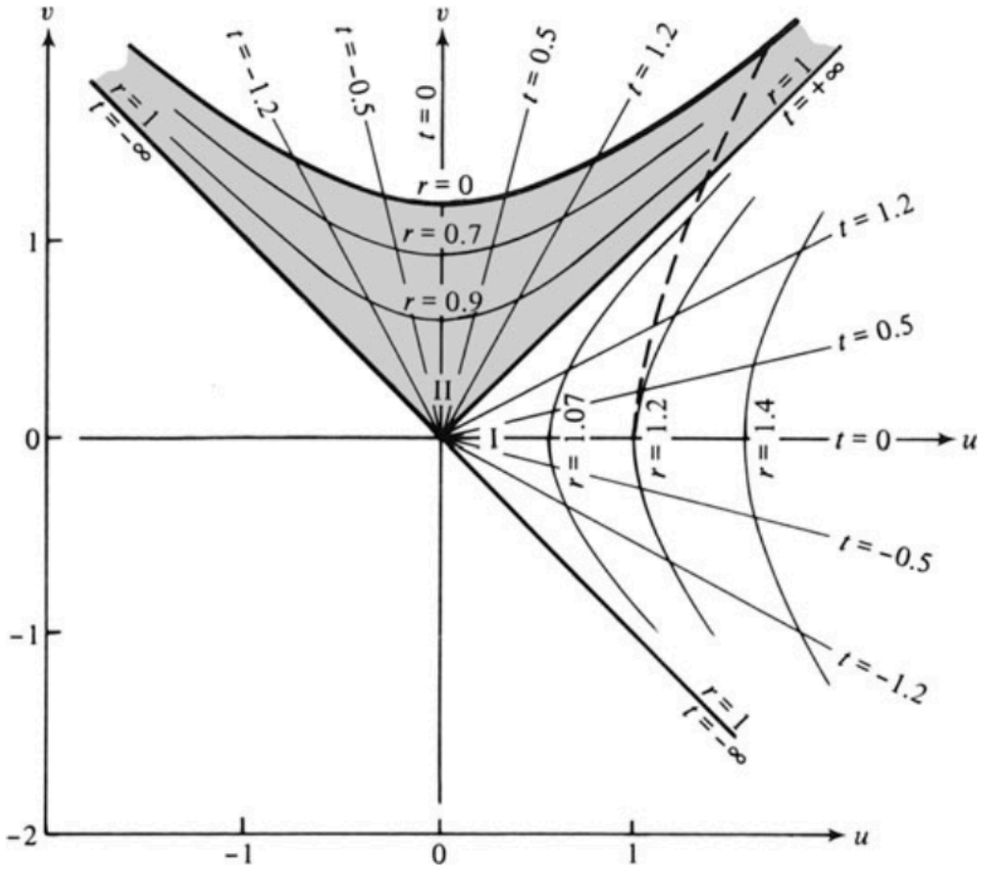


Fig. 8.5 Schwarzschild spacetime in Kruskal coordinates. The u, v coordinates are dimensionless. The r, t coordinates are given in units of r_S. The dashed line is the worldline of a particle falling toward $r = 0$.

It is obvious from the diagram that the coordinate transformation from u, v to r, t is singular at $r = r_S$.

Thus, the line $u = v$ is mapped into the point $r = r_S$, $t = \infty$.

The elimination of the Schwarzschild pseudosingularity is accomplished precisely by the singular character of the coordinate transformation.

One of the nice features of the Kruskal coordinates is that the worldlines of radial light signals (that is, light signals with $d\theta = d\phi = 0$) are straight lines at 45° with respect to the u, v-axes, just like the worldlines of light signals in flat spacetime.

This is immediately obvious from Eq. (8.20), which shows that $ds^2 = 0$ requires $du = \pm dv$.

The worldlines of other freely falling particles are not straight lines.

For example, the dashed line in Fig. 8.5 indicates a typical worldline for a particle (or an astronaut) falling into the black hole.

The surface $r = r_S$ divides spacetime into two regions: the black hole (region II) and the asymptotically flat space surrounding it (region I).

The existence of an event horizon at $r = r_S$ is immediately obvious from inspection of the spacetime diagram; according to Fig. 8.5, a light signal sent out by the astronaut after he crosses $r = r_S$ will proceed at 45° in this spacetime diagram and ultimately intersect $r = 0$.

The signal always moves toward decreasing values of r, and it can never reach $r = r_S$.

Signals other than light must necessarily lie within the forward light cone, and they will be even less able to escape.

Signals from outside the black hole can of course cross $r = r_S$ without trouble; that is, the surface of the black hole acts as a "one-way membrane."

Note that the singularity $r = 0$ displayed in Fig. 8.5 is not a pointlike singularity.

Rather, $r = 0$ is a spacelike surface.

As seen in u, v coordinates, the singularity happens at different points of space at different times.

The singularity happens at the point u when the v time is $v = \sqrt{1 + u^2}$.

Thus, at large values of u, the singularity is delayed.

We recall that, as seen in Schwarzschild coordinates, the singularity happens at the same time ($r = 0$) at all points of space ($-\infty < t < \infty$).

This difference in behavior simply reflects a difference in the definition of simultaneity.

In Fig. 8.5, the spacetime diagram has been left blank below the line $v = -u$.

It is obvious that the only region that can be explored by an astronaut who starts his expedition in the exterior (region I) of the black hole is the region $v > -u$.

The region $v < -u$ is necessarily outside of his light cone.

However, this region could still have a physical significance if signals emitted by something in it cross the line $v = -u$ and reach an astronaut in the region $v > -u$.

Examination of the Kruskal metric shows that in fact the line $v = -u$ is in no way singular, and geodesics can cross it without any trouble.

This indicates that the diagram in Fig. 8.5 is incomplete.

A more complete diagram is given in Fig. 8.6.

This diagram is as complete as can be.

Technically, a manifold is said to be *maximal* if every geodesic either is of infinite extent in both directions (has no beginning and no end) or else ends or begins on a singularity.

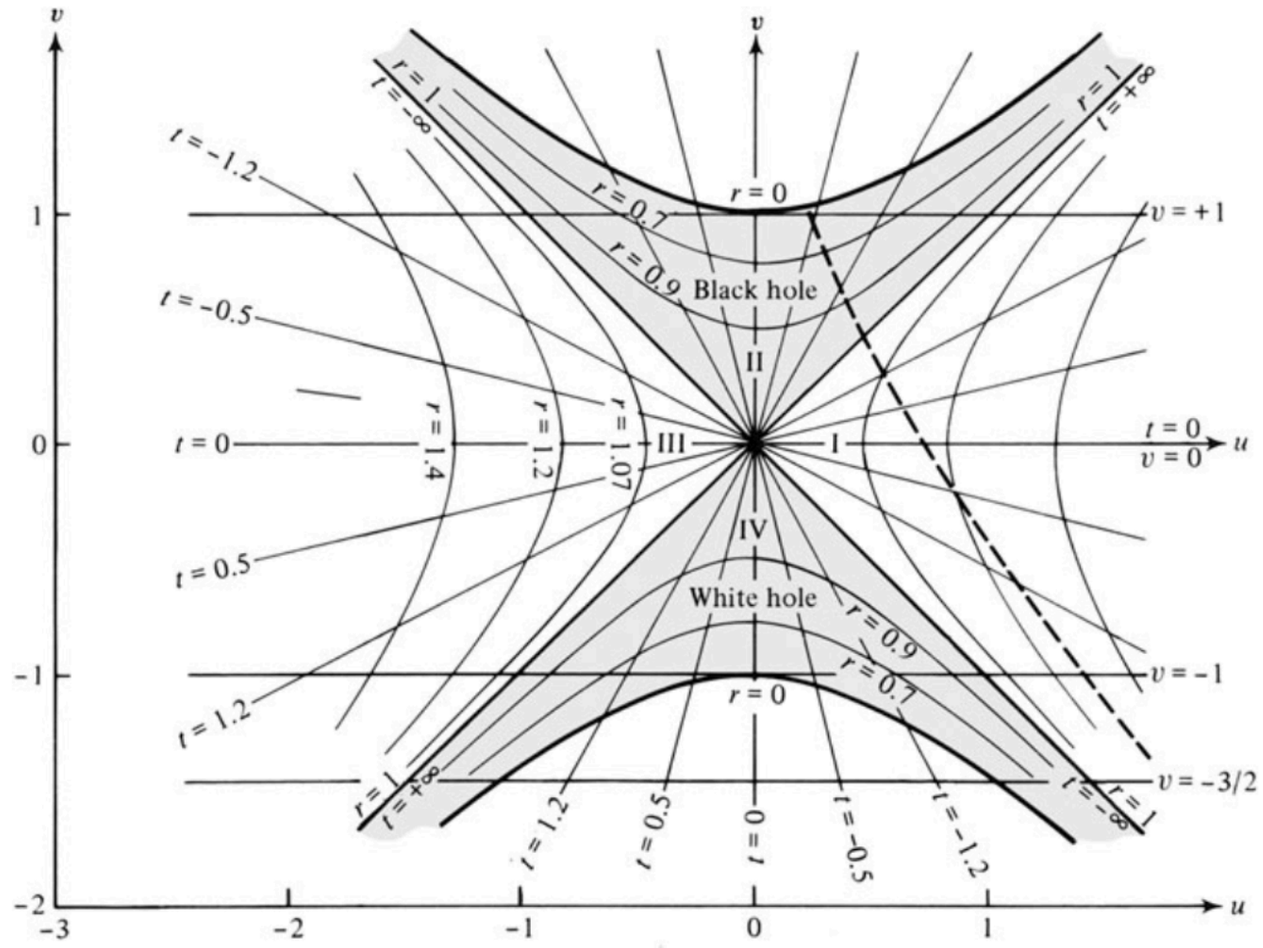


Fig. 8.6 The maximal Schwarzschild spacetime in Kruskal coordinates.

A manifold is said to be geodesically complete if all geodesics are of infinite length. In physical terms, this means that in a complete spacetime manifold, every particle (if stable, and if it does not collide and react with another particle) has been in the spacetime forever and remains in it forever.

In a maximal spacetime, particles can appear or disappear, but only at singularities.

We will always demand that spacetime be maximal, because it makes no sense whatsoever that the geodesic of a particle should suddenly come to an end (that is, leave or enter our universe) in the middle of nowhere.

The spacetime shown in Fig. 8.6 is maximal, but not complete; that shown in Fig. 8.5 is neither maximal nor complete.

Note that several sets of r, t coordinates are needed to cover all parts of spacetime in Fig. 8.6, but one set of u, v coordinates is sufficient.

For a given value of u, the spacetime has two singularities $r = 0$; one is at $v = +\sqrt{1+u^2}$, the other at $v = -\sqrt{1+u^2}$.

The spacetime also has two asymptotically flat regions, one at $u \to \infty$, the other at $u \to -\infty$; these two regions are labeled I and III in Fig. 8.6.

By examination of Fig. 8.6, we can gain some insight into the time evolution of the Schwarzschild spacetime, using v as our time variable.

Several lines v = constant have been drawn through Fig. 8.6.

For instance, at time $v = -3/2$, we have a space that is asymptotically flat (for $u \to \infty$) with a singularity at $u = \sqrt{5}/2$.

At the same time we have a second asymptotically flat space ($u \to -\infty$) with a singularity at $u = -\sqrt{5}/2$.

This second space is in a different universe; alternatively, it may be regarded as a region in our own universe, if this second region of our universe is very far away from the first.

At the time $v = -1$, both of these singularities coalesce, and we obtain two asymptotically flat spaces joined together by a region in which the geometry deviates greatly from flatness.

Such a region is called a *wormhole* (also called an Einstein-Rosen bridge, or a Schwarzschild throat).

Whether we regard this wormhole as connecting two universes, or as connecting two remote parts of the same universe, the topology in either case is non-Euclidean.

Between the times $v = -1$ and $v = +1$ there is no singularity, only a wormhole in which the curvature, although perhaps large, is everywhere finite.

At time $v = +1$ the connection between the two asymptotically flat spaces disappears, and we are left with two separate spaces, each with a singularity.

These spaces continue to evolve parallel in time.

Note that the time evolution described here is perfectly symmetric under time reversal (reversal of v to $-v$).

The second asymptotically flat space (region III) is outside of the forward light cone of the first asymptotically flat space (region I), and conversely.

This implies that communication between the two universes is impossible.

However, if astronauts from both universes jump into the black hole (region II), then they can meet, embrace, and perish together.

What prevents signals from passing through the wormhole is the time evolution of the interior geometry.

The wormhole "pinches off" so fast that not even a light signal can pass through before the singularity happens.

If we want to interpret the two asymptotically flat regions as remote parts of the same universe, then the impossibility of signal transmission through the wormhole is crucial for the preservation of causality.

It is obvious that if remote parts of the universe could communicate via a short wormhole, the effective signal velocity would exceed the speed of light.

Region IV is called a *white hole*.

According to the preceding discussion such a white hole may be regarded as the time reverse of a black hole.

Signals from region IV can go out to the exterior regions I and III, but no signal from the latter regions can ever be sent into region IV.

This must not be interpreted as implying that a white hole is intrinsically luminous.

We are dealing with vacuum solutions of Einstein's equations; light signals are to be regarded as test signals.

Light will come out of a white hole only if somebody places a light source in the white hole.

There is no reason to suspect that anybody has done this or could do this.

The white-hole portion of the complete Kruskal solution is probably as irrelevant to our universe as are the advanced solutions of Maxwell's equations.

In electromagnetism, the mathematical existence of these solutions is required by the time-reversal symmetry of the Maxwell equations, but their physical existence would require some very peculiar boundary conditions at the beginning of the universe (incoming waves that converge on particles and are exactly absorbed by them).

So far we have examined the spacetime geometry in the u-v (or *r*-*t*) plane.

Let us now concentrate on the purely spatial part of the geometry that is, the geometry at a given fixed time.

We are then dealing with a curved 3-D space with coordinates r (or u), θ, and ϕ.

Since it is impossible to make a drawing of 3-D curved space, we prefer to look at the geometry in the 2-D curved space $\theta = \pi/2$ (because of spherical symmetry, the geometry is the same in all such 2-D spaces passing through the center of symmetry).

Figure 8.7 shows this curved surface at the time v = 0.

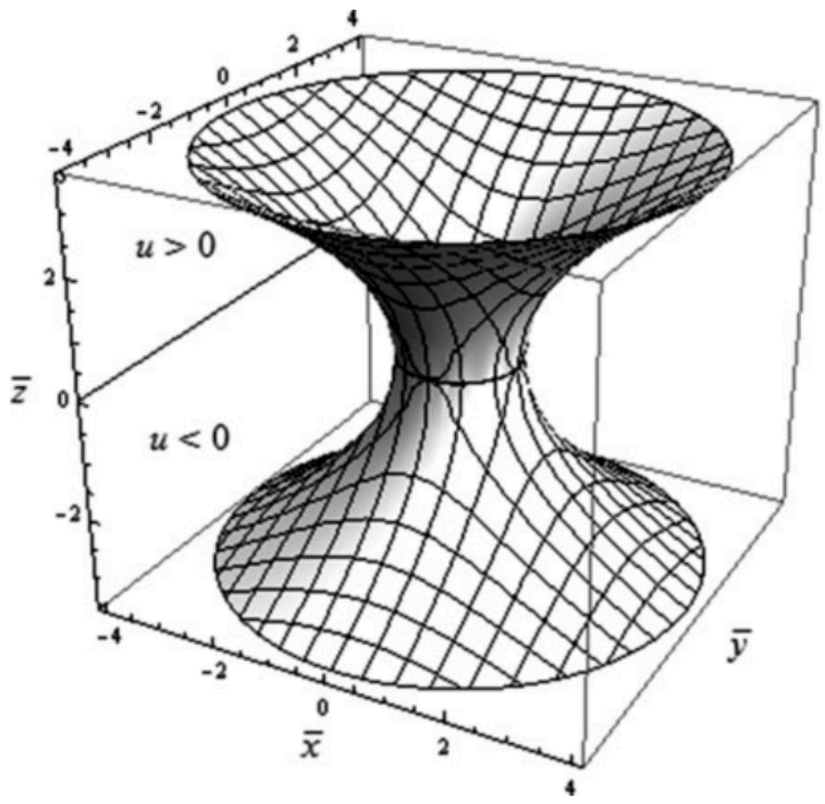


Fig. 8.7 The geometry of the surface $v = 0$, $\theta = \pi/2$ (or, equivalently, $t = 0$, $\theta = \pi/2$) of the Schwarzschild geometry. Only r and ϕ vary along this surface. The r-ϕ surface is represented as a curved 2-D surface embedded in flat 3-D space, and this embedding has been constructed so that the distances measured along the surface (with, say, a string) directly give the true distances that would be measured in the Schwarzschild geometry. Asymptotically, the upper and the lower parts of this Schwarzschild throat spread out and become flat.

The curved 2-D space r, ϕ is shown embedded in a 3-D space x,y,z.

Here, x = r cosϕ,y = r sinϕ are related to the Schwarzschild spherical coordinates, whereas z is an artificial coordinate that has nothing to do with the z-coordinate of real space and is introduced to permit us to represent the curved geometry of the r-ϕ space as a curved 2-D surface.

This 2-D space is a paraboloid of revolution; the shape of the surface has been chosen in such a way that the distances measured along the curved surface with, say, a string agree with the distances measured in the r-ϕ plane of the Schwarzschild geometry.

For v = t = 0 and $\theta = \pi/2$, it can be shown that the Schwarzschild geometry in the r-ϕ plane is described by the distance interval

$$dl^2 = \frac{dr^2}{1 - r_S/r} + r^2 d\phi^2$$

We can embed this curved 2-D surface in flat 3-D space by introducing the extra coordinate z and writing the distance interval as

$$dl^2 = d\bar{z}^2 + d\bar{x}^2 + d\bar{y}^2 = d\bar{z}^2 + dr^2 + r^2 d\phi^2 = dr^2 \left[1 + \left(\frac{d\bar{z}}{dr}\right)^2\right] + r^2\phi^2 \qquad (8.21)$$

By comparing these two expressions for dl^2, we obtain the following differential equation for the surface:

$$dl^2 = \frac{dr^2}{1 - r_S/r} + r^2 d\phi^2$$

If we solve this equation, we can show that the surface is a paraboloid of revolution, $r = r_S + z^2/4r_S$.

The upper half of the figure ($u > 0$ in Kruskal coordinates) can be thought of as showing the deformation of the flat r-ϕ plane with which we start out in the absence of gravitation.

To be precise, imagine that the r-ϕ plane of flat 2-D space has the hole $r < 1$ cut out of it (the region $r < 1$ is nonexistent at "time" v = 0; as we can see from Fig. 8.6, the surface v = 0 does not contain values of r smaller than $r = 1$).

Gravitation deforms this plane minus hole into the funnel-like structure that is the upper part of the curved surface shown in Fig. 8.7.

The lower half (u < 0) arises from a similar deformation of the r-ϕ plane of the second universe.

The two universes of Fig. 8.7 are joined along the circle $0 \le \phi \le 2\pi$.

If we want to take into account the θ coordinate, then we must imagine similar circles for different values of θ.

This means that the two universes are actually joined along a spherical surface $0 \le \phi \le 2\pi, 0 \le \theta \le \pi$ (a "two-sphere").

At later times (v > 0) or earlier times (v < 0) the wormhole is narrower (in Fig. 8.6, the surface v > 0 contains values of r smaller than $r = 1$), and at time $v = \pm 1$, the wormhole pinches off at the center and we are left with two separate surfaces, each with a cusp.

Figure 8.8 gives schematic diagrams that take us through a sequence of values of v, from v = 0 to v > 1.

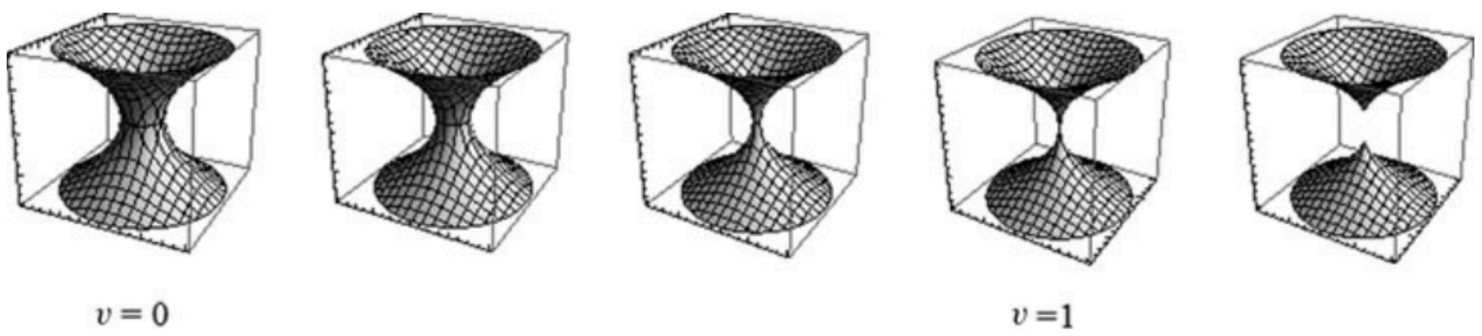


Fig. 8.8 Evolution of the geometry in v-time. The Schwarzschild throat becomes narrower and pinches off at the critical value $v = 1$. The time-reversed evolution (negative v) is the same – the Schwarzschild throat again becomes narrower and pinches off, at $v = -1$.

The time-reversed evolution, from $v = 0$ to $v < -1$, is similar, that is, the evolution is time symmetric.

These diagrams give us a clear picture of how the wormhole pinches off.

Any signal sent into the wormhole from one side or the other fails to pass through because it is trapped in the cusp.

The suggestion has been made that an advanced civilization might keep the wormhole open and thereby avoid the formation of a horizon and a singularity by artificial means.

But this can be achieved only if the wormhole is constructed out of a distribution of matter with negative energy density.

Ordinary matter does not have a negative energy density; quantum fields might perhaps have the required negative energy density, but even if they do, they are not likely to extend over a macroscopically large volume.

If wormholes could be kept open, and travel through them were possible, then they could be exploited for time travel.

A particle that takes a shortcut through a wormhole to travel from one point in space to a distant point effectively attains a speed in excess of the speed of light.

As is well known, such a superluminal, or tachyonic, speed results in travel backward in time when observed from a new reference frame relative to which the wormhole is in translational motion.

Finally, it must be kept in mind that although the Kruskal diagram describes a complete (or at least as complete as possible) solution of Einstein's equations, only a part of this solution will usually be relevant in the real world.

For example, in the formation of a black hole by gravitational collapse of a star, the Schwarzschild vacuum solution applies only to those regions of spacetime outside of the star.

If the dashed line in Fig. 8.6 is the worldline of the surface of the collapsing star, with spherical symmetry, then the Schwarzschild solution is valid everywhere to the right of the dashed line.

The region to the left of the dashed line is occupied by the matter of the star and is described by some new solution of Einstein's equations with $T^{\mu\nu} \neq 0$; the Kruskal diagram does not apply in this region.

Obviously regions III and IV are never of any relevance to the problem of gravitational collapse.

Since the complete Kruskal diagram is irrelevant for black holes formed by gravitational collapse, it can apply only to black holes created jointly with the universe, that is, black holes that have existed ever since the beginning of time.

An annoying feature of the Kruskal coordinates is that the metric tensor depends on time (it depends on v), both inside and outside of the black hole, whereas the Schwarzschild coordinates avoid a dependence on time outside of the black hole.

For some purposes it is convenient to introduce a set of coordinates that achieves a compromise between Kruskal coordinates and Schwarzschild coordinates – the metric tensor does not depend on time and is free of some, but not all, of the pseudosingularities at $r = r_S$.

Such a set of coordinates are the *Eddington-Finkelstein coordinates*, often used for the solution of the wave equation in the Schwarzschild geometry.

The timelike Eddington-Finkelstein coordinate $\tilde{v}$ is defined as follows:

$$\tilde{v} = t + r + r_S \ln |r/r_S - 1| \tag{8.22}$$

The other coordinates r, θ,ϕ remain the same as in the Schwarzschild case.

The spacetime interval in these coordinates $\tilde{v}$, r, θ , ϕ is

$$ds^2 = \left(1 - \frac{r_S}{r}\right) d\tilde{v}^2 - 2d\tilde{v}dr - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \tag{8.23}$$

Note that the spatial part of the metric tensor has no singularity, but the time part of the metric retains the pseudosingularity at $r = r_S$.

In these coordinates, the worldline of a radially outgoing light signal has the equation $d\tilde{v}/dr = 2/(1 - r_S/r)$, and the worldline of an ingoing light signal has the equation $\tilde{v}$ = constant.

Thus, the surfaces $\tilde{v}$ = constant are simply families of worldlines of ingoing light signals.

Figure 8.9 displays the Schwarzschild spacetime in $\tilde{v}$, r coordinates.

The lines $\tilde{v}$ = constant have been drawn at 45°, so that the light cones at large *r* have the usual orientation.

The existence of the event horizon at $r = r_S$ is, again, obvious by inspection of the light cones.

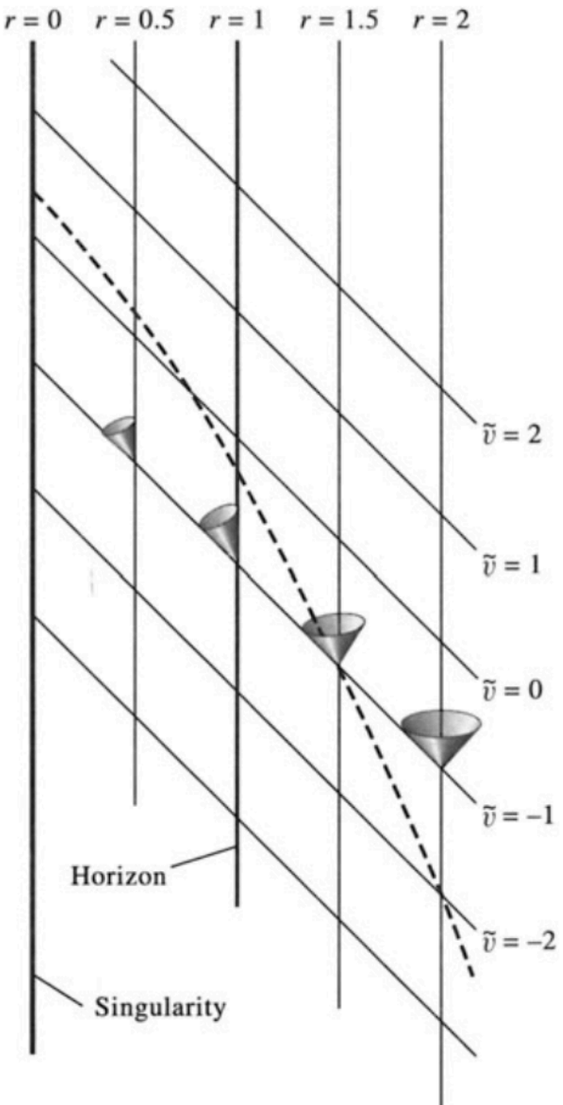


Fig. 8.9 The Schwarzschild spacetime in Eddington-Finkelstein coordinates $\tilde{v}, r$. The coordinates are given in units of r_S. The dashed line is the worldline of an infalling particle.

8.4 Kerr solution and Reissner-Nordstrøm solution

Next to the Schwarzschild solution, the two most important solutions of Einstein's equations are the Kerr solution and the Reissner-Nordstrøm solution.

These two solutions represent the curved spacetime geometries surrounding a rotating mass and an electrically charged mass, respectively.

In the linear approximation, the stationary solution for the gravitational field surrounding a rotating mass was given in Section 4.7.

The corresponding exact solution of the nonlinear Einstein equations is given by the following expression for the spacetime interval due to Kerr, Boyer and Lindquist:

$$ds^2 = dt^2 - \frac{\rho^2}{\Delta}dr^2 - \rho^2 d\theta^2 - (r^2 + a^2)\sin^2\theta d\phi^2 - \frac{2GMr}{\rho^2}(dt - a\sin^2\theta d\phi)^2 \tag{8.24}$$

Here ϱ^2 and Δ are functions of r and θ,

$$\rho^2 \equiv r^2 + a^2\cos^2\theta\ , \quad \Delta \equiv r^2 - 2GMr + a^2 \tag{8.25}$$

and a and M are constants.

It is straightforward, and very tedious, to verify that Einstein's equations in vacuo are satisfied by this metric tensor.

The Kerr solution is stationary, but not static; that is, the metric tensor is not a function of t, but fails to be invariant under time reversal because of the presence of the term $\propto d\phi dt$.

The solution is rotationally symmetric about the z-axis; that is, it is independent of the angle ϕ.

For large r, Eq. (8.24) reduces to

$$ds^2 \cong \left(1 - \frac{2GM}{r}\right) dt^2 - \left(1 + \frac{2GM}{r}\right) dr^2 - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 - \frac{2GMa^2}{r} \sin^4\theta \, d\phi^2 + \frac{4GMa}{r} \sin^2\theta d\phi dt \tag{8.26}$$

In this expression, all terms $\propto 1/r$ have been kept, and all terms $\propto 1/r^2$ have been omitted.

The next to last term on the right side of Eq. (8.26) can be neglected, because it is smaller by a factor $2GMa^2\sin^2\theta/r^3$ than the term that precedes it.

To recognize the last term, we express it in rectangular coordinates.

With $x = r \sin\theta \cos\phi$, $y = r \sin\theta \sin\phi$, we obtain $r^2 \sin^2\theta \, d\phi = xdy - ydx$, and therefore Eq. (8.26) becomes

$$ds^2 \cong \left(1 - \frac{2GM}{r}\right) dt^2 - \left(1 + \frac{2GM}{r}\right) dr^2 - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 + \frac{4GMa}{r^3}(xdy - ydx)dt \tag{8.27}$$

The first four terms in Eq. (8.27) are exactly the same as in the linear approximation for the Schwarzschild solution, and they permit us to identify M as the total mass of the system.

The last term corresponds to off-diagonal components in the metric.

Expressed in rectangular coordinates, these off-diagonal terms are

$$g_{01} = g_{10} = -2GMa\frac{y}{r^3}, \qquad g_{02} = g_{20} = 2GMa\frac{x}{r^3} \tag{8.28}$$

Comparison with the weak field solution for a rotating mass [Eq.(4.93)] shows that the quantity Ma is to be identified as the spin angular momentum of the system,

$$S_z = Ma \tag{8.29}$$

Hence the parameter a is the spin angular momentum per unit mass.

If $a > 0$, then the direction of rotation is counterclockwise around the +z-axis.

Note that in Eq. (8.24), the time reversal $t \rightarrow -t$ has exactly the same effect as the change $a \rightarrow -a$; this simply corresponds to reversal of the direction of rotation.

The Kerr solution is not the only solution for the field surrounding a rotating mass.

In general, solutions for rotating masses depend on the exact shape of the mass.

Different mass distributions will generate different solutions, and these differences will show up at large r as different multipole moments in the gravitational fields.

However, the Kerr solution is the unique stationary solution that represents a rotating black hole, that is, a rotating system that has collapsed inside its horizon.

As we will see in Section 8.8, the gravitational collapse of a rotating, electrically neutral star will ultimately lead to the Kerr geometry.

The coordinates t , r, θ , ϕ used in Eq. (8.24) are Schwarzschild-like.

In fact, in the limit $a \rightarrow 0$, the expression (8.24) for the Kerr solution agrees exactly with the expression (8.1) for the Schwarzschild solution expressed in the usual spherical coordinates.

However, if a $\neq 0$, then the coordinates r, θ , ϕ are not spherical coordinates, but quasi-spheroidal coordinates.

These coordinates are defined in terms of rectangular coordinates by:

$$x = (r^2 + a^2)^{1/2} \sin\theta \cos[\phi - f(r)] \tag{8.30}$$

$$y = (r^2 + a^2)^{1/2} \sin\theta \sin[\phi - f(r)] \tag{8.31}$$

$$z = r \cos\theta \tag{8.32}$$

where

$$f(r) = a \int_{\infty}^{r} \frac{dr'}{(r'^2 - 2GMr' + a^2)} + \tan^{-1} \frac{a}{r} \tag{8.33}$$

If $M = 0$, then $f(r) = 0$, and the quasi-spheroidal coordinates defined by Eqs. (8.30)–(8.32) become ordinary spheroidal coordinates.

In the limit r $\rightarrow \infty$, the coordinates r, θ, ϕ approach spherical coordinates.

Plotted in the rectangular axes x, y, z, the surfaces r = constant are confocal ellipsoids, and the surfaces θ = constant are hyperboloids of one sheet.

The surfaces ϕ = constant are rather complicated "bent planes" that gradually become flat at large values of r (in ordinary spheroidal coordinates, these surfaces would be ordinary planes).

Figure 8.10 shows the lines r = constant and θ = constant in the x-z plane.

At large values of r, the ellipses tend to circles and the hyperbolas straighten out to give us ordinary polar coordinates.

Note that Fig. 8.10 adopts the convention that if the upper part of a given hyperbola is characterized by the parameter θ, then the lower part is characterized by $\theta - \pi$.

This makes the correspondence between spheroidal and spherical coordinates more obvious; the discontinuity in θ at $r = 0$ is of course a purely mathematical device with no physical implications.

Finally, it must be emphasized that Fig. 8.10 is intended to display only the relationships among x, y, z and the r, θ, ϕ coordinates; this figure does not in any way display the geometry of the curved space.

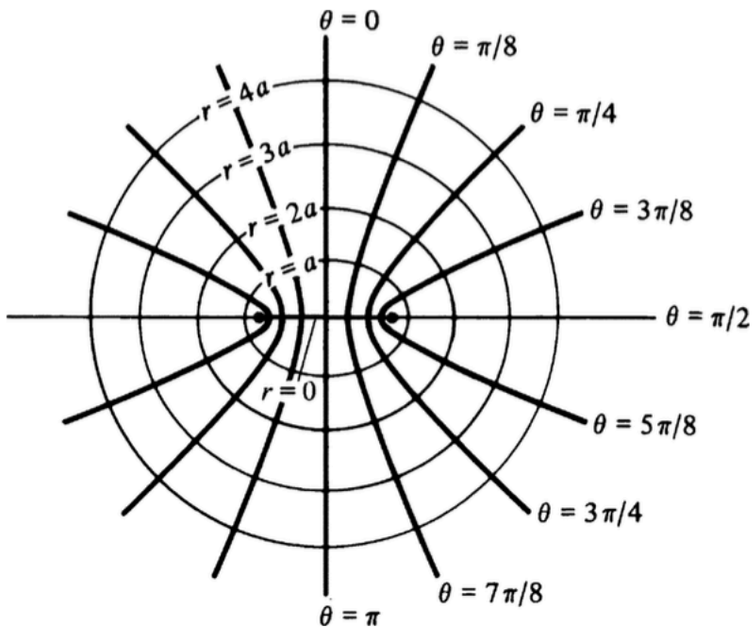


Fig. 8.10 Spheroidal coordinates in the x-z plane. The curves $r =$ constant are confocal ellipses with foci at $x = \pm a$; the curves $\theta =$ constant are confocal hyperbolas. The locus $r = 0$ is a disk spanning the space between the foci. The spheroidal coordinates in three dimensions are obtained by rotating this pattern about the z-axis.

The geodesic equations for the motion of a particle in the Kerr geometry are more complicated than for the Schwarzschild geometry.

It is, however, possible to gain some insight into the general features of the motion by examining the first integral, $g_{\mu\nu}u^{\mu}u^{\nu} = 1$, of the equation of motion, as in the Schwarzschild case [see Eq. (7.116)].

For orbits in the equatorial plane of the Kerr geometry, the equation $g_{\mu\nu}u^{\mu}u^{\nu} = 1$ can be expressed in terms of the energy per unit mass $\mathcal{E}$ and the orbital angular momentum ℓ per unit mass, by a calculation similar to that in Eqs. (7.96)–(7.100).

This leads to the result

$$\left[1+\frac{a^2}{r^2}\left(1+\frac{2GM}{r}\right)\right]\mathcal{E}^2-\frac{4GMa}{r^3}\ell\mathcal{E}-\left(1-\frac{2GM}{r}\right)\frac{\ell^2}{r^2}$$
$$-\left(1-\frac{2GM}{r}\right)-\frac{a^2}{r^2}-\dot{r}^2=0 \tag{8.34}$$

For a = 0, this reduces to the result we obtained for motion in the Schwarzschild geometry, Eq. (7.116).

However, in contrast to the Schwarzschild equation, Eq. (8.34) cannot be separated into a sum of a potential energy and a (radial) kinetic energy, and therefore it is not possible to construct an effective potential of the usual form, that is, a function of r and l (an attempt to do so gives a "potential" that is a function of r, ℓ , and $\mathcal{E}$).

Nevertheless, we can use Eq. (8.34) to determine the turning points *r* of the radial motion, by setting $\dot{r}=0$:

$$\left[1+\frac{a^2}{r^2}\left(1+\frac{2GM}{r}\right)\right]\mathcal{E}^2-\frac{4GMa}{r^3}\ell\mathcal{E}-\left(1-\frac{2GM}{r}\right)\frac{\ell^2}{r^2}$$
$$-\left(1-\frac{2GM}{r}\right)-\frac{a^2}{r^2}=0 \tag{8.35}$$

Figure8.11 shows a plot of these turning points *r* as a function of ℓ and $\mathcal{E}$ for the case of a Kerr black hole of maximum spin angular momentum, that is, a = GM.

The values of the orbital angular momentum ℓ can be positive or negative; these positive or negative values of ℓ represent, respectively, orbital angular momentum parallel or antiparallel to the spin of the black hole.

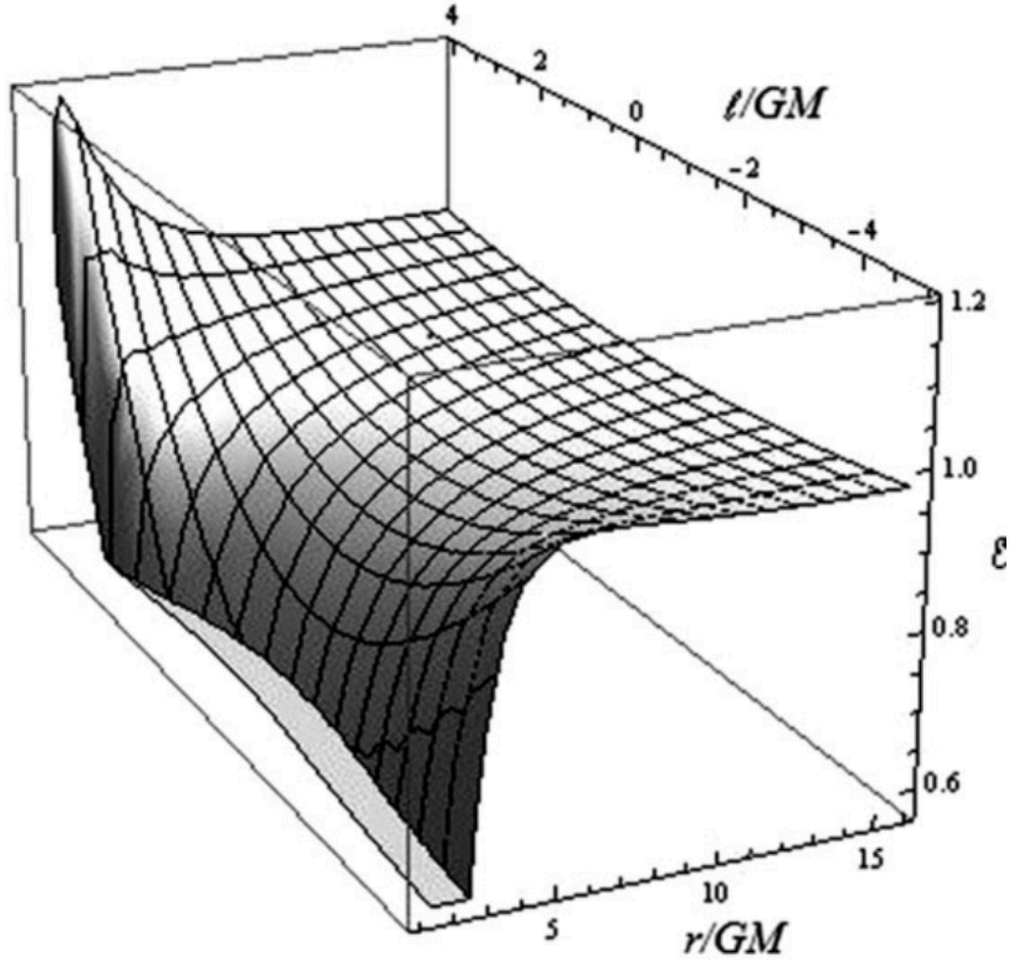


Fig. 8.11 Turning points r for the radial motion plotted as a function of the energy $\mathcal{E}$ per unit mass and the orbital angular momentum ℓ per unit mass for an orbit in the equatorial plane of an extreme Kerr black hole, that is, a black hole with spin parameter $a = GM$. The orbital angular momentum ℓ can be positive (parallel to the spin) or negative (antiparallel to the spin) (Adapted from Rees, Ruffini, and Wheeler, 1974).

As in the Schwarzschild case, stable circular orbits correspond to minima in the energy E as a function of radius, for a given value of l.

At these minima, the outer and inner turning points coincide, so the particle is constrained to remain at the fixed radius of the minimum while it moves around the black hole.

In the plot of Fig. 8.11, stable minima are found within ranges of positive and negative angular momenta, extending, respectively, from $\ell = 2\sqrt{3}\ \mathrm{GM} = 3.46\mathrm{GM}$ to ∞, and from $\ell = -22\mathrm{GM}/\sqrt{27} = -4.23\mathrm{GM}$ to $-\infty$.

Figure 8.12 shows a plot of the radii of the circular orbits as a function of the angular momentum ℓ .

The smallest stable circular orbit with the smallest positive value $\ell = 2\sqrt{3}$ GM has a radius r = GM and an energy E = m/ $\sqrt{3}$.

Thus, the binding energy of this orbit, relative to a particle at rest at infinity, is m – m/ $\sqrt{3}$ = 0.423m, almost one-half of the rest-mass energy!

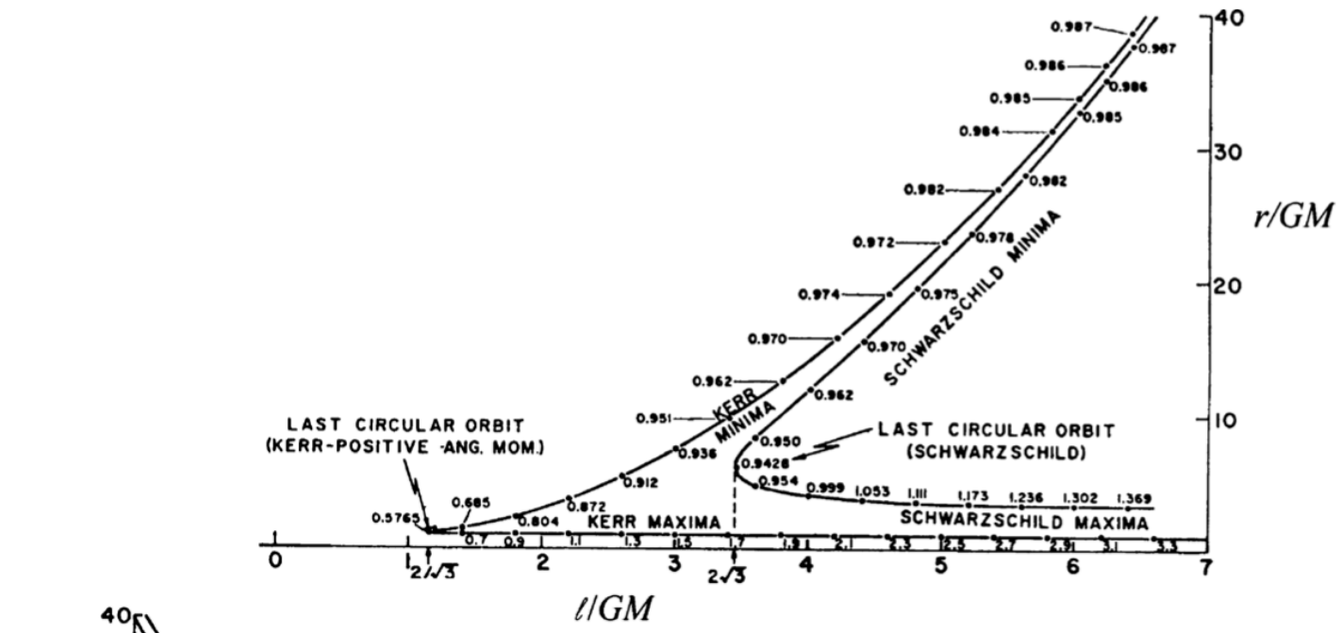


Fig. 8.12 Radii of circular orbits as a function of the orbital angular momentum ℓ, for motion in the field of an extreme Kerr black hole and also a Schwarzschild black hole. The upper branch of each curve gives the radius of the stable orbits, the lower the unstable orbits. The numbers marked along the curves give the orbital energy in units of the particle mass *m* (Rees, Ruffini, and Wheeler, 1974).

The Schwarzschild and Kerr solutions are vacuum solutions of Einstein's equations; in contrast, the Reissner-Nordstrøm solution, for an electrically charged mass, is not a vacuum solution.

Such a charged mass will be surrounded by an electric field, which gives rise to a nonzero energy-momentum tensor throughout space.

We must then solve the coupled Einstein and electromagnetic equations [Eqs. (7.16), (7.17)]:

$$R_\mu{}^\nu - \frac{1}{2}\delta_\mu^\nu R = -8\pi G T_\mu{}^\nu \tag{8.36}$$

$$F^{\mu\nu}{}_{;\nu} = 0 \tag{8.37}$$

$$F^{\mu\nu;\alpha} + F^{\nu\alpha;\mu} + F^{\alpha\mu;\nu} = 0 \tag{8.38}$$

Here $T_\mu{}^\nu$ is the energy-momentum tensor (2.134) of the electromagnetic field.

The spherically symmetric solution of these coupled equations leads to a spacetime interval of the general form of Eq. (7.60), but the functions N (r, t) and L(r, t) differ from those we found in the Schwarzschild case.

The electric field is in the radial direction, and the electromagnetic field tensor therefore has only two nonzero components,

$$F^{01} = -E(r), \qquad F^{10} = E(r) \tag{8.39}$$

The solution of Eqs. (8.36)–(8.38) can then be performed in a straightforward manner, as in the Schwarzschild case.

The results for the spacetime interval and the electric field are

$$ds^2 = \left(1 - \frac{2GM}{r} + \frac{G^2Q^2}{r^2}\right)dt^2 - \frac{dr^2}{1 - 2GM/r + G^2Q^2/r^2} - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \tag{8.40}$$

and

$$E(r) = \frac{Q}{r^2} \tag{8.41}$$

This solution contains two constants of integration, M and Q.

These constants must be interpreted as the mass and the electric charge [in Eqs. (8.40) and (8.41), these constants have been written in a form that anticipates this interpretation].

This interpretation becomes obvious if we compare Eq. (8.40) with the Schwarzschild solution, and if we compare Eq. (8.41) with the usual expression for the electric field of a spherical charge in flat spacetime.

Surprisingly, the expression (8.41) for the electric field in the curved spacetime has exactly the same inverse-square dependence on r as in flat spacetime, even in the strongly curved region near $r = 0$ (however, in curved spacetime, r does not represent the measured radial distance).

This inverse-square dependence is actually a direct consequence of Gauss's law.

The electric flux through all spherical surfaces r = constant must be the same.

Since the area of such a surface is $4\pi r^2$, as in flat spacetime, the electric field strength must be inversely proportional to r^2, as in flat spacetime.

The Kerr and Reissner-Nordstrøm solutions are special cases of a more general solution that represents the curved spacetime geometry surrounding a rotating electrically charged mass.

This is the Kerr-Newman solution.

In addition to the electric field, such a rotating charged mass has a magnetic field, with a magnetic dipole moment.

Remarkably, the ratio of the magnetic dipole moment to the spin angular momentum (the "gyromagnetic ratio") has the same form as the gyromagnetic ratio of the electron, Q/Mc.

8.5 Horizons and singularities of the rotating black hole

The Kerr geometry is much more complicated than the Schwarzschild geometry and depends drastically on whether $a < GM$, $a = GM$, or $a > GM$.

We will first discuss the case $a < GM$.

One of the least serious complications of the Kerr geometry is the presence of an infinite-redshift surface, where $g_{00} = 0$.

According to Eq. (8.24) this corresponds to

$$g_{00} = 1 - \frac{2GMr}{\rho^2} = 0 \tag{8.42}$$

that is,

$$1 - \frac{2GMr}{r^2 + a^2 \cos^2\theta} = 0$$

or

$$r = GM \pm \sqrt{G^2M^2 - a^2 \cos^2\theta} \tag{8.43}$$

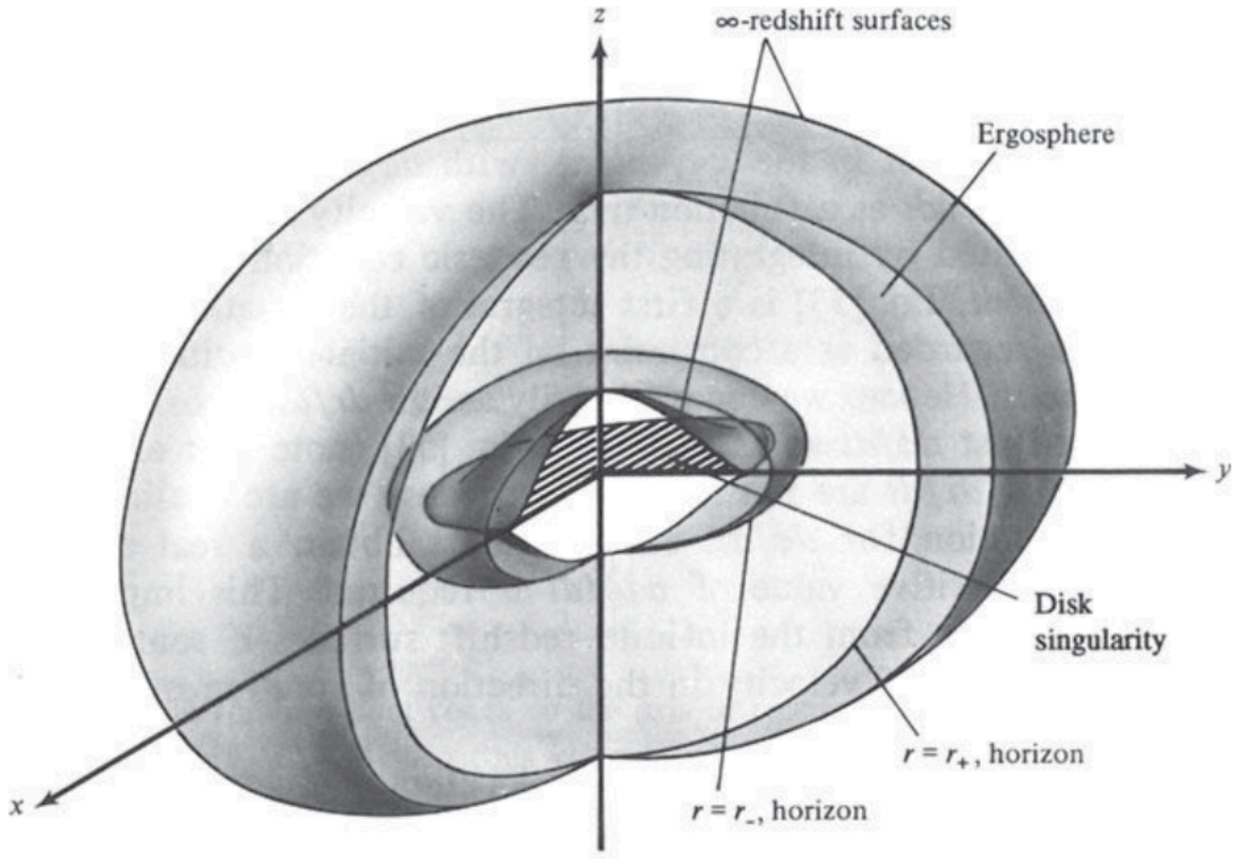
There are two distinct infinite-redshift surfaces; these infinite-redshift surfaces are drawn (in quasi-spheroidal coordinates) in Fig. 8.13.

As in the Schwarzschild case, these surfaces do not correspond to any physical singularity.

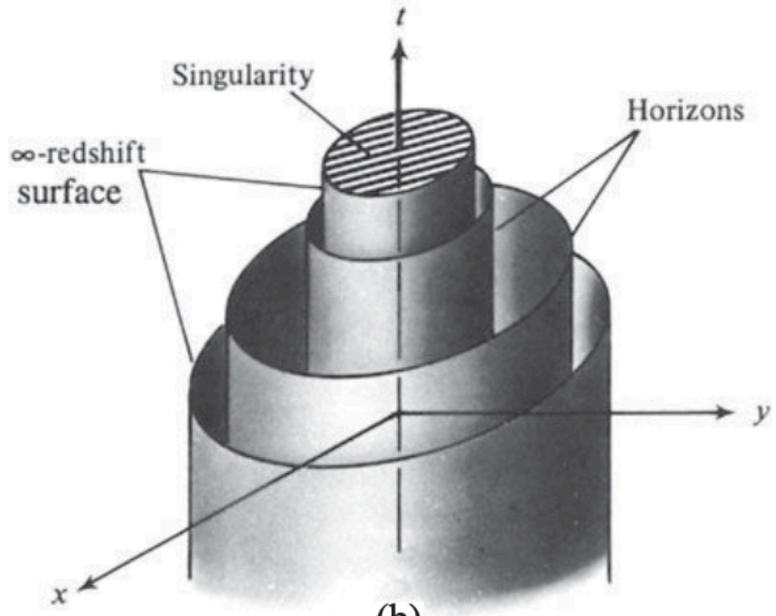
The vanishing of g_{00} merely tells us that a particle cannot be at rest (with $dr = d\theta = d\phi = 0$) at these surfaces; only a light signal emitted in the radial direction can be at rest.

In the region between these two surfaces, the value of g_{00} is negative.

Hence t is not a timelike coordinate, and the t-independence of the metric does not necessarily imply that the geometry is truly time independent.



(a)



(b)

Fig. 8.13 (a) The horizons and infinite-redshift surfaces of a rotating black hole with $GM = 1.2a$, represented in x, y, z coordinates. The horizons $r = r_+$ and $r = r_-$ are ellipsoids. Note that the infinite redshift surfaces do not coincide with the horizons. (b) The horizons and infinite-redshift surfaces of a rotating black hole represented in x, y, t coordinates.

At a radius $r < GM - \sqrt{G^2M^2 - a^2\cos^2\theta}$, the coordinate t again reverts to the character of a time coordinate.

In the Kerr geometry, the infinite-redshift surfaces do not coincide with event horizons.

We can calculate the (coordinate) velocity of a light signal by setting $ds^2 = 0$ in Eq. (8.24).

For a signal in the equatorial plane ($\theta = \pi/2$, $d\theta = 0$), we obtain

$$\left(\frac{dr}{dt}\right)^2 = \frac{\Delta}{\rho^2}\left\{1 - (r^2+a^2)\left[1 + \frac{2GMra^2}{\rho^2(r^2+a^2)}\right]\left(\frac{d\phi}{dt}\right)^2 - \frac{2GMr}{\rho^2} + \frac{4GMra}{\rho^2}\frac{d\phi}{dt}\right\} \tag{8.44}$$

If we evaluate this on the infinite-redshift surfaces, we find

$$\left(\frac{dr}{dt}\right)^2 = \frac{\Delta}{2GMr}\left[2a\frac{d\phi}{dt} - (r^2+a^2)\left(1 + \frac{a^2}{r^2+a^2}\right)\left(\frac{d\phi}{dt}\right)^2\right] \tag{8.45}$$

where it is understood that r has one of the values given by Eq. (8.43).

For a signal in the radial direction, $d\phi/dt = 0$, and hence $dr/dt = 0$; such a signal cannot escape.

For a light signal in the x-y plane, with $d\phi/dt/= 0$, Eq. (8.45) does not directly determine dr/dt as a function of r.

The velocity of the signal must then be calculated by integrating the geodesic equation so as to obtain dr/dt and $d\phi/dt$.

Equation (8.45) is a first integral of the equation of motion and may be regarded as a constraint on the initial velocities.

It is the only constraint.

Hence, we can arbitrarily assign dr/dt a positive initial value and adjust dϕ/dt so as to satisfy Eq. (8.45) (note that excessively large values of dr/dt are not acceptable because we are dealing with a quadratic equation for dϕ/dt, and we require a real solution).

According to Eq. (8.45), a and dϕ/dt must have the same sign.

This implies that signals can escape from the infinite-redshift surfaces only if sent outward with a component of velocity in the direction of rotation of the Kerr field.

Where are the event horizons?

Their location is given by the condition that dr/dt = 0 for arbitrary values of dθ and dϕ.

If we set $ds^2 = 0$ in Eq. (8.24) and solve for $(dr/dt)^2$, we obtain an expression similar to Eq. (8.45), but with some extra terms involving θ and dθ/dt.

However, Δ still appears as an overall multiplicative factor on the right side, just as it does in Eq. (8.45).

Hence the condition dr/dt = 0 becomes $\Delta = 0$,

$$r^2 - 2GMr + a^2 = 0 \tag{8.46}$$

or

$$r = GM \pm \sqrt{G^2M^2 - a^2} \tag{8.47}$$

It is customary to define

$$r_+ \equiv GM + \sqrt{G^2M^2 - a^2} \quad \text{and} \quad r_- \equiv GM - \sqrt{G^2M^2 - a^2} \tag{8.48}$$

At $r = r_+$ and at $r = r_-$ a light signal necessarily has zero velocity in the radial direction; thus, the light cones lie along the surfaces r = constant, and light cannot escape from these surfaces.

These horizons are shown in Fig. 8.13.

At the poles ($\theta = 0, \theta = \pi$), the infinite-redshift surfaces meet the horizons.

The one-way membrane $r = r_+$ defines the outer boundary of the rotating black hole; the one-way membrane $r = r_-$ is a second, inner boundary.

In the region between the two event horizons, $g_{11} > 0$ and hence r is a timelike coordinate.

As in the Schwarzschild case, the r-dependence therefore implies that the metric tensor in this region is necessarily time dependent.

Figures 8.14 (a) and 8.14 (b) show the light cones at different values of the radius: on the axis of rotation and in the equatorial plane, respectively.

These figures should be compared with Fig. 8.3 for the Schwarzschild case.

Neither at the infinite-redshift surfaces nor at the horizons does the Kerr geometry develop a true singularity; in local geodesic coordinates, the curvature remains finite.

The only true singularity occurs at $r = 0$; at this location the curvature tensor diverges.

To understand how this singularity arises, let us examine g_{00} in the vicinity of $r = 0$.

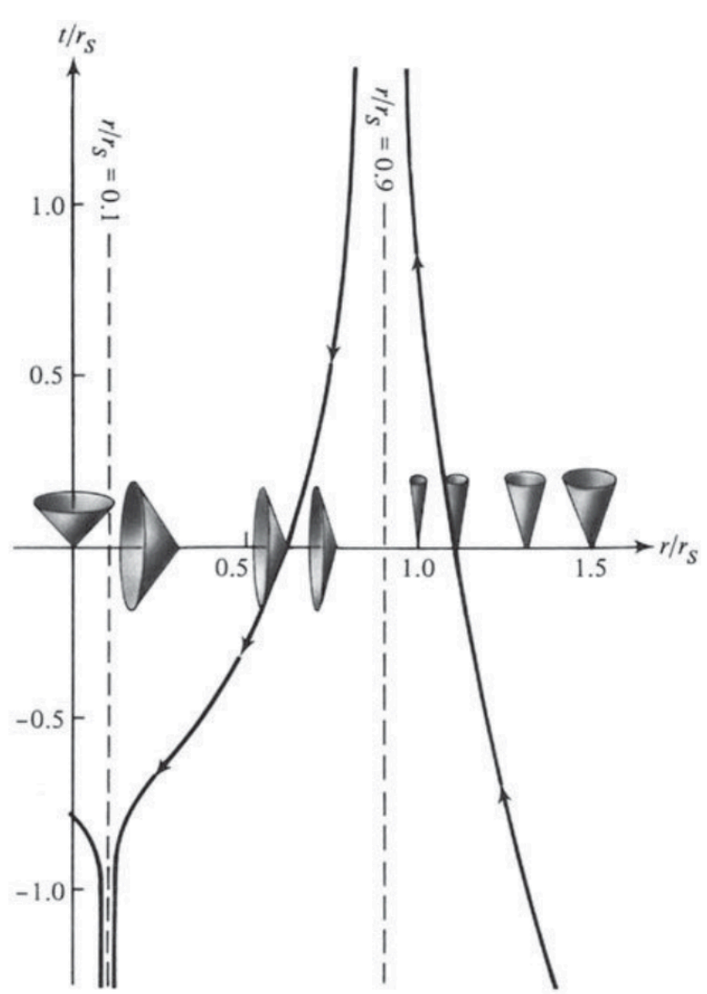
According to Eq. (8.24), we have

$$g_{00} = 1 - \frac{2GMr}{r^2 + a^2 \cos^2 \theta} \tag{8.49}$$

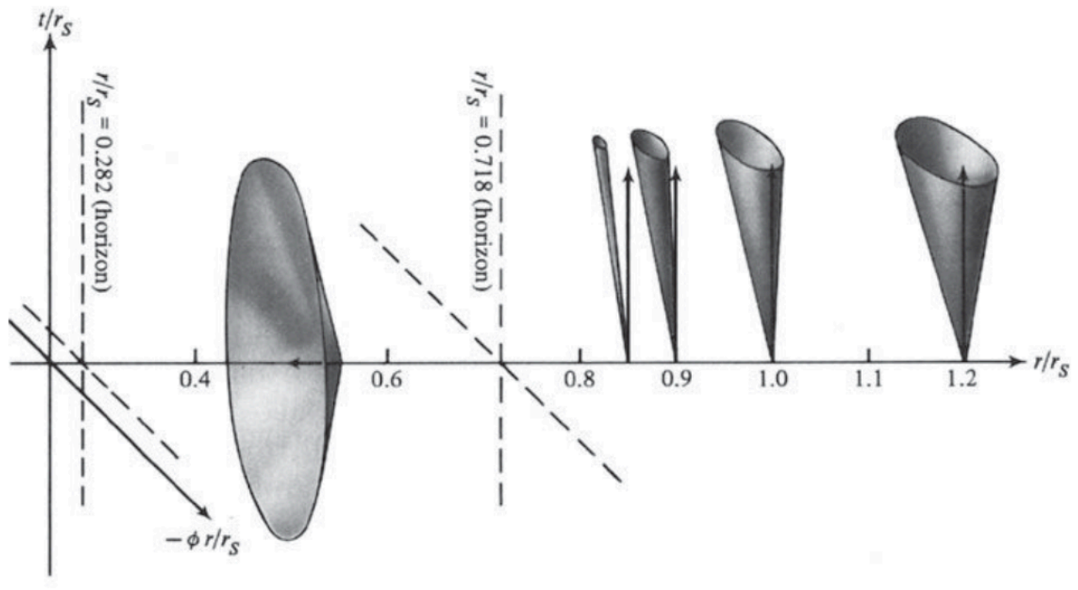
For $\theta = 0$, this expression can be regarded as a function of z,

$$g_{00} = 1 - \frac{2GM\,|z|}{z^2 + a^2} \tag{8.50}$$

In Fig. 8.15, this function is plotted vs. z.



(a)



(b)

Fig. 8.14 (a) The forward light cones on the axis of rotation near and inside a rotating black hole with $GM = 1.666a$ (this value of GM has been chosen for convenience in plotting; it has no special significance). The curve is the worldline of an ingoing light signal. (b) The forward light cones in the equatorial plane near and inside a rotating black hole with $GM = 2.222a$ (this different value of GM has been chosen for convenience in plotting; it has no special significance). The rotation is counterclockwise. The light cones are tilted in the direction of the rotation.

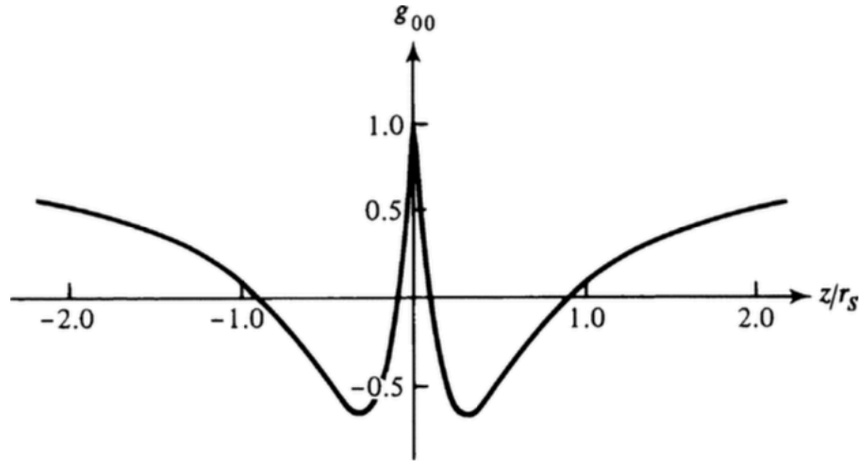


Fig. 8.15 The function g_{00} evaluated along the z-axis.

Note that at $z = 0$, $\partial g_{00}/\partial z$ is not zero.

This implies that the value of $\partial g_{00}/\partial z$ depends on how we approach the point $z = 0$.

If we approach from above ($z > 0$), then $\partial g_{00}/\partial z < 0$; if we approach from below ($z < 0$), then $\partial_{00}/\partial z > 0$.

Thus, $\partial g_{00}/\partial z$ has a discontinuity at $z = 0$.

It follows that the second derivative $\partial^2 g_{00}/\partial z^2$ does not exist at $z = 0$; more precisely, the second derivative diverges at $z = 0$.

The behavior of the second derivatives of g_{11}, g_{33}, and g_{03} is similar.

Since the curvature tensor involves these second derivatives, we conclude that the curvature tensor is singular at $r = 0$.

This singularity is a physical singularity, which cannot be blamed on our choice of coordinates; the spheroidal coordinates r, θ, ϕ and the metric tensor (8.24) are perfectly regular when $r = 0$ (except when $\theta = \pi/2$).

Note that in our spheroidal coordinates, $r = 0$ is a disk centered on the origin.

This can be seen from Eqs. (8.30)–(8.32); in rectangular coordinates, $r = 0$ corresponds to $x^2 + y^2 = a^2 \sin^2 \theta$, $z = 0$.

Since θ varies between 0 and $\pi/2$, this is equivalent to $x^2 + y^2 \leq a^2$, $z = 0$.

The edge of the disk is at $x^2 + y^2 = a^2$; that is, it is at the foci of the ellipses in Fig. 8.10.

If we take the limit $a \to 0$, then the disk singularity shrinks and ultimately reduces to the Schwarzschild singularity $r = 0$.

What is the physical meaning of the disk singularity?

Einstein's equations tell us that an infinite value of the curvature and of $R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R$ corresponds to an infinite value of the energy-momentum density.

Hence, the disk $r = 0$ contains an infinite density of matter.

The Kerr solution can then be interpreted as giving the spacetime geometry in the region surrounding a thin rotating disk of matter.

The Kerr solution also has an alternative interpretation without such a disk of matter, but this interpretation requires that our asymptotically flat space be joined, at the disk $r = 0$, to another asymptotically flat space.

We will deal with this extension of the Kerr geometry in Section 8.6.

The region between the outer surface of infinite redshift and the outer horizon (see Fig. 8.11) is called the *ergosphere*.

A procedure discovered by Penrose makes it possible to extract energy from the rotating black hole by means of particles that are moved in and out of the ergosphere.

The key to this procedure is that some orbits in the ergosphere have a negative total energy; that is, the gravitational binding energy exceeds the sum of rest-mass and kinetic energies.

We recall that the energy is related to the Lagrangian by $P_0 = \partial L/\partial u^0$.

With $L = m\sqrt{g_{\mu\nu}u^\mu u^\nu}$, we obtain

$$P_0 = \frac{m g_{0\alpha} u^\alpha}{\sqrt{g_{\mu\nu}u^\mu u^\nu}} = m g_{0\alpha} u^\alpha \tag{8.51}$$

For a freely falling particle, this energy is conserved, because the gravitational field is independent of x^0 [see the comments following Eq. (3.73)].

With the values of $g_{0\alpha}$ from Eq. (8.24), we find

$$P_0 = m\,(g_{00}u^0 + g_{03}u^3) \tag{8.52}$$

In the ergosphere, $g_{00} < 0$ and $g_{03} > 0$.

Furthermore, $u^0 > 0$ and $u^3 > 0$ (see Fig. 8.14 and note that u^μ must be in the forward light cone).

Hence the two terms in Eq. (8.52) are of opposite signs, and the energy will be negative if the first term dominates.

To extract energy from a black hole, we proceed as follows: We let a spacecraft fall from infinity into the ergosphere along an orbit of positive energy and there, by means of a spring-loaded device, eject a brick from the spacecraft in such a way that the brick enters an orbit of negative energy while the spacecraft, by recoil, enters an orbit of increased positive energy.

The brick then falls into the black hole, and the spacecraft returns to infinity.

Since the total energy of the system consisting of black hole, spacecraft, and brick is conserved, and the brick carries some amount of negative energy into the black hole, the spacecraft must carry a corresponding amount of positive energy away to infinity; that is, the spacecraft emerges with more energy than it had initially – the spacecraft extracts energy.

(Note that the energy budget for the spacecraft includes the rest-mass and the initial kinetic energy of the brick and the energy initially stored in the spring-loaded device; the Penrose process also converts these energies into kinetic energy of the spacecraft, but this conversion does not alter the net energy of the spacecraft and does not represent an energy extraction.)

The extra positive energy with which the spacecraft returns to infinity arises from the rotational kinetic energy of the black hole.

When the black hole captures the negative-energy brick, the rate of rotation of the black hole decreases slightly, and the mass of the black hole decreases slightly, by an amount equal to the extracted energy.

Christodolou and Ruffini have shown that there is an upper limit on the amount of energy that can be extracted from a black hole.

The limit is

$$\Delta M = M - M_{ir} \tag{8.53}$$

where M_{ir} is the *irreducible mass*

$$M_{ir} = \frac{1}{2}\left[\left(M + \sqrt{M^2 - a^2/G^2}\right)^2 + a^2/G^2\right]^{1/2} \tag{8.54}$$

or, more concisely,

$$M_{ir} = \frac{1}{2G}\sqrt{r_+{}^2 + a^2} \tag{8.55}$$

where $r_+ = GM + \sqrt{G^2M^2 - a^2}$ is the radius of the outer horizon of the Kerr black hole.

The upper limit given by Eq. (8.53) corresponds to a complete extraction of the rotational energy; we are then left with a Schwarzschild black hole of mass M_{ir}.

These results follow directly from *Hawking's theorem*: In any process involving one or more black holes, the total area of the horizon surfaces can never decrease.

The proof of this theorem uses some clever global topological techniques that are beyond the scope of our discussion.

We will not give the proof, but we will use the theorem.

In the case of the Kerr black hole, the area of the horizon r = r_+ is

$$A = 4\pi\left(r_+{}^2 + a^2\right) = 16\pi G^2 M_{ir}{}^2 \tag{8.56}$$

From Eq. (8.56) it is obvious that any process that changes the area *A* also changes the mass M_{ir}:

$$M_{ir}\delta M_{ir} = \frac{1}{32\pi G^2}\delta A \tag{8.57}$$

Hawking's theorem tells us that $\delta A \geq 0$, and consequently $\delta M_{ir} \geq 0$.

The black-hole mass *M* can be written in the form

$$M^2 = M_{ir}{}^2 + \frac{S^2}{(2GM_{ir})^2} \tag{8.58}$$

where $S = aM$ is the spin of the black hole.

Since M_{ir} must either remain constant or increase, Eq. (8.58) tells us that any decrease in *M* must come from a decrease of *S*; furthermore, Eq. (8.58) tells us that the minimum attainable value of *M* is M_{ir}.

Similar results apply to a Kerr-Newman black hole, that is, a black hole endowed with both spin angular momentum and electric charge.

The irreducible mass then includes a term involving the electric charge,

$$M_{ir} = \frac{1}{2}\left[\left(M + \sqrt{M^2 - a^2/G^2 - Q^2/G}\right)^2 + a^2/G^2\right]^{1/2} \tag{8.59}$$

However, Eqs. (8.55) and (8.56) remain valid, provided we redefine the horizon radius as $r_+ = GM + \sqrt{G^2M^2 - a^2 - GQ^2}$, whereas Eq. (8.58) is altered to

$$M^2 = \left(M_{ir} + \frac{Q^2}{2\sqrt{{r_+}^2 + a^2}}\right)^2 + \frac{S^2}{(2GM_{ir})^2} \tag{8.60}$$

Any decrease of M must now come from a decrease of S or Q, and the minimum attainable value of M is, again, M_{ir}.

The irreducible mass can be used to set limits on the energy released for processes involving not only one black hole but also collisions and mergers of several black holes.

For such processes, Hawking's area theorem implies that the sum of the squares of the irreducible masses of the final black holes must be at least as large as this sum of squares before the process.

The Penrose process is rather contrived, and we do not expect that it will occur spontaneously in nature.

However, there is a closely related process that is somewhat less contrived and might occur spontaneously.

This process involves the scattering of a wave – electromagnetic, gravitational, or whatever – by a black hole.

When such a wave is incident on a rotating black hole, the ergosphere selectively scatters the partial waves of different angular momenta (different multipoles) contained in the incident wave.

Calculations of the scattering indicate that the partial waves that have angular momentum in the same direction as the black hole are reinforced by the scattering, and they emerge with more energy than they had initially.

The extra energy arises from a slight reduction of the rate of rotation of the black hole.

This process is called *superradiance*.

The energy gain depends on the type of wave; for example, a gravitational wave can emerge with up to 39% more energy.

This increase of the intensity of a wave by scattering is analogous to the Klein paradox of relativistic quantum mechanics, according to which an electron wave incident on a potential step of height in excess of $2m_ec^2$ can be reflected with an increased amplitude.

The explanation is that the incident electron wave triggers the formation of electron-antielectron pairs when the height of the step is large enough to supply the energy for this pair formation.

Likewise, the quantum-mechanical explanation for superradiance is that the incident wave triggers pair formation in the ergosphere.

One member of the pair can enter an orbit of negative energy and the other an orbit of positive energy in such a way that the total energy is conserved.

The former then falls into the black hole while the latter escapes to infinity.

The net result is that particles steadily stream out of the ergosphere while the energy of the black hole steadily decreases.

Note that the energy loss of the black hole is not due to particles emerging from the horizon; rather, it is due to the capture by the black hole of negative-energy particles that originate in the ergosphere.

8.6 Maximal Kerr geometry

In Section 8.3 we saw that the Schwarzschild geometry, spanned by the *r*-*t* coordinates, can be extended to a maximal Schwarzschild geometry, spanned by two sets of *r*-*t* coordinates (see Fig. 8.6).

This maximal Schwarzschild geometry is motivated by the criterion that any geodesic that is not intercepted and terminated by a singularity should be continued indefinitely.

We can also apply this criterion to the Kerr geometry, and we find that this geometry has a maximal extension that comprises not only two asymptotically flat universes but also an infinite sequence of pairs of asymptotically flat universes.

Figure 8.16 shows a part of this infinite sequence.

To fit all the universes on the page, the asymptotically flat regions have been compressed so as to appear finite.

This compression has been accomplished by a conformal coordinate transformation that maps the infinite spacetime into a finite region, but leaves the worldlines of light as straight lines at 45° with respect to the axes.

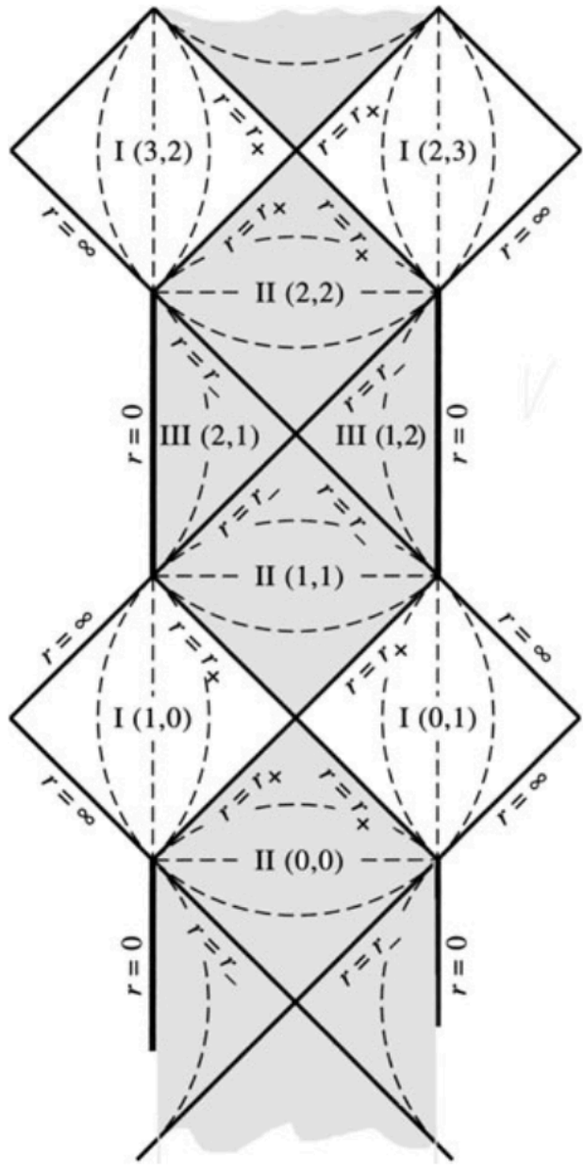


Fig. 8.16 Conformal representation of the maximal Kerr geometry, with $a < GM$. The repeated black or white holes are shown shaded. The infinite, asymptotically flat regions outside these black holes have been compressed into the square boxes on the left and the right of the diagram.

For comparison, Fig. 8.17 shows the Kruskal diagram subjected to the same kind of conformal transformation.

Such diagrams, showing the compressed spacetime, are called *Penrose diagrams*.

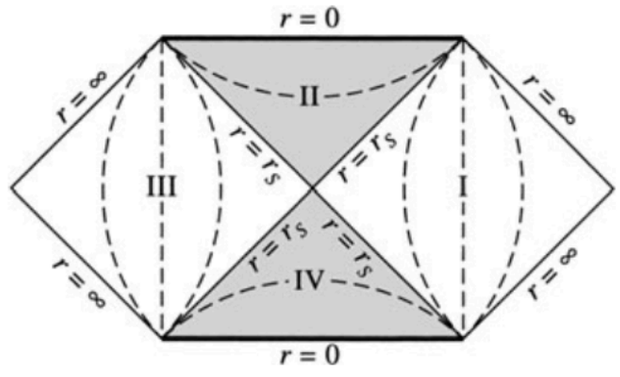


Fig. 8.17 Conformal representation of the maximal Schwarzschild geometry. The two infinite, asymptotically flat regions outside the black hole have been compressed into the two square boxes on the left and the right of the diagram. The dashed lines roughly indicate surfaces of constant *r*.

The squares labeled I(m,n) in Fig. 8.16 are asymptotically flat universes in which the coordinate *r* lies in the region $r_+ < r < \infty$ (in this region, *r* is spacelike).

The squares labeled II(k,k) correspond to the region $r_- < r < r_+$ that lies between the inner and outer horizons (in this region, *r* is timelike).

As in the Schwarzschild case, II(1,1) is a black hole as seen from the universes I(1,0) and I(0,1), whereas II(0,0) is a white hole.

The triangles III(m,n) correspond to the region $0 < r < r_-$ inside the inner horizon.

Note that there are two singularities r = 0 in this diagram.

Both of these singularities are timelike – that is, they happen at a fixed place, for an extended time – in contrast to the Schwarzschild singularity in Fig. 8.17, which happens at one time over an extended region of space (however, as we will see later, this picture of the Kerr singularities is subject to an important correction, and the physically admissible extension of the Kerr solution and the actual physical singularities are quite different from the naıve mathematical construction shown in Fig. 8.16).

The inner horizon r = r_ is called a *Cauchy horizon*, because it is the boundary between a region II(1,1) in which the initial-value problem (or Cauchy problem) for the Einstein equations has a well-defined solution and a region III(1,2) where the intial-value problem has no well-defined solution.

The trouble is that for any point above r = r_, the past light cone intercepts the singularities at $r = 0$.

It is therefore not sufficient to provide initial-value data on, say, the $t = 0$ horizontal surface cutting across the asymptotically flat exterior universes (from r = ∞ on the right to r = ∞ on the left); we also need boundary-value data at the singularities.

Although we have not generated our Kerr solution from initial-value data, the Cauchy horizon is troublesome, because we would like to think of the Kerr solution as generated by the gravitational collapse of a mass, and we expect that there should be an initial-value problem describing this collapse.

Once the collapse reaches the stage where the horizon r = r_ is formed, any further calculation of the evolution of the geometry by step-by-step integration of an initial-value problem becomes impossible – we need extra, unknown information about boundary conditions at the singularities.

Even worse, the Cauchy horizon r = r_ is unstable.

Small, random perturbations traveling into the Kerr black hole from the outside build up to large amplitudes and divergent blueshifted frequencies.

The reason for this weird divergence can be understood by considering an astronaut who travels into the Kerr black hole and, after crossing the outer horizon $r = r_+$, approaches the inner horizon r = r_ and attempts to cross it. The astronaut takes only a short proper time to reach this inner horizon.

However, as can be seen by inspection of the spacetime diagram (Fig. 8.18), *all* the signals or perturbations that travel from the exterior region into the black hole from now until the end of our external time will intercept the worldline of the astronaut.

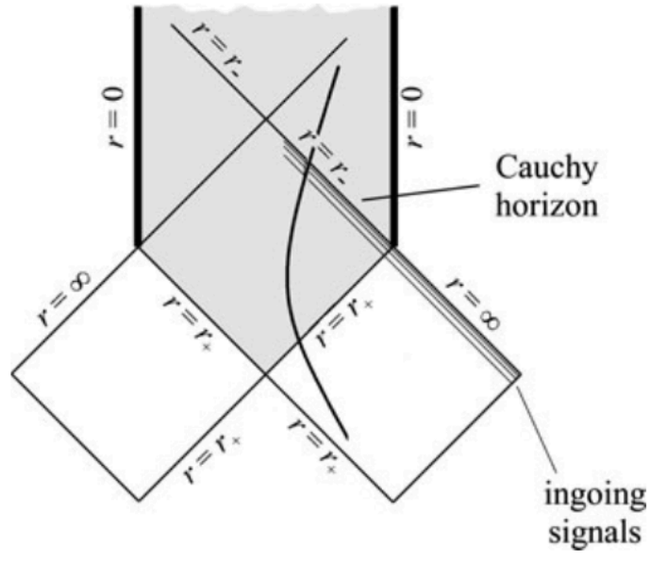


Fig. 8.18 Worldline of an astronaut crossing the Cauchy horizon $r = r_-$.

At the inner horizon, the astronaut receives a catastrophic overload of signals in one instant – "The inner horizon marks the last moment at which our astronaut can still receive news; but then she gets all of the news".

The compression of a long span of emission times into a short (proper) reception time implies an enormous, divergent blueshift and a compression of superposed amplitudes of signals into a short instant, with a large accumulation of amplitudes.

This accumulation of amplitudes will occur even if we shield the black hole from stray waves arriving from its environment.

The gravitational collapse that forms the black hole in itself generates gravitational waves, some of which will travel inward and accumulate near $r = r_-$.

Thus, the instability of the inner horizon is unavoidable.

Because of this instability, the inner horizon is equivalent to a gravitational singularity, and geodesics cannot be extended across it.

The Kerr solution effectively ends at the inner horizon $r = r_-$, and the multitude of extra universes pictured in Fig. 8.16 above the inner horizon are a mathematical fiction.

The net result is that the topology of the extended Kerr solution is essentially the same as that of the extended Schwarzschild solution: two asymptotically flat universes, cut off above and below by a singularity.

The only difference is that whereas the Schwarzschild singularity is spacelike, the Kerr singularity, along the inner horizon $r = r_-$, is lightlike.

The timelike disk singularities $r = 0$ of the Kerr solution are off limits – they lie beyond the inner horizon.

So far, our discussion of the Kerr geometry has dealt only with the case $|a| < GM$.

In the case $|a| = GM$, the two horizons coincide (see Fig. 8.19), and in the case $|a| > GM$, there is no horizon.

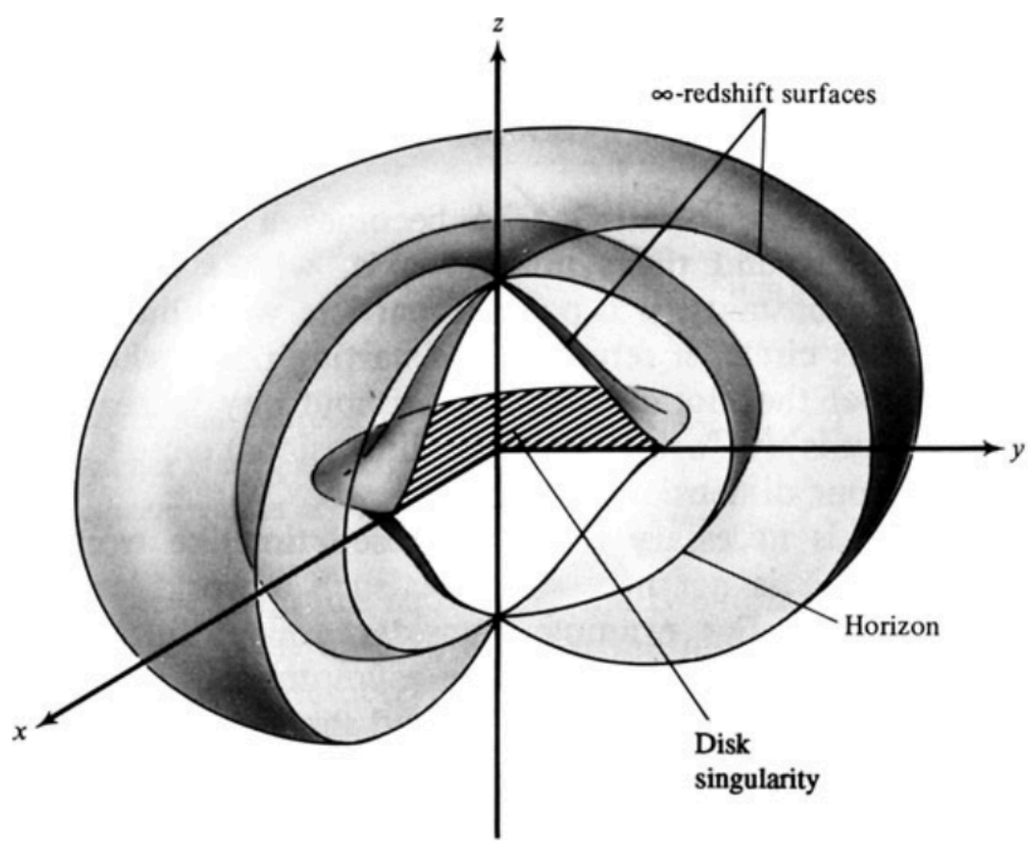


Fig. 8.19 Horizons and infinite-redshift surfaces for a rotation black hole with $a = GM$.

Of course, there still is a disk singularity at $r = 0, \theta = \pi/2$.

A singularity not surrounded by a horizon is called a *naked singularity*.

A rotating "hole" of large spin, $|a| > GM$, has such a naked singularity.

Figures 8.20 and 8.21 show the conformal diagrams for these two cases.

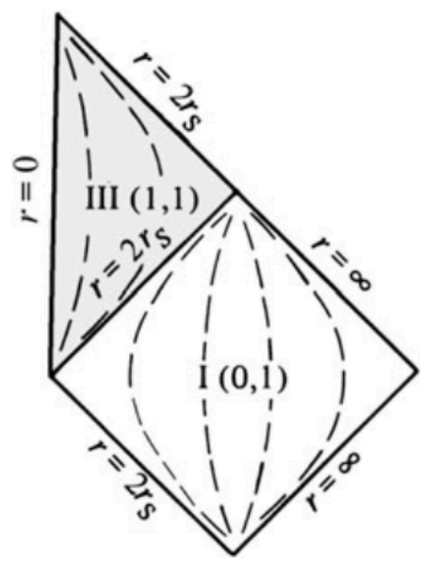


Fig. 8.20 Conformal representation of the maximal Kerr geometry, with $|a| = GM$. The horizon $r = 2r_S$ slanting upward to the left is afflicted with the instability problem discussed earlier and the diagram has been terminated there.

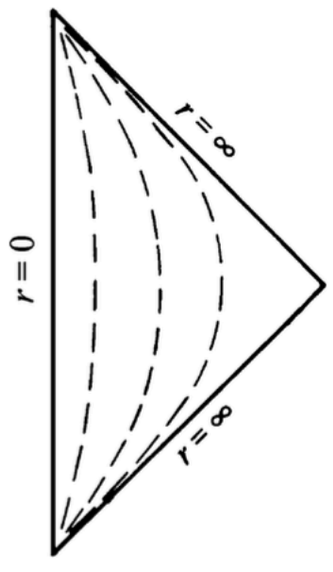


Fig. 8.21 Conformal representation of the maximal Kerr geometry, with $|a| > GM$. There is no horizon, and the singularity at $r = 0$ is exposed to view.

Although naked singularities do exist mathematically, as solutions of the Einstein equations, we do not know whether they are ever found in the real world.

Penrose has conjectured that the complete gravitational collapse of a (nonsingular) mass never results in a naked singularity, but rather in a singularity hidden within a horizon.

This conjecture is known as the *cosmic censorship conjecture*.

This conjecture does not forbid the existence of naked singularities in the mathematical solutions of Einstein's equations; it forbids only the formation of such a naked singularity in gravitational collapse.

The cosmic censorship conjecture remains unproved.

Some attempts have been made to construct counterexamples to the conjecture, by numerical integration of the Einstein equations for the gravitational collapse of various mass distributions.

For instance, the gravitational collapse of an elongated (cigar-shaped) mass distribution appears to generate a naked singularity.

However, the numerical integration becomes inaccurate and breaks down when the curvature tensor becomes large, and this makes it impossible to establish beyond all reasonable doubt that the curvature does indeed become infinite and that no horizon develops at some future time.

The Kerr solution for $|a| \geq GM$ has a naked singularity, but the formation of this singularity in collapse is probably prevented by an instability: the solution is probably unstable for $|a| > GM$, and a collapsing mass with a large spin probably breaks up into several pieces during collapse (whether the Kerr solution is in fact stable or unstable is not yet known).

The Reissner-Nordstrom solution with a large electric charge ($Q^2 > GM^2$) also has a naked singularity.

But a spherical mass with $Q^2 > GM^2$ is indeed unstable, since the large electric repulsive force overcomes the attractive gravitational force, and this means that the mass will not collapse into the Reissner-Nordstrøm configuration.

From a practical point of view, the complete Kerr geometry is as irrelevant to physics as the complete Schwarzschild geometry.

If a rotating black hole forms by gravitational collapse, then the interior region will be occupied by matter, and the pure vacuum Kerr solution does not apply.

Hence the plurality of universes in the Kerr geometry has nothing to do with reality unless a rotating black hole was, somehow, created when the universe began.

And even if such an object was created, it is unlikely that it has survived; as we discussed earlier, the inner horizon of the Kerr geometry is violently unstable.

Finally, note that in the case of collapse of a spherical, nonrotating star, the Schwarzschild solution fails to hold in the interior, but the Birkhoff theorem guarantees that it will at least hold in the exterior. In the case of collapse of a rotating star, there is no similar theorem, and the exterior solution need not be that of Kerr.

Only a long time after the collapse, when everything becomes stationary, will the exterior solution tend toward the Kerr solution.

8.7 Black-hole thermodynamics; Hawking process

The horizon, or one-way membrane, of a black hole acts as a perfect absorber – it permits anything to enter, but does not permit anything to leave.

We therefore expect that it could serve as an ideal heat sink for the operation of a thermodynamic engine.

We expect that when we dump heat on the horizon, in the form of thermal radiation at some given temperature, it will be completely absorbed.

With such an ideal heat sink, we can operate a thermodynamic engine with 100% efficiency; that is, we can accomplish the complete conversion of heat into work.

The following Gedankenexperiment describes a simple machine that attempts to accomplish this goal.

At large distance from a black hole, fill a box with thermal radiation at some temperature T (see Fig. 8.22).

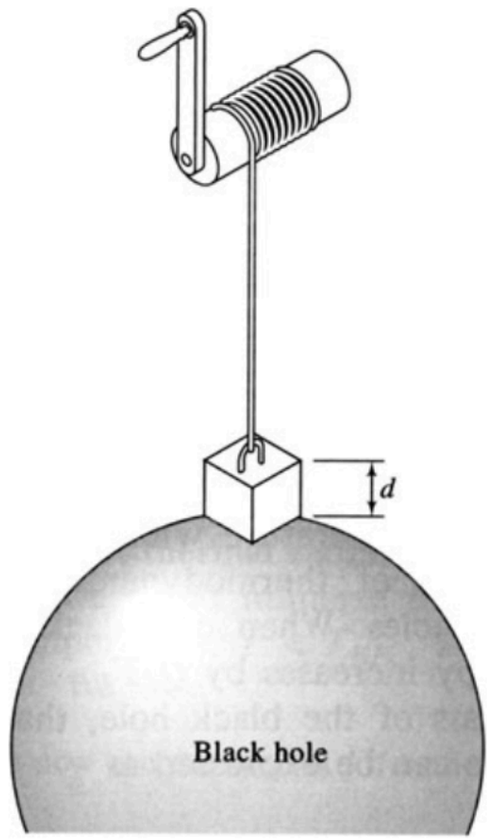


Fig. 8.22 A box attached to a winch is filled with thermal radiation and then lowered until its bottom reaches the horizon of the black hole.

Close the box and slowly lower it toward the horizon, by means of a rope attached to a winch.

As the box descends, the gravitational potential energy of the radiation is converted into useful mechanical work (the changes in the gravitational potential energies of the box and the rope are irrelevant, since we will ultimately have to reverse these changes when we lift the box and rope back to their initial positions).

At the horizon of a Schwarzschild black hole, the gravitational potential energy of a mass m of radiation is $-mc^2$ [see the comments following Eq. (7.118)]; that is, all of the energy of the radiation will have been converted into useful work.

We now open the bottom of the box and dump the radiation into the black hole.

Then we raise the empty box to its initial position and thereby complete one cycle of the operation of our thermodynamic engine.

The net result of this cycle would seem to be the complete conversion of heat into mechanical work.

Such a complete conversion would lead us to the conclusion that the thermodynamic temperature of the black hole is exactly zero.

However, this argument neglects to take into account the size of the box.

Since thermal radiation of temperature T typically has a wavelength of about $\hbar\,c/kT$ (by Wien's law), a box that contains such thermal radiation must have a size d at least as large as the typical wavelength, $d \cong \hbar\,c/kT$.

When the bottom of the box reaches the horizon, the center of gravity of the box will be at a distance of $d/2$ above the horizon, and the potential energy of the radiation will be $-mc^2 + mgd/2$, where g is the acceleration of gravity at the surface of the black hole.

From Newtonian theory, we can estimate this acceleration of gravity as $g = GM/r_S^2 = c^4/4GM$ (by coincidence, the exact relativistic value of g agrees with this Newtonian value).

The mechanical work we gain by dumping the radiation is therefore not mc^2, but

$$W \cong mc^2 - m\frac{c^4}{4GM}\frac{d}{2} \cong mc^2\left(1 - \frac{\hbar c^3}{8GMkT}\right) \tag{8.61}$$

The efficiency of our thermodynamic engine is then at most

$$e = \frac{W}{mc^2} \cong 1 - \frac{\hbar c^3}{8GMkT} \tag{8.62}$$

According to the definition of the thermodynamic temperature, this efficiency can be expressed as $1-T_{BH}/T$, where T_{BH} is the temperature of the black hole.

Theformula (8.62) therefore leads to an approximate estimate for the temperature of the black hole,

$$T_{BH} \cong \frac{\hbar c^3}{8GMk} \tag{8.63}$$

The exact result [see Eq. (8.68)] includes an extra factor of π in the denominator.

Note that this temperature is directly proportional to the surface gravity, $g = GM/r_S^2 = c^4/4GM$, of the black hole.

Speculations about the temperature of a black hole first emerged from work by Bekenstein, who proposed to associate an entropy with the black hole, so that the laws of thermodynamics could be applied to processes involving black holes.

When our black hole absorbs an amount of heat δQ, its entropy increases by $\delta Q/T_{BH}$.

The heat δQ absorbed equals the increase of mass of the black hole, that is, $\delta Q = \delta Mc^2$.

Hence, if we include the extra factor of π, the entropy increase can be expressed as

$$\delta S = \frac{\delta Mc^2}{T_{BH}} \cong \frac{8\pi kGM}{\hbar c}\delta M \tag{8.64}$$

Alternatively, this can be expressed in terms of the change of the surface area of the black hole, $\delta A = 8\pi r_S \delta r_S = 32\pi G^2 M\delta M/c^4$, so

$$\delta S \cong \frac{kc^3}{4\hbar G}\delta A \tag{8.65}$$

This simple proportionality between δS and δA indicates that the entropy of a black hole must be proportional to its surface area,

$$S \cong \frac{kc^3}{4\hbar G} A \tag{8.66}$$

The formation of a black hole therefore entails a large increase of entropy.

Such an increase can be made plausible by the information-theoretic interpretation of entropy.

When the black hole forms, or whenever we increase the size of the black hole by dumping any kind of matter into it, we lose information about the trapped matter.

This loss of information corresponds to an increase of entropy.

By means of the entropy, we can formulate the first law of thermodynamics for a black hole in the usual way: The increase δMc^2 in the internal energy equals the sum of the heat $\delta Q = T_{BH} dS$ absorbed by the black hole and the mechanical work performed on the black hole by external forces.

Since the black hole behaves like a rigid body, the only way to increase its internal mechanical energy is by spinning it up by an external torque.

In this case $\delta W = \Omega\, \delta J$ where Ω is the angular velocity of the black hole and J its spin angular momentum.

Hence the first law of thermodynamics becomes

$$\delta Mc^2 = T_{BH}\delta S + \Omega \delta J \tag{8.67}$$

The second law of thermodynamics for black holes is a direct consequence of Hawking's theorem for the increase of the surface area in any process involving one or more black holes.

Since the entropy is proportional to the surface area, Hawking's theorem ensures the increase of entropy.

In the preceding discussion, the black-hole temperature T_{BH} was treated as a thermodynamic temperature, and it is not immediately clear to what extent it can be treated as a real temperature.

But it easy to see that it must be regarded as a real temperature for thermal radiation; that is, the black hole must emit thermal radiation of this characteristic temperature.

If there were no such thermal radiation, we could immerse the black hole in a bath of radiation of lower temperature, $T < T_{BH}$, and the black hole would then absorb this radiation without emitting any.

Thus, while the radiation bath loses an amount of entropy $\delta Q/T$, the black hole would gain only a smaller amount of entropy $\delta Q/T_{BH}$, in violation of the second law of thermodynamics.

The direct explicit proof that black holes emit thermal radiation was given by Hawking, who discovered that in the curved spacetime of a black hole radiation is generated by a quantum process.

Hawking found an elegant way to calculate the spectrum of the radiation from the behavior of quantum fields in curved spacetime.

Earlier attempts at such calculations foundered on the ambiguities of what boundary conditions to apply at the horizon and at the central singularity.

Hawking cleverly bypassed these ambiguities by tracing the evolution of the quantum fields in time, from an initial, well-defined vacuum state before the beginning of gravitational collapse and before the formation of the black hole.

He demonstrated that when the black hole forms and settles into its final stationary state, the quantum fields settle into a state that involves a steady outward emission of radiation from the horizon toward infinity.

The energy spectrum of this radiation is thermal, with a temperature

$$T_{BH} = \frac{\hbar c^3}{8\pi GMk} \tag{8.68}$$

This exact result for the temperature of the black hole is consistent with the approximate result (8.63) obtained from approximate thermodynamic arguments.

The Hawking emission process seems to contradict the fundamental property of the horizon – nothing should emerge from the horizon.

Actually the thermal radiation does not come from inside the black hole, but it is created by quantum fluctuations at or near its surface.

We know from quantum field theory that the vacuum is a restless and violent place, where particles are continuously created and destroyed.

Figure 8.23 illustrates such a creation and destruction event: A particle-antiparticle pair, such as an electron-antielectron pair, is spontaneously created at a spacetime point, and this pair is then destroyed at a later point.

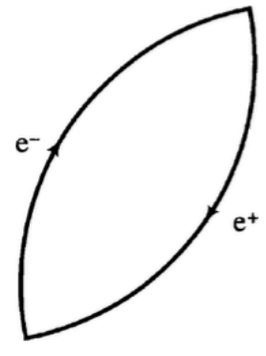


Fig. 8.23 Feynman diagram illustrating the spontaneous creation of a particle-antiparticle pair and its subsequent annihilation.

If this event happens in the normal vacuum, far from a black hole, it merely produces an unobservable, small-scale fluctuation in the electric current density.

But if this event happens just outside the horizon of a black hole, the antiparticle (or the particle) might enter the horizon and fall into the singularity.

If the antiparticle is in a state of negative energy $-E$, the particle will be left behind in a state of positive energy $+E$, and it will then be free to move outward, away from the black hole, and reach the detection instruments of an observer.

The net effect is that a more or less steady stream of particles of positive energy flows outward from the region of the horizon, while the black hole absorbs a stream of antiparticles of negative energy, and therefore decreases its mass.

An explicit calculation by Damour and Ruffini shows that the thermal spectrum of the emerging, liberated, particles arises from a "barrier-penetration" factor.

The region in the immediate vicinity of the horizon strongly attenuates the ingoing antiparticle wave (in the calculation, the ingoing antiparticle wave of negative energy is treated as an outgoing particle wave of positive energy proceeding backward in time; see Fig. 8.24).

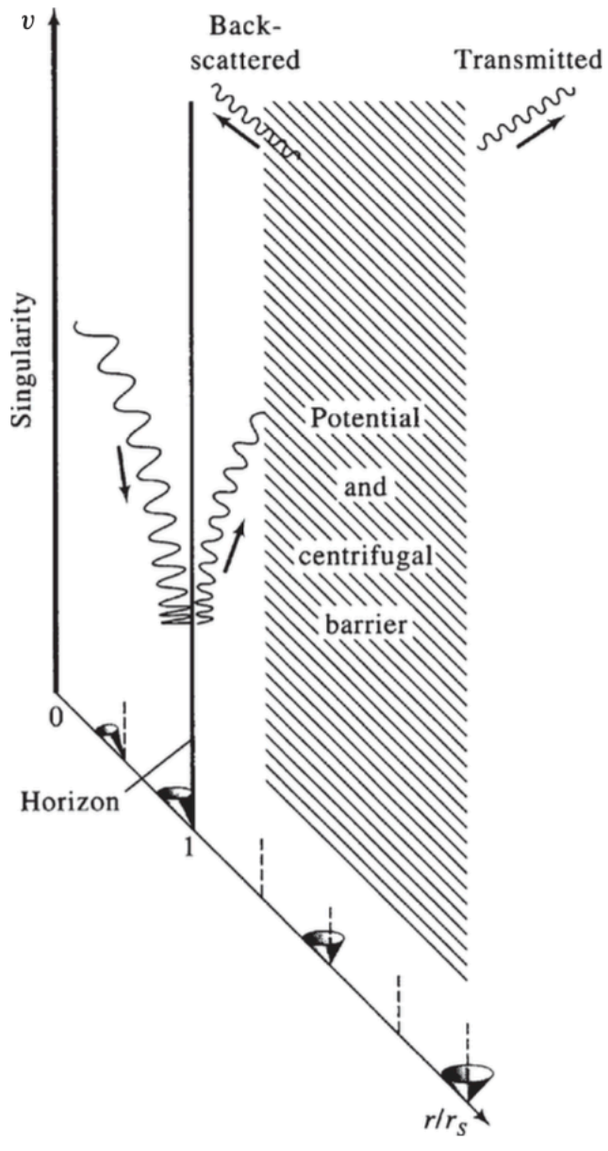


Fig. 8.24 In this diagram, the spacetime inside and outside a black hole is described by Eddington-Finkelstein coordinates. A particle-antiparticle pair is created spontaneously at the horizon. The antiparticle falls into the black hole, and the particle travels away. An antiparticle of negative energy can be treated as a particle of positive energy proceeding backward in time, from the singularity to the horizon. The net wavefunction, indicated schematically, is a wave of positive energy that proceeds from the singularity to the horizon, and from there to infinity. Near the horizon, this wave goes through an infinite number of oscillations, and its amplitude is reduced by a penetration factor. (Damour and Ruffini, 1976).

In the limiting case of large energy (E ≫ $\hbar c^3$GM), it turns out that the attenuation suffered by the antiparticle wave in crossing the horizon is $e^{-4\pi GME/\hbar c^3}$, and thus the probability for the antiparticle to penetrate into the black hole, and leave the particle liberated, is $e^{-8\pi GME/\hbar c^3}$.

From quantum field theory, it is known that all quantum states contribute equally to the particle-antiparticle vacuum fluctuations, so the flux of antiparticles incident on the horizon is proportional to the number of quantum states, that is, $4\pi p^2 dp/h^3$ (for the purposes of this simplified calculation, we ignore spin states).

If we multiply this incident flux by the probability for penetration of the horizon, we obtain the outgoing flux of liberated particles:

$$\text{outgoing flux of particles} \propto \frac{4\pi p^2 dp}{h^3} e^{-8\pi GME/\hbar c^3} \tag{8.69}$$

As expected, in the limiting case of large energy (E ≫ $1/8\pi$ GM $\hbar$ c^3), this has the form of a thermal spectrum. The penetration factor $e^{-8\pi GME/\hbar c^3}$ plays the role of the Boltzmann factor $e^{-E/kT}$, with a temperature $T = T_{BH} = \hbar c^3/8\pi GMk$. If the energy is not large compared with kT_{BH}, then the penetration factor is somewhat more complicated and depends on whether the particles are fermions or bosons; but in any case, the resulting spectrum is thermal.

Actually, not all the particles liberated at the horizon manage to escape to infinity.

Some of these particles are backscattered by the gravitational potential surrounding the black hole, and they disappear into the black hole.

This merely means that the black hole and its surrounding gravitational field form a thermal radiator with a less than perfect emissivity, a thermal “gray” body rather than a thermal blackbody.

The backscattering modifies the spectrum of the radiation that finally escapes to infinity, reducing it considerably below the spectrum of a perfect blackbody (see Fig. 8.25).

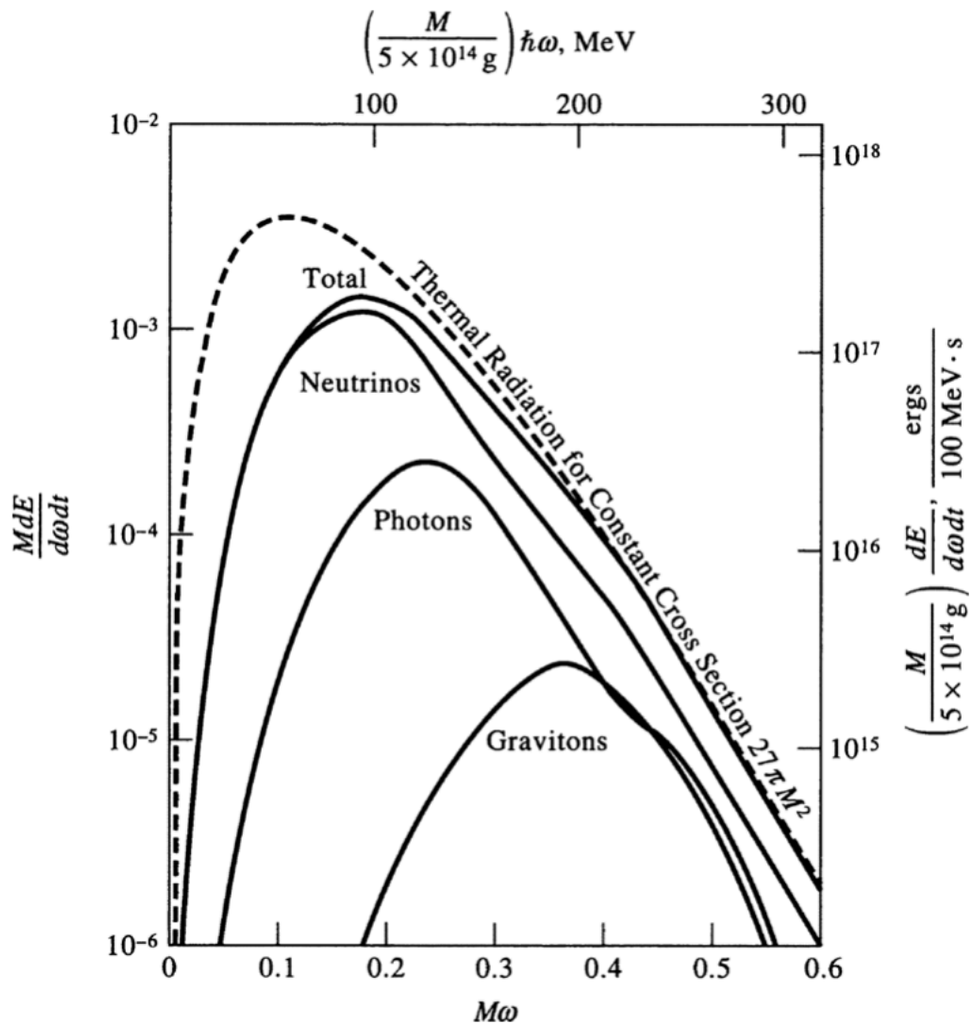


Fig. 8.25 Spectral distribution for different kinds of thermal radiation emitted by a black hole. The scales on the left and the bottom edges of the plot are expressed in units with $G = \hbar = c = 1$ (Page, 1976a).

Numerically, the black-hole temperature can be expressed as

$$T_{BH} = (1.2 \times 10^{26}\,\mathrm{K}) \times \frac{1}{(M/1\mathrm{g})} \tag{8.70}$$

Thus, the temperature of a black hole of a mass equal to that of the Sun ($M = 2 \times 10^{33}$ g) is only 10^{-7} K.

But a very small black hole, or minihole, of a mass of 10^{14} g, would have a temperature of 10^{12} K, and the typical thermal energy of the radiated particles would be about $kT_{BH} \cong 100$ MeV.

In general, the typical thermal energy determines what kinds of particles can be radiated.

Particles of rest mass m cannot be radiated in significant numbers unless the typical thermal energy is of the order of mc^2.

Thus, a black hole of mass equal to that of the Sun can radiate only particles of mass zero, that is, photons, neutrinos, and gravitons; but a minihole of mass 10^{14}g can also radiate electrons, mu mesons, and pions.

Black holes of masses significantly smaller than a solar mass cannot be formed by the gravitational collapse of a star; such miniholes can form only in the early stages of the universe, from fluctuations in the very dense primordial matter.

Figure 8.26 lists the different kinds of particles radiated by a black hole as a function of its mass.

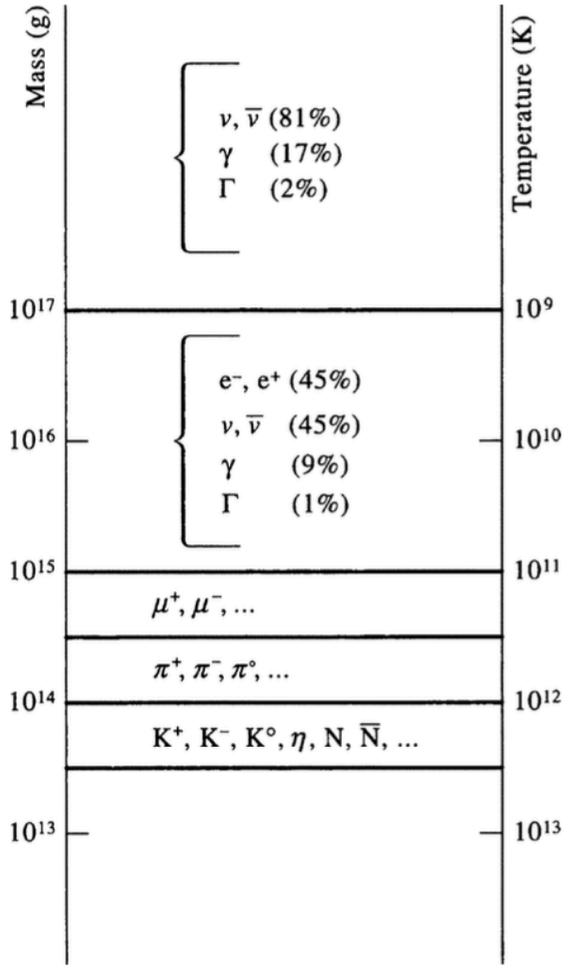


Fig. 8.26 Different kinds of particles radiated by a black hole as a function of its mass.

By Stefan's law the total power radiated by a black hole is of the order of

$$P \cong \text{area} \times \sigma {T_{BH}}^4 \cong 4\pi {r_S}^2 \times \sigma {T_{BH}}^4 \cong \frac{10^{47}\ \text{erg/s}}{(M/1\text{g})^2} \tag{8.71}$$

where σ is the Stefan-Boltzmann constant.

The exact numerical factor in this equation depends on the various species of particles being radiated and on their backscattering.

From Eq. (8.71) we can calculate the rate of decrease of the mass of the black hole:

$$\frac{dM}{dt} = -\frac{P}{c^2} \cong -\frac{10^{26}\ \text{g/s}}{(M/1\text{g})^2} \tag{8.72}$$

As the mass decreases, the rate of radiation increases.

When the mass becomes small, the rate of radiation becomes explosive.

However, it is not known what happens when the mass reaches a value smaller than $\approx \sqrt{\hbar c/G} \cong 10^{-5}$ g (the Planck mass), where the time scale for the change of mass will be smaller than the Schwarzschild time, that is, $|dM/dt| \geq M/(r_S/c)$.

Under these conditions, the spacetime geometry cannot be treated as a stationary or quasistationary background, and the dynamical changes in the geometry will begin to play a crucial role.

Although it is widely believed that the minihole will release all its mass in a final explosion and then disappear, the details are not understood.

From Eq. (8.72) we can estimate the lifetime of a black hole.

If the initial mass is M, the time until it disappears or almost disappears is approximately

$$t \cong \frac{M}{|dM/dt|} \cong 10^{-26} \times (M/1\text{g})^3 \tag{8.73}$$

According to this estimate, a black hole of initial mass 10^{14}g would have a lifetime of about 10^{10} years, equal to the age of the universe.

Thus, such a black hole, formed during the early stage of the universe, would be reaching its final explosive phase today.

The observational search for radiation from black holes has concentrated on gamma rays.

A black hole of 10^{14}g spends most of its lifetime at a temperature of about 10^{12} K, and at this temperature it produces gamma rays of about 100 MeV.

If there is a more or less uniform distribution of such black holes all over the universe, we should see a diffuse background of 100-MeV gamma rays all over the sky.

The observational limit on gamma rays sets an upper limit on the density of black holes of about 10^4 per $(\text{pc})^3$.

A tighter limit has been set by examining the sky for localized bursts of high-energy gamma rays, 1,000 MeV or more, produced during the final explosive phase of a black hole.

Such observations set an upper limit of 0.04 explosion per $(\text{pc})^3$.

These upper limits on the abundance of miniholes imply that their contribution to the overall mass density in the universe is small – no more than 1 part in 10^8 of the total mass is in the form of miniholes.

Although the negative observational evidence for radiation from miniholes is disappointing, we can draw some interesting conclusions from the absence of such miniholes.

The preponderance of normal matter over miniholes tells us that conditions in the early universe were not favorable for the formation of miniholes.

Since the miniholes arise from fluctuations, this means that either the fluctuations were not very violent or else the primordial matter offered strong resistance to compression (had a “stiff” equation of state, that is, high pressure as function of density).

8.8 Gravitational collapse and formation of black holes

In normal stars, such as the Sun, the inward gravitational pull is held in equilibrium by the thermal pressure of the gas.

This thermal pressure will be sufficient to resist the gravitational pull only if the star is hot enough.

The star can therefore remain in equilibrium as long as the thermonuclear reactions in its core supply enough heat; that is, as long as the energy released in these reactions compensates for the energy lost by radiation at the surface.

In a star that has exhausted its supply of nuclear fuel, the thermal pressure will ultimately disappear, and the star will collapse under its own weight.

The collapse may be sudden (implosion) or gradual (contraction), but in any case it can be halted only if an alternative mechanism for generating sufficient pressure becomes available at high density.

In white-dwarf stars and in neutron stars such an alternative mechanism is available:

These stars are so dense that the quantum mechanical zero-point pressure becomes dominant.

Essentially, a degenerate Fermi gas of electrons, packed into the lowest available quantum-mechanical energy states, supplies the equilibrium pressure in a white dwarf and a Fermi gas of neutrons that in a neutron star.

At white-dwarf densities ($\varrho \geq 10^5$ g/cm^3) the electrons are detached from their nuclei and move quite freely throughout the volume of the star.

The star consists of interpenetrating gases of electrons and nuclei.

The zero-point pressure of the electron gas gives the main contribution to the pressure, and the nuclei give the main contribution to the mass density.

The equation of state (pressure as function of density) based on this model permits equilibrium configurations, provided that the total mass is below a critical upper limit.

If the mass exceeds this critical limit, then the electron pressure cannot support the star.

The trouble is that the Fermi energy increases with density, and ultimately the electron gas becomes relativistic; such a relativistic Fermi gas, in contrast to the nonrelativistic gas, cannot supply enough pressure for equilibrium.

The following rough calculation helps make this clear.

Let us pretend that the gas has uniform density throughout the star.

The Fermi momentum for a gas of N electrons in a volume V is

$$p_F = (3\pi^2\hbar^3 N/V)^{1/3} \tag{8.74}$$

In the nonrelativistic case, the Fermi energy equals $p_F^2/2m_e$ (where m_e is the electron mass), whereas in the extreme relativistic case, the Fermi energy equals cp_F.

The average energy per electron is of the same order of magnitude as the Fermi energy, and hence we may regard the quantity

$$\begin{matrix} N(3\pi^2\hbar^3 N/V)^{2/3}/2m_e & \text{(nonrelativistic)} \\ N(3\pi^2 c^3\hbar^3 N/V)^{1/3} & \text{(extreme relativistic)} \end{matrix} \tag{8.75}$$

as a rough estimate of the total zero-point energy.

The Newtonian gravitational binding energy is of the order of magnitude of

$$-\frac{GM^2}{R} \tag{8.76}$$

where M is the mass of the star and R the radius.

Hence, apart from rest masses, the total energy in the volume $V = 4\pi R^3/3$ is roughly

$$\begin{aligned} &\frac{N^{5/3}(9\pi\hbar^3/4)^{2/3}}{2m_e R^2} - \frac{GM^2}{R} \quad \text{(nonrelativistic)} \\ &\frac{N^{4/3}(9\pi c^3\hbar^3/4)^{1/3}}{R} - \frac{GM^2}{R} \quad \text{(extreme relativistic)} \end{aligned} \tag{8.77}$$

We must now try to adjust the density, or the radius R, so that a star of given N and M is placed in an equilibrium configuration.

For stable equilibrium, the energy (8.77) should have a minimum as a function of R.

In the nonrelativistic case, a minimum exists.

But in the extreme relativistic case, there is no minimum, because both the zero-point energy and the gravitational energy are proportional to the same power of R.

Essentially, the trouble is that, although in the extreme relativistic case the Fermi energy increases with decreasing R, it does not increase strongly enough – the relativistic gas is too soft to resist the gravitational compression.

It is straightforward to show that the energy (8.77) has a minimum at

$$R = \frac{N^{5/3}(9\pi\hbar^3/4)^{2/3}}{GM^2 m_e}$$

in the nonrelativistic case and that it has no minimum in the extreme relativistic case

The critical mass at which the equilibrium disappears may be estimated by asking where the electron gas turns relativistic.

This happens at $p_F \cong m_e c$.

According to Eq. (8.74), this corresponds to

$$R = \frac{(9\pi\hbar^3 N/4)^{1/3}}{m_e c}$$

By comparing Eqs. (8.78) and (8.79), we find a critical mass

$$M_{crit}{}^2 \cong \frac{cN^{4/3}(9\pi\hbar^3/4)^{1/3}}{G}$$

If the nuclei in the white dwarf are helium, then $N \cong M_{crit}/2m_n$, where m_n is the neutron mass.

According to Eq. (8.80), this results in a critical mass

$$M_{crit} \cong \sqrt{\frac{9\pi}{4^3}} \left(\frac{\hbar c}{G m_n{}^2}\right)^{3/2} m_n$$

Numerically, $\sqrt{9\pi/4^3}\,(\hbar c/G m_n{}^2)^{3/2}\, m_n = 2.4 \times 10^{33}\ \mathrm{g} \cong 1\ M_\odot$.

Thus, our rough calculation indicates that the critical mass should be of the order of one solar mass.

For a realistic model of white dwarfs, with nuclear matter in the form of helium nuclei and with a pressure distribution appropriate for hydrostatic equilibrium, a more precise calculation of the critical mass gives a value of $1.44 M_\odot$.

This result was first obtained by Chandrasekhar, and the critical mass, beyond which the zero-point pressure of electrons becomes insufficient for equilibrium, is known as the *Chandrasekhar limit*.

The density of a neutron star is much higher than that of a white dwarf.

The density of a neutron star is as large or larger than nuclear densities ($\approx 2 \times 10^{14}$ g/cm^3), and hence the star may be described as a single giant nucleus.

The bulk of the star consists of a Fermi gas of neutrons, with a small admixture of protons and electrons.

The ratio of neutrons to protons and electrons is determined by beta-decay and inverse beta-decay reactions.

At high density, the equilibrium of these reactions favors neutrons, because the electrons have a high Fermi energy, which encourages their combination with protons and the formation of neutrons.

Near the center of the neutron star there is perhaps a small inner core with heavy particles (hyperons).

The zero-point pressure of the degenerate gas of neutrons holds the star in equilibrium.

Like the white dwarf, the neutron star has a critical value of the mass, beyond which the zero-point pressure of the neutrons becomes insufficient for equilibrium.

Our rough estimate for the critical mass of a white dwarf also applies to a neutron star.

Note that the mass of the electron, or the mass of whatever particle generates the zero-point pressure, cancels in Eq. (8.80), so the values of the critical masses of white dwarfs and neutron stars should be about the same, about 1 $M_\odot$.

But when we take into account the pressure distribution required for hydrostatic equilibrium and the effects of nuclear forces, these critical masses will differ.

According to a first calculation by Oppenheimer and Volkoff, in which the nuclear forces between neutrons were neglected, the critical mass of a neutron star is about 0.7 $M_\odot$.

Calculations by Wheeler and colleagues sought to take into account the nuclear forces and found a slightly larger critical mass.

Various other extrapolations of nuclear forces to neutron-star densities later gave critical masses of up to 2.7 $M_\odot$.

The most recent, more sophisticated, quantum-mechanical calculation adapted by Ruffini and colleagues from the Fermi-Thomas method of atomic physics yields the same critical mass, about 2.7 $M_\odot$.

In contrast to the traditional calculations – which treated the neutrons, protons, and electrons in a neutron star as a hydrostatic equilibrium configuration of a fluid characterized by an equation of state and a local density derived from quantum statistics – the Fermi-Thomas model relies on a more direct quantum-mechanical description of the degenerate particle distribution by the local value of the Fermi momentum, that is, the momentum of the highest occupied quantum state.

This new calculation has led to some surprising insights, such as the presence of a strong electric field at the interface between the neutron star's core and the outer crust (see Fig. 8.27).

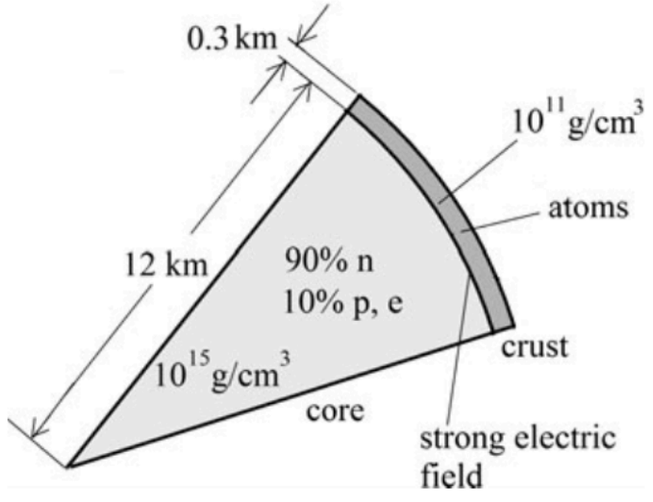


Fig. 8.27 Schematic diagram showing the structure of an angular segment in the midplane of a neutron star of mass 1.35 $M_\odot$.

The core of the neutron star has a density of about 10^{15}g/cm^3, and it consists of about 90% neutrons and 10% protons and electrons (by mass).

The electron density in the core is about 10^{37}/cm^3, and at such a high density, the electron Fermi energy is large, about 100 MeV.

This exceeds the magnitude of the electron gravitational potential energy (about – 0.1 MeV), so electrons tend to escape from the core.

But such an incipient escape leaves the core with a positive electric charge, and the resulting large electric field ($\approx 10^{19}$ V/cm) quickly blocks the continued outflow of electrons.

The Fermi-Thomas calculation imposes a constraint of overall, global neutrality on the electric charge distribution, but – in contrast to the hydrostatic calculations – it does not impose a constraint of local neutrality and therefore permits a local charge separation and an associated electric field.

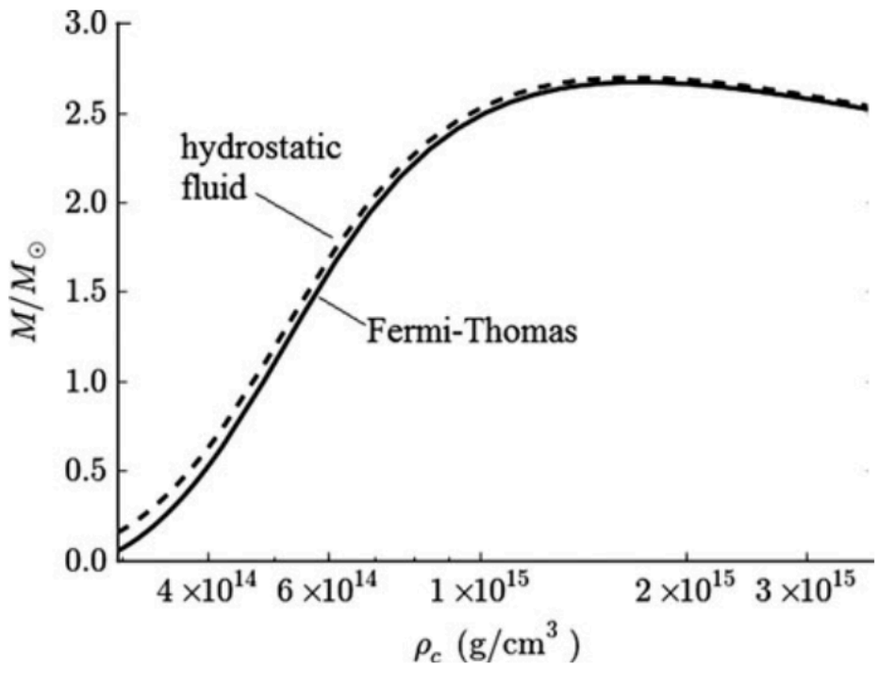
The calculation shows that the resulting electric field is concentrated within a thin layer at the core-crust boundary, and this electric field binds the electrons to the core.

Outside of the core, in the crust, the binding of electrons involves a different mechanism.

There, the electrons are bound in individual neutral atoms in the same way as electrons are bound in samples of ordinary materials.

Figure 8.28 gives a plot of the mass of a neutron star as a function of central density (the central density ϱ_c is a convenient parameter for describing neutron stars, because this value serves as the starting point for the numerical integration that generates the Fermi-Thomas particle distribution and the total mass).

Fig. 8.28 Mass as a function of central density for the possible equilibrium configurations of neutron stars, according to the Fermi-Thomas model (solid line) and according to a traditional hydrostatic equilibrium calculation. The maximum mass is about 2.7 $M_\odot$ (Belvedere et al., 2012).



This figure includes the results of calculations according to the Fermi-Thomas model and also according to a traditional hydrodynamic model.

Both of these calculations assume that the matter has reached the endpoint of thermonuclear evolution, so the star is completely burnt out and cold.

In this context, "cold" merely means that the thermal pressure is negligible compared with the degeneracy pressure – even a temperature of 10^9K would be regarded as cold for a neutron star.

Note that the high-density configurations in Fig. 8.28 (on the right side of the plot, beyond the maximum mass) are unstable.

Their energy decreases with increasing central density, which means that such neutron stars will collapse under their own weight.

Figure 8.28 shows that the results for the masses calculated from the Fermi-Thomas and hydrostatic models differ only slightly.

However, the results for the radius of the neutron star differ significantly.

The Fermi-Thomas calculation gives a smaller radius than the hydrostatic calculation (see Fig. 8.29).

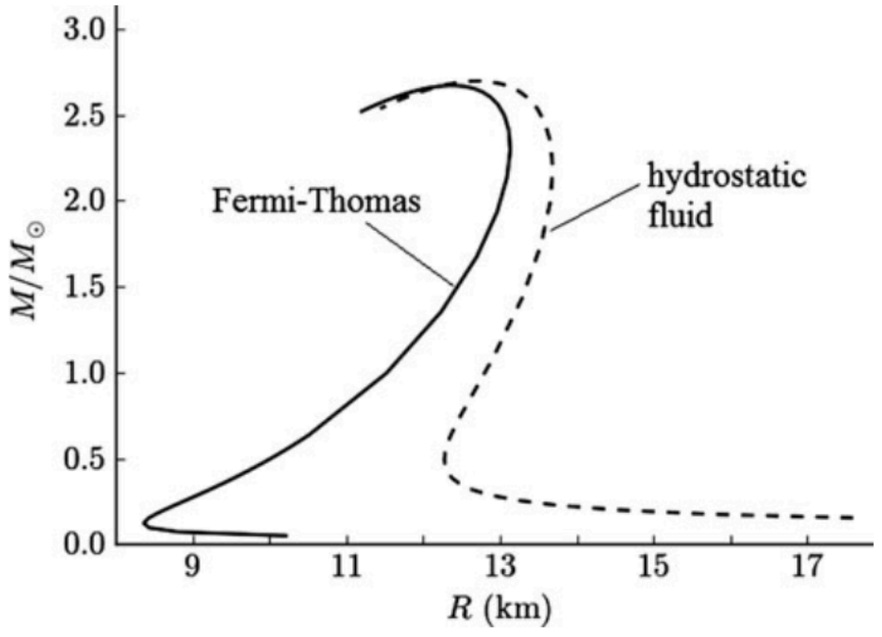


Fig. 8.29 Radius of a neutron star as a function of mass. The sizes calculated from the Fermi-Thomas model are significantly smaller than those calculated according to hydrostatic equilibrium (Belvedere et al., 2012).

The difference arises mainly from the presence of an extra layer of crust (the "inner crust"), which in the hydrostatic model provides a gradual transition from the high density of the core to the low density of the outer crust.

In the Fermi-Thomas model, this inner crust is absent, and the transition occurs suddenly, at the core-crust boundary, where the strong electric field acts as a "wall" that separates the high- and low-density regions.

With the exception of the early calculations of Oppenheimer-Volkoff and Harrison-Wheeler, all calculations of the critical mass of neutron stars give values somewhere above 1 $M_\odot$.

The highest values of the critical mass arise from including a hard repulsive core in the nuclear interactions, which makes the nuclear material very stiff at high densities and resistant to collapse.

But regardless of how much the equation of state at high densities is modified by nuclear interactions, it can be shown that there always is an upper limit to the stiffness and to the mass that can be supported.

The reason for this is that the stiffness of a material is directly related to the speed of sound waves in the material – the stiffer the material, the higher the speed of sound.

The requirement that the speed of sound be no more than the speed of light then sets an upper limit on the stiffness of the material.

According to Rhoades and Ruffini, this leads to the conclusion that the mass of a neutron star can never exceed 3.2 $M_\odot$, independently of any assumption about the details of the equation of state at high densities.

The absence of equilibrium configurations with masses exceeding these limits has drastic consequences for stellar evolution.

A star with a mass below the Chandrasekhar limit evolves along a sequence of equilibrium configurations.

When the nuclear fuel is exhausted it settles down in a white-dwarf state.

The luminosity of such a white dwarf is entirely due to the residual thermal energy.

As the energy is lost by radiation, the star grows dimmer and dimmer, ultimately becoming a “black star” (the cooling time is typically of the order of 10^9 years).

In contrast, when the nuclear fuel is exhausted in a star with a mass above the Chandrasekhar limit, the pressure will become insufficient to prevent gravitational collapse.

At late stages of its evolution, such a massive star will have a dense core, and the collapse begins in this core – the core implodes.

If the mass of the core is sufficiently small, then the implosion can be halted when neutron-star densities are reached.

The sudden halt of the implosion creates a violent outward shock wave, which carries away the kinetic energy of the infalling material.

This shock wave is analogous to the shock wave observed in a water pipe when the flow of water is suddenly halted by quickly shutting off a valve; the kinetic energy of the moving water is then transferred to a shock wave that travels backward along the pipe, halting the flow of water ("water hammer").

As the shock wave of the imploding star travels outward, it blows away the outer layers of the star.

The outward explosion of these layers shows up on the sky as a supernova (a type II supernova).

For example, the famous Crab pulsar is a neutron star that was formed in such a supernova explosion; the Crab nebula, which surrounds it, consists of hot gases that once constituted the outer layers of the original massive star (of a mass of $\approx 4\ M_{\odot}$).

If the mass of the collapsing core exceeds the critical mass for a neutron star, then the implosion cannot be halted.

The complete gravitational collapse of the core produces a black hole.

Supernova explosions are one possible mechanism for the formation of black holes.

Mass exchange in a compact binary star system is another mechanism.

In such a system, a neutron star is in an orbit near the surface of a more or less normal companion star.

If the neutron star is near enough to the companion, it will pull material off its surface and gradually increase in mass.

Once the mass increases beyond the critical limit, the neutron star will collapse to a black hole.

The first calculations of complete gravitational collapse, with formation of a Schwarzschild horizon, were performed by Oppenheimer and Snyder, who assumed that the collapsing material consists of a sphere of gas without pressure (a cloud of dust), so each particle is in free fall along a radial geodesic.

Although this is an unrealistic assumption, their calculation illustrates the general behavior of the space-time geometry during the approach to the horizon.

In a sense, the collapse never ends.

An outside observer would say that the surface of the collapsing star asymptotically approaches the Schwarzschild radius, but never reaches it in a finite time.

In this sense, the object may be appropriately called a “frozen star” rather than a black hole.

However, the “frozen star” very soon becomes practically indistinguishable from a black hole.

Calculations of the intensity of the light emitted by the surface of a collapsing star show that the brightness seen by an outside observer decreases sharply while the redshift increases (the brightness decreases exponentially with time, with a characteristic half-life of the order of GM).

Furthermore, the gravitational and electromagnetic fields surrounding the collapsing mass asymptotically approach those corresponding to a black hole.

The evolution of the initial gravitational and electromagnetic fields into the final black-hole fields will in general be accompanied by radiation, because the collapsing mass has to shed any gravitational or electromagnetic multipole fields in excess of those characteristic of a black hole.

For example, detailed calculations show that if the initial fields deviate slightly from the spherically symmetric Schwarzschild field, then these multipole fields are radiated away as the collapse proceeds.

Some of the fields are radiated downward, into the black hole; some are radiated outward, forming a pulse of gravitational and electromagnetic radiation.

Within a finite time, the electromagnetic and gravitational fields surrounding the collapsing mass approach so closely to those characteristic of a black hole that the quantum uncertainties in the measurement process prevent an outside observer from detecting any difference at all.

When the external gravitational and electromagnetic fields of a collapsing mass ultimately reach a stationary state, the fields will be uniquely characterized by three parameters: mass (M), spin angular momentum (S), and electric charge (Q).

This general theorem was established by a sequence of several complementary theorems on the uniqueness of static and stationary black-hole solutions of Einstein's equations.

For instance, the Kerr solution is the unique black-hole solution with $S \neq 0$, $Q = 0$; the Reissner-Nordstrøm solution is the unique black-hole solution with $S = 0$, $Q \neq 0$; and the Kerr-Newman solution is the unique solution with $S \neq 0$ and $Q \neq 0$ (as already mentioned in Section 8.4, the Kerr solution and the Reissner-Nordstrøm solution are special cases of the Kerr-Newman solution).

As far as an outside observer is concerned, any two black holes with the same mass, angular momentum, and charge are therefore absolutely identical, regardless of how the two black holes were created.

When a black hole engulfs some matter, only its mass, angular momentum, and charge change.

All other properties of matter, such as intrinsic multipole moments, baryon number, lepton number, and so on, are "forgotten" by the black hole.

If we drop anything into a black hole, there is no way of telling from the fields surrounding the black hole at a later time what was dropped.

The absence of distinctive fields – such as independent gravitational or electromagnetic multipole fields, or nuclear force fields – in the region surrounding a black hole led Wheeler to remark that "a black hole has no hair".

Although the collapse of a mass never quite ends for an outside observer, it does end very soon for an observer who rides along with the collapsing mass.

Within a finite proper time, this observer reaches the central singularity.

In the special case of spherical symmetry or axial symmetry, the existence of the Schwarzschild singularity ($r = 0$) or of the Kerr singularity ($r = 0, \theta = \pi/2$) follows directly from the solution of Einstein's equations.

In the general case, the existence of singularities is indicated by the Hawking-Penrose theorem, which roughly says that a singularity will develop whenever the spacetime contains a closed, two-dimensional, spacelike, *trapped surface*.

By a trapped surface is meant a surface such that the wavefronts generated by wavelets of light emitted from the surface become smaller, regardless of whether light propagates in the inward or the outward direction.

In flat spacetime, there can be no such trapped surface.

For example, Fig. 8.30a shows wavelets propagating inward and outward from a surface; the wavefront propagating inward has a smaller area than the original surface, but the wavefront propagating outward has a larger area.

In contrast, Fig. 8.30b shows what can happen in a strongly curved spacetime; both the wavefronts propagating inward and outward have smaller areas.

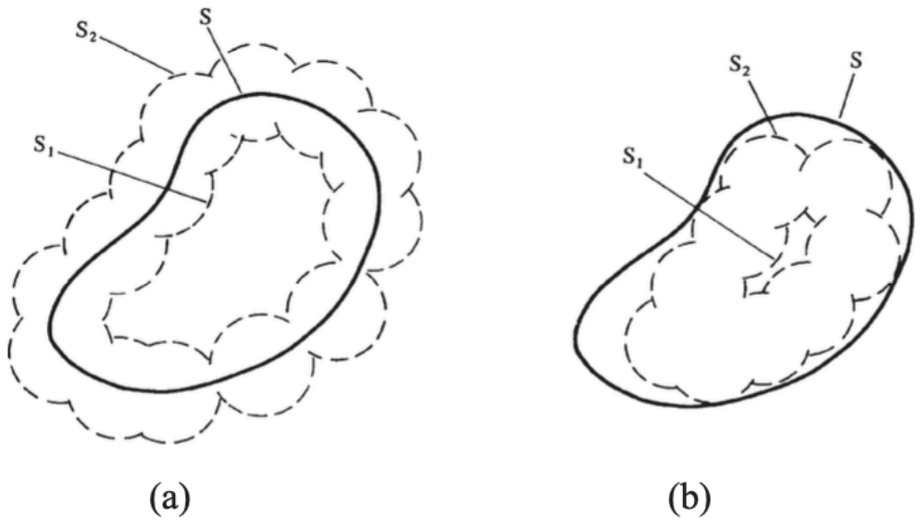


Fig. 8.30 The solid line represents a closed, 2-D surface. The broken lines show wavefronts of emitted light at a later time. In (a), the surface is not trapped; in (b) it is trapped.

In the Schwarzschild geometry, the spherical surface r = constant $< r_S$, $0 \le \theta \le \pi, 0 \le \phi \le 2\pi$ is a spacelike trapped surface.

Wavefronts emitted from this surface, in the inward or outward directions, necessarily move toward smaller values of r (see Fig. 8.30), and therefore their area $4\pi r^2$ decreases.

It is known that trapped surfaces will form in all cases of gravitational collapse that are reasonably close to spherical collapse, and it is believed that trapped surfaces will form even if the collapse is very different from a spherical collapse.

The Hawking-Penrose theorem then tells us that singularities will also form.

What is the nature of these singularities?

In the context of this theorem, a singularity is simply a region in which geodesics terminate.

What is not known is how they terminate.

The most obvious possibility is that the curvature becomes infinite; this is, of course, exactly what happens at $r = 0$ in the Schwarzschild geometry.

In any case, the presence of a singularity signals the breakdown of classical physics, and quantum effects will then presumably play a crucial role.

The presence of quantum fields spoils the Hawking-Penrose theorem, because this theorem hinges on special restrictive conditions imposed on the energy-momentum tensor, such as the "energy-condition" $T^{00} + T/2 \geq \Lambda/8\pi G$.

Although this condition is readily satisfied by the energy-momentum tensor of an ordinary fluid and by electromagnetic fields, it is not generally satisfied by quantum fields on small length scales.

It is tempting to conjecture that the Hawking-Penrose theorem of classical geometrodynamics is as irrelevant to the real world as the Earnshaw theorem of classical electrodynamics, according to which atoms are unstable and must collapse to a singularity.

It may well be that when quantum effects are taken into account, the collapsing matter attains some nonsingular final state.

If a singularity develops, it is expected to be accompanied by the formation of a horizon around the singularity; that is, the formation of a black hole.

This expectation is based on Penrose's cosmic censorship conjecture (see Section 8.6), which forbids the formation of naked singularities, that is, singularities not surrounded by horizons.

For an outside observer, the horizon that forms around the singularity then hides the singularity from view.

8.9 In search of black holes

Because black holes are an end product of stellar evolution, they could be quite abundant.

It is even possible that in some galaxies most of the mass is found in the form of black holes.

For example, in elliptical galaxies, the observed ratio of total mass to luminosity is about 70 times that of the Sun.

Since the Sun is a fairly typical star, most of the mass in these galaxies must be "dark mass" of very low luminosity.

Several forms of dark mass have been hypothesized (see Table 10.3), among them black holes.

The observed light distribution and the observed velocity dispersion in bright elliptical galaxies may be used to set limits on the amount of mass that can be in the form of black holes.

If there is a large central black hole, then its mass cannot exceed $10^{10}\ M_{\odot}$.

However, a much larger amount of mass could exist in the form of a large number of smaller black holes, of about $10^{3}\ M_{\odot}$ each, distributed over the galaxy, or maybe a much larger number of black holes of a mass much smaller than $1\ M_{\odot}$.

An isolated black hole does not emit light and would not be directly observable.

If the black hole is located on or very near the line of sight from the Earth to a star, then the gravitational-lensing effect brings about a change in the apparent position of the star and also a brief enhancement of brightness, when the black hole passes directly in front of the star and generates an Einstein ring (see Section 4.5).

However, the probability for such an alignment is low, and searches for MACHOs (*m*assive *c*ompact *h*alo *o*bjects) have detected only very few events of this kind that might possibly be attributed to black holes.

Favorable conditions for the observation of a black hole occur if there is a nearby source of gas or plasma, such as a companion star or an interstellar cloud.

The black hole can then capture gas and plasma, and while this material falls down toward the horizon, it heats by viscous dissipation of its kinetic energy and emits thermal radiation (the radiation is emitted by the infalling material before it reaches the horizon; any radiation emitted after the material has crossed the horizon remains trapped in the black hole).

This accretion process results in an exceptionally efficient conversion of gravitational potential energy into radiation.

For maximum efficiency of radiation, the gas ought to approach the black hole with some angular momentum.

This will prevent the gas from falling straight in on a radial line; instead, the gas will first be captured into a circular orbit, and then gradually spiral downward into smaller and smaller circular orbits while it sheds its angular momentum.

Gas in such circular orbits around a black hole (or a compact star, such as a neutron star) forms an *accretion disk* (see Fig. 8.31).

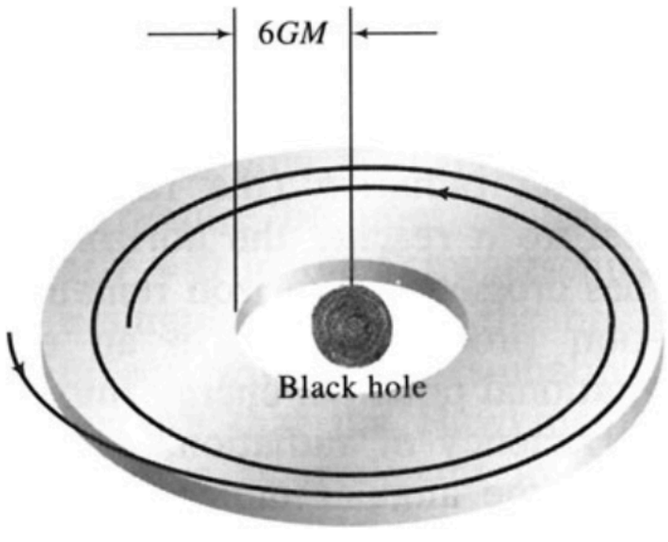


Fig. 8.31 Accretion disk around a black hole. New material enters the accretion disk at its outer rim and gradually spirals inward, toward the black hole. The accretion disk ends at $r = 6GM$, which is the radius of the smallest stable circular orbit around a black hole. Beyond this radius, the material plunges in quickly, and therefore the density of material is drastically reduced.

Because the orbital angular velocities for adjacent circular orbits are different, parcels of gas in such adjacent orbits experience a viscous drag.

This friction generates heat and also transfers orbital angular momentum from the inner regions of the disk to the outer (ultimately, this orbital angular momentum must be transferred from the outer edge of the accretion disk to the surrounding material).

The mechanism that generates the viscous drag is poorly understood; it probably involves a combination of turbulence that tends to mix adjacent parcels of gas and magnetic fields whose field lines link adjacent parcels.

Accretion by a Schwarzschild black hole can convert up to 5.7% of the rest-mass energy of the infalling material into radiation (see the later discussion); accretion by a Kerr black hole can convert up to 42%.

This is to be compared with the maximum energy released in nuclear reactions in stars, which is only about 0.7% of the rest-mass energy.

To see how much energy is converted into heat in the accretion disk of a Schwarzschild black hole, consider the total change in orbital energy of a parcel of gas as it spirals inward from the outer rim of the disk to the inner. This change is $E(r_2) - E(r_1)$, where $E(r) = mV(r)$ is the energy for a circular orbit, which includes the rest-mass energy m of the parcel.

At the outer radius, we can approximate $E(r_1) \cong m$, because the outer radius of the disk is much larger than the inner radius, and the binding energy at the outer radius is insignificant.

At the inner radius, $r_2 = 6GM$, which is the radius of the smallest stable circular orbit around a black hole (see Section 7.5). For this radius, Eq. (7.118) gives $E(r_2) = \sqrt{8/9}\ m$.

Hence the amount of energy that is converted into heat is $-E(r_2) + E(r_1) = (1 - \sqrt{8/9}\ )m = 0.057m$, or 5.7% of the rest-mass energy of the parcel.

The temperature attained by the accretion disk depends on several properties of the disk, such as its thickness and opacity.

Detailed calculations indicate that, for a black hole of $M \cong M_\odot$, temperatures of 10^8 or 10^9 K are attained.

At these temperatures the disk will radiate X-rays of a characteristic energy of 10 to 100 keV.

The high efficiency for energy release by the accretion of a black hole suggests that this mechanism could provide a suitable power source for quasars.

As is well known, a quasar releases energy at a prodigious rate, typically more than 10^{48} erg/s, or 10 $M_\odot$ per year.

To account for such a large radiated power, we need a source of exceptional efficiency.

A black hole of about 10^7 or 10^8 $M_\odot$ at the center of the quasar could satisfy the power requirements if it accretes gas, dust, and maybe entire stars at the rate of about a few hundred $M_\odot$ per year.

Alternative power sources, such as dense star clusters or multiple supernova explosions, have been proposed, but their lower efficiency for the conversion of rest mass into radiation would demand improbably large masses at the cores of quasars (Blandford, 1987).

Some galaxies with exceptionally active nuclei, such as radio galaxies and Seyfert galaxies, which emit almost as much radiation as quasars, might also be powered by accreting black holes.

It has even been proposed that ordinary galaxies might contain black holes in their nuclei. Figure 8.32 shows a large disk of dust and gas orbiting the nucleus of the galaxy NGC 4261; the bright spot at the center of this disk is believed to indicate a hot accretion disk belonging to a large black hole in the galactic nucleus.

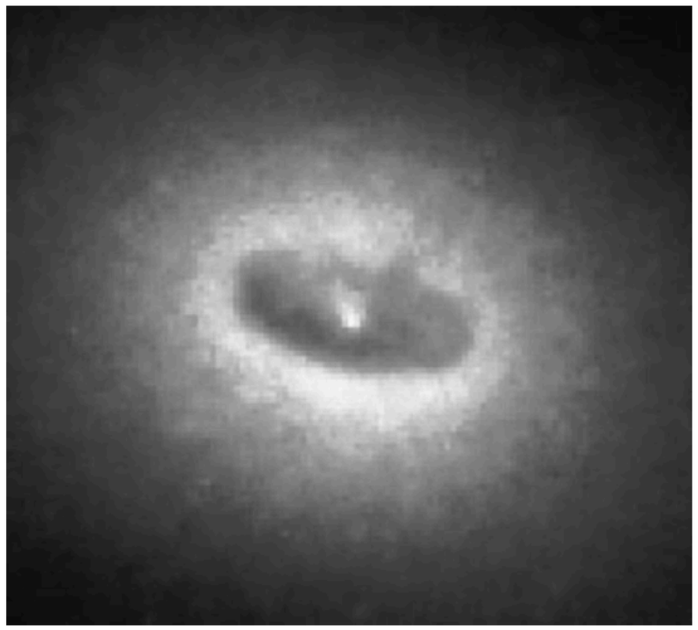


Fig. 8.32 A disk of dark dust encircles a possible supermassive black hole at the center of NGC 4261 seen in this photograph from the Hubble Space Telescope. The disk spans about 800 ly. [NASA/ W. Jaffe (Leiden Observatory)/H.Ford/JHU/STScI]

Observations of the center of our own Galaxy have revealed a radio and X-ray source (Sgr A*), which has been interpreted as an accreting black hole (see Fig. 8.33).

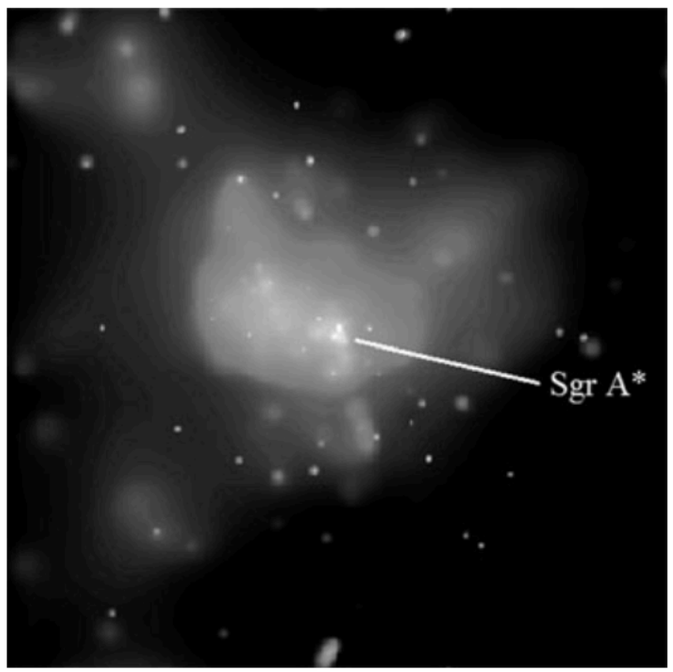


Fig. 8.33 This image shows the central region of our Galaxy as seen by NASA's Chandra X-ray satellite. The bright spot shows a huge X-ray flare that occurred in the vicinity of the compact X-ray and radio source Sgr A*, the probable supermassive black hole at the center of our galaxy. The X-ray source brightened drastically in a few minutes and, after about three hours, rapidly declined to the pre-flare level. The rapid variation in X-ray intensity indicates that the flare was produced in a region of small size, no more than the size of the Earth's orbit. (NASA/ MIT/ F. Baganoff et al.)

The radio observations indicate that this source is quite small, no more than 1 pc across.

Measurements of orbital velocities of stars in the vicinity show that the position of this radio source coincides with the position of a mass concentration of about $4 \times 10^5 M_\odot$.

On the assumption that the radio source and the mass concentration are the same object (and not an accidental, improbable overlap on our line of sight), we can identify this radio source as a supermassive black hole, because its mass is much too large for a star and its size is much too small for a star cluster.

Similar observations of the central parts of other galaxies have revealed several dozen cases of such objects of large masses, 10^6 to $10^9 M_\odot$, and of small sizes.

All of these are likely to be supermassive black holes.

However, the evidence is somewhat circumstantial.

The observational limits on the sizes are small by astronomical standards, but much larger than the Schwarzschild radius (for example, in Sgr A^*, the observational limit is 1 pc, but the Schwarzschild radius is 4×10^{-8} pc).

To obtain conclusive evidence for black holes, it is better to rely on observations of binary systems consisting of a compact body – a neutron star or a black hole – and a more or less normal star in orbit around each other.

The compact body will capture gas from its normal companion and form an accretion disk that emits a copious flux of X rays (see Fig. 8.34).

Such X-ray emission is a distinctive feature of compact binary systems – observationally they are seen as X-ray binaries.

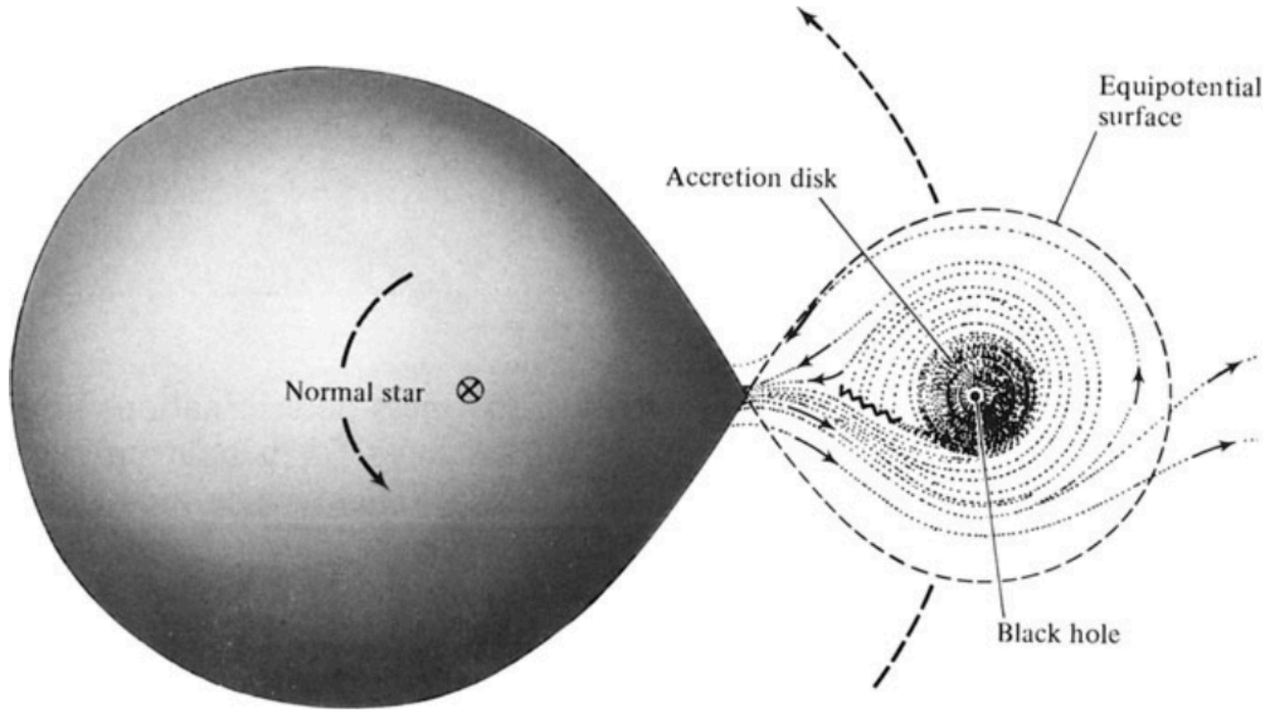


Fig. 8.34 Binary system consisting of a black hole with a normal star as a companion. The black hole draws gas from the normal star and forms an accretion disk. (After Eardley and Press, 1975.)

The observation of X-ray sources from the Earth is complicated by atmospheric absorption.

The X-ray telescopes must be lifted above the atmosphere in high-flying rockets or in satellites.

The first X-ray sources were discovered by Giacconi et al. with detectors flown on rockets.

Later, X-ray telescopes were installed on satellites, such as the UHURU satellite, the Einstein satellite, and the Chandra satellite.

By now, hundreds of X-ray sources have been discovered, most lying within our galaxy.

In some cases, the accurate measurement of the position of the source has made it possible to identify the optical companion of the compact star.

Measurements of the Doppler shifts of the spectral lines of the optical companion then give us the orbital period and the orbital velocity (projected along the line of sight).

The orbital data, in conjunction with other data – such as the intensity modulation of the light (the light curve), the spectral classification, the luminosity, and the distance – permit us to determine the masses of the compact star and the companion star.

In most cases, the mass determination yields a mass of about 1 $M_\odot$ or less, which is consistent with the mass expected for a neutron star.

The neutron-star interpretation for the compact star in these binary systems is supported by the observational evidence in two ways: A large X-ray luminosity of the sources (typically 10^{37} or 10^{38} erg/s for X rays in the range of 2 to 10 keV), which requires a large mass of small size, with a deep gravitational potential well; and pulsations in the X-ray luminosity of some of the sources, with a pulse period of the order of seconds, in agreement with the typical periods of rotating neutron stars (pulsars).

However, in about 40 exceptional cases, the mass determinations give results in excess of the maximum permitted mass for a neutron star.

These exceptional cases are interpreted as black holes – about half of these are now regarded as "confirmed" black holes and the other half as "candidate" black holes.

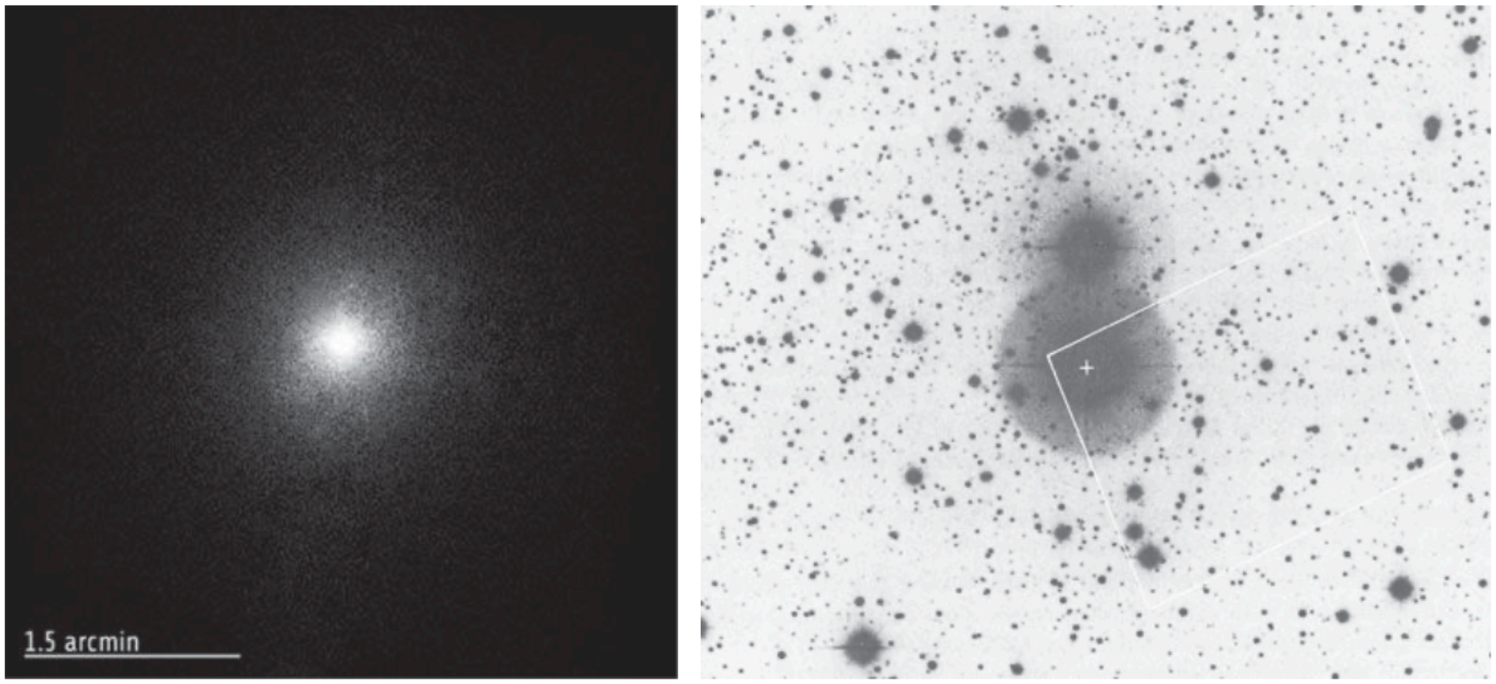
Some examples of confirmed black holes are listed in Table 8.1. This table includes the first black holes that were identified: Cygnus X-1, LMC X-3, and 6200–003.

Table 8.1 Some examples of confirmed black-hole binary systems*

X-Ray Source	Orbital Period (hr)	Distance (kpc)	M_{BH} ($M_\odot$)
Cyg X-1	134.4	2.0	6.9–13.2
LMC X-3	40.9	50	5.9–9.2
LMC X-1	101.5	50	4.0–10.0
0620–003	7.8	1.2	3.3–12.9
0422+32	5.1	2.6	3.2–13.2
1550–564	37.0	5.3	8.4–10.8

* Based on the review by McClintock and Remillard (2004).

Among these, the strongest case for a black hole is provided by Cygnus X-1 (see Fig. 8.35).



(a) (b)

The data available from spectroscopic observations of the optical companion include the orbital period, amplitude of periodic Doppler shifts (arising from orbital motion), reddening of light by interstellar dust (an indicator of distance), effective temperature, brightness, and periodic variation in brightness (arising from the tidal distortion of the shape of the star).

These data allow several more or less independent calculations of a lower limit for the mass of the compact body.

All these calculations agree that its mass is several times the solar mass.

A comprehensive analysis of the data sets a lower limit of 6.9 $M_\odot$ on the mass of the compact body.

This means that the mass is considerably in excess of the mass limit for neutron stars – about 2 $M_\odot$ for a reasonable equation of state, and 3.2 $M_\odot$ for the most extreme equation of state (see Section 8.8).

Accordingly, the compact body in this system is not made of neutrons or of any other kind of normal matter. The only remaining alternative is a black hole.

Calculations of the X-ray flux and spectrum emitted by gas in the accretion disk surrounding a black hole give results roughly consistent with the observed flux and spectrum of the identified black holes.

The accretion model can also account, in a qualitative way, for the observed quasiperiodic oscillations seen in the X-ray flux of some of these black holes on a time scale of a few milliseconds (these oscillations are "quasiperiodic" in that they occur erratically; that is, they are liable to fade away and then reappear).

This time scale is the orbital period for a parcel of gas near the inner edge of the accretion disk.

In some of the identified black holes, the quasiperiodic oscillations occur in multiples of a fundamental frequency; for instance, in 1550–564 we see quasiperiodic oscillations at 92, 184, and 276 Hz, that is, a fundamental frequency and two multiples.

How these multiples of the fundamental frequency arise is not quite clear; probably this process involves some resonance phenomenon within the accretion disk.

9 - Cosmology

We are all in the gutter,
but some of us are looking at the stars . . . Oscar Wilde, *Lady Windermere's Fan*.

As we begin the study of the universe, we have to confront a fundamental question: Are the laws of physics that hold on or near the Earth also valid in distant regions of the universe?

And are these laws of physics also valid at all times?

When cosmologists look at distant regions of the universe, they see them as they were a long time ago – they see a galaxy at a distance of, say, 10^{10} light-years as it was 10^{10} year ago.

To interpret the data collected in such observations, we need to know the laws of physics that govern these regions, far away in space and in time.

Newton set a precedent for the universal validity of physical laws in his cosmological speculations.

He conjectured that the (apparently) static distribution of the "fixed" stars was the result of an equilibrium of their mutual gravitational forces.

He assumed that his inverse-square law for the gravitational force was also valid for distant stars, and that "the fixed stars, being equally spread out in all points of the heavens, cancel out their mutual pulls by opposite attractions".

Einstein intended to follow Newton's precedent by applying to the entire universe the field equation he had posited for the Solar System.

But when he found that with these field equations he could not obtain a static solution for the mass distribution of the universe, he introduced new physics, in the form of a cosmological term added to the field equations.

After the discovery of the expansion of the universe, Einstein promptly retracted his adoption of the cosmological term, and he thereafter strictly followed the example of Newton in treating the universe by the same laws as apply within the Solar System.

In this chapter we will see that recent observational data on the acceleration of the expansion of the universe compel us to resurrect Einstein's cosmological term.

We will see that the expansion of the universe today is dominated by this cosmological term, also called "dark energy."

But this was not always true.

During most of the history of our universe, the expansion was dominated by the familiar gravitational effects of gravitating masses, that is, ordinary matter.

The dominance of the cosmological term began only fairly recently, in the last quarter of the history of our universe.

In the next chapter we will see that the cosmological term also made a brief appearance in the very early universe, where it supposedly became active for a short fraction of a second and caused an enormous, inflationary, expansion of the early universe.

In this chapter, we will ignore the physics and the evolution of the very early universe, and we will deal only with the "mature" universe, from about 10^{12} s after the Big Bang and onward.

Besides the new and unknown physics involving the behavior of the cosmological term, we are also confronted with new and unknown physics involving the "dark mass" contained in galaxies.

This dark mass shows up as an enhancement of the gravitational interactions of galaxies, but it has so far been impossible to determine what it consists of, and it has sometimes been attributed to new exotic particles.

Thus, the study of cosmology on large distance scales is leading us to deep modifications of the laws of physics that we have learned from our explorations of our immediate neighborhood on smaller scales.

9.1 Large-scale structure of the universe

For the purposes of cosmology, distances of less than 100 Mpc are regarded as small.

On such "small" distance scales, the mass distribution in our universe is irregular and lumpy.

It consists of galaxies (such as our Milky Way Galaxy, 3×10^4 pc across), clusters of galaxies (such as our Local Group, with about 30 galaxies, including Andromeda, about 1 Mpc across), and superclusters (such as our Local Supercluster, about 50 Mpc across).

On a large distance scale the lumps in the mass distribution average out, and the universe becomes fairly uniform.

We can discern this large-scale uniformity in the plot of galaxies in Fig. 9.1, which shows the distribution of galaxies in a patch of the sky, at an average distance of about 340 Mpc.

Except for fine detail, all regions more than about 200 Mpc across seem to have about the same density – on a scale of a few hundred Mpc the universe is fairly homogeneous.

We can also infer this large-scale homogeneity from the large-scale isotropy of the galaxy distribution over the sky.

The reason is that the observed isotropy about the position of the Earth requires that the density of galaxies be some function $n(r)$, independent of angular variables.

Unless $n(r)$ is a constant, this would imply that the universe has a unique center of symmetry ($r = 0$) and that the Earth is at this center.

In accord with the Copernican tradition, we will make the reasonable assumption that the Earth does not occupy such a privileged position.

Then the function $n(r)$ must be a constant, which means that, on a large scale, the distribution of galaxies is homogeneous.

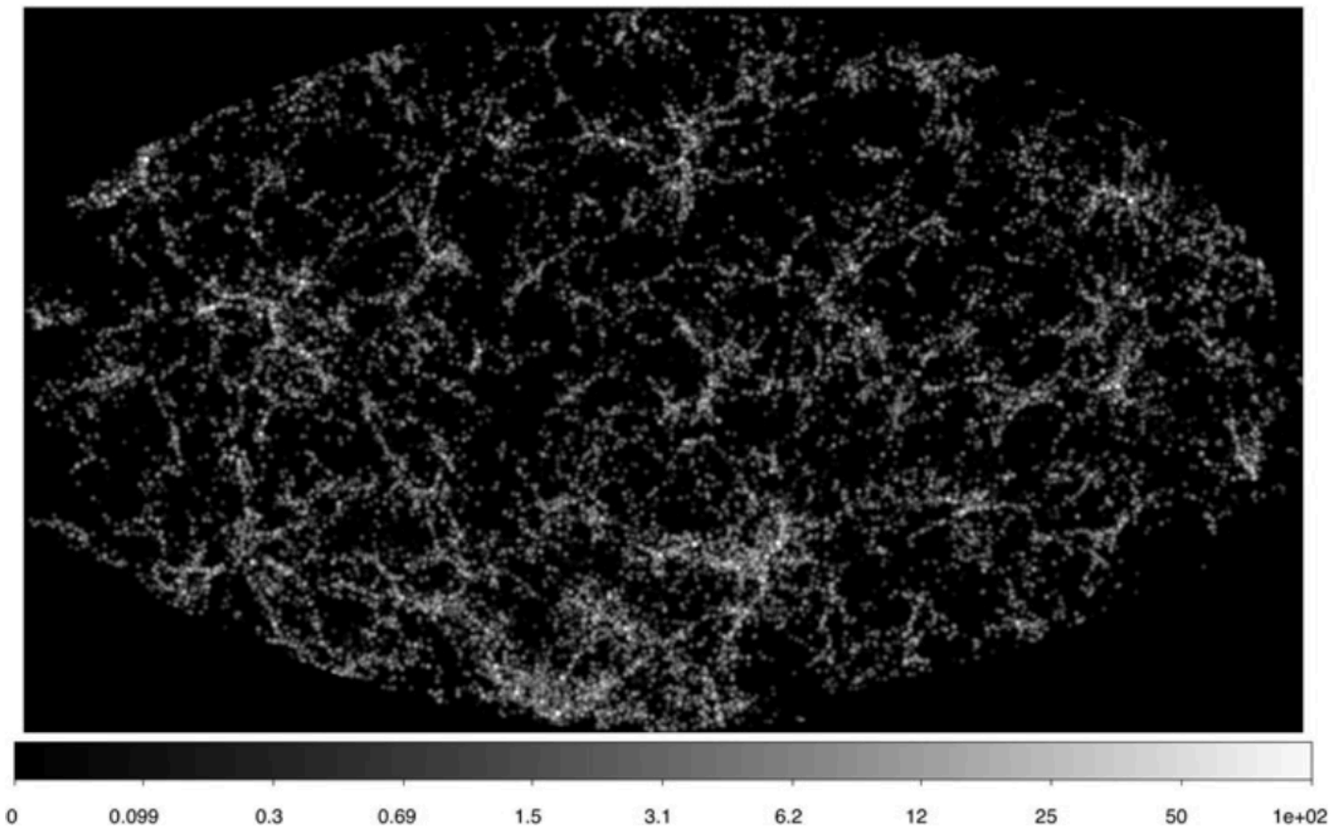


Fig. 9.1 The distribution of galaxies in a patch of the sky, according to the Sloan Digital Sky Survey. This patch spans about $130^\circ \times 70^\circ$ on the sky and a range of radial distances between 330 and 360 Mpc. There are about 30,000 galaxies in this patch. On the small to medium scales of this picture, we can see a cellular distribution of superclusters and filaments of various richnesses. The exceptionally strong concentration of galaxies in the lower area of the picture is part of the "Sloan Great Wall," consisting of two exceptionally rich superclusters. On the largest scales of distance included in this picture – a few hundred Mpc – the distribution is fairly uniform. (Courtesy J. Einasto, Tartu Observatory and ICRANet, Pescara)

Further evidence for large-scale isotropy (and, therefore, homogeneity) comes from observations of distant radio sources and of the diffuse X-ray background.

The distant radio sources – radio galaxies and quasars – are distributed randomly over the sky.

Some of these radio sources are about 3,000 Mpc away, and their random distribution therefore confirms isotropy on a very large scale.

Likewise, the diffuse X-ray flux originating from distant sources is distributed uniformly over the sky, with fluctuations of only a few percent on a scale of a few degrees.

But the strongest evidence for isotropy is supplied by the observations on the cosmic microwave background radiation, which is a relic from the early, hot stage of the universe, or the Big Bang.

As we will see in Section 9.5, this radiation is distributed uniformly over the sky to better than a few parts in 10^5, and this indicates that the universe was extremely uniform in its early stages.

The large-scale isotropy and homogeneity of the universe imply that all observers, wherever they are located in the universe, see the same large-scale density of galaxies.

Thus, on a large scale, the most remarkable feature of the universe is its lack of distinguishing features. The assertion that all positions in the universe are essentially equivalent, except for local irregularities, is known as the *Cosmological Principle*.

For a direct test of homogeneity of the galaxy distribution plotted in Fig. 9.1, we would need distance measurements of these galaxies.

But such distance measurements (either by direct methods or via the redshift; see Section 9.3) have been performed on only a small fraction of the galaxies.

Alternatively, we can test the large-scale homogeneity of the universe by simply counting the number of galaxies of a given brightness.

If the number of galaxies per unit volume is constant, then the total number within a distance d increases as d^3.

Consider those galaxies that have a given intrinsic luminosity L (in units of erg/s); the apparent brightness (erg/cm^2·s) of such a galaxy, or the incident flux measured at the Earth, is then

$$S = \frac{L}{4\pi d^2} \tag{9.1}$$

Accordingly, $d \propto S^{-1/2}$ and the volume within this distance is $(4\pi/3)d^3 \propto S^{-3/2}$.

For a homogeneous distribution, the number of such galaxies with an apparent brightness in excess of some given value S is then proportional to $S^{-3/2}$,

$$N(> S) \propto S^{-3/2} \tag{9.2}$$

This proportionality does not depend on the value of L and hence applies to the totality of all galaxies.

Thus, a count of numbers of galaxies as a function of brightness can be used as a direct observational test of homogeneity.

Surveys by Hubble established that for the brighter, nearer, galaxies, the counts were in good agreement with Eq. (9.2), which provided the first direct observational evidence for the large-scale homogeneity of the universe.

Later surveys extended the counts to fainter, faraway, galaxies. Over a wide range of brightnesses, the counts remain in good agreement with Eq. (9.2).

However, noticeable deviations occur for the faintest observable galaxies.

These deviations are interpreted as a combined effect of the curvature of spacetime and evolution of the luminosity of the galaxies – we see these distant galaxies as they were a long time ago, when their luminosity was different from what it is now, and such a change of luminosity with time invalidates Eq. (9.2).

For radio sources, such as radio galaxies and quasistellar objects (quasars), the counts are not in agreement with Eq. (9.2).

There is an excess of radio sources of low brightness.

This is believed to be an evolutionary effect – if there was a higher density of radio sources in the past, then the counts will exceed $S^{-3/2}$ at small S.

The evidence for evolutionary effects in radio sources can be used as an argument against the steady-state theory of the universe that was proposed by Bondi and Gold and by Hoyle.

In this theory, the universe has no beginning and no end, and the average properties of the universe are constant not only in space but also in time (the "Perfect Cosmological Principle").

In such a universe, new radio sources must be continually created to replace those that fade away or move away, so the average luminosity and density remain constant.

The conclusion that radio-source counts contradict the steady-state theory was first reached by Ryle and Scheuer.

Later, more precise counts by Ryle and others have confirmed this conclusion.

For quasars, the analysis of counts has also been carried out using their optical, rather than radio, brightness.

According to Schmidt, the counts indicate that the density of quasars must have been much higher in the past than it is now.

An analysis of counts with respect to redshift (which is a measure of distance; see Section 9.3) also confirms this conclusion.

This, again, contradicts the steady-state theory.

9.2 Cosmic distances

The methods used by astronomers to measure distances to objects in the sky vary with the distance involved.

For the nearest stars, at distances of up to 30 pc, triangulation with the diameter of the Earth's orbit as a baseline is adequate.

Some other geometric methods give the slightly larger distances to nearby clusters of stars, for instance, the Hyades cluster, at 41 pc.

Such trigonometric and other geometric methods are called *parallax* methods, because they hinge on measuring an angular displacement (apparent or real) of the stars across the sky.

Beyond a few hundred pc, distances are usually determined by photometric methods – by the apparent brightness of a standard light source, or "standard candle."

If a light source, such as a star, has an intrinsic luminosity L (erg/s), then the apparent brightness S (erg/cm^2·s) as measured at the Earth is inversely proportional to the square of the distance, $S = L/4\pi d^2$, as already stated in Eq. (9.1).

A measurement of the apparent brightness S therefore gives the distance d provided the intrinsic luminosity L is known.

For distances within our Galaxy and within our Local Group, ordinary stars can serve as standard light sources.

But for distances beyond 10 Mpc, ordinary stars are too faint to be distinguished, and larger and more luminous standard light sources are needed, such as supernovas, clusters of stars, or entire galaxies.

The more luminous standard light sources used for large distances must be calibrated by comparing them with the less luminous standard light sources used at shorter distances.

For such a calibration, we must find some galaxy at an intermediate distance that contains both the more luminous and the less luminous standard light sources.

This step-by-step scheme for the calibration of standard light sources is called the *cosmological distance ladder*.

Until recently, the cosmological distance ladder relied on five or more steps, with progressively brighter standard candles at larger and larger distances.

But the 21st century has brought a significant simplification of this distance ladder.

The data now available from the Hubble Space Telescope give us a simplified distance ladder, with just two steps that take us all the way to distances as far as several thousand Mpc.

The two kinds of standard light sources used in this two-step distance ladder are Cepheid stars at the shorter distances (out to 30 Mpc) and type Ia supernovas at the longer distances (out to 5,000 Mpc).

Although not all Cepheids have the same luminosity, they can be used as standard light sources, because the luminosity is in direct correspondence with their pulsation periods.

Cepheids are variable supergiant stars, whose luminosity pulsates with a period of between 2 and 150 days.

Their time-average luminosity is correlated with the period: The long-period Cepheids are intrinsically the most luminous, and the short-period Cepheids are the least luminous.

A simple measurement of the period of a Cepheid therefore indicates the luminosity and permits us to adopt the Cepheid as a standard light source.

Likewise, not all Ia supernovas have the same luminosity.

The type Ia supernovas result from the thermonuclear explosion of a white dwarf star in a binary system, when this white dwarf accumulates mass by accretion and exceeds the Chandrasekhar limit of 1.44 $M_{\odot}$.

The star then collapses, and this triggers thermonuclear reactions that burst the star apart (in contrast, type II supernovas involve the spontaneous explosion of a large-mass star, when it exhausts its nuclear fuel).

The peak luminosity of the Ia supernova explosion is correlated with the time-dependence of its light curve – supernovas with a slow rise and fall of their light curve have the highest luminosity, whereas supernovas with a fast rise and fall are less luminous.

Thus, as in the case of Cepheids, a simple measurement of the time dependence of the light curve indicates the peak luminosity and permits us to adopt the supernova as a standard light source.

The Ia supernovas typically attain peak luminosities of more than a billion times the luminosity of the Sun, which makes them detectable at distances almost as large as the distance to the boundary of the observable universe (the “horizon”).

The Hubble Space Telescope has proved invaluable for the calibration of the Cepheids and Ia supernovas.

With its Fine Guidance Sensor, this telescope is able to measure the parallax of Cepheids in our Galaxy and therefore establish their distance by triangulation and calibrate their luminosities.

Furthermore, the Hubble Space Telescope can identify Cepheids in distant galaxies that contain type Ia supernovas and thereby establish the distance to these supernovas and calibrate their luminosities.

This gives us a two-step distance ladder, which avoids the complications of the older, many-step distance ladders and the accumulation of calibration errors from these many steps.

The distance d determined by the method of standard candles according to Eq. (9.1) is called the *luminosity distance*, usually designated by d_L.

In principle, it is also possible to determine the distance from the angular size of a target of standard linear size.

In a flat three-dimensional geometry, the linear size D of the target, its observed angular size $\delta\theta$, and the distance are related by

$$\delta\theta = \frac{D}{d_A} \tag{9.3}$$

The distance d_A calculated from this equation is called the *angular-diameter distance*.

In practice, this method of distance determination is of limited use, because it is more difficult to identify targets of standard size than targets of standard luminosity.

In an expanding and curved universe, neither the luminosity distance d_L nor the angular-diameter distance d_A is equal to the "true" distance, or the "proper" distance, as it would be determined with meter sticks laid end to end between target and observer.

Furthermore, in an expanding universe such as ours, the calculated luminosity distance is always larger than the angular-diameter distance [see Eq. (9.153)].

In an expanding but spatially flat universe (a universe that has a curved 4-D spacetime, but a flat 3-D space), the angular-diameter distance has a remarkable property: It is always equal to the true, or proper, distance of the emitter at the time of emission; that is, the angular-diameter distance preserves the information about the distance at the time of emission.

This becomes immediately obvious if we consider that two light rays emitted toward the Earth from opposite ends of the emitter's diameter continue to propagate with a fixed angular separation, and therefore the distance d_A calculated from Eq. (9.3) is not affected by the subsequent motion of the emitter (of course, the light emitted at a later time will indicate a larger distance d_A, because the later emission occurs when the emitter has moved to a larger distance).

9.3 Expansion of the universe; Hubble's law

Observed from the Earth, the spectral lines of almost all galaxies display a shift toward lower frequencies, that is, a redshift.

This redshift is conventionally described by a redshift parameter z defined as the fractional increase of wavelength,

$$z = \frac{\lambda_0 - \lambda}{\lambda} \tag{9.4}$$

where λ_0 is the wavelength received at the Earth and λ is the initial the wavelength emitted by the atoms in the remote galaxy.

The physical interpretation of this cosmological redshift is as a Doppler shift produced by the motion of the galaxies away from the Earth.

If the speed of the galaxy is small compared with the speed of light, the fractional change in wavelength of the light is given by the simple Doppler formula

$$\frac{\lambda_0 - \lambda}{\lambda} = \frac{v}{c} \tag{9.5}$$

Thus, the speed of recession of the galaxy, in units of c, equals the redshift parameter z.

For example, the galaxies in the Virgo cluster have a redshift z = 0.004, and therefore a speed of recession of 0.004c, or 1,200 km/s.

Many galaxies have much larger redshifts.

The galaxy with the largest redshift discovered so far has $z = 8.6$.

For such a large value of z, the simple Doppler formula (9.5) fails – it becomes necessary to take into account relativistic effects and the large-scale geometry of the universe.

The redshift of those galaxies that are nearby is small, and the redshift of those galaxies that are far away is large.

Figure 9.2 is a plot of the measured redshift of some galaxies vs. their distance.

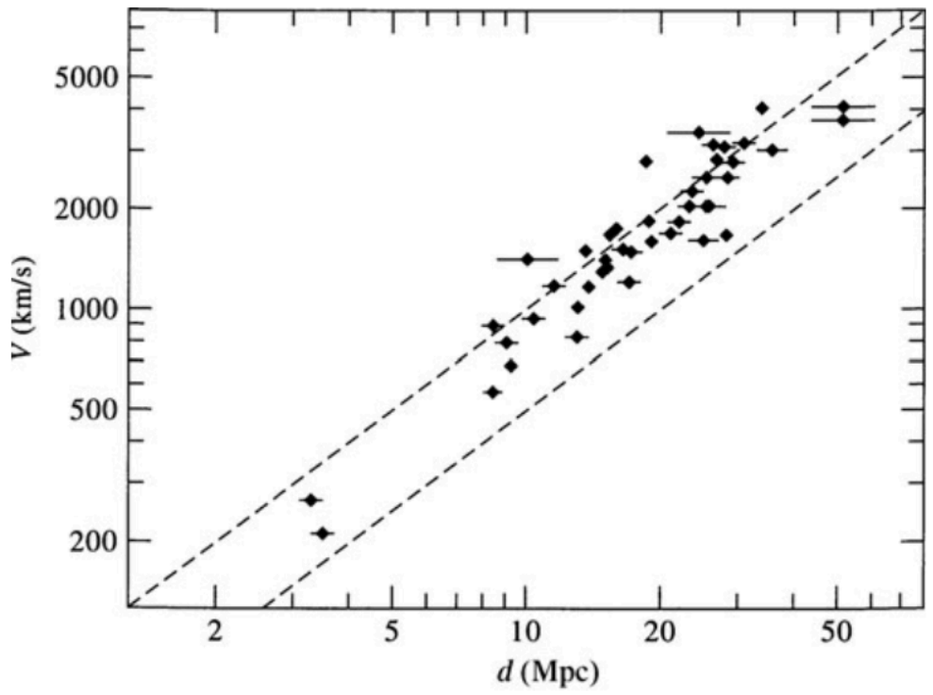


Fig. 9.2 Measured redshift vs. distance for some galaxies. The redshift increases with distance. (From Jacoby et al., 1992)

From this plot we see that the magnitude of the redshift is directly proportional to the distance, but the large scatter in the observational data makes it difficult to determine the constant of proportionality.

For the galaxies plotted in Fig. 9.2, $z \ll 1$, and hence the simple Doppler formula (9.5) is applicable.

A redshift directly proportional to the distance then indicates a recession velocity directly proportional to the distance:

$$v = H_0 l \tag{9.6}$$

The proportionality of redshift and distance was first noticed in 1917 by de Sitter in his study of a theoretical model of the universe (the empty de Sitter model), but he did not interpret the redshift as an indication of velocity.

This interpretation was adopted later, and it became the favored interpretation through the observational work of Hubble in the late 1920s.

Equation (9.6) became known as *Hubble's law*.

The constant of proportionality H_0 is called Hubble's constant.

It is usually expressed in units of kilometers per second per megaparsec [km/(s·Mpc)].

The search for an accurate value of H_0 has a long and checkered history.

To determine the value of H_0 astronomers must examine the recession velocities and the distances of remote galaxies.

The measurement of the recession velocities, via the redshift, is nearly free of error.

However, galaxies have peculiar velocities, that is, irregular deviations from the Hubble motion of recession, caused by the gravitational attraction of other galaxies or clusters in their vicinity.

For instance, our own Local Group is thought to have acquired a peculiar "infall" velocity toward the large Virgo cluster, caused by the gravitational attraction of the large mass in Virgo, and also a peculiar velocity toward the even larger Hydra-Centaurus cluster ("the Great Attractor"), farther away than Virgo.

Because of this peculiar velocity, the recession velocity of Virgo relative to us is somewhat smaller than the Hubble recession.

For an accurate determination of H_0, we must correct all the observed recession velocities according to our peculiar velocity.

Alternatively, we can try to minimize the influence of the peculiar velocity by dealing exclusively with galaxies much farther away than Virgo.

The Hubble recession velocities of such galaxies are much larger than any peculiar velocity, and this permits us to neglect the latter.

The measurements of the distances are afflicted with large errors because of the uncertainties in the calibrations of the standard light sources, as described in the preceding section.

The large errors in the distance measurements lead to correspondingly large errors in H_0.

Hubble originally found a value of about 500 km/(s·Mpc) for H_0.

But his distance measurements contained several errors, and recent results for H_0 are about ten times smaller.

Extensive series of distance determinations were completed by Sandage and Tammann and by de Vaucouleurs et al. in the 1980s.

Their results were $H_0 = 40$ to 60 km/(s·Mpc) (Sandage and Tamman) and $H_0 = 80$ to 100 km/(s·Mpc) (de Vaucouleurs).

The difference between these results arose from different choices of methods used for distance determinations, different ways of calibrating the standard light sources, and different corrections of the observed brightnesses of these sources for absorption by interstellar dust.

More recent distance determinations based on observations made with the Hubble Space Telescope gave results intermediate between those of Sandage and de Vaucouleurs.

The best available recent distance determinations, summarized by Weinberg, all agree on a value of the Hubble constant of about

$$H_0 = 71 \text{ km/(s} \cdot \text{Mpc)} \tag{9.7}$$

with an uncertainty of ±6 km/(s·Mpc).

An entirely different new method for the determination of the Hubble constant relies on small anisotropies observed in the distribution of the cosmic microwave background radiation over the sky.

These anisotropies have been measured by the Wilkinson Microwave Anisotropy Probe (WMAP), a satellite launched in 2001 and parked in an orbit around the Sun, about 10^6 km farther out than the Earth.

The development of the anisotropies during the expansion of the universe depends on various cosmological parameters – among them the Hubble constant – and by analysis of the WMAP anisotropy data it is possible to deduce the same value of 71 km/(s · Mpc) for H_0, but with a smaller uncertainty of ±3 km/(s·Mpc).

However, this method hinges on a specific model of the early universe: a spatially flat universe, a particularly simple spectrum of initial anisotropies (proportional to a power of the distance scale), and dark mass made of fairly heavy particles (cold dark mass).

So the result might be regarded as evidence in favor of the adopted model of the universe, and not as an independent determination of H_0.

Determinations of other cosmological parameters by means of other special assumptions about the early universe will be discussed in Section 10.4.

Another method for the determination of the Hubble constant exploits observations of gravitational lenses that form multiple "images" of distant quasars.

The light from the quasar to the Earth then travels along different paths through the lens, with a different path length and travel time for each image.

The path differences depend on the distance scale, and therefore a measurement of the time delay between light signals reaching us along different paths can be used to determine the distance scale, if the geometry of the paths is known.

As already mentioned in Section 4.5, the results of determinations of the Hubble constant by this method are in complete agreement with Eq. (9.7).

In view of the remaining uncertainties in the value of H_0, cosmologists have found it expedient to write H_0 in the form

$$H_0 = 100h \text{ km/(s} \cdot \text{Mpc)} \tag{9.8}$$

where h is a fudge factor whose value lies somewhere between 0.6 and 0.8.

Although according to Eq. (9.6) galaxies are moving radially away from us, this does not mean that the Earth occupies a preferred position in the universe.

Let us write Eq. (9.6) in vector notation:

$$\mathbf{v} = H_0\mathbf{r} \tag{9.9}$$

Consider now an observer on some extragalactic galaxy at a position $\mathbf{r}'$. Relative to the Earth this observer has a velocity

$$\mathbf{v}' = H_0\mathbf{r}' \tag{9.10}$$

Hence, relative to this observer, another galaxy at a position $\mathbf{r}$ has a velocity

$$\mathbf{v} - \mathbf{v}' = H_0(\mathbf{r} - \mathbf{r}') \tag{9.11}$$

Since $\mathbf{r} - \mathbf{r}'$ is simply the position vector of the other galaxy relative to this observer, Eq. (9.11) shows that there is a general motion of galaxies radially away from this observer, just as there is such a motion away from an observer on Earth.

Hence Eq. (9.9) indicates a general expansion of the universe, with distances increasing in proportion.

However, note that this increase of distances applies only to distances between galaxies, and only when they are not gravitationally bound to each other.

The distances within a galaxy or the distances within a (bound) cluster of galaxies do not increase.

Incidentally, a nonlinear relation between $\mathbf{v}$ and $\mathbf{r}$ would lead us to the conclusion that the Earth does occupy a preferred position in the universe, because then the right sides of Eqs. (9.9) and (9.10) could not combine by simple vector addition, and expansion of the universe would look different from the Earth and from some extragalactic galaxy.

This is a strong argument in favor of the linear Hubble law.

An important question is how the recession velocities change in time.

Since the attractive gravitational forces oppose the expansion, we expect a decrease in the expansion rate.

Let us calculate the expected deceleration in the expansion rate.

For the calculation of the average motion of the galaxies, we can approximate the mass distribution of the universe by an average uniform mass density ϱ.

Consider a spherical volume of radius l, which is small compared with the size of the observable universe, but large enough so irregularities in the mass density can be neglected.

The mass distribution that surrounds this volume and extends throughout the rest of the universe can be regarded as a spherical shell of very large thickness; such a shell produces no gravitational force in its interior.

Hence, the gravitational force acting on a particle on the surface of our spherical volume depends only on the mass m in this spherical volume, and the acceleration of this particle is

$$\frac{dv}{dt} = -\frac{Gm}{l^2} \tag{9.12}$$

With $m = 4\pi l^3 \varrho/3$, we then obtain

$$\frac{dv}{dt} = -\frac{4\pi G\rho}{3} l \tag{9.13}$$

This equation shows that the change in the relative velocity of two galaxies is proportional to the distance between them.

Since Hubble's law tells us that the present velocities are proportional to the distances, it follows that the velocities remain proportional to the distance at all times.

We can therefore express the expansion velocity at any time t in the form

$$v(t) = H(t)l(t) \tag{9.14}$$

where H(t) is the Hubble "constant" at that time [in contrast, $H(t_0)$, or H_0, is the Hubble constant at the present time t_0].

This expression for the expansion velocity in terms of a time-dependent Hubble constant remains valid if we include a cosmological constant in our equations.

We know from Section 7.3 that the cosmological term has the gravitational effect of a uniform mass density $-\Lambda/4\pi G$ [see Eq. (7.51), but remember that the contribution of the cosmological term to the energy density is $+\Lambda/8\pi G$, that is, the contributions to gravitation and to energy density have opposite signs].

Thus, the mass density ϱ in Eq. (9.13) must be replaced by $\varrho - \Lambda/4\pi G$, so

$$\frac{dv}{dt} = -\frac{4\pi G}{3}\left(\rho - \frac{\Lambda}{4\pi G}\right) l \tag{9.15}$$

This, again, shows that the change of the velocity is proportional to the distance, so the velocity remains proportional to the distance at all times, in accord with Eq. (9.14).

For comparison with observations, it is convenient to rewrite the acceleration equation (9.15) in terms of the Hubble constant and its first derivative.

Differentiation of Eq. (9.14) gives

$$\frac{dv}{dt} = Hv + \frac{dH}{dt} l = \left(H^2 + \frac{dH}{dt}\right) l \tag{9.16}$$

Here we can define a *deceleration parameter*,

$$q \equiv -\left(1 + \frac{1}{H^2}\frac{dH}{dt}\right) \tag{9.17}$$

so Eq. (9.16) takes the compact form

$$\frac{dv}{dt} = -qH^2 l \tag{9.18}$$

The acceleration equation (9.15) then implies

$$-qH^2 = -\frac{4\pi G}{3}\left(\rho - \frac{\Lambda}{4\pi G}\right) \tag{9.19}$$

or, at the present time,

$$q_0 = \frac{4\pi G}{3{H_0}^2}\left(\rho_0 - \frac{\Lambda}{4\pi G}\right) \tag{9.20}$$

It is customary to define two dimensionless density parameters $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ that

$$\Omega_{m,0} = \frac{8\pi G}{3{H_0}^2}\rho_0 \tag{9.21}$$

$$\Omega_{\Lambda,0} = \frac{\Lambda}{3{H_0}^2} \tag{9.22}$$

and a combined density parameter

$$\Omega_0 = \Omega_{m,0} + \Omega_{\Lambda,0} \tag{9.23}$$

The quantity $3H_0^2/8\pi G$ inserted in the definition (9.21) has the dimensions of a mass density, and it is called the *critical density*.

As we will see later, in a universe without a cosmological constant, this critical density discriminates between the case of an expansion that continues forever ($\varrho_0 < \varrho_{crit}$) and the case of an expansion that comes to a stop and then reverses, ending in a re-collapse ($\varrho_0 > \varrho_{crit}$).

The numerical value of the critical density is

$$\rho_{crit} = \frac{3{H_0}^2}{8\pi G} = 1.88 \times 10^{-29}\, h^2\ \mathrm{g/cm^3} \tag{9.24}$$

In terms of these density parameters the acceleration equation (9.20) becomes

$$q_0 = \tfrac{1}{2}(\Omega_{m,0} - 2\Omega_{\Lambda,0}) \tag{9.25}$$

Note that Eq. (9.25) is equally valid in Newton's theory and in Einstein's theory.

The derivation of this equation depends on the absence of gravitational fields in a spherical cavity surrounded by a spherical mass distribution.

According to Birkhoff's theorem (see Section 7.4), this is just as true in general relativity as it is in Newtonian theory.

Furthermore, if we assume that the cavity is of small dimension, then the velocity and acceleration will be small, and therefore the nonrelativistic equation of motion will be an adequate approximation.

The direct observational determination of the deceleration parameter is very difficult.

In principle, it is possible to determine q_0 by looking for systematic deviations from Hubble's law at large distances.

Since we see remote galaxies as they were at the time when they emitted the light that reaches us now, such galaxies indicate the state of the universe a long time ago.

If q_0 is, say, positive and the expansion rate has been

In practice, because of the uncertainties in the distance scale, it is not feasible to detect the systematic deviations from Hubble's linear law in a simple plot of redshift vs. distance.

Such deviations, if any, occur only at very large distances, 100 Mpc or more, where the deviations of the geometry of the universe from flat spacetime become significant.

Since the distance in Hubble's law is determined by the brightness of a standard light source, we must take into account the modification of the inverse-square law (9.1) relating brightness to distance in a curved spacetime, and also the decrease of the energy of the light brought about by the reduction of the frequency of the light wave.

In general, for a given cosmological model, the distance can be expressed as a nonlinear function of z, which takes into account the geometry of the universe and its dynamic evolution.

Two independent teams of astronomers – the High-z Supernova Search Team and the Supernova Cosmology Project – have accumulated redshift vs. distance data from type Ia supernovas (see Fig. 9.3).

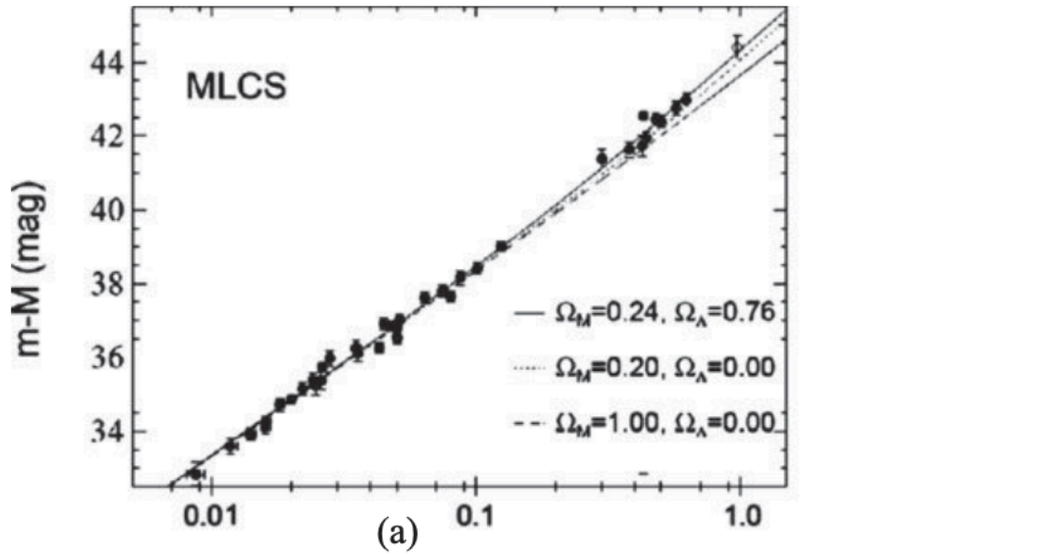


Fig. 9.3 Distance vs. redshift for type Ia supernovas measured by (a) the High-z Supernova Search Team and (b) the Supernova Cosmology Project. The vertical axis is proportional to the logarithm of the distance, minus an additive constant. (Riess et al., 1998; Perlmutter et al., 1999)

They then used these data for a best-fit determination of the density parameters $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ that characterize the standard Friedmann-Lemaitre model of our universe (a homogeneous, isotropic model, with a mass density and a nonzero cosmological constant; see Section 9.10).

With the additional assumption that the spatial geometry is flat (which, as we will see in Section 9.11, implies the constraint $\Omega_0 = 1$, or $\Omega_{m,0} + \Omega_{\Lambda,0} = 1$), this gave a reasonably good determination for $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$.

The data displayed in Fig. 9.4 yielded the result

$$\Omega_{m,0} = 0.3 \quad \text{and} \quad \Omega_{\Lambda,0} = 0.7 \tag{9.26}$$

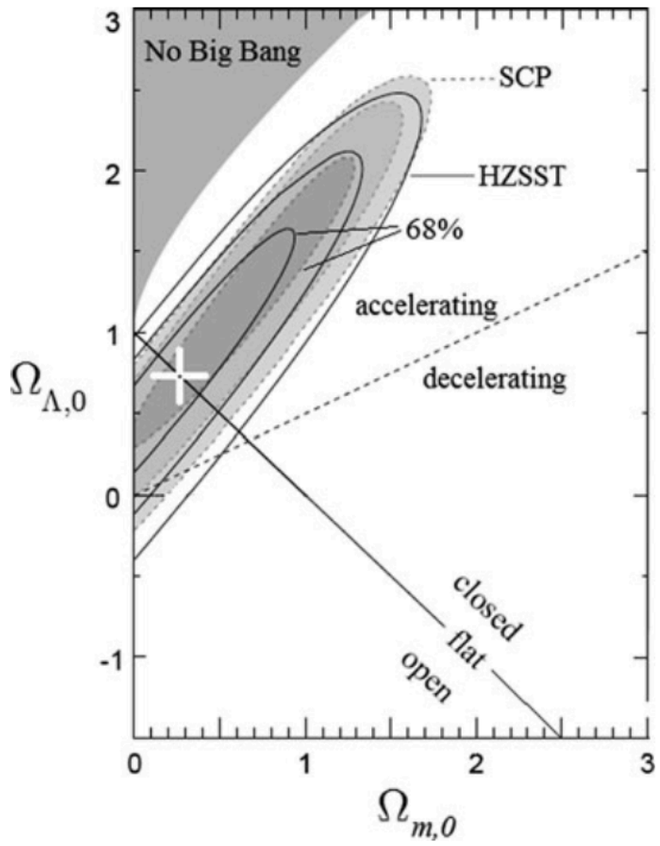


Fig. 9.4 Best-fit determination of $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$. The contours indicate the reliability of the fit (68%, 95%, or 99.7%), with solid lines for High-z Team data, and dashed lines for Supernova Project data, respectively. For a flat, zero-curvature universe, with $\Omega_{m,0} + \Omega_{\Lambda,0} = 1$ (see slanted straight line), the best fit is at $\Omega_{m,0} = 0.3$, $\Omega_{\Lambda,0} = 0.7$, but with uncertainties of ± 0.1. The white cross indicates a recent determination from an analysis of anisotropies in the cosmic background radiation (see Section 10.4). This yields more precise values, $\Omega_{m,0} = 0.27 \pm 0.01$ and $\Omega_{\Lambda,0} = 0.727 \pm 0.003$ (the uncertainties are smaller than the size of the dot marked at the center of the cross). (HZSST and SCP data based on Perlmutter and Schmidt, 2003)

The estimated uncertainties in these numbers are about ± 0.1.

If we do not impose the a priori constraint $\Omega_0 = 1$, then the uncertainties are considerably larger (see the elliptical contours in Fig. 9.3).

However recent determinations of the cosmological parameters from data on slight fluctuations observed in the distribution of blackbody radiation over the sky confirm the numbers (9.26), with drastically reduced uncertainties (see Section 10.4).

The surprising conclusion from the supernova data was that q_0 is negative. With the values of the density parameters obtained from the supernova data, the value of q_0 is

$$q_0 = \frac{1}{2}(\Omega_{m,0} - 2\Omega_{\Lambda,0}) = \frac{1}{2}(0.3 - 2 \times 0.7) = -0.55 \tag{9.27}$$

Accordingly, the acceleration (9.18) is positive – the expansion of our universe is accelerating!

This cosmological acceleration was the biggest surprise in cosmology in the last 20 years.

It used to be thought that the universe is decelerating, as expected from the normal gravitational attraction between galaxies and clusters of galaxies, and that only the magnitude of this deceleration was in question.

But the supernova data establish that, at large distances, the repulsive effect of the cosmological term overcomes the normal gravitational attraction and produces an acceleration.

In time, this acceleration will increase further, because the gravitational attraction will continue to decrease as the galaxies recede, whereas the repulsion associated with the cosmological constant will increase – ultimately, the expansion rate will increase exponentially.

9.4 Age of the universe

If we extrapolate the observed expansion of the universe backward in time, we find that the early universe must have been extremely dense and extremely hot.

This suggests an explosive origin of the universe.

It seems that initially all the matter was concentrated in a very dense fireball (the fireball had no preferred position in the universe; rather, it filled the entire universe).

This fireball blew apart in a violent, primordial explosion, or *Big Bang*. As the fireball expanded and cooled, slight fluctuations in density grew into local condensations that evolved into gravitationally bound systems, such as galaxies and bound clusters of galaxies.

These bound systems continued the average motion of expansion, away from each other.

Thus, the general motion of recession of the galaxies is directly related to the initial velocities that these galaxies had when they were first formed.

Hubble's law gives us an immediate estimate of how long ago the universe began. For an order-of-magnitude estimate of the age of the universe, assume that the recession velocity v in Eq. (9.6) remains constant for any given galaxy.

Then the backward extrapolation of the motion of such a galaxy tells us that the time needed for complete collapse is $l/v = H_0^{-1}$.

This time is called the *Hubble age*.

The value of H_0 given in Eq. (9.7) implies a Hubble age

$$H_0^{-1} = 14 \times 10^9 \text{ years} \tag{9.28}$$

More generally, with the fudge factor h included, the Hubble age is

$$H_0^{-1} = \frac{1}{100h}\text{s} \cdot \text{Mpc/km} = h^{-1} \times 9.78 \times 10^9 \text{ years} \tag{9.29}$$

The actual age of the universe is presumably somewhat less than the Hubble age, because the recession velocities in the past were somewhat larger, on average, than they are now.

For instance, in a cosmological model of zero curvature, the actual age can be shown to be 13 billion years (see Section 9.12).

It is therefore of considerable interest to compare the Hubble age with estimates of the actual age as directly determined from observational data.

There are several methods that permit us to estimate the actual age of the universe.

By radioactive dating methods, the age of solid bodies in the Solar System (rocks on Earth and Moon, meteorites) is known to be $(4.6 \pm 0.1) \times 10^9$ years.

But the age of solid bodies in the Solar System sets only a lower limit on the age of the universe.

A better limit is set by the age of the chemical elements.

Almost all elements, with the exception of hydrogen, some of the helium, and a few other light elements, were formed by nuclear reactions in the interior of stars.

For instance, the uranium found on the Earth was produced by massive stars that lived, and died, before the Solar System was born.

Astrophysical theory allows us to calculate the relative abundance $[^{235}\mathrm{U}]_i$ and $[^{238}\mathrm{U}]_i$ with which the isotopes ^{235}U and ^{238}U were produced in the nuclear reactions in stars:

$$[^{235}\mathrm{U}]_i/[^{238}\mathrm{U}]_i = 1.16 \tag{9.30}$$

At present, the measured abundance ratio of these isotopes is

$$[^{235}\mathrm{U}]_0/[^{238}\mathrm{U}]_0 = 7.25 \times 10^{-3} \tag{9.31}$$

The drastic decrease of the relative abundance is due to the higher decay rate of the isotope ^{235}U as compared with the isotope ^{238}U; the decay rate of the former isotope is $1/(1.015 \times 10^9$ yr), whereas that of the latter is $1/(6.446 \times 10^9$ yr).

If we assume that all of the uranium and thorium were produced suddenly and changed only by radioactive decay ever since, it is a simple matter to calculate that the time needed to change (9.30) into (9.31) is

$$t_0 = \frac{\ln([^{235}\mathrm{U}]_i/[^{238}\mathrm{U}]_i) - \ln([^{235}\mathrm{U}]_0/[^{238}\mathrm{U}]_0)}{1/1.015 - 1/6.446} \times 10^9\ \mathrm{yr} = 6.1 \times 10^9\ \mathrm{yr} \qquad (9.32)$$

However, for a more accurate calculation, we must take into account that in the early stars in our Galaxy the production of isotopes occurred gradually, and for a while, the abundance was determined by the balance between generation and decay.

This complicates the analysis considerably and increases the age implied by Eqs. (9.30) and (9.31) by a factor of 2 or more.

The ages of some other elements can be calculated by the same method, with similar results.

Calculations of ages from the radioactive decay of uranium and thorium have also been performed with isotope abundance ratios measured on some stars, by analysis of the spectral lines seen in their emitted light.

This yields calculated ages of about $(14 \pm 2) \times 10^9$ years, consistent with the ages of elements found on Earth.

The ages of globular star clusters in our Galaxy provide us with another estimate of the age of the universe.

The stars in such a cluster were all born at once, presumably soon after the Galaxy formed.

The stars initially burn hydrogen into helium by thermonuclear reactions, and, as long as the supply of nuclear fuel lasts, they lead a placid, stable existence.

Within a population of such stable, hydrogen-burning stars, the hottest (bluest) stars are always the brightest; that is, the surface temperature increases monotonically with the luminosity.

In a plot of luminosity vs. temperature (a Hertzsprung-Russell diagram) for such stars, we find a curve that slants from low temperature and low luminosity to high temperature and high luminosity.

This curve is called the *main sequence*.

When the stars reach old age and run low on nuclear fuel, they depart from the main sequence in the Hertzsprung-Russell diagram.

The first stars to depart are the most luminous, most massive stars at the upper end of the main sequence.

Thus, as a globular cluster ages, its main sequence is progressively vacated from the top down (the main sequence "turns off" or "burns down").

From astrophysical calculations of stellar evolution we know how long a star of given mass remains on the main sequence, and we can therefore predict how far the main sequence will burn down as a function of time.

Figure 9.5 shows a Hertzsprung-Russell diagram for stars in the globular cluster M92, the oldest globular cluster near our Galaxy.

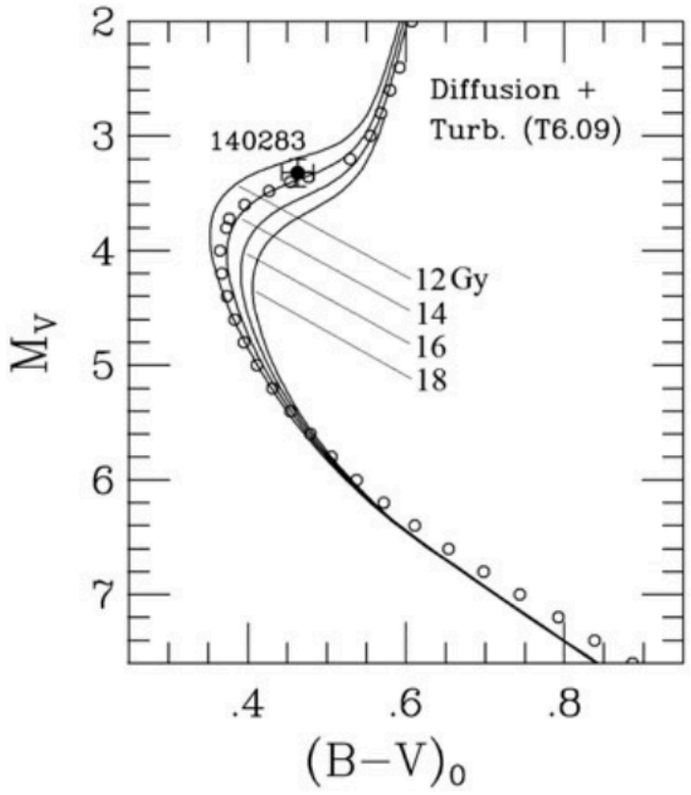


Fig. 9.5 Hertzsprung-Russell diagram for the globular cluster M92. The magnitude M_V is a measure of luminosity, and the $B-V$ parameter is a measure of temperature (in this plot, the temperature decreases from left to right). Only the metal-poorest stars in the cluster are included in this plot, because these are the oldest stars. Theoretical predictions for luminosity and temperature are plotted for assumed ages of 12, 14, 16, and 18 billion years. (From VandenBerg et al., 2002)

Instead of the monotonic increase of temperature with luminosity expected for a young population of stars, this cluster shows a drastic *decrease* of temperature with luminosity in the upper part of the diagram, which is an effect of aging (note that along the horizontal axis of this plot, the temperature decreases from left to right).

The several curves plotted in this diagram display the theoretical predictions for luminosity and temperature for different assumed ages for the cluster.

Comparison of these theoretical curves with the observational points indicates an age of 13.5×10^9 years for this cluster.

Similar age determinations have been made for several other globular clusters.

Some are old, some are young; but for the oldest globular clusters, the available age determinations give about $(13 \pm 2) \times 10^9$ years.

In principle, the ages of stars and of chemical elements give us only a lower limit for the age of the universe.

The universe could be much older than the stars and the chemical elements.

However, according to the standard model of the Big Bang, star formation and nucleosynthesis began soon ($\approx 10^8$ years) after the Big Bang.

Hence the ages of stars and of chemical elements may be regarded as equal to the age of the universe, to within the stated uncertainties.

The age of the universe arrived at by these observational data is consistent with the Hubble age, just as the Big Bang theory leads us to expect.

9.5 Cosmic background radiation

At a very early epoch, the dense and hot universe must have contained thermal blackbody radiation at a very high temperature.

As the universe expanded, the density of this thermal radiation decreased, but the radiation is still present, filling the interstellar and intergalactic space around us even now.

The expansion of the universe, and the expansion of the radiation, leads to a reduction of the temperature of the radiation.

To see what happens to the thermal radiation in some expanding region of space, imagine that this region, of size much smaller than the size of the universe, is enclosed in a box with perfectly reflecting walls.

Suppose that this box expands at the same rate as the universe.

Under these assumptions, the radiation in the region inside the box will behave in the same way as the radiation in some region outside the box.

The reason is that whereas in a region surrounded by walls, photons are reflected and cannot escape, in a region without walls, any photons that do escape are, on the average, simply replaced by similar photons that enter the region from the surrounding space.

It is then easy to calculate what happens to the radiation temperature in the expanding box.

At temperature T, the number of photons in some normal mode of oscillation of the box is given by statistical mechanics as

$$n(\nu) = \frac{1}{e^{h\nu/kT} - 1} \tag{9.33}$$

where ν is the frequency of the mode.

If the linear dimension of the box increases by a factor α, then the wavelength of the mode must increase by the same factor, and hence the frequency decreases to the value $\nu' = \nu/\alpha$.

If the expansion proceeds sufficiently slowly (adiabatically), the number of photons remains constant as their frequency changes.

Hence after expansion by the factor α, Eq. (9.33) will give the number of photons at the frequency ν':

$$n(\nu') = \frac{1}{e^{h\nu/kT} - 1} = \frac{1}{e^{\alpha h\nu'/kT} - 1} \tag{9.34}$$

Obviously Eq. (9.34) still corresponds to a thermal distribution of photons, but with the new temperature $T' = T/\alpha$.

This shows that expansion by, say, a factor of 2 changes thermal radiation of a given temperature into thermal radiation of one-half the initial temperature.

In an attempt at a detailed calculation of nucleosynthesis in the early universe, Gamow and his collaborators estimated the temperature of the primordial fireball as $\approx 10^9$ K at a time of ≈ 200 s after the Big Bang.

From this Alpher and Herman estimated the present temperature of the remaining thermal radiation at $\approx$ 5 K (for details, see Section 10.2).

Such a low temperature implies that the maximum in the spectral distribution of the radiation (according to Planck's law) is at a wavelength of a few millimeters, which is the wavelength of microwaves.

Some early evidence for a nonzero temperature of interstellar space had been uncovered by astronomers by spectroscopic examination of interstellar clouds of cyanogen gas.

But at the time, this was not recognized as indicating the presence of microwave radiation.

The first direct observational evidence for a uni-form distribution of microwave radiation over the sky was found by Penzias and Wilson with a very sensitive horn antenna originally constructed for a satellite communication system.

This microwave radiation was identified as the cosmic fireball radiation by Dicke et al.

These first measurements were performed at a wavelength of several centimeters, and they gave a radiation temperature of about 3 K.

Later measurements of the intensity of the radiation were performed over a wide range of wavelengths.

For wavelengths between ≈ 100 and 0.3 cm, ground-based radiometers can be used, but for wavelengths below 0.3 cm the radiometers must be carried above the atmosphere by a rocket, balloon, or satellite because the atmosphere absorbs radiation at these wavelengths.

Ground-based measurements and rocket and balloon measurements between 1965 and 1989 indicated a Planck spectrum, but with some possible deviations at short wavelengths.

All of these measurements were superseded by the new measurements performed with the COBE satellite (Cosmic Background Explorer) launched in 1989, which showed that the radiation has a perfect Planck spectrum, with a temperature of 2.73 K (see Fig. 9.6).

The distribution of the radiation over the sky is nearly isotropic, but displays two large-scale systematic deviations (Fig. 9.7).

The first of these deviations is of the form cos θ (a “dipole” term), which can be attributed to the motion of the Earth relative to the blackbody radiation.

Because of Doppler shifts, such a motion of the Earth increases the temperature of the radiation incident on the Earth from the front, and it reduces the temperature of the radiation incident from behind.

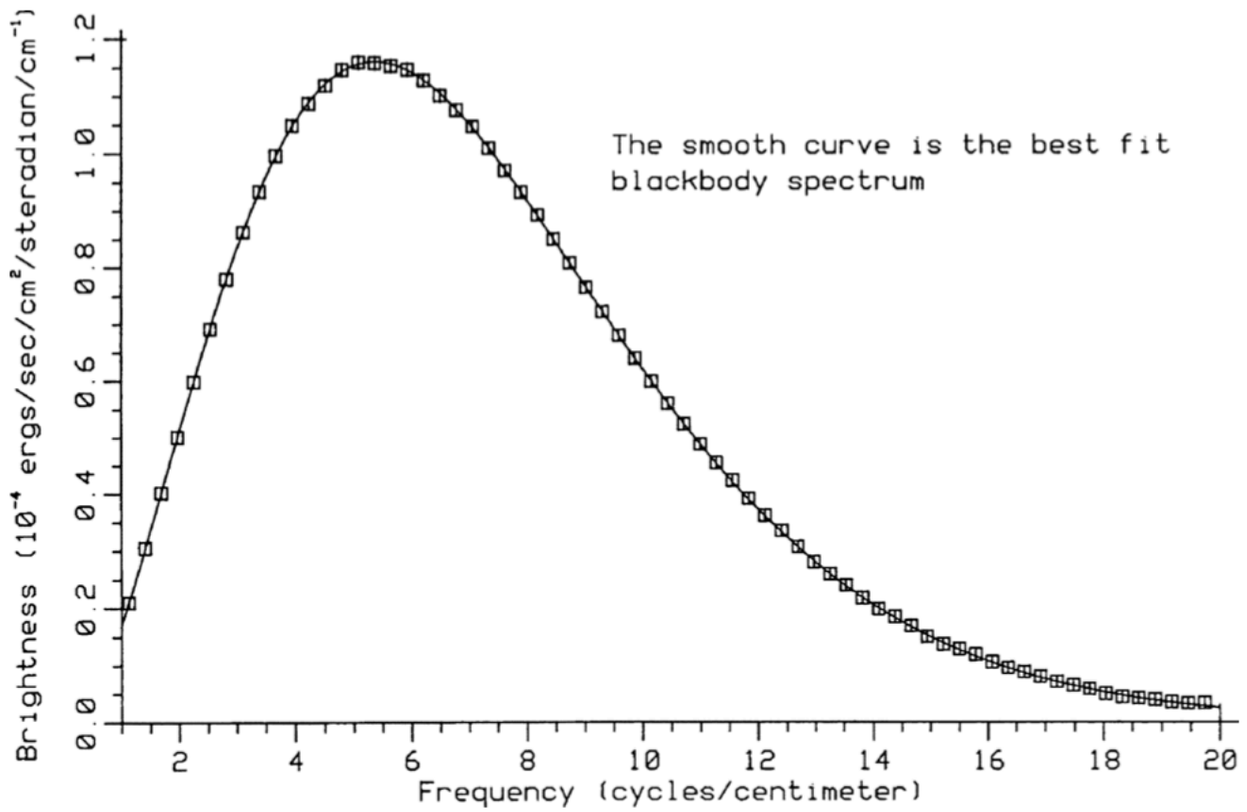


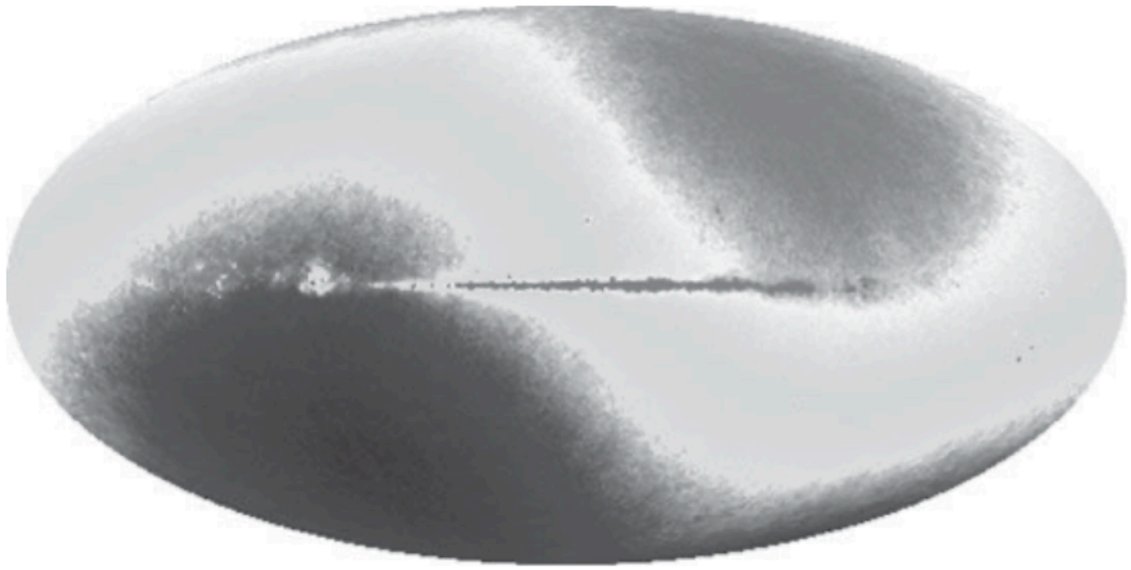
Fig. 9.6 The spectrum of the cosmic background radiation as measured by the COBE satellite. The smooth curve is the Planck blackbody spectrum for $T = 2.73$ K. More recent measurements and data analysis have reduced the error boxes to about 1/300 of what is shown here, so they are now smaller than the width of the line used for plotting the curve. The best available value for the temperature is now 2.7255 ± 0.0006 K (From Mather et al., 1990; see also Fixsen et al., 1996; Fixsen, 2009).

This results in a small additive perturbation in the temperature, of the form $\cos\theta$, where the angle θ is measured from the line of motion of the Earth.

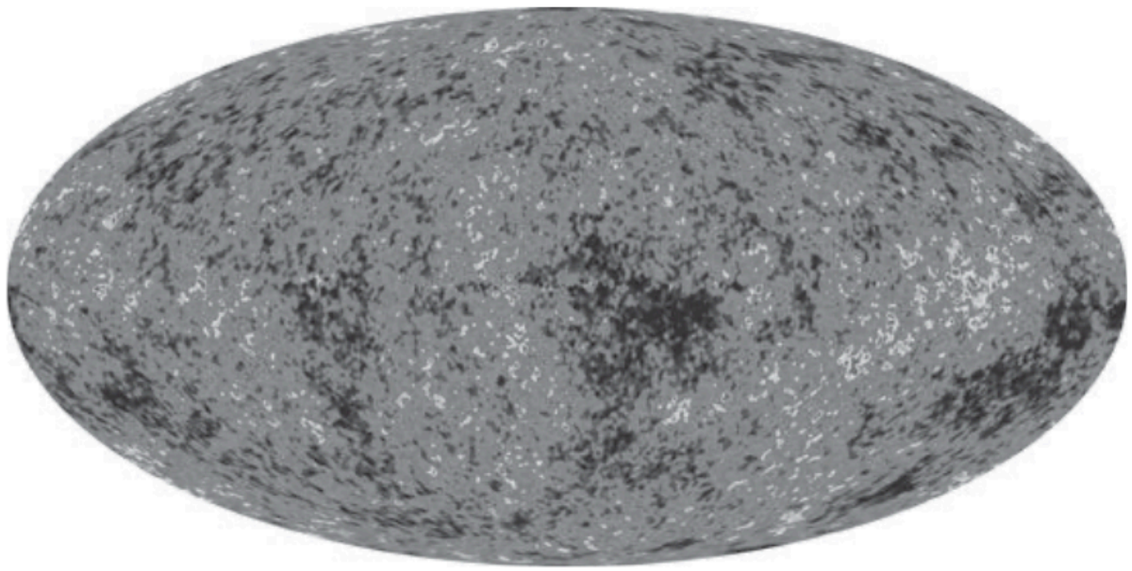
The observed temperature variation is about $\pm 3.3 \times 10^{-3}$ K and indicates a velocity of about 370 km/s.

Note that this determination of the velocity of the Earth in no way contradicts the principle of relativity, since the measurement is not made relative to empty space, but relative to the photon gas of the blackbody radiation.

The other large-scale deviation from isotropy is concentrated along the Milky Way and is attributed to radiation emitted by gas and by electrons (from ionized hydrogen) in our Galaxy.



(a)



(b)

Fig. 9.7 (a) Map of cosmic background radiation for the whole sky according to the survey by the WMAP satellite. The yin-yang pattern is a dipole anisotropy resulting from the motion of the Earth relative to the cosmic background radiation. This motion generates a difference of 0.007 K between the regions of the sky at the front and the rear. (b) In this high-contrast map, both the dipole contribution from the motion of the Earth and the contribution from thermal sources in our Galaxy have been subtracted. The residue reveals small temperature fluctuations in the CBR, with hot and cold regions differing by ± 0.0002 K from the average temperature (the map also includes some small amount of instrumental noise). (WMAP/NASA)

By means of a differential radiometer, with two matched antennas pointed in different directions, the COBE satellite has also detected various additional anisotropies in the radiation, on a smaller angular scale and of smaller magnitudes.

Some of these anisotropies arise from scattering of the radiation by electrons in the intergalactic space within galaxies in intervening clusters of galaxies (Sunyaev-Zel'dovich effect).

But the anisotropies of most interest to cosmologists are primary fluctuations associated with perturbations in the density and the gravitational potential at the point of origin of the radiation, where the photons last interacted with the matter in the early universe (the surface of "last scattering") before beginning their long unimpeded travel to the Earth.

These anisotropies arise from hills or wells in the gravitational potential, from which photons acquired small increments or decrements in frequency (Sachs-Wolfe effect).

The Sachs-Wolfe effect is the dominant contribution to anisotropies on a scale of angles of several degrees or larger.

However, the cosmic background radiation also displays anisotropies on an angular scale of a degree and smaller, which were explored in detail by the WMAP satellite.

These small-scale anisotropies originate from density oscillations ("acoustic" oscillations) in the proton-electron-photon plasma that existed before the surface of last scattering.

As we will see in Section 10.4, such oscillations reveal important information about the early universe.

For the purposes of this chapter, the crucial fact about cosmic background radiation is that all the primary deviations from isotropy are extremely small, less than 1 part in 10^4 (see Fig. 9.7b).

This indicates an almost perfect isotropy of the mass distribution of the early universe, at the time of the last interaction between a typical photon in the radiation and normal matter (electrons and protons), at an age of about 200,000 years.

This is conclusive evidence for the overall isotropy (and homogeneity) of the universe.

9.6 Mass density; dark mass

To apply Einstein's equations to the dynamics of the universe, we need to know the energy-momentum tensor of the universe and, in particular, the mass density.

In view of the large-scale homogeneity of the universe, it will be sufficient to deal with the average mass density; that is, the mass density averaged over distances of 100 Mpc and volumes of $(100\ \text{Mpc})^3$ or more.

Astronomers find it convenient to express the average mass density as a product of two factors, the average luminosity density and the average mass-to-light ratio for galaxies:

$$\rho_0 = \text{luminosity density} \times M/L \tag{9.35}$$

The luminosity density is obtained from a direct count of galaxies per unit volume and their luminosity.

According to a recent survey, the value of this factor is $2 \times 10^8\ \text{hL}_\odot/\text{Mpc}^3$ (the factor h enters here because it determines the distance scale).

To find the average mass-to-light ratio M/L, we need to determine the masses of various kinds of galaxies.

The masses of spiral galaxies, in which stars and gas move in nearly circular orbits about the galactic center, can be determined by measuring the orbital velocities of stars and gas as a function of the radius (the "rotation curve").

If the mass distribution of the galaxy can be approximated as spherical or ellipsoidal, the orbital velocity at a radius r is given by Newton's equation of motion

$$v^2 r = GM(r) \tag{9.36}$$

where $M(r)$ is the mass included within the sphere or the ellipsoid of radius r.

Thus, the mass of a galaxy can be determined by a measurement of the orbital velocity of a star or a gas cloud at the outer edge of the galaxy.

This orbital velocity can be readily measured by means of the Doppler shifts of spectral lines of the star or gas cloud.

From such measurements, astronomers find that the visible, luminous portions of galaxies have mass-to-light ratios of the order of $M/L = (10h \text{ to } 20h) \times M_\odot/L_\odot$.

If all of the mass of a galaxy were in its visible, luminous portions, then we would expect that orbital velocities beyond the visible outer edge of the galaxy would decrease according to $v^2 \propto 1/r$, because M(r) would remain constant beyond the outer edge.

However, measurements on isolated stars and gas clouds in this region show that the orbital rotation velocity at and beyond the visible outer edges of galaxies tends to a constant value, with no indication of a decrease with radius, not even at the largest radii for which data are available (see Fig. 9.8).

Such a flat rotation curve indicates that there is extra, invisible, dark mass beyond the visible outer edges of the galaxies. (Boccio - or Newtonian equation fails for small accelerations - MOND theory)

This dark mass is believed to form spherical halos around galaxies.

To give a flat, constant rotation curve, the mass of this halo must increase with radius, $M(r) \propto r$, and the density of the halo, as a function of radius, must therefore be proportional to $1/r^2$.

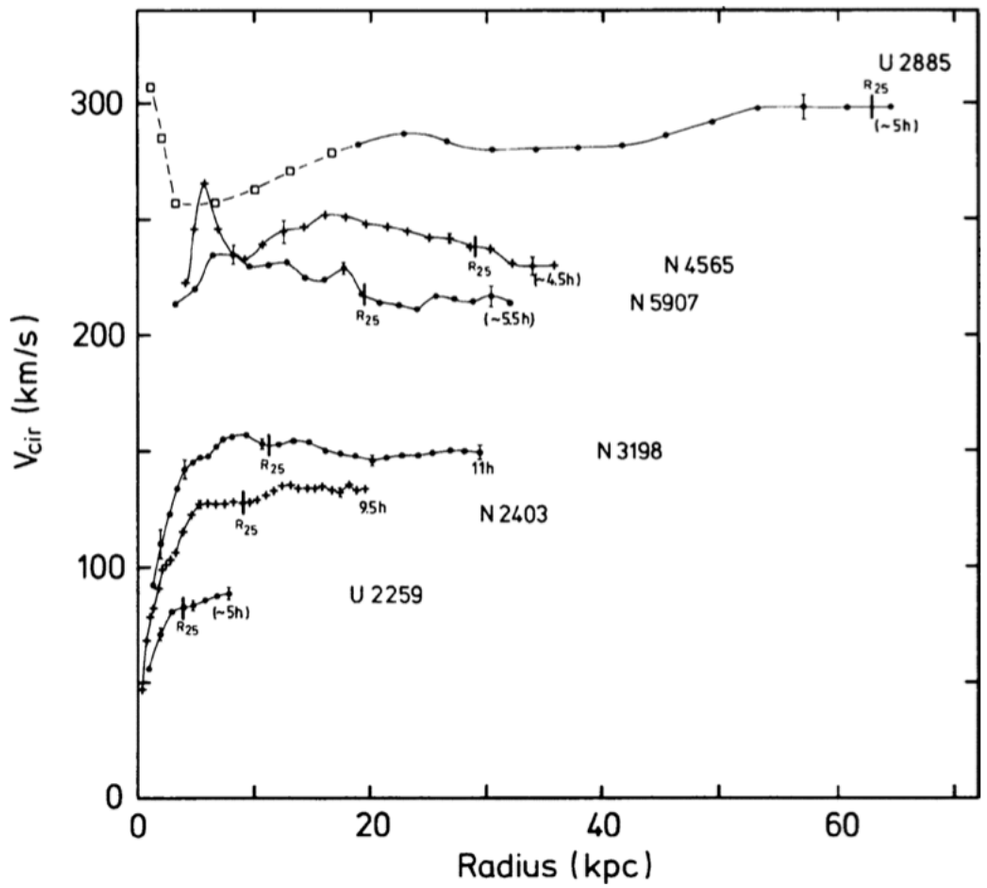


Fig. 9.8 Rotation curves of several spiral galaxies. These curves show the orbital velocities of clouds of hydrogen gas; the velocities were measured by the Doppler shifts of the 21-cm line. The small vertical markers indicate the outer edges of the visible portions of these galaxies. (From Sancisi and van Albada, 1987)

We do not know what the dark mass, or the "missing mass," in halos consists of.

It has been conjectured that it could consist of black holes, cold white dwarfs, or brown dwarfs (bodies like Jupiter, too small to support thermonuclear reactions).

Alternatively, it could consist of clouds of elementary particles, such as neutrinos (since the neutrino mass is nonzero), or some kind of exotic particles.

We will consider some of these speculations regarding the nature of the dark mass in Section 10.3 (see Table 10.3).

For our determination of the average mass density of the universe, we need the total masses of galaxies, including the masses of their halos.

Rotation curves cannot be measured far enough outward beyond the edges of galaxies to detect the extent of the halos.

However, we can determine the total masses of galaxies by measuring the velocities of galaxies in binary systems, small groups, and clusters.

In such systems, the rms velocity of the galaxies and the mean inverse distance between the galaxies are related to the total mass M by the virial theorem of statistical mechanics.

If, for the sake of simplicity, we assume that all the galaxies in the system have equal masses, the virial theorem can be put in the form

$$\langle v^2 \rangle = \tfrac{1}{2} G M \langle r^{-1} \rangle \tag{9.37}$$

This equation is a generalization, for a large number of particles, of the well-known result that for one particle moving in a circular orbit in a 1/r potential, the kinetic energy equals one-half of the absolute value of the potential energy.

In the case of binary systems or small groups containing only a few galaxies, this equation is valid only in a statistical sense, when averaged over an ensemble consisting of a large number of such systems.

In the case of clusters containing many galaxies, the equation is valid when averaged over an ensemble of clusters, but it is also valid for each individual cluster, provided the galaxies in the cluster have orbited around each other long enough for their velocities to attain statistical equilibrium.

With the measured velocities and positions of galaxies, the virial theorem leads to mass-to-light ratios of about 100h $M_\odot/L_\odot$ in binary systems and about 300h $M_\odot/L_\odot$ in groups and in large, rich clusters, such as the Coma cluster.

Since most of the galaxies are found in groups, we can adopt 300h $M_\odot/L_\odot$ as a representative value of the mass-to-light ratio for the entire population of galaxies.

If we multiply this average mass-to-light ratio by the average luminosity density, we obtain the average mass density contributed by galaxies and the halos of dark mass associated with them:

$$\begin{aligned}\rho_0 &= \text{luminosity density} \times M/L = 2 \times 10^8\, h\ L_\odot/\text{Mpc}^3 \times 300h\ M_\odot/L_\odot \\ &= 6 \times 10^{10} h^2\ M_\odot/\text{Mpc}^3 = 4 \times 10^{-30} h^2\ \text{g/cm}^3\end{aligned} \tag{9.38}$$

The corresponding density parameter is

$$\Omega_{m,0} = \frac{\rho_0}{3H_0{}^2/8\pi G} = \frac{4 \times 10^{-30} h^2\ \text{g/cm}^3}{1.88 \times 10^{-29} h^2\ \text{g/cm}^3} = 0.2 \tag{9.39}$$

Note that the fudge factor h cancels in this equation; thus, the uncertainty in H_0 does not affect the density parameter [although it affects the density ϱ_0; see Eq. (9.38)].

However, the uncertainties in the mass determinations of galaxies and clusters affect $\Omega_{m,0}$, and a rough estimate of the probable error is about ±0.5, which means that $\Omega_{m,0}$ lies somewhere between 0.15 and 0.25.

These determinations of the average mass density refer only to the luminous mass in galaxies and the dark mass in the halos of galaxies.

Besides this, there could be extra dark mass uniformly distributed throughout the universe.

The only observational limit on such a uniform dark mass distribution comes from the redshift-distance data on supernovas.

As described in Section 9.3, this gives the result $\Omega_{m,0} = 0.3$.

The discrepancy between this number and Eq. (9.39) suggests that even the deep intergalactic space between clusters contains some dark mass.

A different constraint on the mass density can be extracted from a detailed analysis of the process of nucleosynthesis in the early universe.

As we will see in Section 10.2, the early universe cannot produce the observed amount of deuterium found in interstellar and intergalactic clouds unless the baryon contribution to $\Omega_{m,0}$ is set at $\Omega_B h^2 = 0.02$, which implies $\Omega_B = 0.04$ if $h = 0.7$.

Accordingly, ordinary baryonic matter (protons and neutrons) contributes only a small fraction of the mass content of the universe.

Most of the mass is in the form of nonbaryonic matter of unknown composition, perhaps consisting of gravitational radiation, black holes, neutrinos, or exotic, unknown elementary particles.

Theoretical considerations concerning the early universe strongly suggest that the most plausible value of Ω_0 is exactly 1 (see Section 10.4).

If our universe actually has this value of Ω_0, then the upper limit on baryonic matter set by deuterium production implies that by far most of the matter in the universe must be something very different from the familiar form of matter we find in our immediate environment.

Table 9.1 lists the known contributions to the average mass density of the universe.

Table 9.1 Average mass density in the universe*

Form of Matter	Contribution to Ω_0	ρ_0 (g/cm^3)
Luminous mass in galaxies	0.01	0.1×10^{-30}
Luminous and dark mass in galaxies and clusters	0.2	2×10^{-30}
Luminous and dark mass, total	0.3	2.8×10^{-30}
Baryonic mass	0.04	0.36×10^{-30}
Dark energy (cosmological constant)	0.7	6.4×10^{-30}
All forms	1	9.2×10^{-30}

* Assumes $h = 0.7$.

Besides the forms of matter listed in Table 9.1, the universe contains some amount of electromagnetic energy.

For example, the cosmic background radiation at 2.74 K contributes 4×10^{-34} g/cm^3, and radio waves, starlight (optical), and X rays contribute smaller amounts.

There are also cosmic rays with a density no larger than 10^{-35} g/cm^3.

9.7 Comoving coordinates; Robertson-Walker geometry

In our calculations of the large-scale dynamics of the universe, we will ignore any small-scale deviations from uniformity.

We want to concentrate on the average features of the geometry, and we will assume that on average the universe is isotropic and homogeneous.

The mass in the universe can then be treated as a fluid, with a constant density throughout.

The galaxies, or clusters of galaxies, may be regarded as the particles out of which this fluid is made.

This means that we ignore local irregularities in the density and the motion of galaxies, and we take into account only the motion of uniform expansion.

By appealing to the uniformity of the universe, we can simplify the spacetime interval and the metric tensor considerably.

We begin by specifying the coordinates we will use: they will be comoving coordinates.

By this is meant that the spatial coordinates partake of the uniform motion of expansion of the universe.

If we ignore the small irregularities in the motion of galaxies (local deviations from uniform expansion), we may say that each galaxy carries its spatial coordinates with it; the coordinate points move with the galaxies as the latter move freely in the background gravitational field of the universe.

The coordinate interval between any two galaxies then remains forever constant, and the expansion of the universe results not from a change in the coordinate position of the galaxies, but rather from a change in the metric tensor of spacetime.

For the time coordinate x^0 we will use the proper time measured by clocks carried by the galaxies.

We will assume that these clocks not only run at the same rate but also that they are synchronized.

In general, synchronization is a rather tricky problem, but in a uniform universe a simple operational procedure for synchronization is available.

An observer on a galaxy *A* sends out a flash of light when his clock reads time t_0; an observer on galaxy *B* also sends out a flash of light when her clock reads t_0.

The flash from *A* is received at *B* when the clock at the latter point reads t_B; the flash from *B* is received at *A* when the clock there reads t_A.

Theclocks are then synchronized if $t_A = t_B$.

This method depends on the uniformity of the universe; isotropy and homogeneity ensure that both light signals and both clocks behave in the same way, and if we adjust the synchronization on one occasion, the clocks will remain synchronized forever.

In principle, an alternative synchronization method is also available: All observers set their clocks so they read zero at the instant of the Big Bang. But in practice, this is impossible to implement with any precision.

We can show that with the above choice of coordinates,

$$g_{00} = 1 \tag{9.40}$$

$$g_{0k} = 0 \tag{9.41}$$

This means that our comoving coordinates are time orthogonal [see Eq. (7.34)].

Equation (9.40) follows from the assumption that the clocks used to measure x^0 are at rest in our comoving coordinates.

For such a clock $dx^1 = dx^2 = dx^3 = 0$, and hence $d\tau^2 = g_{00}(dx^0)^2$.

Thi sestablishes Eq.(9.40) because, by definition of x^0, we have $dx^0 = d\tau$.

To establish Eq. (9.41), we note with a nonzero value of g_{0k} light signals in opposite direction would have different speeds, which is excluded by the symmetry of our synchronization procedure.

Since we have chosen our coordinates so galaxies are forever at rest, we must check that this is consistent with the geodesic equation of motion.

To see that a galaxy at rest at some initial time will remain at rest, consider the geodesic equation of motion for a particle at rest, with $dx^\mu/d\tau = (1, 0, 0, 0)$,

$$\frac{d^2x^k}{d\tau^2} = -\Gamma^k{}_{00} \tag{9.42}$$

But, according to Eqs. (9.40) and (9.41), $\Gamma^k{}_{00} = \frac{1}{2} g^{kl}(g_{l0,0} + g_{0l,0} - g_{00,l}) = 0$.

Hence the particle has no acceleration, and it remains at rest.

In consequence of Eqs. (9.40) and (9.41), the spacetime interval reduces to

$$ds^2 = (dx^0)^2 + g_{kn}dx^k dx^n \tag{9.43}$$

It is customary to write this as

$$ds^2 = dt^2 - dl^2 \tag{9.44}$$

where $dt = dx^0$ is simply the time interval and

$$dl^2 \equiv -g_{kn}dx^k dx^n \tag{9.45}$$

is the distance interval.

To emphasize the 3-D nature of the distance interval, we will write Eq. (9.45) as

$$dl^2 = {}^{(3)}g_{kn}dx^k dx^n \tag{9.46}$$

where

$${}^{(3)}g_{kn} \equiv -g_{kn} \tag{9.47}$$

The 3×3 tensor ${}^{(3)}g_{kn}$ describes the geometry of 3-D space at a given instant of time.

As a first step toward the construction of the geometry of spacetime, we construct the geometry of 3-D space.

This space is supposed to be homogeneous and isotropic: The geometry should not distinguish between different points or between different directions about a point.

To discover the consequences of these symmetry requirements, it is helpful to concentrate on the curvature tensor ${}^{(3)}R_{mnkl}$ of the three-geometry.

It follows from the isotropy and homogeneity of the three-geometry that ${}^{(3)}R_{mnkl}$ must have the form

$${}^{(3)}R_{mnkl} = K\left({}^{(3)}g_{mk}\,{}^{(3)}g_{nl} - {}^{(3)}g_{ml}\,{}^{(3)}g_{nk}\right) \tag{9.48}$$

where K is some constant.

We can justify Eq. (9.48) as follows: At a given point, introduce local geodesic coordinates so the metric tensor becomes ${}^{(3)}g'_{ml} = \delta^{l}_{m}$.

The isotropy of space demands that it must not be possible to distinguish between different directions by their curvature; that is, the curvature tensor ${}^{(3)}R_{mnkl}$ must be unchanged by rotations of the geodesic coordinates.

Since the unit tensor δ^{l}_{k} is the only tensor unchanged by rotations, the curvature tensor in these coordinates must be some combination of unit tensors,

$$ {}^{(3)}R'_{mnkl} = K\delta^k_m\delta^l_n + K_1\delta^l_m\delta^k_n + K_2\delta^n_m\delta^k_l \tag{9.49} $$

The antisymmetry relation ${}^{(3)}R'_{mnkl} = -{}^{(3)}R'_{mnlk}$ requires that $K_1 = -K$ and $K_2 = 0$; hence

$$ {}^{(3)}R'_{mnkl} = K\left(\delta^k_m\delta^l_n - \delta^l_m\delta^k_n\right) \tag{9.50} $$

Transformation of this equation from geodesic coordinates back to the original coordinates then gives the form (9.48).

It is easy to show that ${}^{(3)}R_{nk} = -2K{}^{(3)}g_{nk}$ and ${}^{(3)}R = -6K$.

The quantity K must be a constant (independent of position) to satisfy the requirement of homogeneity.

This conclusion also follows from isotropy: If there were some point at which the gradient of K is different from zero, then the direction of this gradient would define a preferred direction in space, in contradiction with isotropy.

Thus, isotropy at all points implies that K is a constant, which means it implies homogeneity.

Because the curvature tensor ${}^{(3)}R_{mnkl}$ can be expressed in terms of the metric tensor ${}^{(3)}g_{nk}$ and its first and second derivatives, Eq. (9.48) may be regarded as a differential equation for ${}^{(3)}g_{nk}$.

We want to find a solution of this equation; for this purpose we consider the cases of positive, negative, and zero curvature separately.

Positive Curvature. Rather than try to solve Eq. (9.48) by brute force, we will exploit geometric arguments.

First consider the analogous 2-D problem.

What 2-D space has uniform (that is, homogeneous and isotropic) curvature?

Obviously, the surface of an ordinary sphere has this property; in mathematics, this surface is usually called a two-sphere.

Correspondingly, to obtain a 3-D space of uniform curvature, we take the surface of a 4-D hypersphere; this surface is called a three-sphere.

The equation of the surface of a 4-D hypersphere is, in rectangular coordinates,

$$(x^1)^2 + (x^2)^2 + (x^3)^2 + (x^4)^2 = a^2 \tag{9.51}$$

where a is the radius of the sphere.

The distance between any two nearby points on the surface is

$$dl^2 = (dx^1)^2 + (dx^2)^2 + (dx^3)^2 + (dx^4)^2 \tag{9.52}$$

Note that the coordinate x^4 has nothing to do with time; it is simply an extra, unphysical coordinate that must be introduced if we want to pretend that the curved 3-D space is a subspace of a flat 4-D Euclidean space.

Equation (9.51) can be used to eliminate differential $(dx^4)^2$ of the unphysical coordinate x^4.

This yields a distance interval

$$dl^2 = (dx^1)^2 + (dx^2)^2 + (dx^3)^2 + \frac{(x^1dx^1 + x^2dx^2 + x^3dx^3)^2}{a^2 - (x^1)^2 - (x^2)^2 - (x^3)^2} \tag{9.53}$$

In this expression there appear only the physical coordinates x^1, x^2, x^3.

This expression gives us the desired metric tensor for a space of uniform curvature.

We can easily check that the curvature tensor for the distance interval (9.53) does satisfy Eq. (9.48), with a particular value of K,

$$K = \frac{1}{a^2} \tag{9.54}$$

It is sufficient to check Eq. (9.54) near the origin, because the geometry described by Eq. (9.53) is uniformly curved (by construction), and what holds at one point must hold at all. In the vicinity of the origin it is sufficient to approximate

$$ {}^{(3)}g_{kn} \cong \delta_k^n + x^k x^n / a^2] \tag{9.55} $$

■

A geometry with a positive value of K is said to have positive curvature.

To study this geometry further, we begin by introducing "polar" coordinates in the usual way:

$$ x^1 = r \sin\theta \cos\phi, \quad x^2 = r \sin\theta \sin\phi, \quad x^3 = r \cos\theta \tag{9.56} $$

In terms of these coordinates, the distance interval (9.53) takes the form

$$ dl^2 = \frac{dr^2}{1 - r^2/a^2} + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \tag{9.57} $$

Note that for each set of values of the "rectangular" coordinates x^k and for each value of the "radial" coordinate r there are actually two distinct points on the three-sphere (the coordinates do double duty).

If we start at the "top" of the three-sphere ($x^4 = a$), move in the x^1-direction and continue straight ahead, we reach the equator ($x^4 = 0$) and then the bottom of the sphere ($x^4 = -a$).

If we continue straight ahead, we again pass through the equator and finally return to the top. The values of x^1 for the top, the first equatorial crossing, the bottom, and the second equatorial crossing are, respectively, $x^1 = 0$, $x^1 = a$, $x^1 = 0$, and $x^1 = -a$; and the values of r are, respectively, $r = 0$, $r = a$, $r = 0$, and $r = a$ (Fig. 9.9).

That there are two distinct points with the same values of x^1 and r need not bother us; we can resolve the ambiguity by, say, writing coordinates for points in the upper hemisphere in red ink, and points in the lower in blue.

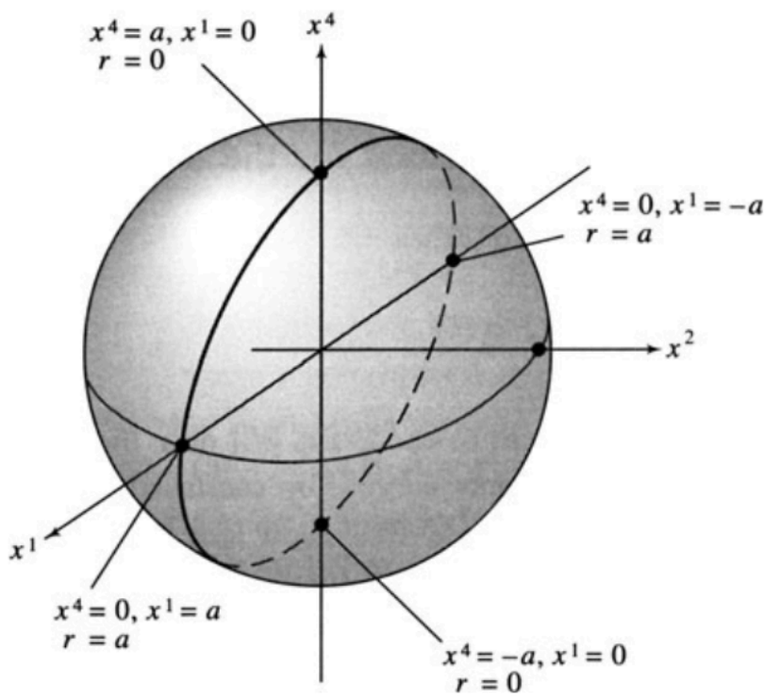


Fig. 9.9 Section through the three-sphere in the plane $x^3 = 0$. The equation for this section is $(x^1)^2 + (x^2)^2 + (x^4)^2 = a^2$; thus, this section is the surface of an ordinary sphere. The heavy line shows a path that goes from $x^4 = a$ to $x^4 = 0$ to $x^4 = -a$ to $x^4 = 0$ to $x^4 = a$.

To gain some feeling for our space of positive curvature, let us look at the radius and the circumference of a circle placed in this space.

For convenience, take the circle defined by a radial coordinate r = constant around the origin.

This circle has as its radius the measured distance between 0 and r,

$$l = \text{radius} = \int_0^r \frac{dr'}{\sqrt{1 - r'^2/a^2}} = a \sin^{-1} \frac{r}{a} \tag{9.58}$$

The circumference of the circle is obtained in the usual way.

For example, if the circle is in the plane $\theta = \pi/2$,

$$\text{circumference} = \int_0^{2\pi} r \sin\theta d\phi = 2\pi r \tag{9.59}$$

The ratio of the radius l to the circumference is therefore more than $1/2\pi$; this is a familiar property of spaces of positive curvature.

Note that for a radius larger than $\pi a/2$, the circumference $2\pi r$ decreases as the radius l increases.

The surface area of the sphere r = constant surrounding the origin has the usual value

$$\text{area} = \int_0^{2\pi} \int_0^{\pi} r^2 \sin\theta d\theta d\phi = 4\pi r^2 \tag{9.60}$$

Hence the ratio of the radius squared l^2 to the area is more than $1/4\pi$.

For a radius l larger than $\pi a/2$, the area decreases as the radius l increases.
The volume inside the sphere r = constant is

$$\begin{aligned} \text{volume} &= \int_0^{2\pi} \int_0^{\pi} \int_0^{r} \frac{r'^2}{\sqrt{1 - r'^2/a^2}} dr' \sin\theta d\theta d\phi \\ &= 4\pi \left(\frac{a^3}{2} \sin^{-1} \frac{r}{a} - \frac{ra^2}{2} \sqrt{1 - r^2/a^2} \right) \end{aligned} \tag{9.61}$$

To obtain the total volume of the three-sphere we must take $r = 0$ but $\sin^{-1}r/a = \sin^{-1}0 = \pi$ ("bottom" of sphere). The total volume is then $2\pi^2a^3$.

Our three-sphere is a closed space; it has a finite volume even though it has no boundaries.
It is sometimes convenient to replace the radial coordinate r by a new "angular" coordinate χ such that

$$r = a \sin\chi, \quad 0 < \chi < \pi \tag{9.62}$$

The coordinate χ has an advantage over the coordinate r: it stands in unique correspondence to the points of space. In terms of χ, the distance interval becomes

$$dl^2 = a^2[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta d\phi^2)] \tag{9.63}$$

In this expression, a plays the role of an overall scale factor that characterizes distances in the three-geometry.

For instance, according to Eq. (9.58), the radial distance from the origin to the point χ is simply

$$l = \text{radial distance} = a\chi \tag{9.64}$$

Negative Curvature. Next we want to find a homogeneous and isotropic geometry of negative curvature ($K < 0$).

The spacetime interval for such a geometry can be obtained from Eq. (9.53) by simply replacing a^2 by $-a^2$:

$$dl^2 = (dx^1)^2 + (dx^2)^2 + (dx^3)^2 + \frac{(x^1 dx^1 + x^2 dx^2 + x^3 dx^3)^2}{-a^2 - (x^1)^2 - (x^2)^2 - (x^3)^2} \tag{9.65}$$

The geometric arguments that precede Eq. (9.53) are, of course, not applicable when this replacement is made.

Because the metric tensor corresponding to Eq. (9.53) satisfies the condition (9.48) on the curvature tensor, the metric tensor corresponding to Eq. (9.65) will also satisfy this condition, the algebra being the same in both cases.

The only difference is that the value of K is now negative,

$$K = -\frac{1}{a^2} \tag{9.66}$$

In "polar" coordinates, Eq. (9.65) becomes

$$dl^2 = \frac{dr^2}{1 + r^2/a^2} + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \tag{9.67}$$

The circle r = constant around the origin has as its radius the measured distance

$$l = \text{radius} = a \sinh^{-1} \frac{r}{a} \tag{9.68}$$

This makes the ratio of radius to circumference smaller than $1/2\pi$.

Likewise, for a sphere, the ratio of radius squared to area is smaller than $1/4\pi$.

The volume of a sphere with r = constant is

$$\text{volume} = 4\pi\left(-\frac{a^3}{2}\sinh^{-1}\frac{r}{a} + \frac{ra^2}{2}\sqrt{1+r^2/a^2}\right) \tag{9.69}$$

For r $\rightarrow \infty$, this volume diverges.

The space of negative curvature is open and infinite.

We can introduce a dimensionless coordinate χ such that

$$r = a\sinh\chi \tag{9.70}$$

In terms of this coordinate,

$$dl^2 = a^2[d\chi^2 + \sinh^2\chi(d\theta^2 + \sin^2\theta d\phi^2)] \tag{9.71}$$

This is the analog of Eq. (9.63)

Zero Curvature. Obviously, Eq. (9.48) also has the trivial solution $K = 0$ and ${}^{(3)}\mathrm{R}_{\mathrm{mnkl}} = 0$.

This corresponds to flat, Euclidean space.

At first sight, such a flat geometry seems of no interest since we certainly expect that the matter distribution in the universe will produce some curvature.

However, we must remember that matter generates curvature in the four-geometry; it need not generate curvature in the three-geometry.

As we will see, it is possible to construct solutions of Einstein's equations such that the three-geometry is flat at each instant of time, but the four-geometry is curved.

The flat space is of course open and infinite.

Table 9.2 summarizes the properties of our homogeneous, isotropic three-geometries of constant curvature.

Table 9.2 Homogeneous, isotropic three-geometries

Curvature	Distance Interval*	Radial Distance	Three-volume
Positive, $K = \frac{1}{a^2}$	$dl^2 = \frac{dr^2}{1 - r^2/a^2} + r^2 d\Omega^2$	$l = a \sin^{-1} r/a$	finite, closed
Negative, $K = -\frac{1}{a^2}$	$dl^2 = \frac{dr^2}{1 + r^2/a^2} + r^2 d\Omega^2$	$l = a \sinh^{-1} r/a$	infinite, open
Zero, $K = 0$	$dl^2 = dr^2 + r^2 d\Omega^2$	$l = r$	infinite, open

* $d\Omega^2 \equiv d\theta^2 + \sin^2\theta d\phi^2$

For these three-geometries we can now construct the four-geometry, according to $ds^2 = dt^2 - dl^2$.

For the cases of positive and negative curvature this leads to, respectively,

$$ds^2 = dt^2 - \frac{dr^2}{1 - r^2/a^2} - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \tag{9.72}$$

$$ds^2 = dt^2 - \frac{dr^2}{1 + r^2/a^2} - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \tag{9.73}$$

These spacetime intervals define the *Robertson-Walker* geometries of positive and of negative curvature.

These spacetime intervals are often written in the form

$$ds^2 = dt^2 - a^2 \left(\frac{dr^2/a^2}{1 \mp r^2/a^2} + \frac{r^2}{a^2} d\theta^2 + \frac{r^2}{a^2} \sin^2\theta d\phi^2 \right) \tag{9.74}$$

where the term in parentheses depends only on the dimensionless quantity r/a.

Although the 3-D geometry, of positive or negative curvature, can be described equally well by the coordinates x^1, x^2, x^3 or r, θ, ϕ, or χ, θ, ϕ, these coordinates are not equally good for describing the expanding universe.

Of these three sets of coordinates, only χ , θ , ϕ may be regarded as comoving with the galaxies.

The reason is that isotropy and homogeneity of the expansion require that if dl(t) is the distance between two galaxies at time t, then the distance at time t + Δt must be proportional to the initial distance,

$$dl(t + \Delta t) = f(t)dl(t) \tag{9.75}$$

where the proportionality factor f(t) depends on time, but not on the position in space.

Equation (9.75) simply says that when the distance between a pair of galaxies increases by a factor of f, then the distance between any other pair also increases by the same factor.

In the expressions (9.53), (9.57), and (9.63) for the distance interval, the only quantity that can change in time is a; an expanding universe will have a time-dependent value of a.

Only in the last of these expressions does a enter in such a way that Eq. (9.75) is satisfied.

Thus, the coordinates χ , θ , ϕ may be regarded as comoving coordinates with respect to which the galaxies are at rest.

Note that χ,θ,ϕ are not the only possible comoving coordinates.

Any time-independent functions of χ,θ,ϕ can be regarded as comoving coordinates; for instance, r/a = sinχ, θ,ϕ are comoving coordinates, and they are sometimes used instead of χ,θ,ϕ.

9.8 Friedmann models ($\rho \neq 0, \Lambda = 0$)

The time dependence of the scale factor a is determined by the Einstein equations.

We will solve these equations for the separate cases of the Robertson-Walker geometries of positive, negative, and zero (spatial) curvature.

To write down the Einstein equations, we must know the energy-momentum tensor of the matter.

We will assume that the energy-momentum tensor in the universe today is that of a uniform gas with zero pressure.

The galaxies may be regarded as the "particles" out of which this gas is made, and since the velocities of the galaxies do not deviate much from uniform expansion, we can neglect the "pressure" of the gas of galaxies (in general, kinetic pressure is an indication of an rms deviation of the velocities from uniform velocity).

However, in the very early, very hot universe, the pressure generated by radiation and by elementary particles must have been important; we will explore this in the next chapter.

The general homogeneous, isotropic models of the universe with mass density, pressure, and, possibly, a cosmological term are called the *Friedmann-Lemaıtre models*.

For the sake of a more detailed classification, we will call the special models with nonzero mass density and zero cosmological constant the Friedmann models, and we will call the models with zero mass density and nonzero cosmological constant the Lemaıtre models (these Lemaıtre models should be regarded as approximations, valid when the mass density is negligible compared with the cosmological constant).

We begin with a study of the Friedmann models.

Positive Curvature and $\Lambda = 0$. In the comoving coordinates χ,θ,ϕ, the spacetime interval for the Robertson-Walker geometry of positive curvature is

$$ds^2 = dt^2 - a^2(t)[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta d\phi^2)] \tag{9.76}$$

The solution of the Einstein equations becomes simplest if we replace the time coordinate t by a dimensionless time parameter η defined by

$$dt \equiv a d\eta \tag{9.77}$$

Then

$$ds^2 = a^2(\eta)[d\eta^2 - d\chi^2 - \sin^2\chi(d\theta^2 + \sin^2\theta d\phi^2)] \tag{9.78}$$

and the metric tensor is

$$g_{\mu\nu} = \begin{pmatrix} a^2 & 0 & 0 & 0 \\ 0 & -a^2 & 0 & 0 \\ 0 & 0 & -a^2\sin^2\chi & 0 \\ 0 & 0 & 0 & -a^2\sin^2\chi\sin^2\theta \end{pmatrix} \tag{9.79}$$

Some of the Christoffel symbols are easy to calculate:

$$\begin{aligned} &\Gamma^0{}_{00} = \frac{\dot{a}}{a} \quad \Gamma^0{}_{k0} = 0 \quad \Gamma^0{}_{kn} = -\frac{\dot{a}}{a^3}g_{kn} \\ &\Gamma^k{}_{00} = 0 \quad \Gamma^k{}_{0n} = \frac{\dot{a}}{a}\delta^n_k \end{aligned} \tag{9.80}$$

where

$$\dot{a} \equiv \frac{da}{d\eta} \tag{9.81}$$

Here the dots stands for the derivatives with respect to the time parameter η, contrary to the usage of Chapter 7 [see Eqs. (7.90)–(7.93)], where the dots stood for the derivative with respect to proper time.

These derivatives da/dt and da/dη are related as follows:

$$\frac{da}{d\tau} = \frac{da}{dt} = \frac{1}{a}\frac{da}{d\eta} = \frac{1}{a}\dot{a} \tag{9.82}$$

It is easy to check that the R_{kn} component of the Ricci tensor is entirely determined by Eq. (9.80):

$$R_{00} = \frac{3}{a^2}(a\ddot{a} - \dot{a}^2) \tag{9.83}$$

The R_{kn} components can be easily evaluated by relating them to ${}^{(3)}R_{kn}$ of the 3-D spatial geometry; this saves some labor.

We begin with the definition of R_{kn}

$$R_{kn} = R^{\alpha}{}_{kn\alpha} = R^{0}{}_{kn0} + R^{l}{}_{knl} \tag{9.84}$$

The term $R^{0}{}_{kn0}$ involves only Christoffel symbols of the type (9.80).

The term $R^{l}{}_{knl}$ equals

$$R^{l}{}_{knl} = -\Gamma^{l}{}_{kn,l} + \Gamma^{l}{}_{kl,n} + \left(\Gamma^{0}{}_{kl}\Gamma^{l}{}_{0n} + \Gamma^{m}{}_{kl}\Gamma^{l}{}_{mn}\right) - \left(\Gamma^{0}{}_{kn}\Gamma^{l}{}_{0l} + \Gamma^{m}{}_{kn}\Gamma^{l}{}_{ml}\right) \tag{9.85}$$

In this expression, the Christoffel symbols with one zero index are known from Eq. (9.80); the remaining symbols, with purely spatial indices, add up to the Ricci tensor of three-space, so

$$R^{l}{}_{knl} = {}^{(3)}R^{l}{}_{knl} + \Gamma^{0}{}_{kl}\Gamma^{l}{}_{0n} - \Gamma^{0}{}_{kn}\Gamma^{l}{}_{0l} \tag{9.86}$$

If we substitute

$${}^{(3)}R^{l}{}_{knl} = \frac{1}{a^2}\,{}^{(3)}g^{lm}\left[{}^{(3)}g_{mn}\,{}^{(3)}g_{kl} - {}^{(3)}g_{ml}\,{}^{(3)}g_{kn}\right] \tag{9.87}$$

according to Eq. (9.48), then everything on the left side of Eq. (9.86) is determined.

The result is

$$R_{kn} = \frac{1}{a^4}(2a^2 + \dot{a}^2 + a\ddot{a})g_{kn} \tag{9.88}$$

The curvature scalar is then given by

$$R = R^{0}{}_{0} + R^{k}{}_{k} = g^{00}R_{00} + g^{kn}R_{kn} = \frac{6}{a^3}(a + \ddot{a}) \tag{9.89}$$

The Einstein equation, including the cosmological term, is

$$R_{\mu}{}^{\nu} - \tfrac{1}{2}\delta^{\nu}_{\mu}R = -8\pi G T_{\mu}{}^{\nu} - \Lambda\delta^{\nu}_{\mu} \tag{9.90}$$

We will concentrate on the 00 component of this equation:

$$R_0{}^0 - \tfrac{1}{2}R = -8\pi G T_0{}^0 - \Lambda \tag{9.91}$$

With Eqs. (9.83) and (9.89) this becomes

$$-\frac{3}{a^4}(a^2 + \dot{a}^2) = -8\pi G T_0{}^0 - \Lambda \tag{9.92}$$

Note that in this equation $\ddot{a}$ has canceled, so we have a first-order differential equation.

This equation suffices to determine $a(t)$ if $T_0{}^0$ is given.

The other components of the Einstein equation can be ignored because it turns out that they tell us nothing different from Eq. (9.92).

For a gas of galaxy "particles" without pressure, the energy-momentum tensor is

$$T_\mu{}^\nu = \rho u_\mu u^\nu \tag{9.93}$$

where ϱ is the proper mass density.

In our comoving coordinates, the galaxies are at rest and hence

$$T_0{}^0 = \rho \tag{9.94}$$

The volume of the universe varies in time as a^3; hence the density varies as a^{-3}, and we can write

$$\rho(t) = \frac{\rho_0 a_0{}^3}{a^3} \tag{9.95}$$

Our general dynamical equation (9.92) now becomes

$$\frac{3}{a^4}(\dot{a}^2 + a^2) = 8\pi G\rho + \Lambda \tag{9.96}$$

or

$$\frac{3}{a^4}(\dot{a}^2 + a^2) = 8\pi G\frac{\rho_0 {a_0}^3}{a^3} + \Lambda \tag{9.97}$$

In the Friedmann models, it is assumed either that the cosmological constant is exactly zero, or else that it is so small that it can be neglected compared with ϱ.

Under this assumption, Eq. (9.97) reduces to

$$\frac{3}{a^4}(\dot{a}^2 + a^2) = 8\pi G\frac{\rho_0 {a_0}^3}{a^3} \tag{9.98}$$

This is the differential equation that describes the Friedmann model of positive curvature.

This differential equation can also be derived without making explicit use of Einstein's equations.

In Section 9.3 we gave a simple argument based on Newtonian physics that led to the following equation of motion for the distance between galaxies [see Eq. (9.15)]:

$$\frac{d^2 l}{dt^2} = -\frac{4\pi G}{3}\rho l + \frac{\Lambda}{3}l = -\frac{4\pi G}{3}\frac{\rho_0 {l_0}^3}{l^2} + \frac{\Lambda}{3}l \tag{9.99}$$

If we multiply this equation by dl/dt, each side becomes a perfect differential,

$$\frac{1}{2}\frac{d}{dt}\left(\frac{dl}{dt}\right)^2 = \frac{d}{dt}\left(\frac{4\pi G}{3}\frac{\rho_0 {l_0}^3}{l} + \frac{\Lambda}{6}l^2\right) \tag{9.100}$$

The integral of Eq. (9.100) is then

$$\frac{1}{2}\left(\frac{dl}{dt}\right)^2 + \text{constant} = \frac{4\pi G}{3}\frac{\rho_0 {l_0}^3}{l} + \frac{\Lambda}{6}l^2 = \frac{4\pi G}{3}\rho l^2 + \frac{\Lambda}{6}l^2 \tag{9.101}$$

To compare this with Eq. (9.98), we note that distances in the universe vary as $a(t)$, that is, $l(\mathrm{t}) \propto \mathrm{a(t)}$.

Furthermore, by Eqs. (9.77) and (9.81),

$$\frac{dl}{dt} \propto \frac{da}{dt} = \frac{da}{d\eta}\frac{d\eta}{dt} = \frac{\dot{a}}{a} \tag{9.102}$$

which results in

$$\frac{1}{2}\left(\frac{\dot{a}}{a}\right)^2 + \text{constant} = \frac{4\pi G}{3}\rho a^2 + \frac{\Lambda}{6}a^2 \tag{9.103}$$

This equation has the same form as Eq. (9.96), except that the value of the constant is left undetermined.

The solution of Eq. (9.98) is

$$a(\eta) = a_*(1 - \cos\eta) \tag{9.104}$$

where

$$a_* = \frac{4\pi G}{3}{a_0}^3\rho_0 \tag{9.105}$$

By integrating Eq. (9.77), we can then find the connection between the time t and the time parameter η,

$$t(\eta) = a_*(\eta - \sin\eta) \tag{9.106}$$

Equations (9.104) and (9.106) may be regarded as parametric equations for $a(t)$.

Figure 9.10 shows a plot of a as a function of t.

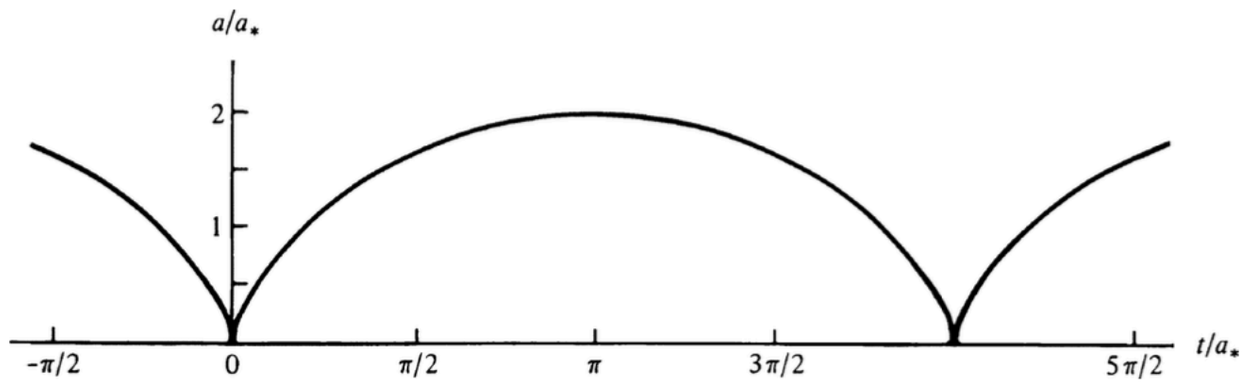


Fig. 9.10 Radius of curvature of the positive-curvature Friedmann universe as a function of time. The curve is a cycloid.

The curve given by this plot is a cycloid. At $t = 0, \pm 2\pi a_*, \pm 4\pi a_*, \ldots$, etc., $a(t)$ vanishes, that is, the universe contracts to a point.

Since the density will become very large when this is about to happen, our approximate expression for the energy-momentum tensor will fail.

For example, it will obviously be necessary to take into account the pressure exerted by matter and radiation in the dense, hot universe.

We should also keep in mind the possibility that the classical Einstein equation becomes inapplicable at very high densities; quantum effects will be one obvious source of difficulties.

It is therefore not clear what happens at the singular points of Fig. 9.10, and we do not know whether the positive-curvature Friedmann model actually has the periodic behavior suggested by this figure.

Besides, thermodynamic considerations (increase of entropy) would imply that each succeeding cycle in Fig. 9.10 lasts a longer time and attains a larger maximum radius, so there cannot be a strict periodicity of the cycles.

For comparison of our model with observation, we would want to concentrate on one expansion-contraction cycle, with the Big Bang at $t = 0$, the moment of maximum expansion at $t = \pi a_*$, and total re-collapse at $t = 2\pi a_*$.

Negative Curvature and $\Lambda = 0$. For the Robertson-Walker geometry of negative curvature, a calculation similar to that carried out in Eqs. (9.78)–(9.97) leads to an equation of motion

$$\frac{3}{a^4}(\dot{a}^2 - a^2) = 8\pi G \frac{\rho_0 {a_0}^3}{a^3} + \Lambda \tag{9.107}$$

Again, we introduce the assumption that Λ is zero, so

Again, we introduce the assumption that Λ is zero, so

$$\frac{3}{a^4}(\dot{a}^2 - a^2) = 8\pi G \frac{\rho_0 {a_0}^3}{a^3} \tag{9.108}$$

The model of the universe described by this equation is called the Friedmann model of negative curvature.

The solution of Eq. (9.108) is

$$a(\eta) = a_*(\cosh \eta - 1) \tag{9.109}$$

and it then follows that

$$t(\eta) = a_*(\sinh \eta - \eta) \tag{9.110}$$

Figure 9.11 shows a as a function of t.

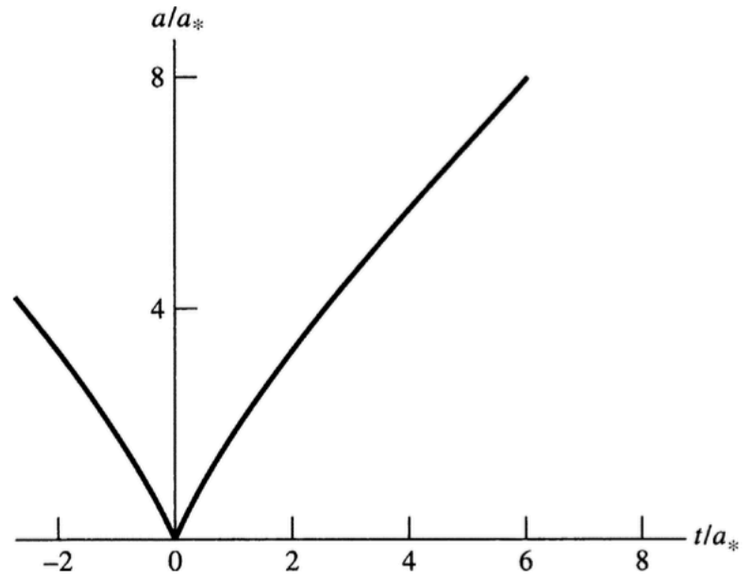


Fig. 9.11 Radius of curvature of the negative-curvature Friedmann universe as a function of time. The plot includes a continuation to negative values of t. Note that for small values of t, the behavior is similar to that shown in Figure 9.10.

The universe begins with a Big Bang and continues to expand forever.

For t $\rightarrow \infty$, the universe gradually becomes flat.

Again, the state near the singularity at $t = 0$ is not adequately described by our equations; some of the required modifications are understood, but the closer we approach $t = 0$, the less we understand.

Zero Curvature and $\Lambda = 0$. In the case of a universe with a flat three-geometry, it is convenient to introduce a dimensionless radial coordinate $\chi = r/a$.

With this coordinate, we can write the spacetime interval in a form analogous to the other cases:

$$ds^2 = a^2(\eta)(d\eta^2 - d\chi^2 - \chi^2 d\theta^2 - \chi^2 \sin^2\theta d\phi^2) \tag{9.111}$$

so

$$g_{\mu\nu} = \begin{pmatrix} a^2 & 0 & 0 & 0 \\ 0 & -a^2 & 0 & 0 \\ 0 & 0 & -a^2\chi^2 & 0 \\ 0 & 0 & 0 & -a^2\chi^2 \sin^2\theta \end{pmatrix} \tag{9.112}$$

Note that the spacetime interval (9.111) differs from flat spacetime only by the overall factor a(η).

A spacetime geometry that differs from the flat spacetime geometry only by such an overall factor is said to be *conformally flat*.

The equation of motion for a(η) can then be derived as in Eqs. (9.78)–(9.97) with the result

$$\frac{3}{a^4}\dot{a}^2 = 8\pi G \frac{{a_0}^3 \rho_0}{a^3} + \Lambda \tag{9.113}$$

In the Friedmann case, $\Lambda = 0$.

With $\dot{a}$ = ada/dt, Eq. (9.113) then reduces to

$$\left(\frac{da}{dt}\right)^2 = \frac{8\pi G}{3}\frac{{a_0}^3 \rho_0}{a} \tag{9.114}$$

which has the solution

$$a(t) = \left(6\pi G {a_0}^3 \rho_0\right)^{1/3} t^{2/3} \tag{9.115}$$

This function is plotted in Fig. 9.12.

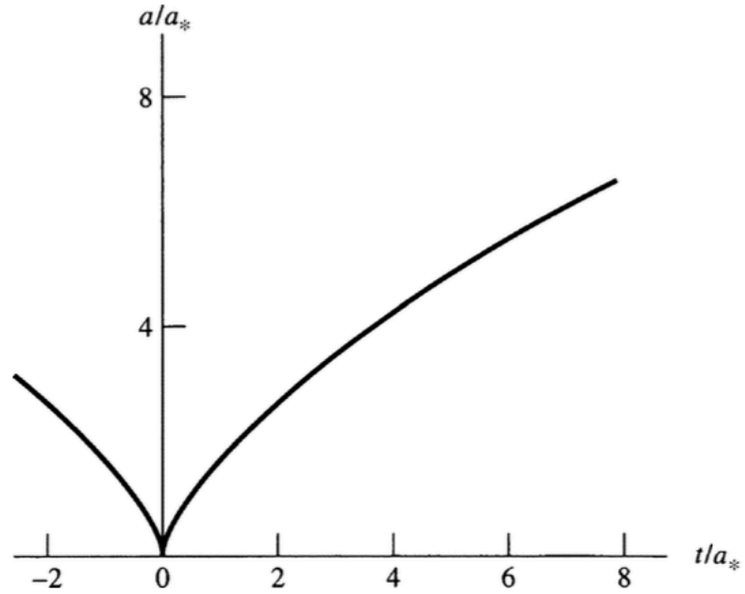


Fig. 9.12 Radius of curvature of the flat Friedmann universe as a function of time. The plot includes a continuation to negative values of t. The behavior at small values of t is, again, similar to that shown in Figure 9.10.

For t $\rightarrow \infty$, the four-geometry becomes flat.

9.9 Empty Lemaître models ($\rho = 0, \Lambda \neq 0$)

The Lemaıtre models differ from the Friedmann models in that the cosmological constant is nonzero and the mass density is assumed to be much smaller than the cosmological constant (more precisely, $\varrho \ll |\Lambda| / 8\pi G$), so the mass density can be neglected completely.

Although such a universe is empty (or nearly empty) of ordinary matter, the cosmological term gives the universe "dark energy," and this controls the dynamical evolution.

The cosmological term roughly has the same effect as a uniform gravitational mass density $\varrho_{eff} = -\Lambda/4\pi G$, constant in space and time [see Eq. (7.51)], and this effective mass density is opposite in sign to the energy density contributed by the cosmological constant.

A positive value of Λ corresponds to a negative effective mass density (repulsion), and a negative value of Λ corresponds to a positive mass density (attraction).

Hence, we expect that in a universe with a positive value of Λ, the expansion will tend to accelerate (monotonic behavior); whereas in a universe with a negative value of Λ, the expansion will slow down, stop, and reverse (oscillatory behavior).

The calculations confirm this.

Positive Curvature and $\Lambda \neq 0$. If we neglect $T_0{}^0$ in Eq. (9.92), we obtain

$$-\frac{3}{a^4}(a^2+\dot{a}^2) = -\Lambda \tag{9.116}$$

With $\dot{a}$ = ada/dt, this becomes

$$\left(\frac{da}{dt}\right)^2 = -1 + \frac{\Lambda a^2}{3} \tag{9.117}$$

It is immediately obvious from this equation that $-1 + \Lambda a^2/3$ cannot be negative.

This implies that $\Lambda > 0$, and that the value of a can never be less than $\sqrt{3/\Lambda}$.

Thus the empty Lemaıtre model with positive curvature requires a positive cosmological constant, and its radius of curvature is never zero (no Big Bang!).

The integration of Eq. (9.117) is elementary and gives

$$a(t) = \sqrt{\frac{3}{\Lambda}}\cosh\sqrt{\frac{\Lambda}{3}}t \tag{9.118}$$

where t has been set to zero at the time at which a has its minimum value.

Figure 9.13a gives a plot of $a(t)$.

For $t > 0$, the universe expands monotonically; and for $t \to \infty$, the universe becomes flat.

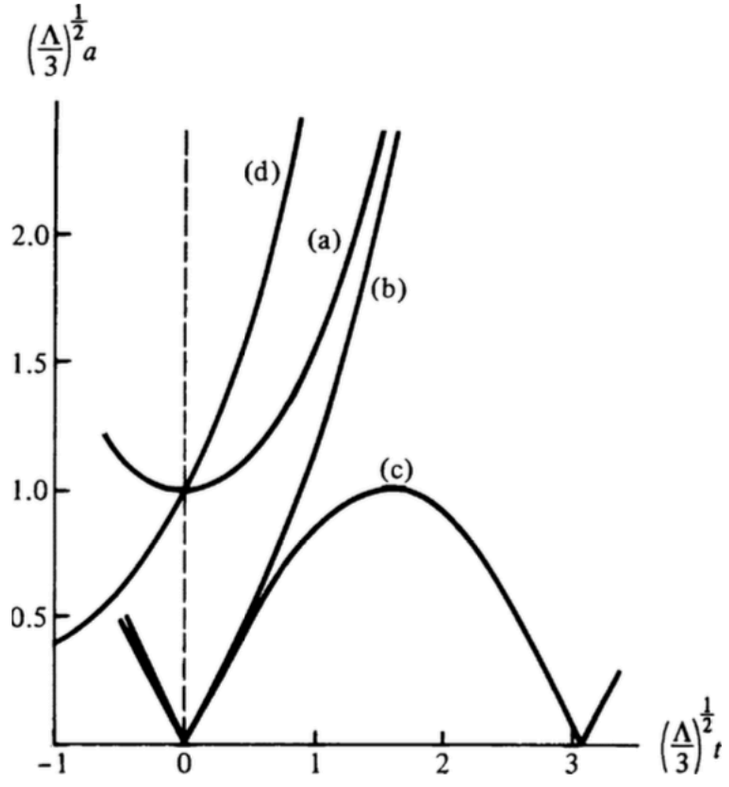


Fig. 9.13 Radius of curvature of the empty Lemaître universes as a function of time.
(a) Positive-curvature model, $\Lambda > 0$.
(b) Negative-curvature model, $\Lambda > 0$.
(c) Negative-curvature model, $\Lambda < 0$.
(d) Flat model, $\Lambda > 0$.

Negative Curvature and $\Lambda \neq 0$. In this case, the equation analogous to Eq. (9.117) is

$$\left(\frac{da}{dt}\right)^2 = 1 + \frac{\Lambda a^2}{3} \tag{9.119}$$

which can be integrated to give

$$a(t) = \sqrt{\frac{3}{\Lambda}} \sinh \sqrt{\frac{\Lambda}{3}} t \qquad \text{for } \Lambda > 0 \tag{9.120}$$

and

$$a(t) = \sqrt{\frac{3}{-\Lambda}} \sin \sqrt{\frac{-\Lambda}{3}} t \qquad \text{for } \Lambda < 0 \tag{9.121}$$

These functions are plotted in Figs. 9.13b and c, respectively.

Note that both these universes begin with a Big Bang at $t = 0$.

The first of these expands monotonically, whereas the second oscillates with a period $2\pi \sqrt{3/(-\Lambda)}$.

Of course, in our actual universe, the mass density near the singularity at $t = 0$ must have been very large, and hence the empty Lemaıtre models cannot be used to describe the behavior near this time.

Zero Curvature and $\Lambda \neq 0$. Our equation of motion is now

$$\left(\frac{da}{dt}\right)^2 = \frac{\Lambda a^2}{3} \tag{9.122}$$

This equation makes sense only if $\Lambda > 0$, and it then has the solution

$$a(t) = a(0)\exp\sqrt{\frac{\Lambda}{3}}t \tag{9.123}$$

This universe expands exponentially with a characteristic doubling time of $0.693\sqrt{3/\Lambda}$ (see Fig. 9.13d).

The model described by Eq. (9.123) is usually called the *de Sitter model*.

Note that Eq. (9.122) also has an exponentially decreasing solution, but this is of no relevance to our (expanding) universe.

Table 9.3 summarizes the characteristics of the Friedmann and empty Lemaıtre models.

Table 9.3 Homogeneous, Isotropic Models of the Universe

Model	Three-Volume	Time Dependence	Big Bang?
Friedmann, $\Lambda = 0$, $\rho > 0$			
(i) positive curvature	finite	oscillatory	yes
(ii) negative curvature	infinite	monotonic (after $t = 0$)	yes
(iii) zero curvature	infinite	monotonic (after $t = 0$)	yes
Lemaître, $\Lambda \neq 0$, $\rho = 0$			
(i) positive curvature, $\Lambda > 0$	finite	monotonic (after $t = t_{min}$)	no
(ii) negative curvature, $\Lambda > 0$	infinite	monotonic (after $t = 0$)	yes
negative curvature, $\Lambda < 0$	infinite	oscillatory	yes
(iii) zero curvature, $\Lambda > 0$	infinite	monotonic	no

9.10 Friedmann-Lemaître models ($\rho \neq 0, \Lambda \neq 0$)

The differential equations (9.97) and (9.107) for the general Friedmann-Lemaıtre model with a matter density and a cosmological term and positive or negative curvature cannot be explicitly integrated in terms of elementary functions.

Only the differential equation for the Friedmann-Lemaıtre model of zero curvature permits an explicit integration (see the later discussion).

Positive or Negative Curvature and $\varrho \neq 0,\ \Lambda \neq 0.$

The models with positive or negative curvature can be roughly described as some combination of the Friedmann and empty Lemaıtre cases discussed in the preceding sections.

Consider a universe that begins with a Big Bang.

In the early universe the mass density is very large, and we can neglect the cosmological term – that is, we have approximately a Friedmann model.

As the universe expands and the mass density decreases, the cosmological term will become more important.

In the Friedmann cases of negative or zero curvature, the expansion and the decrease in mass density are monotonic (see Figs. 9.11 and 9.12), and hence the cosmological term, if any, will ultimately dominate the behavior of the universe.

Thus, the universe gradually turns into an empty Lemaıtre universe of negative or zero curvature.

Note that in the case of negative curvature with $\Lambda < 0$, the expansion of the

Lemaıtre universe will stop at some later time (see Fig. 9.13c); the universe reverses and we finally end up in a re-contracting Friedmann universe of negative curvature.

In the case of an approximate Friedmann universe of positive curvature, the mass density reaches a minimum when one-half the period has elapsed (see Fig. 9.10).

Hence, the cosmological term will dominate the behavior of the universe only if it is sufficiently large compared with this minimum mass density.

The critical value of Λ is

$$\Lambda_E = \left(\frac{1}{4\pi G a_0^3 \rho_0}\right)^2 \tag{9.124}$$

If Λ is larger than Λ_E, then what began as a positive-curvature Friedmann universe (see Fig. 9.10) gradually turns into a positive-curvature, expanding Lemaıtre universe (see Fig. 9.13a).

A typical example of the function $a(t)$ for $\Lambda > \Lambda_E$ is shown in Fig. 9.14.

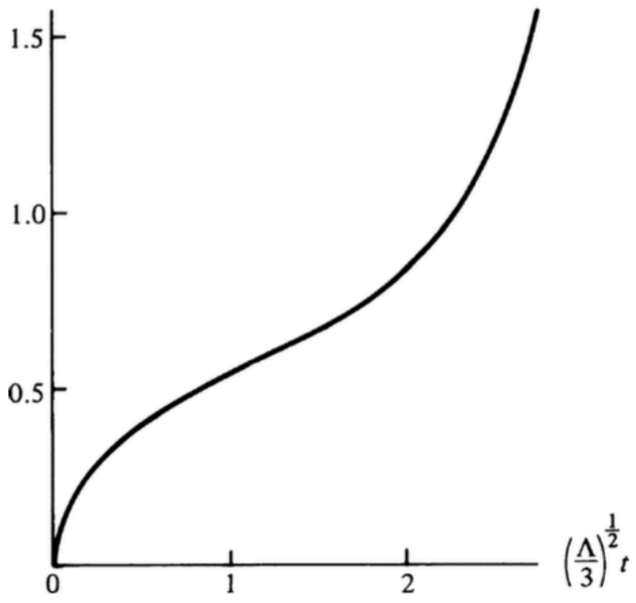


Fig. 9.14 Radius of curvature of a typical nonempty Lemaître universe, with $\Lambda = 1.17\Lambda_E$.

The transition from the nearly Friedmann to the nearly empty Lemaıtre universe can be fast or slow depending on the value of Λ.

In the exceptional case $\Lambda = \Lambda_E$, the transition is never completed; the universe remains suspended at a constant value of a,

$$a = \text{constant} = \frac{1}{\sqrt{\Lambda_E}} \tag{9.125}$$

The static universe with this constant value of a is known as the *Einstein universe*.

This was the model that Einstein proposed 1917, and for which he invented the cosmological constant.

However, the equilibrium at the value a = $1/\sqrt{\Lambda_E}$ is unstable.

Any perturbation in a leads either to monotonic expansion (toward an expanding Lemaıtre universe) or to monotonic contraction (toward a contracting Friedmann universe).

Finally, note that Table 9.3 contains no empty Lemaıtre models that are monotonic and have $\Lambda < 0$.

We can understand the absence of such models if we recall that a negative cosmological constant corresponds to a positive effective mass density of the vacuum; as the volume of the universe grows so does the total effective mass, and ultimately this stops the expansion and initiates a re-contraction.

We can draw an interesting conclusion:

The general Lemaıtre model with a nonzero mass density and $\Lambda < 0$ must necessarily be of the oscillating type.

If this universe were monotonic, it would gradually approach a monotonic empty model, which, as we have seen, does not exist.

Zero Curvature and $\varrho \neq 0$, $\Lambda \neq 0$. The differential equation for the zero-curvature case is Eq. (9.113).

With $\dot{a}$ = ada/dt this becomes

$$\frac{1}{a^2}\left(\frac{da}{dt}\right)^2 = \frac{8\pi G\rho_0}{3}\frac{{a_0}^3}{a^3} + \frac{\Lambda}{3} \tag{9.126}$$

This equation has a simple analytic solution.

For $\Lambda > 0$, the solution is

$$a(t) = a_0 \left(\frac{8\pi G\rho_0}{\Lambda} \right)^{1/3} \sinh^{2/3} \left(\frac{3}{2} \sqrt{\frac{\Lambda}{3}} t \right) \tag{9.127}$$

and for $\Lambda < 0$, the solution is

$$a(t) = a_0 \left(\frac{8\pi G\rho_0}{|\Lambda|} \right)^{1/3} \sin^{2/3} \left(\frac{3}{2} \sqrt{\frac{|\Lambda|}{3}} t \right) \tag{9.128}$$

Because the $\Lambda > 0$ solution (9.127) is thought to describe our actual universe (see Section9.12), it is expedient to rewrite it in terms of the density parameters $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ defined in Eqs. (9.21) and (9.22).

The solution then takes the convenient form

$$a(t) = a_0 \left(\frac{\Omega_{m,0}}{\Omega_{\Lambda,0}} \right)^{1/3} \sinh^{2/3} \left(\frac{3}{2} \sqrt{\Omega_{\Lambda,0}} H_0 t \right) \tag{9.129}$$

Here it is understood that the density parameters are subject to the constraint $\Omega_{m,0} + \Omega_{\Lambda,0} = 1$, required for the flat spatial geometry [see Eq. (9.143) in the next section].

This model of the universe expands monotonically, but it has two stages [see Fig. 9.18 for an example of a plot of a(t)].

In the early stage, the matter density is dominant [the first term on the right side of Eq. (9.126) is larger than the second term], and the expansion of the universe decelerates in the normal way, because of the gravitational attraction of the mass distribution.

In the later stage, the "dark energy" becomes dominant [the second term on the right side of Eq. (9.126) becomes larger that the first], and the expansion of the universe accelerates.

Ultimately, the expansion accelerates exponentially.

Figure 9.15 is a "diagnostic chart" with different shaded regions that permit us to determine from the present values of $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ whether the expansion of the uni- verse is oscillatory or monotonic; that is, whether the universe ends in an imploding Big Bang or fades away into nothingness.

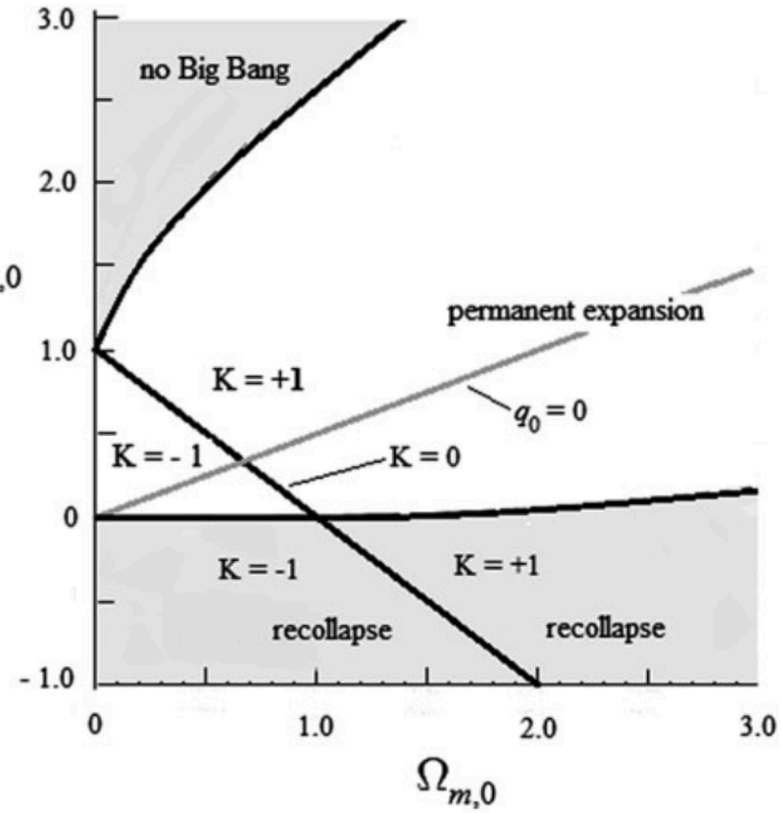


Fig. 9.15 A diagnostic chart for different Friedmann-Lemaître models of the universe. If the characteristic parameters of the universe are today within one of the colored areas (gray or white), then they remain within that area at all times and never cross any of the black boundary lines. If our universe is indeed flat, it evolves along the line marked $K = 0$, toward the upper left.

9.11 Propagation of light; particle horizon

Light signals (or light pulses) propagate along lightlike worldlines, that is, worldlines with $ds^2 = 0$.

If we place the origin of coordinates at the position of the light source, then the light will move outward radially, with $d\theta = d\phi = 0$, and hence for all our models of the universe we obtain

$$0 = a^2(\eta)(d\eta^2 - d\chi^2) \tag{9.130}$$

The equation of the worldline of a light signal emitted at the initial time parameter η(i) is

$$\chi = \eta - \eta(i) \tag{9.131}$$

This means that in χ-η spacetime diagram, the worldlines of light signals are at 45° (see Fig. 9.16).

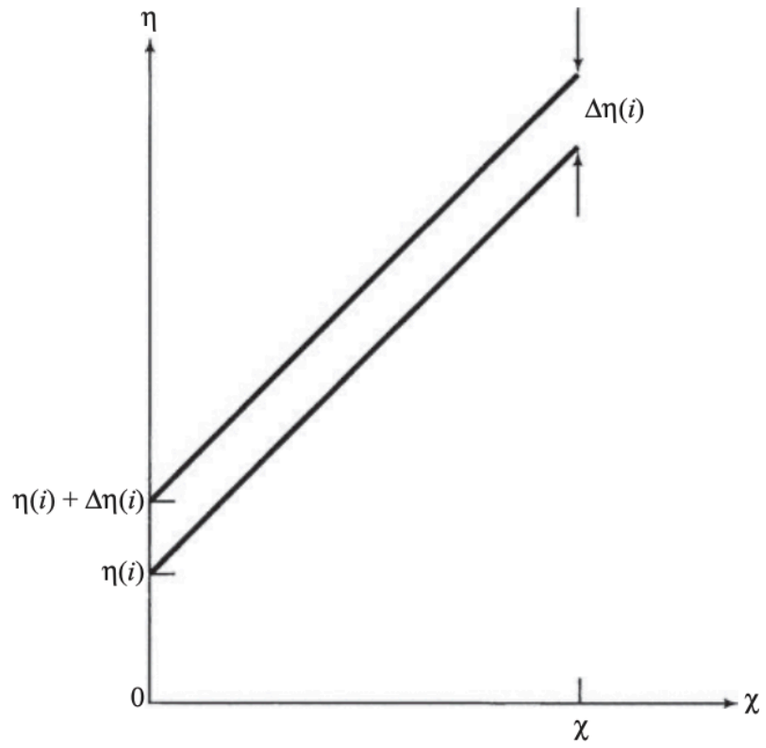


Fig. 9.16 Worldlines of light signals are at an angle of 45° with respect to the χ and η-axes. The two worldlines shown here belong to two light signals sent out from $\chi = 0$ at different times $\eta(i)$ and $\eta(i) + \Delta\eta(i)$.

Suppose that two light signals are sent out from the source at $\chi = 0$, the first when the time parameter has the value η(i) and the second when it has the value η(i) + Δη(i).

The worldlines are, respectively,

$$\chi = \eta - \eta(i) \quad \text{and} \quad \chi = \eta - [\eta(i) + \Delta\eta(i)]$$

The two light signals therefore arrive at any given point χ with a difference

$$\Delta\eta(\chi) = \Delta\eta(i) \tag{9.132}$$

between their time parameters.

This says that $\Delta\eta$ is a constant, independent of χ, and it implies that if time is measured by η, then there is no redshift.

However, the atomic clocks in the galaxies do not measure η time; instead they measure τ time, or *t* time.

Since $\Delta t = a\ \Delta\eta$, it follows from Eq. (9.132) that $\Delta t/a$ remains constant while the two signals propagate.

The time interval Δ t between the signals therefore changes in direct proportion to *a*.

For a continuous wave train with a frequency ν, in which each peak may be regarded as a signal, the time interval between signals is $\Delta t = 1/\nu$, and hence

$$a\nu = \text{constant} \tag{9.133}$$

This gives us the redshift of light – in an expanding universe, *a* increases and hence ν must decrease.

[Incidentally, we have already given an alternative derivation of Eq. (9.133) in terms of photons, by means of the simple argument following Eq. (9.33).]

According to Eq. (9.133), the frequency with which a wave arrives at the point χ is

$$\nu(\chi) = \nu(i)\frac{a(\eta(i))}{a(\eta)} = \nu(i)\frac{a(\eta-\chi)}{a(\eta)} \tag{9.134}$$

where ν(i) is the initial emitted frequency.

If the distance between the emitter and receiver is small ($\chi \ll 1$), then we can approximate $a(\eta - \chi)$ by the first term in a power series, so

$$\nu(\chi) \cong \nu(i)\left[1 - \chi\frac{\dot{a}(\eta)}{a(\eta)} + \ldots\right] \tag{9.135}$$

that is,

$$\frac{\nu(\chi)-\nu(i)}{\nu(i)} \cong -\chi\frac{\dot{a}(\eta)}{a(\eta)} \tag{9.136}$$

In terms of the wavelength $\lambda = c/\nu$, we can write this as

$$\frac{\lambda(\chi)-\lambda(i)}{\lambda(i)} \cong \chi\frac{\dot{a}(\eta)}{a(\eta)} \tag{9.137}$$

The distance between emitter and receiver, at the instant η, is given by [see Eq. (9.64)]

$$l = a(\eta)\chi \tag{9.138}$$

and Eq. (9.137) therefore takes the form of Hubble's law [compare with Eqs. (9.5) and (9.6)]:

$$\frac{\lambda(\chi) - \lambda(i)}{\lambda(i)} \cong \frac{\dot{a}}{a^2} l \tag{9.139}$$

From this we can identify the Hubble constant as

$$H = \frac{\dot{a}}{a^2} \tag{9.140}$$

Since $\dot{a}$ = ada/dt [see Eq. (9.82)], we can also write this as

$$H = \frac{1}{a}\frac{da}{dt} \tag{9.141}$$

In this derivation of Hubble's law we assumed that the distance l is a metric distance, that is, a distance defined according to the metric tensor of spacetime.

In practice, astronomers calibrate distances by luminosities, and this means that the distance appearing in Hubble's law is actually defined as proportional to the square root of the area of a sphere, rather than the radius of a sphere.

For small distances, where the effects of curvature of space are insignificant, the luminosity distance agrees with the metric distance.

However, if we want to explore deviations from linearity in Hubble's law, we must take this distinction into account, and we must also take into account higher order terms in the power series for a(χ − η) in Eq. (9.135).

Note that with the expression (9.141) for H, we can write the Einstein equation as

$$3\left(H^2+\frac{K}{a^2}\right)=8\pi G\rho+\Lambda \tag{9.142}$$

where $K=+1,-1$, or 0 for the cases of positive, negative, or zero curvature, respectively.

This leads to an expression for a in terms of observable parameters today:

$$a_0{}^2=\frac{K}{8\pi G\rho_0/3+\Lambda/3-H_0{}^2}=\frac{K}{H_0{}^2(\Omega_{m,0}+\Omega_{\Lambda,0}-1)} \tag{9.143}$$

We will find this relation between the curvature and the magnitude of the density parameters useful in the next section.

Note that according to this relation, the sign of K is necessarily the same as the sign of $\Omega_{m,0}+\Omega_{\Lambda,0}-1$.

Thus, universes of positive, negative, and zero curvature have, respectively, $\Omega_{m,0}+\Omega_{\Lambda,0}>1$, $\Omega_{m,0}+\Omega_{\Lambda,0}<1$, and $\Omega_{m,0}+\Omega_{\Lambda,0}=1$ [we already made use of the last of these conditions in connection with Eq. (9.129)].

A curious feature of the Friedmann models and other models with an initial Big Bang is that the universe has a part that is visible to us and a part that is invisible.

Consider a light signal sent out from the origin $\chi = 0$ when the universe began, at $\eta(i) = 0$.

At the time η, this light signal will have reached a point $\chi = \eta$ [see Eq. (9.131)].

Conversely, a light signal emitted from this point χ when the universe began will have reached the origin, but any light signal emitted from a more distant point will take a longer time to reach the origin.

Thus, for an observer at the origin, the surface $\chi = \eta$ represents the boundary of the visible universe.

This boundary is called the *particle horizon* or the *object horizon*, because it tells us the location of the most distant particle or object we can see.

In the spacetime diagram, this boundary is the intersection of the past light cone with the initial hypersurface $\eta(i) = 0$ (see Fig. 9.17).

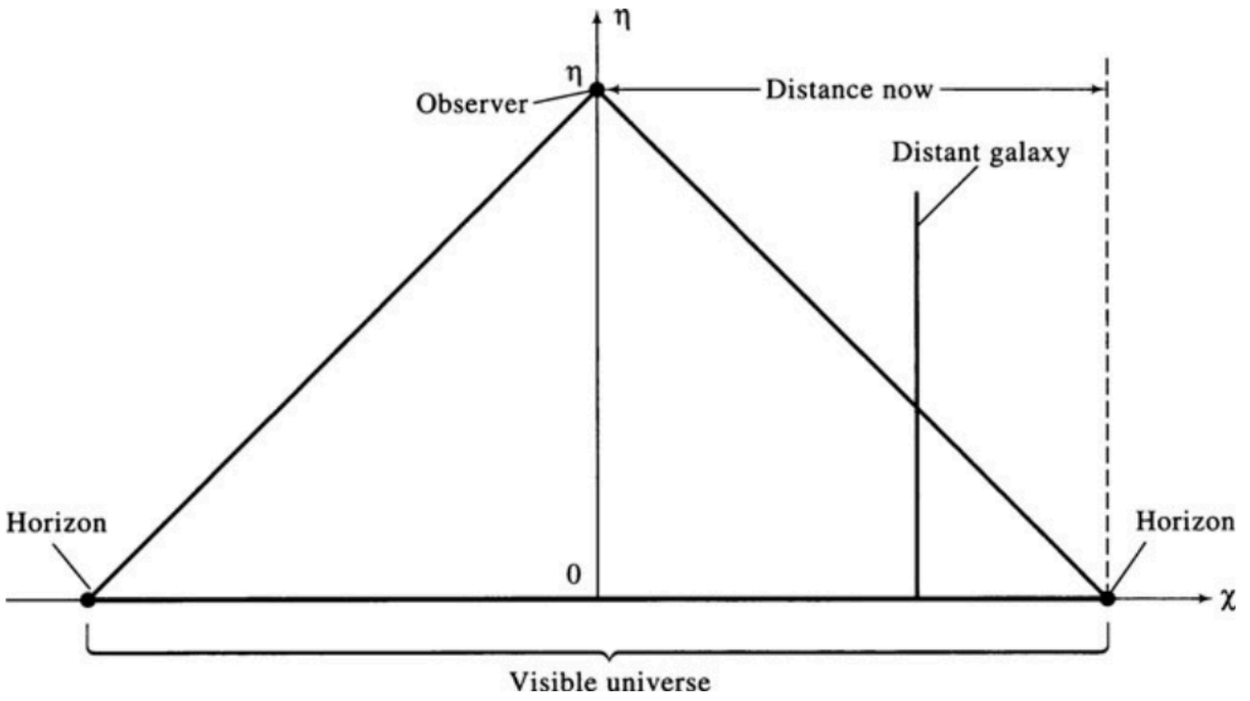


Fig. 9.17 The past light cone of an observer at a time η. The observer can see only the portion of the universe that lies within this light cone. The particle horizon for this observer is the boundary of the observable portion of the initial Big-Bang universe.

Note that an observer in the early universe, near the instant $\eta(i) = 0$, can see only a small fraction of the universe.

Even though the size of the universe at or near the instant $\eta(i) = 0$ is extremely small, light signals take a long time to travel between two points separated by some radial interval χ , because the distance between these points keeps on increasing while the signal attempts to "catch up."

If we look at a faraway region of the universe, we see this region as it was a long time ago, when the light reaching us now began its journey.

If we look at the horizon of our universe, we see the matter as it was at the initial instant; we see the primordial fireball.

As the horizon expands, more and more of this fireball comes into our visible universe.

Of course, at the horizon the redshift is infinite [see Eq. (9.134), with $a(\eta(i)) = 0$], and hence the light will actually be too faint to be seen.

The actual distance measured at time η between the observer and the horizon point from which the light was emitted is $l = a(\eta)\chi$ [see Eq. (9.64)], or

$$l = a(\eta)\eta \tag{9.144}$$

In this form, the equation is valid for all the models of the universe based on the Robertson-Walker geometry.

To express this distance to the horizon as a function of t time we must evaluate a and η in terms of t.

For the Friedmann and Lemaıtre models, the required equations for $a(\eta)$ and $\eta(t)$ were given in the preceding section.

For these models, the horizon distance l is finite, that is, the horizon exists.

For more general models, with some different functions $a(\eta)$ and $\eta(t)$, the existence of a horizon hinges on whether the factor η that appears on the right side of Eq. (9.144) is finite.

Since $d\eta = dt/a$, we can express $\eta(t)$ as an integral over $1/a$:

$$\eta(t) = \int_0^t \frac{dt'}{a(t')} \tag{9.145}$$

and the criterion for the existence of a horizon is that this integral be finite.

Clearly, the integral (9.145) will be finite whenever $a(t)$ near $t = 0$ behaves like a fractional power of t, that is, $a(t) \rightarrow t^n$ with $0 < n < 1$.

For example, in the case of the zero-curvature Friedmann-Lemaıtre model with the function $a(t)$ given by Eq. (9.129), the horizon distance now is

$$l_0 = a_0 \int_0^{t_0} \frac{dt'}{a(t')} = \int_0^{t_0} \frac{dt'}{(\Omega_{m,0}/\Omega_{\Lambda,0})^{1/3} \sinh^{2/3}(3\sqrt{\Omega_{\Lambda,0}} H_0 t'/2)} \tag{9.146}$$

Near $t = 0$, the denominator in the integrand behaves like $1/t^{2/3}$, so the integral is finite.

We will evaluate this integral numerically in Section 9.12, with appropriate values for the various constants.

The horizon expands as a function of time, and the horizon distance $l = a(\eta)\eta$ increases.

The rate of increase of this distance is

$$\frac{dl}{dt} = \frac{da}{dt}\eta + a\frac{d\eta}{dt} = Ha\eta + 1 = Hl + 1 \tag{9.147}$$

Here the second term on the right side (1, or c in cgs units) is the expected contribution from the normal speed of light, and the first term (Hl) represents an additional contribution from the expansion of the universe, which carries the light along and boosts the recession speed.

It is interesting to compare this speed of recession of the horizon with the speed of recession of a comoving particle at the same distance.

At time η, the distance to this particle is $l = a(\eta)\chi$ (where χ is fixed), and the speed of recession is

$$\frac{dl}{dt} = \frac{da}{dt}\chi = \frac{da}{dt}\eta = Hl \tag{9.148}$$

Thus, the speed of recession of the horizon exceeds the speed of recession of the particle by 1, that is, by the speed of light.

Of course, this is in accord with our intuitive expectation that an outward-moving light signal should exceed the speed of a particle that is locally at "rest," but comoving with the expansion.

It also agrees with our expectation that gradually more and more such particles become included within the visible part of our universe.

From our expression for the redshift [Eq. (9.134)] and the expressions for the distance interval [Eqs. (9.63) and (9.71)] we can establish a general relation between the luminosity distance d_L and the angular-diameter distance d_A of some given source.

For instance, for the geometry of positive curvature, Eq. (9.63) tells us that an angular size $\delta\theta$ corresponds to a transverse length $D = a(t_i) \sinh \chi \times \delta\,\theta$, where t_i is the initial time, that is, the emission time.

The angular-diameter distance given by Eq. (9.3) is then

$$d_A = \frac{D}{\delta\,\theta} = a(t_i) \sinh \chi \tag{9.149}$$

To compare this with the luminosity distance, we need to express the apparent brightness S seen at the Earth in terms of the intrinsic luminosity L of the source.

If the source radiates isotropically, the area over which the radiation is distributed at the Earth distance is $4\pi a^2(t_0)\sinh^2\chi$, where t_0 is the reception time.

Naively, we might then suppose that the energy flux per unit area reaching us would be $L/4\pi a(t_0)^2\sinh^2\chi$.

However, this flux must still be corrected for redshift.

Each photon in the light arriving at the Earth has its energy reduced by a factor $a(t_i)/a(t_0)$; furthermore, the rate of arrival of these photons is reduced by another factor of $a(t_i)/a(t_0)$, so the actual energy flux per unit area is reduced to

$$S = \frac{L}{4\pi a^2(t_0)\sinh^2\chi} \times \left[\frac{a(t_i)}{a(t_0)}\right]^2 \tag{9.150}$$

The relation $S = L/4\pi d_L^2$ for the luminosity distance then tells us that

$$d_L = \frac{a(t_0)}{a(t_i)} \times a(t_0)\sinh\chi \tag{9.151}$$

And comparison of Eqs. (9.149) and (9.151) yields the relation

$$d_L = \left[\frac{a(t_0)}{a(t_i)}\right]^2 d_A \tag{9.152}$$

In terms of the redshift parameter z, this can be conveniently written as

$$d_L = (1+z)^2 d_A \tag{9.153}$$

It is easy to see that exactly the same relation is valid in universes of negative or zero curvature (in these cases, the factor $\sinh\chi$ in the preceding equations is replaced by $\sin\chi$ or by 1, respectively, and this does not affect the final result).

Equation (9.153) shows that the luminosity distance d_L is always larger than the angular-diameter distance d_A; and, of course, for nearby targets, with $z \ll 1$, the two distances are approximately equal, as expected in a flat, nonexpanding universe.

9.12 Comparison of theory and observation

In the Friedmann models, the observable cosmological parameters are the Hubble constant, the deceleration parameter, the mass density, and the age of the universe.

To calculate the evolution of the model, we need to measure (or stipulate) two of these parameters; the cosmological model then gives us the other two parameters and completely determines the evolution of the model, forward in time and backward in time.

Thus, cosmology may be described, in Sandage's words, as "a search for two numbers".

In the Friedmann-Lemaıtre models, there is an extra observable parameter, the cosmological constant, and we must carry out a search for three numbers.

If we express the mass density and the cosmological constant in terms of the density parameters $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$, the five observable parameters for the Friedmann-Lemaıtre models are

$$H_0, q_0, \Omega_{m,0}, \Omega_{\Lambda,0}, t_0 \tag{9.154}$$

However, this ignores the complications of the early universe, where we need to include extra parameters to describe the cosmic background radiation and its pressure, and the density and pressure of a multitude of particles created by thermal fluctuations at high temperature.

We will deal with the physics of the early universe in the next chapter; for now we will concentrate on the more recent evolution of the universe, where the five parameters listed above provide an adequate description of the large-scale behavior of our universe.

Besides these basic observable parameters, there are various relations between observable quantities that we can derive for the assumed model of the universe.

For instance, if we pretend that there exists a class of light sources of standard size (instead of standard brightness), then we can derive a redshift vs. angular size relation.

We can also derive a redshift vs. apparent brightness relation (usually called the redshift-magnitude relation). And we can derive a redshift vs. galaxy count relation.

The examination of such detailed relations among all possible observable quantities is beyond the scope of this chapter, and we will base our comparison of theory and observation on the five basic parameters.

Of the five basic parameters listed earlier, the three that are best determined by observation are H_0, $\Omega_{m,0}$, and $\Omega_{\Lambda,0}$ [for more recent, improved numbers, see Eq. (10.57)]:

$$H_0 = 70 \text{ km/(s} \cdot \text{Mpc)}, \quad \Omega_{m,0} = 0.3, \quad \Omega_{\Lambda,0} = 0.7 \tag{9.155}$$

If we take the numbers for $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ at face value, then $\Omega_{m,0} + \Omega_{\Lambda,0} = 1$.

According to Eq. (9.143), this requires $K = 0$, so the 3-D geometry of the universe is flat.

The recent observational determination of the sum of the density parameters has an uncertainty of only ±1%, which leaves little room for doubt about this conclusion concerning the flat geometry.

Furthermore, a flat 3-D geometry is strongly suggested by investigations of the very early universe (see Chapter 10), so we will tentatively accept that a flat geometry is the most probable configuration of our universe.

Other features of the large-scale universe – including the values for the other two parameters, q_0 and t_0 – are then fixed by the model.

The value of q_0 has already been calculated in Eq. (9.27); it is directly related to $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$,

$$q_0 = \tfrac{1}{2}(\Omega_{m,0} - 2\Omega_{\Lambda,0}) = -0.55 \tag{9.156}$$

This relation actually does not depend on the details of the cosmological model; it merely involves homogeneity and isotropy of the mass distribution.

With the parameters specified in Eq. (9.155), the value of the expansion parameter a(t) is given by Eq. (9.129), and this completely determines the evolution of the universe. Figure 9.18 is a plot of the function a(t).

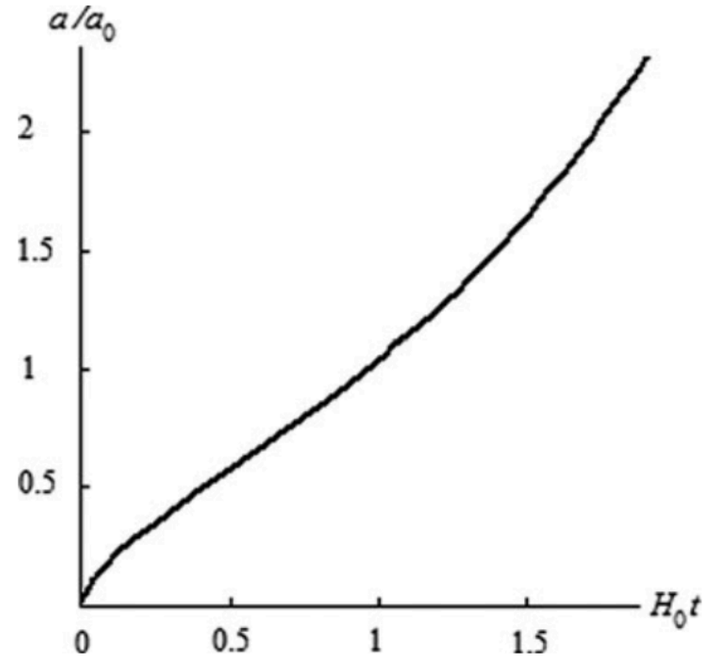


Fig. 9.18 Radius of curvature of a zero-curvature Friedmann-Lemaître universe as a function of time, according to Eq. (9.129). The density parameters are $\Omega_{m,0} = 0.3$ and $\Omega_{\Lambda,0} = 0.7$. The vertical axis is in units of a_0 and the horizontal axis in units of H_0^{-1}.

The expansion begins with a very large initial speed, then slows down because of the gravitational attraction of normal matter, but then accelerates when the cosmological term (the "dark energy") becomes dominant.

The time of the switchover from deceleration to acceleration is $0.70 \times H_0^{-1} = 0.70 \times 14 \times 10^9$ yr $= 9.8 \times 10^9$ yr.

To find the value of t_0, we need to invert Eq. (9.129) and obtain t as a function of a:

$$t = \frac{2}{3\sqrt{\Omega_{\Lambda,0}}H_0} \sinh^{-1}\left(\frac{a^3\Omega_{\Lambda,0}}{{a_0}^3\Omega_{m,0}}\right)^{1/2} \tag{9.157}$$

At the present time, this gives

$$t_0 = \frac{2}{3\sqrt{\Omega_{\Lambda,0}}H_0} \sinh^{-1}\left(\frac{\Omega_{\Lambda,0}}{\Omega_{m,0}}\right)^{1/2} = \frac{2}{3\sqrt{0.7}H_0} \sinh^{-1}\left(\frac{0.7}{0.3}\right)^{1/2} = 0.964 H_0^{-1} \tag{9.158}$$

With the Hubble age $H_0^{-1} = 1/(70\ \text{km/s} \cdot \text{Mpc}) = 14 \times 10^9$ yr, this implies a true age

$$t_0 = 13.5 \times 10^9\ \text{yr} \tag{9.159}$$

This theoretical calculated age is in reasonable agreement with the measured ages given in Section 9.4.

We can also calculate the horizon distance for this universe.

According to Eq. (9.146), with $t_0 = 0.964 H_0^{-1}$, the horizon distance now is

$$l_0 = \int_0^{0.964 H_0^{-1}} \frac{dt'}{(0.3/0.7)^{1/3} \sinh^{2/3}\left(3\sqrt{0.7} H_0 t'/2\right)} \tag{9.160}$$

Numerical integration gives the result $l_0 = 3.27 H_0^{-1} = 4.6 \times 10^{10}$ ly $= 4 \times 10^4$ Mpc.

This is the radius of our observable universe.

For comparison, it is instructive to recalculate the evolution of the universe with somewhat different values of the density parameters.

Figures 9.19a and b show the results of numerical integrations with density parameters $\Omega_{m,0} = 0.5$ and $\Omega_{\Lambda,0} = 0.9$, respectively.

The first case has $\Omega_{m,0} + \Omega_{\Lambda,0} = 0.3 + 0.5 = 0.8$, which, by Eq. (9.143), requires $K < 0$ (a three-geometry of negative curvature).

The second case has $\Omega_{m,0} + \Omega_{\Lambda,0} = 0.3 + 0.9 = 1.2$, which requires $K > 0$ (a three-geometry of positive curvature).

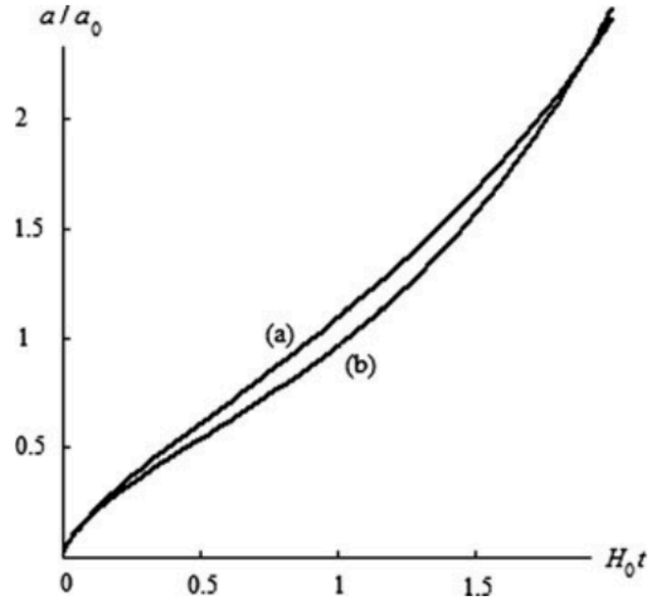


Fig. 9.19 Radius of curvature of Friedmann-Lemaître universes as a function of time, obtained by numerical integration of Eqs. (9.97) and (9.107). (a) Negative curvature; the density parameters are $\Omega_{m,0} = 0.3$ and $\Omega_{\Lambda,0} = 0.5$. (b) Positive curvature; the density parameters are $\Omega_{m,0} = 0.3$ and $\Omega_{\Lambda,0} = 0.9$. Because of observational uncertainties, the difference between these two models of the universe is not directly detectable.

Table 9.4 summarizes information for our three different Friedmann-Lemaıtre models of the universe.

All these models are consistent with the available high-z redshift and distance data, but the zero-curvature model is preferred, because of its implications for the early universe (which we will discuss in the next chapter).

Table 9.4 Friedmann-Lemaître models of our universe

Parameter	Flat	Negative Curvature	Positive Curvature
Mass-density parameter, $\Omega_{m,0}$	0.3	0.3	0.3
Mass density, ρ_0	2.8×10^{-30} g/cm^3	2.8×10^{-30} g/cm^3	2.8×10^{-30} g/cm^3
Cosmological density parameter, $\Omega_{\Lambda,0}$	0.7	0.5	0.9
Deceleration parameter, q_0	-0.55	-0.35	-0.75
Hubble constant, H_0	70 km/(s · Mpc)	70 km/(s · Mpc)	70 km/(s · Mpc)
Hubble age, H_0^{-1}	1.4×10^{10} yr	1.4×10^{10} yr	1.4×10^{10} yr
True age, t_0	1.35×10^{10} yr	1.27×10^{10} yr	1.45×10^{10} yr
Horizon distance, l_0	4.46×10^{10} ly	4.37×10^{10} ly	4.83×10^{10} ly
Radius of curvature, a_0	N.A.*	3.1×10^{10} ly	3.1×10^{10} ly

* The value of a_0 for the flat universe is undefined. Only the ratio a/a_0 is meaningful.

Finally, it may be well to keep in mind that past experience in cosmology suggests that the observable cosmological parameters could be subject to further substantial revisions, in excess of the quoted uncertainties.

For instance, the value of the Hubble constant has suffered several drastic revisions, and the recent discovery of the acceleration of the expansion of the universe came as a total surprise.

Despite the remarkable progress in the determination of the cosmological parameters during the last 20 years, we cannot yet predict with confidence the future of our universe.

10 - The Early Universe

FIAT LUX. *Genesis, 1.3*

Extrapolating the present motion of expansion of the universe backward in time, we conclude that the early universe must have been very dense.

And extrapolating the (adiabatic) expansion of the cosmic background radiation backward in time, we conclude that the early universe must have been very hot.

Thus, at an early time, the universe must have been very different from what it is now.

There were no stars and no galaxies, but only a uniform hot plasma, consisting of free electrons and free nuclei.

The chemical composition of the early universe must also have been different.

The heavy elements (that is, elements other than hydrogen, deuterium, helium, and lithium) in our immediate environment were formed by nuclear reactions in the cores of stars, so these elements did not exist in the early universe.

At very early times, the violent thermal collisions would have prevented the existence of any kind of nuclei, and the matter in the universe must have been in the form of free electrons, protons, and neutrons.

At the earliest times, even the protons and neutrons would have been disrupted, and the universe must have contained a mix of quarks, gluons, and other elementary particles.

The observed expansion of the universe and the observed cosmic background radiation provide the empirical basis for a Friedmann-Lemaıtre model of the universe with a Big Bang, sometimes called the *Standard Model*.

Further evidence supporting this model is provided by calculations of the synthesis of helium in the universe.

Although stars make helium by the thermonuclear burning of hydrogen, most of the helium in the universe must be primordial, since it is found even in stars that have not yet burned long enough to accumulate a significant amount of helium.

This primordial helium was formed by nuclear reactions in the early universe at about 100s, and the abundance of this helium (relative to hydrogen) can be calculated by examining the thermal equilibrium attained by protons and neutrons in reactions in the early, hot universe.

The numbers obtained by such calculations of the helium abundance are in excellent agreement with the observational data.

The abundances of other light elements formed in the early universe can be calculated similarly.

At the earlier times, the universe must have contained a mixture of many kinds of elementary particles.

Thus, particle physics has come to play a central role in the study of cosmology.

Grand Unified Theories (GUTs) predict that at energies of the order of 10^{14} GeV, the weak, electromagnetic, and strong interactions merge into a single interaction, endowed with a high symmetry.

Energies of this order are far beyond anything attainable with accelerators; however, at a sufficiently early time, particles in the early universe could easily have had thermal energies of this order of magnitude.

Hence theories of elementary particles and their interactions at extreme energies help us understand the behavior of the early universe; conversely, the behavior of the early universe, as revealed by the presence (or the absence) of relic particles, helps us understand particle physics.

Particle physicists have come to regard the early universe as a testing ground for their theories.

Entirely new physics needs to be introduced to deal with the behavior of the universe at the earliest times, at about 10^{-32}s and earlier, when the universe apparently suffered a quick and very large inflation.

Strong circumstantial evidence consistent with such an inflation has now been discovered by analysis of miniscule temperature fluctuations in the observed distribution of the cosmic background radiation over the sky.

But the mechanism underlying inflation remains purely conjectural, and a large variety of possible theoretical scenarios are still under investigation.

10.1 Temperature of the early universe

At present, the universe is cold and it is matter dominated.

At a temperature of only 2.73 K, the contribution of the cosmic background radiation to the mass density of the universe is negligible compared with the contribution of massive particles.

However, at an earlier time, the universe was much hotter and it was radiation dominated.

To see this, let us compare the densities of matter and radiation.

Since the volume of the universe is directly proportional to a^3, the density of matter is inversely proportional to a^3,

$$\rho_m = \rho_0 \left(\frac{a_0}{a}\right)^3 \tag{10.1}$$

According to the Stefan-Boltzmann law, the (mass) density of the radiation is proportional to the fourth power of the temperature,

$$\rho_{rad} = 4\left(\frac{\pi^2 k^4}{60\hbar^3}\right) T^4 \tag{10.2}$$

where $\pi^2k^4/60h^{-3}$, or, in cgs units, $\pi^2k^4/60h^{-3}c^2 = 5.67 \times 10^{-5}$ g/s^{-3}·K^4, is the Stefan-Boltzmann constant [in cgs units, besides the factor of $1/c^2$ in the Stefan-Boltzmann constant, an extra factor of 1/c is required on the right side of Eq.(10.2)].

With T ∝ 1/a, we obtain

$$\rho_{rad} = \rho_{rad,0}\left(\frac{a_0}{a}\right)^4 \tag{10.3}$$

By comparing Eqs. (10.1) and (10.3), we see that in the early universe, when a is small, the density of radiation dominates.

The transition from a radiation-dominated universe to a matter-dominated universe occurs when $\varrho_m = \varrho_{rad}$, that is, when

$$\rho_0\left(\frac{a}{a_0}\right)^3 = \rho_{rad,0}\left(\frac{a}{a_0}\right)^4 \tag{10.4}$$

The present density of the radiation, at a temperature $T_0 = 2.73$ K, is $\varrho_{rad,0} = 4.68 \times 10^{-34}$ g/cm^3.

The present density of matter is somewhat uncertain, and it is convenient to write it as $\varrho_0 = \Omega_{m,0} \times 3H_0^2/8\pi G = \Omega_{m,0} \times 1.88h^2 \times 10^{-29}$ g/cm^3 (see Eq. 9.21), so that the uncertainties in the matter density are contained within $\Omega_{m,0}$ and h.

The condition (10.4) then leads to

$$\frac{a_0}{a} = \frac{\rho_0}{\rho_{rad,0}} = 4.02 \times 10^4 \Omega_{m,0} h^2 \tag{10.5}$$

This shows that matter and radiation had equal densities when the universe was about $4.02 \times 10^4\ \Omega_{m,0}h^2$ times smaller than it is now.

Accordingly, the temperature of the cosmic background radiation was then larger than now by the same factor.

With the probable values $\Omega_{m,0} = 0.3$ and h = 0.7 for our universe (see the preceding chapter), we find that the temperature was then

$$T = \frac{a_0}{a} T_0 = 4.02 \times 10^4 \Omega_{m,0} h^2 \times 2.73\,\mathrm{K} = 1.6 \times 10^4\,\mathrm{K} \tag{10.6}$$

However, when cosmologists compare the densities of matter and radiation they usually include extra cosmic background radiation in the form of neutrinos.

Although this neutrino background radiation, or neutrino gas, has not been detected directly, it must be present as a relic from the Big Bang, and it increases the total amount of radiation in the universe today by a factor of almost 1.7 (three kinds of neutrinos participate, but their contribution is smaller than that of photons, because their temperature today is lower than that of photons).

Correspondingly, in Eq. (10.6), the temperature T_{eq} for equal matter and radiation densities is reduced by a factor of about 1.7, with the final result

$$T_{eq} = 9700\,\mathrm{K} \tag{10.7}$$

To calculate the age at which this temperature is reached, we need to examine the radius of curvature as a function of time.

For our Friedmann-Lemaitre model with $\Omega_{m,0} = 0.3$ and h = 0.7, this function is given approximately by Eq.(9.129), from which we find that the critical time is about 5×10^4 years.

This time may be said to mark the beginning of the matter era and the end of the radiation era of the universe.

However, the criterion of equal matter and radiation densities does not provide a sharp dividing line, because radiation has continued to contribute a significant fraction of the mass density and of the pressure for many thousands of years after the instant of equal densities.

Some cosmologists prefer to place the end of the radiation era at the time when the ionized hydrogen plasma in the universe changed into neutral hydrogen gas ("recombination" of hydrogen) and the cosmic background radiation ceased to interact with matter ("photon decoupling").

Before recombination, there was a high density of free electrons, and the scattering of photons by these free electrons maintained thermal equilibrium between matter and radiation.

After recombination, the scattering of photons was rare – neutral hydrogen is almost completely transparent to radiation and hardly ever scatters a photon.

Thus, after recombination, the photons ceased to interact with matter, and their temperature became independent of the temperature of matter.

At the prevailing density, this decoupling occurred at a temperature $T_{dec} = 3000$ K.

The age at which this transition occurred was about 3.5×10^5 years.

To formulate the Einstein equations for a radiation-filled universe, we need the energy-momentum tensor for the radiation.

If we regard the radiation as a fluid with a proper mass density ϱ_{rad} and a pressure P, its energy-momentum tensor is [see Eq. (2.83)]

$$T_\mu{}^\nu = (\rho_{rad} + p)u_\mu u^\nu - \delta_\mu^\nu p \tag{10.8}$$

Here, u^μ is the flow velocity of the fluid.

In our comoving coordinates, the fluid is at rest, so $u^0 = 1/a$, $u_0 = a$.

Hence
$$T_0{}^0 = (\rho_{rad} + p)u_0 u^0 - p = (\rho_{rad} + p) - p = \rho_{rad} \tag{10.9}$$

Because we are contemplating a universe much younger than the present universe, the cosmological term is insignificant compared with the energy-momentum of matter, so we can neglect the cosmological term and pretend that we are dealing with a Friedmann universe with $\Lambda = 0$.

The Einstein equation (9.142) then becomes

$$\frac{3}{a^4}(\dot{a}^2 + Ka^2) = 8\pi G\rho_{rad} \tag{10.10}$$

where $K = +1, -1$, and 0 for the cases of positive, negative, and zero curvature, respectively.

Note that according to this equation, the pressure of the radiation has no effect whatsoever on the expansion of the universe – the pressure cancels in (10.9) and does not appear in Eq. (10.10).

The reason why the pressure of a fluid cannot affect the dynamics of the universe is that the fluid is homogeneous.

The pressure is the same at all points in the fluid, and there are no pressure gradients.

Hence each element of fluid experiences exactly the same pressure forces from all sides, and there is no net force on the fluid that might accelerate or decelerate it.

Although the Big Bang is sometimes described as a primeval "explosion," it differs from ordinary explosions in a crucial way – its outward expansion is the result of initial conditions, not the result of outward pressure forces.

The mass density of the radiation is proportional to $1/a^4$ [see Eq. (10.3)]; hence we can write the Einstein equation as

$$\frac{3}{a^4}(\dot{a}^2 + Ka^2) = 8\pi G\rho_{rad,0}\frac{{a_0}^4}{a^4} \tag{10.11}$$

where a_0 is the radius at some reference time (an early time, not today).

In the early universe, a is small and the term $\propto 1/a^4$ on the right side of this equation is much larger than the term $\propto K/a^2$ on the left side.

We can therefore neglect the latter term, and write the equation in the approximate form

$$\frac{3}{a^4}\dot{a}^2 = 8\pi G\rho_{rad,0}\frac{{a_0}^4}{a^4} \tag{10.12}$$

or in terms of the time derivative da/dt = $\dot{a}$ /a,

$$\frac{3}{a^2}\left(\frac{da}{dt}\right)^2 = 8\pi G\rho_{rad,0}\frac{{a_0}^4}{a^4} \tag{10.13}$$

The term we have neglected in Eqs. (10.12) and (10.13) is the "curvature" term, which distinguishes between the Friedmann universes of positive, negative, or zero curvature.

Thus, the approximate equation (10.13) is equally valid for all these Friedmann models – near the Big Bang, the curvature of the space geometry makes no difference to the dynamics of the universe.

We already noticed this insensitivity of the early universe to the curvature of space in Figs. 9.10–9.12, where we saw that near $t = 0$, the behavior of $a(t)$ is the same for all the Friedmann-Lemaıtre models.

The solution of Eq. (10.13) is trivial.

With the initial condition a(0) = 0, we find

$$a(t) = \left(\frac{32\pi G}{3}{a_0}^4\rho_{rad,0}\right)^{1/4} t^{1/2} \tag{10.14}$$

Accordingly, the mass density of the radiation is

$$\rho_{rad}(t) = \rho_{rad,0}\frac{{a_0}^4}{a^4} = \frac{3}{32\pi G}\frac{1}{t^2} \tag{10.15}$$

Note that the constant $\varrho_{rad,0}{a_0}^4$ cancels in the final expression, and the dependence of the density on time does not involve the reference time we adopted in Eq. (10.11).

Likewise, the dependence of the temperature of the radiation on the time is

$$T = \left(\frac{\rho_{rad}}{4\pi^2 k^4/60\hbar^3}\right)^{1/4} = \left(\frac{45\hbar^3}{32\pi^3 k^4 G}\right)^{1/4} \frac{1}{t^{1/2}} \tag{10.16}$$

This shows that the temperature becomes arbitrarily large as $t \to 0$.

Equation (10.16) can be expressed concisely in terms of the *Planck mass*,

$$m_{Pl} = \sqrt{\hbar/G} = \sqrt{\hbar c/G} = 2.18 \times 10^{-5}\,\mathrm{g} = 1.22 \times 10^{19}\,\mathrm{GeV}/c^2 \tag{10.17}$$

So

$$kT = \left(\frac{45}{32\pi^3}\right)^{1/4} \sqrt{\frac{\hbar m_{Pl}}{t}} = 0.461\sqrt{\frac{\hbar m_{Pl}}{t}} \tag{10.18}$$

It is easy to check that for the radiation-dominated universe, with $a \propto t^{1/2}$, the Hubble age is $H^{-1} = 2t$.

As a function of the temperature this becomes

$$H^{-1} = 2t = \sqrt{\frac{45}{8\pi^3}} \frac{\hbar m_{Pl}}{k^2 T^2} = 0.426 \frac{\hbar m_{Pl}}{k^2 T^2} \tag{10.19}$$

It is straightforward to verify that if $K = 1$, the exact solution of Eq. (10.11) is

$$a(\eta) = b_* \sin(\eta) \tag{10.20}$$

$$t(\eta) = b_*(1 - \cos\eta) \tag{10.21}$$

where $b_* = \sqrt{8\pi G \rho_{rad,0} a_0{}^4/3}$.

At very early times, the universe was so extremely hot that the typical thermal energies of the blackbody photons were large enough to permit the creation of electron-positron pairs.

The required temperature is given by $kT \cong m_e c^2$, that is, $T \cong 6 \times 10^9$ K.

According to Eq. (10.16), this temperature corresponds to an age of ≈ 10 s.

At higher temperatures, the electrons, positrons, and also neutrinos form a relativistic gas of leptons.

Such a relativistic gas of particles has properties that are very similar to those of the gas of photons making up the cosmic background radiation.

The energy density of the gas is proportional to the fourth power of the temperature, but because there are several kinds of leptons, the constant of proportionality is somewhat larger.

Thus, the universe was lepton dominated at very early times.
The dynamical equation for the lepton-dominated universe is essentially the same as for the radiation-dominated universe.

Only some of the constants of proportionality are somewhat altered; for rough estimates of the thermal history of the universe, we can ignore these alterations.

From Eq. (10.16) we find that at an age of 10^{-4} s, the universe had a temperature of 10^{12} K, which was sufficient to permit the creation of muons and pions by thermal fluctuations.

At slightly earlier times even the creation of protons, neutrons, and other baryons and antibaryons, as well as mesons and antimesons, was possible, and the universe was dominated by strongly-interacting particles, or hadrons.

The earliest time at which it is meaningful to speak of a more or less well-defined spacetime geometry that can serve as background for the motion of particles is the *Planck time*,

$$t_{Pl} = \sqrt{\frac{\hbar G}{c^5}} = 5.39 \times 10^{-44}\,\mathrm{s} \tag{10.22}$$

Before this time, the universe was dominated by quantum fluctuations in the geometry, about which we know nothing.

We can therefore distinguish five eras in the chronology of the universe (see Fig. 10.1):

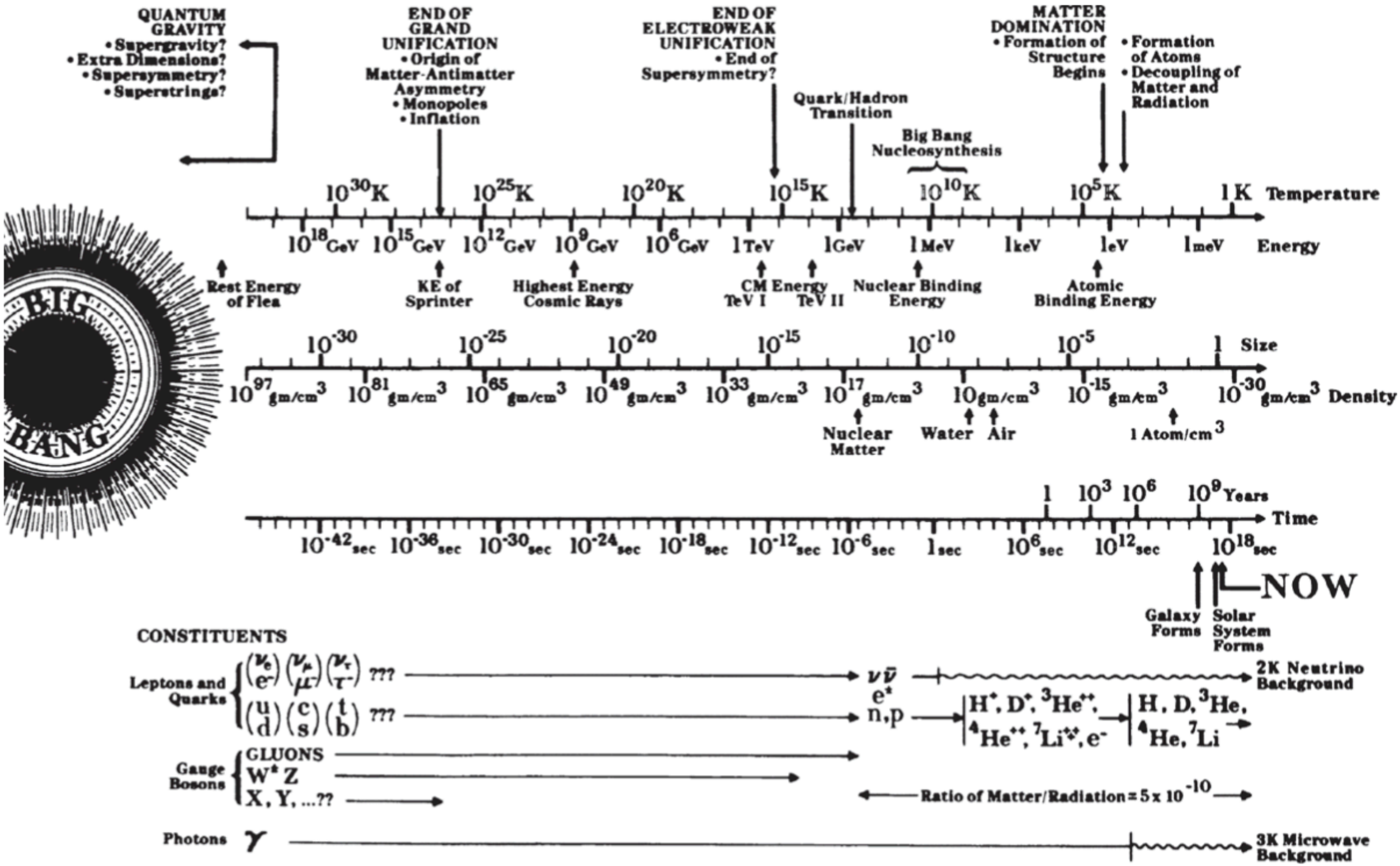


Fig. 10.1 Chronology of the universe. (Courtesy Fermilab)

(i) **The quantum era** ($t < 10^{-43}$ s; $T > 10^{31}$ K; $kT > 10^{19}$ GeV). In these earliest times, the universe was dominated by quantum fluctuations, especially fluctuations in the geometry. A simple estimate suggests that on a time scale t, the fluctuations in the metric tensor are of the order of $\Delta g_{\mu\nu} \approx t_{Pl}/t$, which means that in the quantum era, the fluctuations in the spacetime geometry were large, $\Delta g_{\mu\nu} > 1$, so there was no measurable geometry at all. Even the topology of spacetime would have been subject to quantum fluctuations, with an abundance of black holes and wormholes of all kinds. All this is *terra incognita*.

(ii) **The hadron era** ($t = 10^{-43}$ to 10^{-4} s; $T = 10^{31}$ to 10^{12} K; $kT = 10^{19}$ to 1 GeV). During this era, the universe contained a highly relativistic plasma of elementary particles and antiparticles of all kinds – quarks, gluons, electrons, muons, taus, photons, W and Z bosons, neutrinos, and gravitons. The numbers of baryons and antibaryons in this primordial fireball were nearly equal, but there was a small excess of baryons (representing a small fraction of the whole). As the temperature dropped toward 10^{12} K, the baryons and antibaryons annihilated, leaving only the small excess of baryons that accounts for the number of baryons found in the present universe. Near the end of this era, at $t \cong 10^{-6}$ s, the quarks combined into nucleons (such as protons and neutrons) and mesons (such as pions).

(iii) **The lepton era** ($t = 10^{-4}$ to 10 s; $T = 10^{12}$ to 4×10^{9} K; $kT = 1$ GeV to1 MeV). In this era, only a much smaller number of particles survived. The remaining particles were electrons, muons, taus, neutrinos, and their corresponding antiparticles, and also photons, gravitons, protons, and neutrons. Even today the neutrinos and antineutrinos from this era should still be with us and fill our intergalactic space with a neutrino gas whose temperature is 71% of the cosmic background radiation temperature. Since the interactions of neutrinos are extremely weak, there is not much hope for the direct detection of this cosmic neutrino gas.

(iv) **The radiation era** ($t = 10$ to 10^{12} s; $T = 4 \times 10^9$ to 10^4 K; $kT = 1$ MeV to10 eV). This era began when electrons annihilated, leaving only a relatively small residue that balanced the positive charge density of the protons. The density of the remaining electrons, and that of protons and neutrons, was negligible compared with the densities of photons and neutrinos. During this era, at a temperature of $\cong 10^9$ K, the neutrons fused with the protons and formed ^{4}He nuclei. About 25% (by mass) of the baryonic matter was made into helium, and some very small amounts were made into deuterium, lithium, and other light nuclei (see Section 10.2).

(v) **The matter era** ($t > 10^{12}$ s; $T < 10^4$ K; $kT < 10$ eV). The most recent era began when the density of radiation dropped below the density of matter. The exact birthdate of this era depends on the present mass density of the universe and is therefore somewhat uncertain. Near the beginning of this era, electrons combined with protons and formed neutral atoms of hydrogen and helium. This happened when the temperature dropped to 3000 K. Once matter became neutral, the interaction between matter and radiation was drastically reduced, and there was insufficient exchange of energy to maintain thermal equilibrium. Matter and radiation expanded and cooled independently, each with its own temperature. The temperature of matter decreased as $T \propto 1/a^2$,[2] while the temperature of radiation decreased as $T \propto 1/a$. Hence, the temperature of matter ultimately dropped below that of the radiation.

As long as matter and radiation were coupled, any small irregularities in the mass distribution were held in check by the diffusion of photons, which provides viscous damping.

But once the mass was on its own, a region with a slight excess of mass tended to grow by the gravitational attraction it exerted on its surroundings.

By some such condensation process galaxies began to form at $t \cong 10^{16}$ s.

And within one of these galaxies the Solar System formed at $t \cong 3 \times 10^{17}$ s after the Big Bang.

At about the same time, the overall mass density in the universe became lower than the mass density associated with the cosmological term.

Accordingly, the expansion of the universe began to accelerate, and it continues to accelerate today, at $t \cong 4 \times 10^{17}$ s after the Big Bang.

10.2 Nucleosynthesis; abundance of primordial helium

Helium is a ubiquitous component of stars.

This is not surprising, because stars make helium by the thermonuclear burning of hydrogen.

However, even quite young stars, which have not had time to make much helium, are found to contain a substantial amount of helium – in the youngest stars, the helium abundance is about 25%, or 1/4, by mass, and all the youngest stars have this same abundance.

This uniform helium abundance in young stars indicates that their helium is primordial – the helium must be a relic from the Big Bang, and it must have been present in the gas out of which the stars formed.

Helium is synthesized in the early universe when the temperature drops sufficiently for neutrons to bind with protons and form deuterium nuclei.

In turn, these bind with each other or with free protons or neutrons to form tritium and helium.

The first calculations of such nucleosynthesis were performed by Gamow and his collaborators.

Gamow proposed that the universe initially contained only neutrons, and that the radioactive decay of these primordial neutrons gradually produced protons.

These protons then captured one or more neutrons in succession and thereby built up heavier nuclei step-by-step.

If these neutron-capture processes had continued indefinitely, all of the mass of the universe would have been converted into helium and heavier nuclei.

But Gamow recognized that the resulting abundances of the various nuclei formed by neutron capture hinge on an interplay between the relevant reaction rates and the expansion rate of the universe.

The reaction rate for a particle exposed to an incoming flux of other particles is nvσ, where *n* and v are, respectively, the density and the speed of the incoming particles and σ is the reaction cross section.

The expansion of the universe is described quantitatively by the fractional rate of increase of the radius *a*, that is, (1/a)(da/dt), which equals the Hubble constant *H*.

Gamow pointed out that if the reaction rate nvσ is smaller than the fractional expansion rate (1/a)(da/dt) = H, then the expansion of the universe arrests the reaction, because the reactants disperse before they have a chance to react.

Thus, the condition for the arrest, or the "freeze out," of the reaction is

$$nv\sigma < H \tag{10.23}$$

Gamow assumed that the protons produced by radioactive decay would engage in neutron-capture reactions until the neutron density decreased to the limit set by Eq. (10.23) and the reactions froze out.

Any protons produced after this time would have stayed as protons and ultimately formed hydrogen gas.

Gamow was delighted to find that the ratio of the abundances of hydrogen and helium that emerged from his calculation agreed roughly with the observed abundance ratio.

Alpher and Herman improved Gamow's calculations and obtained more accurate results by numerical integrations of the equations for reaction rates.

They also extrapolated the temperature of the universe to the present time, starting with the temperature and the density at an early time.

This extrapolation is an immediate consequence of the dependence of T and n on the radius of the universe, $T \propto 1/a$, and $n \propto 1/a^3$, which implies $T^3/n = \text{constant}$ and $T^3/n = T_0^3/n_0$.

Thus, the temperature T and the density n at the time of nucleosynthesis determine the present temperature T_0, if the present density n_0 is assumed known.

Alpher and Herman deduced a present temperature of about 5 K, but a more careful analysis showed that the uncertainty in such a prediction of the present temperature is rather large – the predicted temperature could be anywhere in the range from 1 to 30 K.

Gamow's scheme for nucleosynthesis had a fundamental defect.

He assumed that the early universe was filled with a primordial gas of neutrons.

But, as noticed by Hayashi, the actual composition of the early universe must have been a mixture of neutrons and protons, because at the high prevailing temperatures, there must have been an abundance of leptons (the lepton era), and these triggered reactions [see Eqs. (10.26) and (10.27)] that converted neutrons to protons and viceversa with sufficiently fast rates to maintain an equilibrium between neutrons and protons, with nearly equal concentrations of both.

In contrast, the radioactive-decay reaction of the neutron, which is central to Gamow's scheme, played next to no role in the early universe, because it is a much slower reaction than the reactions involving leptons.

From the work of astrophysicists who explored Gamow's ideas, we now know that the formation of heavy elements did not take place in the early universe, but much later, in the interior of stars.

Only the lightest elements – deuterium, helium, lithium, beryllium – were formed in the early universe, by the neutron-capture reactions envisioned by Gamow.

But despite its errors, Gamow's work gave valuable guidance for the nucleosynthesis calculations of his successors.

In general, the calculations of element formation in the early universe require integrations of reaction rates; but in the case of helium, we can make a simple estimate of the resulting abundance, as follows.

Because of the large binding energy of helium, formation of this nucleus is strongly favored over formation of other nuclei, and almost all the free neutrons initially available in the early universe are ultimately bound into helium nuclei.

Thus, the abundance of primordial helium hinges on the abundance of neutrons relative to protons.

At early times, say, $t < 10^{-2}$ s, in the lepton era, neutrons and protons are in thermal equilibrium.

Since the neutron is slightly more massive than the proton, thermal equilibrium favors protons over neutrons.

The ratio of the neutron and proton abundances is given by the Boltzmann factor

$$\frac{[\mathrm{n}]}{[\mathrm{p}]} = \mathrm{e}^{-Q/kT} \tag{10.24}$$

where $Q = 1.29$ MeV is the difference between the neutron and proton rest-mass energies.

The thermal equilibrium between neutrons and protons is maintained by the following reactions, which involve the weak interaction:

$$\mathrm{n} \leftrightarrow \mathrm{p} + \mathrm{e}^- + \overline{\nu} \tag{10.25}$$

$$\mathrm{e}^+ + \mathrm{n} \leftrightarrow \mathrm{p} + \overline{\nu} \tag{10.26}$$

$$\nu + \mathrm{n} \leftrightarrow \mathrm{p} + \mathrm{e}^- \tag{10.27}$$

The first of these reactions is by far the slowest, and it does not contribute much to the equilibrium.

Proceeding in the forward direction, this reaction is neutron decay, with a half-life of about 10 min, which is much longer than the typical time scale, or Hubble age, of the early universe; proceeding in the backward direction, this reaction requires a simultaneous collision among three particles, which is much less likely than a collision between two particles.

Hence, the important equilibrium reactions are (10.26) and (10.27).

The cross sections for these reactions can be calculated from the theory of the weak interactions, and these cross sections are found to be of the order of magnitude of $\sigma \cong \mathrm{G_F}^2 \hbar^{-4} \mathrm{E}^2$, where $\mathrm{G_F} = 1.4\times10^{-49}$ erg·cm^3 is the Fermi constant and E is the lepton energy.

The characteristic reaction time for a neutron is then $1/n\sigma$ v, where n is the density of incident leptons and v their speed.

The lepton density is $n \cong (kT/\hbar)^3$ and the speed is ≈ 1, because the leptons are relativistic at the high temperatures of interest; furthermore, the lepton energy is $\cong kT$.

Combining these factors, we find a characteristic reaction rate per neutron or proton of

$$n\sigma = \frac{G_F{}^2(kT)^5}{\hbar^7} \tag{10.28}$$

The reactions were able to maintain thermal equilibrium as long as the reaction rate was larger than the expansion rate H of the universe.

For the lepton era, H is given approximately by Eq. (10.19), and the condition for thermal equilibrium is approximately

$$\frac{G_F{}^2(kT)^5}{\hbar^7} > \frac{(kT)^2}{\hbar m_{Pl}} \tag{10.29}$$

from which we obtain

$$(kT)^3 > (0.8\ \text{MeV})^3 \tag{10.30}$$

From this we see that in the early universe, when the temperature was high, the reaction rate was large compared with the expansion rate, which confirms that the reactions (10.26) and (10.27) were able to maintain thermal equilibrium between neutrons and protons.

While thermal equilibrium lasted, the ratio of neutrons to protons was given by the Boltzmann factor, as in Eq. (10.24).

But when the universe cools to a lower temperature, the reaction rate decreases and becomes smaller than the expansion rate, that is, the reaction freezes out and thermal equilibrium of neutrons and protons fails.

For a rough estimate of the residual abundance of neutrons, we can assume that the neutron-proton ratio becomes frozen at the fixed value corresponding to the instant when the expansion rate equals the reaction rate.

According to Eq. (10.30), this happens at a temperature of $kT \cong 0.8$ MeV (at a time $t \cong 1$ s).

The neutron-proton ratio at this critical time is

$$\frac{[n]}{[p]} \cong e^{-1.29/0.8} \cong \frac{1}{5} \tag{10.31}$$

The abundance of neutrons, expressed as a fraction of the total amount of baryonic matter, is therefore

$$\frac{[n]}{[n]+[p]} \cong \frac{1}{6} \tag{10.32}$$

After the temperature drops to about 0.1 MeV (at $t \cong 1$ min), the neutrons quickly bind with protons to make deuterium nuclei, and further reactions of these deuterium nuclei with each other and with protons and neutrons lead to the formation of ^{4}He nuclei.

Within 2 or 3 minutes, almost all the neutrons are incorporated into helium nuclei.

The only neutrons to avoid this fate are a few that decay during the time interval between 1 s and 3 min (since the half-life of the neutron is 10 min, there is not enough time for a large amount of decay), a few that remain as deuterium, and a few that become incorporated into lithium.

If we ignore these small losses, we see that the fraction of ^{4}He of the total amount of baryonic matter, by mass, must be $\cong 1/3$ (of which 1/6 is due to the neutrons and 1/6 is due to the equal number of protons in ^{4}He).

The result of this rough estimate is in surprisingly good agreement with the observed helium abundance, $\cong 1/4$.

Note that we were able to obtain this result without specifying detailed initial conditions for the early universe – a high initial temperature is the only initial condition that we need to specify, and thermal equilibrium then determines all the relevant initial conditions for our calculation.

Also note that our result is independent of the mass density in the universe.

Although our rough calculation gives us a clear picture of the basic physics involved in the synthesis of primordial helium, in an accurate calculation we need to take into account that the cross sections for the reactions (10.26) and (10.27) are somewhat suppressed, because some of the final states are already occupied by the leptons of the fermion gas, and these states are then forbidden to the leptons emerging from the reactions (10.26) and (10.27).

Furthermore, the thermal equilibrium does not cease at one instant, but fails gradually over a time interval around $t \cong 1$ s.

This means that to evaluate the final neutron abundance we need to integrate the rates for the reactions (10.26) and (10.27) numerically.

And to evaluate the helium abundance, we need to integrate the rates of all the reactions involved in helium synthesis, as well as the competing reactions involved in the synthesis of light elements other than helium.

Figure 10.2 shows the results of such a numerical calculation of the synthesis of ^{4}He and other light elements.

The final abundance of each element is plotted as a function of the density of baryon matter.

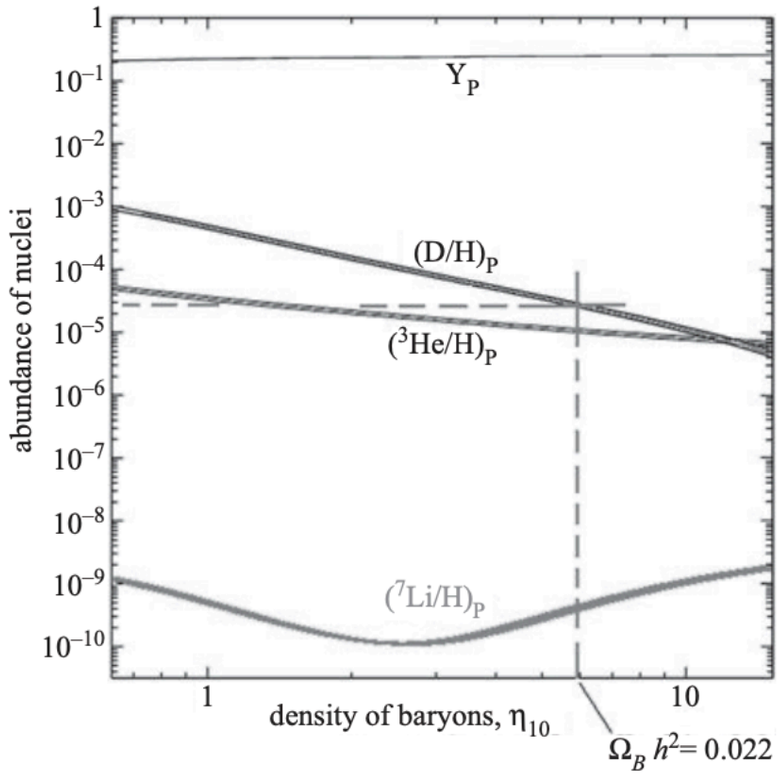


Fig. 10.2 Calculated abundances of light elements as a function of the density parameter η_{10}, which is proportional to the baryon density, $\eta_{10} = 274\Omega_B h^2$. The abundance of ^{4}He (marked Y_p) is expressed as a fraction of the baryon mass. Other abundances are expressed as fractions of the H abundance. For agreement with the measured deuterium abundance of 2.8×10^{-5} (see Table 10.1), the baryon density must be $\Omega_B h^2 = 0.022 \pm 10\%$. (From Steigman, 2006.)

Note that in the case of ^{4}He, the result is nearly independent of the baryon density – the abundance is between 0.20 and 0.25 over a wide range of values of ϱ_0.

This insensitivity to the baryon density was already suggested by our rough calculation.

The slight decrease of the abundance at small values of ϱ_0 revealed in Fig. 10.2 is a consequence of neutron decay; a universe of small ϱ_0 takes a longer time to cool to the temperature at which neutrons can bind with protons, and this delay permits more neutrons to decay.

For other light nuclei, the abundance is quite sensitive to the baryon density.

For instance, Fig. 10.2 shows that if the baryon density is high, the abundance of deuterium will be low.

The reason for this inverse relationship is that, although the rate of formation of deuterium increases when there are more neutrons and protons, the rate of destruction of the deuterium increases even more.

The abundances of ^{3}He and ^{7}Li are also sensitive to the baryon density.

Heavy elements cannot be formed in any significant amounts in the early universe.

This was first recognized in 1949 by Fermi and his colleague Turkevich, who tried to manufacture all the elements by nuclear reactions starting with neutrons as the initial ingredient.

They carried calculations of nuclear reaction rates forward for the first 30 minutes of the life of the universe, and obtained reasonable results for the accumulated H, He, and H isotopes.

But the lack of stable nuclei at mass numbers 5 and 8 constitutes a "barrier" that inhibits the further growth of a nucleus by stepwise capture of neutrons or protons.

Fermi and Turkevich could find no way to manufacture heavier elements, and therefore never published their calculations.

We now know that the heavier nuclei are formed much later, in the cores of stars, by fusion of lighter nuclei.

Observational data on the abundances of light elements come from a large variety of sources.

Intensities of spectral lines of ^{4}He can be measured in stars, planetary nebulas, and H II regions of galaxies.

Spectral lines of ^{7}Li have been measured in metal-poor stars (a low abundance of heavy elements indicates that the star was formed from primordial, uncontaminated gas).

Deuterium abundances have been measured in the Solar System and in molecular gas in the interstellar medium.

And ^{3}He abundances have been measured in H II regions and in planetary nebula.

All of these numbers suffer from uncertainties, ranging from a few percent to more than 10%.

Table 10.1 summarizes recent observational data on the abundances of light elements.

Table 10.1 Observed abundances of light elements*

Element	Abundance
^{4}He	0.25 as fraction of baryon mass
^{3}He	$< 1 \times 10^{-5}$ as fraction of H abundance
^{2}D	2.8×10^{-5} as fraction of H abundance
^{7}Li	2×10^{-10} as fraction of H abundance

* From a summary of recent data by Weinberg (2008).

The observed abundances are consistent with the calculations displayed in Fig. 10.2, provided that the baryon density parameter today is $\Omega_B h^2 = 0.022$, which means $\Omega_B = 0.044$ if $h = 0.7$.

This consistency of the theoretical calculations of nucleosynthesis and observation must be reckoned as a triumph of the Standard Model of the Big Bang.

In essence, in Fig. 10.2 we are attempting to fit four distinct pieces of data by adjusting one single parameter, Ω_B .

The Standard Model of the Big Bang meets this challenge with spectacular success.

Furthermore, the required value of Ω_B is consistent with observational data on the (baryon) mass density of the universe.

Another test of the Standard Model is provided by an examination of the number of neutrino families, characterized by distinct "flavors."

Although only the electron type of neutrino participates in the reactions (10.26) and (10.27), the presence of other kinds of neutrinos affects the abundance of ^{4}He because it changes the temperature at which the neutron-proton ratio freezes out of equilibrium.

The presence of other neutrinos increases H^{-1} somewhat; hence the freeze-out occurs at a higher temperature, with a higher abundance of neutrons, which leads to a higher abundance of helium.

For consistency with the observed helium abundance, the Standard Model demands that the number of neutrino families be no more than 4.

This is in good agreement with data from high-energy accelerator experiments, according to which the most likely number of neutrino families is 3.

10.3 Density perturbations; Jeans mass

Although the universe is fairly smooth on a large scale, it has prominent clumps on a small scale, where we find galaxies and clusters of galaxies whose mass density is much in excess of the average density of the universe.

The development of such clumps in the universe is due to a gravitational instability: If there is some small initial enhancement of the density in some volume, the gravitational attraction tends to contract this volume, which leads to a further enhancement of the density.

Thus, density perturbations tend to grow.

However, the gravitational contraction is opposed by pressure forces, which tend to maintain the density uniform.

We will have to investigate under what conditions a density perturbation is unstable, so it will continue to grow.

For the sake of simplicity we begin with an analysis of the density perturbations in a uniform, static fluid governed by Newtonian physics.

This means we ignore the expansion of the universe and the curvature of space and time.

Our result is therefore only an approximation, applicable only to a small region of our expanding and curved universe.

However, as we will see, the qualitative features of our result are of wider applicability.

The Newtonian equations that determine the motion of the fluid under the influence of pressure forces and gravity are

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{10.33}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} + \frac{1}{\rho}\nabla p + \nabla \Phi = 0 \tag{10.34}$$

$$\nabla^2 \Phi = 4\pi G \rho \tag{10.35}$$

The first of these equations is simply the continuity equation; the second is Newton's second law (the first two terms are the "total" derivative of the velocity, and the other two terms are the negatives of the pressure force and the gravitational force per unit mass); and the third is Poisson's equation for the gravitational potential.

Suppose that $\bar{\rho}$, $\bar{p}$, $\bar{\mathbf{v}}$, and $\bar{\mathbf{\Phi}}$ are the quantities appropriate for the static, uniform fluid.

If the fluid suffers a small perturbation, these quantities are changed to $\bar{\rho} + \delta\rho$, $\bar{p} + \delta p$, $\bar{\mathbf{v}} + \delta\mathbf{v}$, and $\bar{\mathbf{\Phi}} + \delta\mathbf{\Phi}$.

To first order in small quantities, Eqs. (10.33)– (10.35) become

$$\frac{\partial \delta\rho}{\partial t} + \bar{\rho}\nabla \cdot (\delta\mathbf{v}) = 0 \tag{10.36}$$

$$\frac{\partial \delta\mathbf{v}}{\partial t} + \frac{{v_s}^2}{\bar{\rho}}\nabla\delta\rho + \nabla\delta\Phi = 0 \tag{10.37}$$

$$\nabla^2 \delta\Phi = 4\pi G \delta\rho \tag{10.38}$$

where

$$v_s = \sqrt{\frac{\delta p}{\delta \rho}} \tag{10.39}$$

is the speed of sound in the fluid.

If we take the divergence of Eq. (10.37) and use the other two equations, we obtain a single differential equation for δϱ:

$$\frac{\partial^2 \delta\rho}{\partial t^2} - {v_s}^2 \nabla^2 \delta\rho = 4\pi G \overline{\rho}\, \delta\rho \tag{10.40}$$

To investigate the character of the solutions of this equation, consider a plane-wave solution of the form

$$\delta\rho(\mathbf{x}, t) = A e^{i\mathbf{k}\cdot\mathbf{x} - i\omega t} \tag{10.41}$$

Upon substituting this into Eq. (10.40), we find that ω and **k** must satisfy the dispersion relation

$$\omega^2 - {v_s}^2 k^2 = -4\pi G \overline{\rho} \tag{10.42}$$

This tells us that if k is large, ω will be real, and the perturbation will oscillate harmonically.

But if *k* is small, ω will be imaginary, and the perturbation will grow exponentially.

The critical value of *k* below which the perturbation will grow is $k_J = \sqrt{4\pi G\bar{\rho}/v_s^2}$, and the critical wavelength is

$$\lambda_J = \frac{2\pi}{k_J} = \sqrt{\frac{\pi {v_s}^2}{G\overline{\rho}}} \tag{10.43}$$

This is called the *Jeans wavelength*.

If we want to construct a localized contracting perturbation of finite extent, we must form a wave packet by superposition of such plane waves.

Since the minimum wavelength for a contracting perturbation is $\cong \lambda_J$, the volume of the wave packet cannot be smaller than $\cong (\lambda_J)^3$, and the amount of mass in this volume is

$$M_J = (\lambda_J)^3 \overline{\rho} = \pi^{3/2} \frac{{v_s}^3}{\sqrt{G^3 \overline{\rho}}} \tag{10.44}$$

This is called the *Jeans mass*; it is the minimum mass that leads to growth of a density perturbation.

We can understand the growth of a density perturbation by a simple physical argument.

Suppose that the initial density perturbation occurs in a spherical volume of radius λ within which the density is slightly in excess of the normal density.

The mass M in this volume will then begin to fall toward the center under the influence of its own gravitational attraction.

The typical speed of fall is $\sqrt{GM/\lambda}$, and the characteristic "free-fall" time for the process of contraction equals the radius λ divided by the typical speed, $\cong \lambda/\sqrt{GM/\lambda} \cong 1/\sqrt{G\bar{\rho}}$.

The pressure forces that oppose the contraction are propagated in the form of sound waves, so the characteristic response time for the pressure forces is $\cong \lambda/v_S$.

If this characteristic time is longer than the free fall time, the pressure forces will respond too slowly to prevent the contraction.

Hence, the condition for instability and for growth of the density perturbation is $\lambda/v_S > 1/\sqrt{G\bar{\rho}}$, which says that λ must exceed the Jeans wavelength and the mass must exceed the Jeans mass, in agreement with our perturbation analysis based on Eq. (10.40).

Figure 10.3 is a plot of the magnitude of the Jeans mass vs. temperature, for a universe dominated by baryons and dark mass at late times.

The Jeans masses for the baryonic and the dark-mass components can be treated separately, because the dark mass consists of weakly interacting massive particles (WIMPs), which do not interact with baryons, so the perturbations of baryons and of dark mass remain independent, except when the perturbations grow so large that they have mutual gravitational effects.

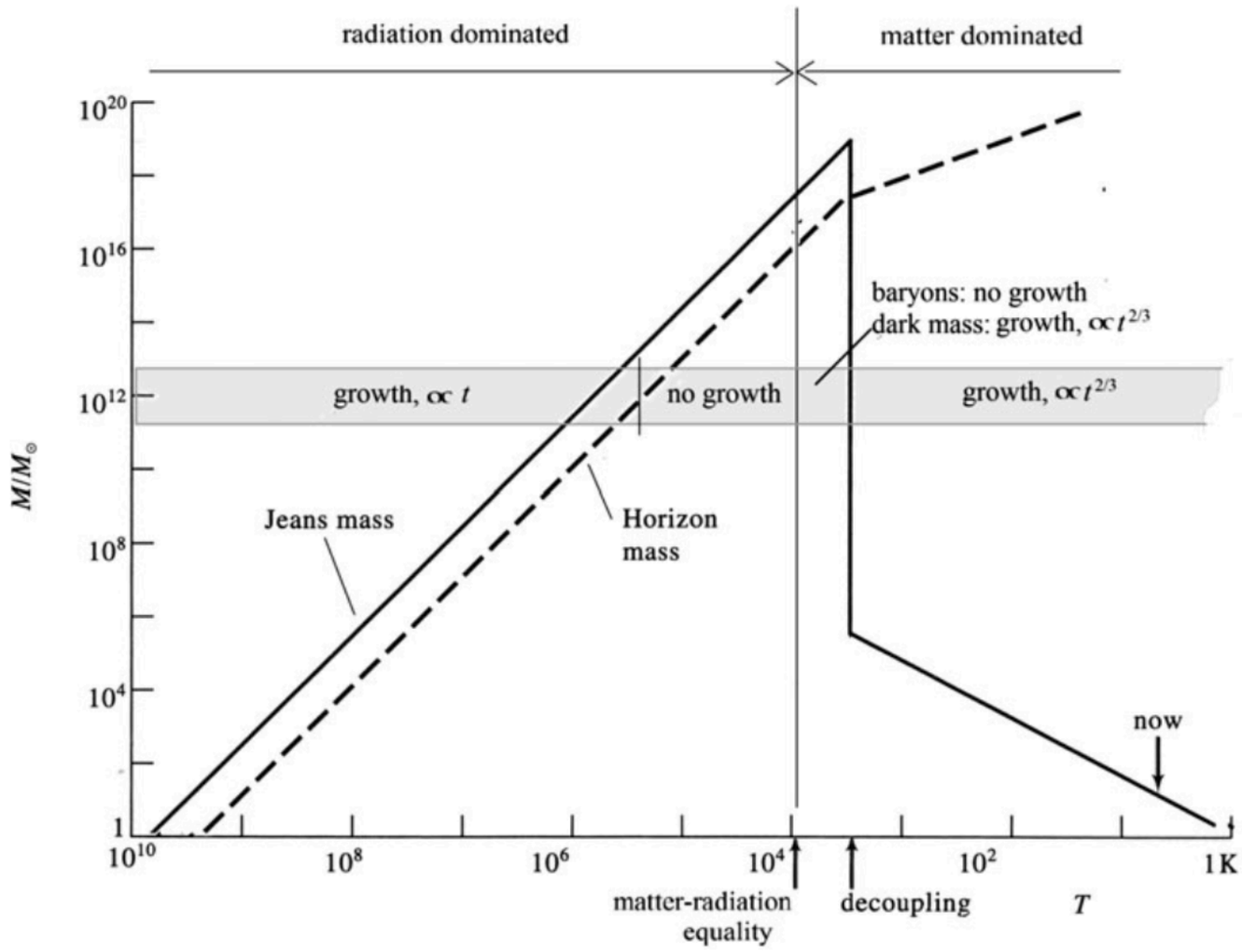


Fig. 10.3 Magnitude of the baryon Jeans mass as a function of the temperature T of the universe for a model with $\Omega_B = 0.04, \Omega_{dark,0} = 0.26, \Omega_{\Lambda,0} = 0.70$ (see Table 9.1). The temperature T decreases monotonically with time; hence this plot also indicates the time dependence of the Jeans mass. The temperature for decoupling is $T_{dec} = 3000$ K, and the temperature for matter-radiation equal density is $T_{eq} = 9700$ K. The dashed line shows the horizon mass. The horizontal band summarizes the different perturbation regimes for a clump of baryon mass of about $10^{12} M_\odot$ and its associated radiation and dark mass components.

During the radiation era, the pressure in the universe is supplied by radiation, which is strongly coupled to the baryonic plasma.

The pressure of the radiation is high (compared with the mass density), and this leads to a high and constant sound velocity of $c/\sqrt{3}$, nearly equal to the speed of light.

According to Eq. (10.44), the Jeans mass for baryons therefore increases $\propto 1/\sqrt{\rho}$ when the density ϱ decreases.

The baryon Jeans mass reaches a maximum value of $10^{19}M_{\odot}$ at the time of decoupling.

After that, the radiation pressure ceases to affect baryons, and the only pressure acting on the baryons is then their own, relatively low, kinetic pressure.

This leads to a drastic drop in the sound velocity and a corresponding drop in the Jeans mass.

Immediately after decoupling, the Jeans mass drops to $10^{5}M_{\odot}$, and it continues to decrease thereafter.

Our analysis of perturbations is defective in that it does not take into account the expansion of the universe and general-relativistic effects arising from the existence of a horizon.

Calculations in an expanding background reveal that the growth of perturbations is always much smaller than in a static background, because any increase of density arising from the Jeans instability is partially canceled by the decrease of density arising from the expansion – the expansion “dilutes” any increase of density.

The net result is that the growth of perturbations is not exponential, but instead follows a power law, and if the expansion is fast (as in the case of a radiation-dominated universe), the growth of perturbations is almost completely quenched.

Furthermore, over most of the range of temperatures and times displayed in Fig. 10.3, the Jeans mass exceeds the horizon mass, that is, the total mass that exists within the distance to the particle horizon.

This horizon mass is shown in Fig. 10.3 by the dashed line.

Since perturbations in the gravitational potential cannot propagate beyond this horizon, the Newtonian analysis for the behavior of perturbations becomes invalid.

We can conclude from Fig. 10.3 that masses below the horizon mass will not develop a Jeans instability, with growing perturbations, but we cannot conclude that masses above the nominal Jeans mass plotted in this figure will actually develop the growing perturbations suggested by the Newtonian analysis.

To verify the growth of perturbations, we need to perform a relativistic analysis, by means of Einstein's gravitational field equations.

Table 10.2 summarizes the behavior of perturbations in an expanding universe in different epochs, for mass scales below and above the horizon mass.

In this table, the perturbations are characterized by the fractional increment in density, called the density contrast δ,

$$\delta = \frac{\delta\rho}{\rho} \tag{10.45}$$

Table 10.2 Growth of perturbations in the expanding universe*

Epoch	Mass < Horizon Mass	Mass > Horizon Mass
Before T_{eq} (radiation dominated)	no growth	growth, $\delta \propto t$
Between T_{eq} and T_{dec} (matter dominated)	baryonic mass: no growth cold dark mass: growth, $\delta \propto t^{2/3}$	growth, $\delta \propto t^{2/3}$
After T_{dec} (matter dominated)	growth, $\delta \propto t^{2/3}$, if mass > M_J	growth, $\delta \propto t^{2/3}$

* Based on Kolb and Turner (1990).

The time t used in the power laws given in the table is the usual cosmological time, which starts at t = 0, when the universe is at infinite density.

The results listed in Table 10.2 were obtained by a perturbation analysis that includes both the expansion of the universe and general-relativistic effects.

The full analysis of the behavior of perturbations that extend over distances beyond the horizon distance is somewhat complicated, but the results in the last column of Table 10.2 can be understood by examining the simple and special case of a uniform density perturbation in a Friedmann universe of zero curvature.

In such a universe, the unperturbed Einstein equation is [see Eq. (9.113)]

$$3H^2 = 8\pi G\rho + \Lambda \tag{10.46}$$

If we increase the density slightly, to a new value $\varrho + \delta\varrho$, the universe changes to a Friedmann universe of positive curvature with an Einstein equation

$$3\left(H^2 + \frac{1}{a^2}\right) = 8\pi G(\rho + \delta\rho) + \Lambda \tag{10.47}$$

From Eqs. (10.46) and (10.47) with equal values of the Hubble constant, we find that the density contrast $\delta = (1/\varrho)\delta\varrho$ obeys the equation

$$\delta = \frac{3}{8\pi G\rho a^2} \tag{10.48}$$

The time dependence of a and ϱ therefore determines the time dependence of δ.

In the radiation-dominated era, $a \propto t^{1/2}$ and $\varrho \propto t^{-2}$ [see Eqs. (10.14) and (10.15)]; hence

$$\delta \propto t \text{ (radiation dominated)} \tag{10.49}$$

In the matter-dominated era, $a \propto t^{2/3}$ and $\varrho \propto t^{-2}$ [see Eqs. (9.115) and (9.95)]; hence

$$\delta \propto t^{2/3} \quad \text{(matter dominated)} \tag{10.50}$$

Although we obtained these results for the time dependence of the growing perturbations from a special model, the general analysis leads to the same results: The density contrast for a perturbation involving a mass much larger than the horizon mass grows in proportion to t in the radiation era, and in proportion to $t^{2/3}$ in the matter era.

Thus, in Fig. 10.3, the mass scales above the plotted lines actually do correspond to growing perturbations, as suggested by the simple nonrelativistic Jeans analysis.

With the results of Table 10.2 on hand, it is instructive to consider the evolution of a perturbation consisting of a baryon mass of $10^{12}M_{\odot}$, which is approximately the mass of a typical galaxy.

During the early radiation era, this mass was in excess of the horizon mass, and the perturbation therefore grew; that is, the density of the clump of mass began to increase relative to the density of the background (because the background density of the expanding universe is decreasing, such a growth in the density contrast of a clump does not imply an increase in the density itself, which actually decreases, but more slowly than the background density, $\delta\varrho = \varrho \times \delta \propto t^{-2} \times t \propto t^{-1}$ and therefore $\varrho + \delta\varrho \propto t^{-2} + \text{constant} \times t^{-1}$).

At a time corresponding to about $T \sim= 2 \times 10^5$ K, the horizon mass reaches $10^{12}M_{\odot}$ (see the horizontal band in Fig. 10.3).

At this point, the character of the perturbation in our clump changed from a growing to an oscillatory nongrowing behavior, and our clump is said to have "entered the horizon."

The perturbation then oscillated until the time of decoupling, where the Jeans mass dropped to about $10^5M_{\odot}$, and the character of the perturbation changed back from oscillatory to growing.

In the matter-dominated universe after decoupling, the density contrast of the clump of mass grew in proportion to $t^{2/3}$, and finally it evolved into a galaxy.

The dark-mass component of the clump entered the horizon at the same time as the baryon component ($T \cong 2 \times 10^5$ K).

It persisted in oscillations until T_{eq}, at which point the expansion slowed sufficiently to permit the growth of perturbations.

For "cold" dark mass, with particle rest masses much larger than 2.7 eV, the pressure and the sound velocity are very small, so the Jeans mass is also very small, and the perturbations in the dark mass resumed their growth at T_{eq} (see Table 10.2) and continued to grow thereafter.

Thus, the growth of perturbations in the dark mass began earlier than that in the baryon mass, and the dark mass developed gravitational potential wells earlier than the baryon mass.

The baryons settled into these potential wells; that is, they became bound to the dark-mass distribution.

The baryonic density contrast thereby became dependent on the dark-mass density contrast, and it quickly built up to the level of this stronger, more developed density contrast, which helped galaxy formation.

Accordingly, the structure of galaxies and of clusters of galaxies is mainly governed by the properties of the dark mass.

The hypothesized forms of dark mass fall into two categories: hot dark mass and cold dark mass.

Hot dark mass consists of particles of fairly small mass (say, $m \leq 100$ eV), such as neutrinos of small but nonzero mass, which would form a relativistic gas at the temperatures prevailing toward the end of the radiation era.

Cold dark mass consists of particles of large mass (say, $m \geq 10$ keV), which would form a nonrelativistic gas at these temperatures.

Here, the words "hot" and "cold" do not really describe the actual temperature, but whether the particles experience this temperature as a high temperature (particles of small mass, $m \leq kT$) or as a low temperature (particles of large mass, $m \gg kT$).

Table 10.3 lists some possible dark-mass candidates.

Table 10.3 Dark-mass candidates (*W*eakly *I*nteracting *M*assive *P*articles, or WIMPs)*

WIMP	Mass	Abundance, now (number per cm^3)
Hot Dark Mass		
Invisible axion	10^{-5} eV	10^9
Light neutrino	30 eV	10^2
Warm Dark Mass		
Photino, gravitino	$\approx$keV	10
Cold Dark Mass		
Photino, Sneutrino, neutralino, axino, heavy neutrino	$\approx$GeV	10^{-5}
Magnetic monopoles	10^{16} GeV	10^{-21}
Pyrgons, Maximons, Newtorites	10^{19} GeV	10^{-24}
Quark nuggets	$\approx 10^{15}$ g	10^{-44}
Primordial black holes	$>10^{15}$ g	$<10^{-44}$

* From Kolb and Turner, 1990.

The perturbations that ultimately grow into galaxies and clusters of galaxies begin as primeval fluctuations in the very early universe.

For instance, in the inflationary model (see Section 10.4), such fluctuations arise from quantum processes.

But the evolution of such perturbations into galaxies and clusters of galaxies proceeds in different ways in a universe filled with hot dark mass and a universe filled with cold dark mass.

In the hot-mass-dominated universe, the large structures form first, and the small structures form later, by fragmentation of the larger objects (such an evolution of structure is called “from the top down”).

In contrast, in the cold-mass-dominated universe, the small structures form first, and the larger structures form later, by the merging of small objects (such an evolution of structure is called “from the bottom up”).

This difference in the evolution of structures arises from a difference in the behavior of the small-scale density perturbations in these two kinds of universes.

In the hot-mass-dominated universe, any small-scale density perturbations tend to be wiped out by the free-streaming motion of the particles.

The hot-mass particles are relativistic ($E \cong kT \geq m$), and their high-speed motion tends to disperse any initial density perturbation.

The only density perturbations that survive this free-streaming motion are those of a diameter so large that the particles do not have enough time to stream out.

In the absence of small-scale density perturbations, the large-scale perturbations will be the first to contract gravitationally, and they will form large-scale structures.

In the cold-mass-dominated universe, the particles move slowly, and small-scale density perturbations do not tend to disperse.

According to the inflationary model of the early universe, the small-scale density perturbations predominate, and when they contract gravitationally, they will form small-scale structures.

The evolution of large-scale structures in a universe filled either with hot mass or with cold mass has been explored by means of numerical simulations that track the motion of a large number of galaxies, treated as mass points interacting gravitationally.

In the hot-mass-dominated universe, the initial gravitational contraction of each large cloud of galaxies is strongly anisotropic, usually with one preferential direction of collapse.

The result is that the initial clouds squash into quite thin “pancakes”, which subsequently fragment into smaller objects.

Large voids are left between the “pancakes.”

Although some such voids have been observed in the actual universe, the simulations indicate that in the hot-mass-dominated universe the large-scale structures tend to overdevelop, with excessive contrasts between the voids and the thin, but dense, pancakes.

The simulations for the cold-mass-dominated universe gives better results (see Fig. 10.4).

The calculations are in fair agreement with what we see in the actual universe, except that large-scale structures tend to underdevelop somewhat.

The best results are obtained if the dark mass is mostly cold mass, but with a modest admixture of hot mass, maybe 10% or so.

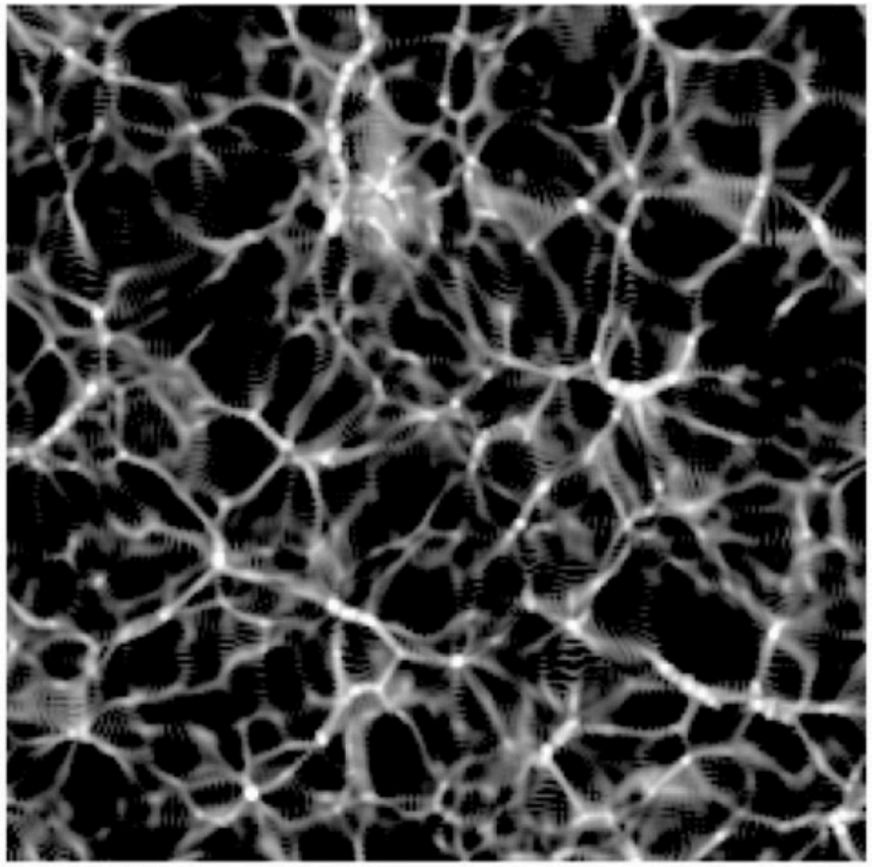


Fig. 10.4 Density distribution of mass calculated from a numerical simulation involving nearly 20 million mass points, with a spectrum of initial density perturbations typical of cold dark mass. The picture shows the final density distribution, after an evolution starting at an age of about 10^7 years. The filamentary structures in this evolved distribution are quite similar to the observed structures in our universe today (compare Fig. 9.1). (From Suhonenko et al., 2011. Courtesy J. Einasto, Tartu Observatory and ICRANet, Pescara)

10.4 Inflationary model

The Standard Model of the early universe, described in the previous sections, is quite successful in dealing with the development of the universe from an age of about 10^{-5} s onward.

Thus, the model gives us an account of the history of our universe over a range of ages spanning 23 powers of 10.

Its most impressive success lies in the calculations of the abundances of helium and other light elements.

However, when we seek to extrapolate the Standard Model to earlier times, the model presents us with several puzzles for which it offers no answers:

The Flatness Puzzle. Even years before astronomers acquired the new observational evidence from high-z supernovas, it was known that the sum of the density parameters $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ is of the order of magnitude $\Omega_0 \approx 1$.

Thus, Ω_0 is of the order of magnitude of the value required for a spatially flat universe.

What makes this puzzling is that an aged universe, such as ours, can achieve a value close to $\Omega_0 \approx 1$ only by an extremely delicate tuning of the value of Ω at an early time.

At an early time, the value of Ω was necessarily even closer to 1 than it is now.

We can see this from Eq. (9.143):

$$a^2 = -\frac{K}{H^2(1-\Omega)}$$

This shows that $|1-\Omega|$ is proportional to $1/H^2a^2$, and therefore

$$|1-\Omega| = |1-\Omega_0|\,\frac{H_0{}^2 a_0{}^2}{H^2 a^2} \tag{10.51}$$

If today $|1-\Omega_0|$ is of the order of magnitude of 1, then in the early universe $|1-\Omega|$must have been much smaller, because H^2a^2 was much larger than it is now.

We can estimate how much smaller it was by examining the changes in the factor $1/H^2a^2$ in Eq. (10.51) as we go back in time.

In a matter-dominated universe, as ours was over most of its history, $a \propto t^{2/3}$ and $H = (1/a)(da/dt) \propto 1/t$ [see Eq. (9.115) for a flat universe; the negatively or positively curved universes have approximately the same behavior until they reach a mature age].

The factor $1/H^2a^2$ is then $\propto t^{2/3}$, and between now and the beginning of the matter era (at 10^{12} s), this factor decreases by about $(10^{12}\ \text{s}/1.4 \times 10^{10}\ \text{y})^{2/3} = 2 \times 10^{-4}$.

At earlier times, in the radiation-dominated universe, $a \propto t^{1/2}$ and $H \propto 1/t$ [see Eqs. (10.14) and (10.19)].

The factor $1/H^2a^2$ is then $\propto t$, and between the beginning of the matter era and the beginning of the hadron era (at 10^{-43} s), this factor decreases by about $(10^{-43}\ \text{s}/10^{12}\ \text{s}) = 10^{-55}$.

Thus, the net decrease of $1/H^2a^2$ between now and the earliest stage of the universe is a factor of about 10^{-59}; that is, $|1 - \Omega|$ must have had an initial value of about 10^{-59} to attain a value of about 1 today!

If the early universe had an initial value of Ω that was not so extremely close to 1, the difference would have grown drastically (proportional to t), and the universe would quickly have attained either a very large value of Ω (if $K > 0$ and the initial value of Ω was larger than 1) or else a very small value (if $K < 0$ and the initial value of Ω was smaller than 1).

This means the universe either re-contracts very quickly, or else it expands very quickly and disperses.

Only if the initial value of Ω is *extremely* close to 1 will the universe evolve into an era where Ω remains of the order of magnitude of 1 for a fairly long time.

Thus, our universe is a very exceptional universe: It must have started with an initial value of Ω almost exactly equal to 1 in order to avoid re-contraction or dispersion by now.

Expressed another way, the initial universe must have been an almost exactly flat universe to permit our present universe to be not too far from a flat universe.

Although such a fine-tuning of the initial value of Ω seems puzzling, it may not have any special significance.

A priori, any initial value of Ω is as likely as any other value, and maybe at $t = 10^{-43}$ s the universe had an initial value of, say, $\Omega = 1 - 10^{-60}$ by coincidence.

This, of course, is a very lucky coincidence for us, because if Ω had been outside the range 1 ± 10^{-59}, then the universe would not have permitted us to exist.

But the coincidence $\Omega = 1 - 10^{-60}$ would require an explanation only if we had several universes before our eyes, and we found that in all of them Ω has this value.

In observing only a single universe, and in finding some value of Ω, we are not faced with a repetitive pattern, and we require no explanation – the universe is what it is.

Furthermore, the puzzle disappears completely if we assume that our universe is an exactly flat universe.

In such a universe, Ω is not a free parameter, but is permanently fixed at $\Omega = 1$.

Maybe the above arguments indicating that the early value of Ω was extremely close to 1 should be taken as a strong hint that our universe is indeed flat.

Since the flatness puzzle is more imagined than real, it does not indicate much of a deficiency in the Standard Model.

However, if we can invent an alternative model of the universe that gives us a value of Ω of the order of 1 at the present time without invoking any special conditions, then we might judge such an alternative model somewhat superior to the Standard Model.

The Smoothness Puzzle. On a large scale, the present universe is fairly smooth, that is, homogeneous and isotropic.

We relied on this smoothness of the universe in our treatment of the Friedmann models in the preceding chapter.

The best observational evidence for this smoothness is the isotropy of the cosmic background radiation – the temperature of this radiation is uniform over the sky to better than 1 part in 10^4.

This uniformity poses a true puzzle, because it cannot be attributed to thermal contact and thermal equilibrium in the early universe.

Different regions of the early universe were out of contact with each other; they could not have communicated by thermal signals or even by light signals, because not enough time had elapsed to permit signals to reach from one region to the other.

Consider, for instance, two opposite patches in the sky, separated by 180°.

The cosmic background photons reaching the Earth now were emitted by the hot hydrogen plasma, just before decoupling.

Since the photons from one of these patches are reaching the Earth only now, they have not yet reached the opposite patch – even today the two patches are out of contact.

The maximum distance that a signal can reach from one patch is limited by the available travel time; as we saw in Section 9.11, this maximum distance defines the particle horizon.

Because the smoothness puzzle hinges on the existence of the particle horizon, it is also called the horizon puzzle.

To appreciate how much two opposite patches in the sky are out of contact, let us express the distance between them as a multiple of the horizon distance.

The coordinate distance χ of each patch from the Earth can be calculated by examining the worldline of a light signal from the patch to the Earth,

$$\chi = \eta_0 - \eta_{dec} \tag{10.52}$$

where η_0 is the time parameter now and η_{rec} is the time parameter at decoupling.

The coordinate distance between the two opposite patches is therefore

$$\Delta\chi = 2(\eta_0 - \eta_{dec}) \tag{10.53}$$

The horizon distance associated with each patch at the decoupling time equals η_{dec}.

(This is the distance reached by a light signal that has left the patch at $t = 0$. Actually, the earliest time at which it is meaningful to speak of the travel of light signals is the Planck time, 10^{-43} s; but since the decoupling time is much larger than the Planck time, it hardly matters whether we take η_{dec} or $\eta_{dec} - \eta_{Pl}$ as our horizon distance.)

Hence the number of horizon distances lying between the two patches is

$$\text{number of horizon distances} = \frac{2(\eta_0 - \eta_{dec})}{\eta_{dec}} \tag{10.54}$$

The values of η_0 and η_{dec} that appear in this ratio can be obtained by calculation of the integral $\eta(t) = \int dt'/a(t')$.

For times t' that are within the matter-and- Λ -dominated era, the appropriate expression for the function a(t) is given by Eq. (9.129); for times t' that are within the radiation-dominated era, a(t) is given by Eq. (10.14).

The result for the number of horizon distances lying between the two patches is about 120.

This means that at decoupling, the two patches were separated by a diameter that spanned 60 causally disconnected regions, each of diameter equal to two horizon distances.

Under the assumption that the spatial geometry is flat, a diameter of 60 implies a circumference of $\pi \times 60 \cong 180$; that is, there were 180 causally disconnected regions around the circumference of the sky.

The angular diameter of each such causally dis- connected region is therefore $\cong 2^{\circ}$.

Accordingly, when we look at the sky, we would expect to see that the temperatures in patches separated by 2° or more are completely independent – yet the observational data show that these temperatures are almost exactly the same!

An accidental coincidence of the temperatures in all such patches of the sky is beyond belief.

There ought to be some other explanation for the smoothness of the temperature distribution of the universe and, likewise, for the smoothness in the mass distribution, the smoothness of the expansion rate, and so on.

One possible explanation is to suppose that the present expanding universe was preceded by a contracting universe, which passed through a moment of greatest contraction.

Such a behavior of a(t) is illustrated in Fig. 9.12.

However, to make the continuation from one universe to the next meaningful, a(t) must remain finite at the moment of maximum contraction, that is, there must be no singularity.

The absence of a singularity at the first moment of the Big Bang would seem to be in contradiction with general singularity theorems that have been proved for the solutions of Einstein's equations.

But these theorems deal only with the classical regime; they can be circumvented in the very early universe, where matter is in the form of quantum fields and where even the geometry is quantized.

Furthermore, the presence of a positive cosmological constant can lead to a violation of the energy condition for the Hawking- Penrose theorem even in the classical regime.

As we will see, in the inflationary epoch of the early universe, the cosmological constant was large and positive, and singularity theorems were inapplicable.

And in the pre-inflationary, very dense epoch, quantum fields dominated the universe, so singularity theorems were, again, inapplicable.

Besides, if the spacetime geometry develops a singularity of the kind envisioned in the singularity theorems, this merely indicates a breakdown in the equations for geodesic motion, which is pretty much irrelevant on the short distance scales that are of interest in connection with singularities.

As a last resort, we could adopt some modification of the Einstein equations that avoids singularities altogether, and some attempts at such modifications have been made.

Thus, the question of whether there is an actual singularity at the first moment of the Big Bang still remains open.

A contracting and re-expanding universe, without a singularity, has no horizon, and no horizon problem.

This explanation of the horizon problem has not found much favor among cosmologists, because we do not (yet) know how to calculate the evolution of the universe from the contracting stage to the expanding stage.

The inflationary explanation (see the later discussion) is popular because some of its details can be calculated.

The monopole puzzle. The generally accepted theory of fundamental interactions between particles (the "Standard Model" of electromagnetic, weak, and strong interactions) predicts that, at high energies, the interactions become unified.

Such a unification means that different interactions become merely different aspects of a single underlying interaction.

Thus, at an energy of about 300 GeV, the electromagnetic and weak interactions become unified into an electroweak interaction, and at an energy of about 10^{14} GeV, the strong interaction should become unified with the electroweak interaction (Grand Unified Theory, or GUT).

In the unified phase, the particles and the vacuum are endowed with special symmetries.

At the highest energies, the particle states have the maximum symmetry; at lower energies, the symmetry decreases ("spontaneous breakdown" of the symmetry), and concurrently, the unification of the interactions is destroyed.

The extreme temperatures in the very early universe are more than sufficient to bring about the complete unification of the strong and electroweak interactions.

This grand unification lasts until the universe cools to 10^{26} K ($\cong 10^{14}$ GeV), where there is a first spontaneous breakdown of symmetry.

The unification of the electromagnetic and weak interactions lasts until the universe cools to 10^{15} K ($\cong$ 300 GeV), where there is a further spontaneous breakdown of the remaining symmetry.

Figure 10.1 includes a summary of the unification eras of the universe.

Unfortunately, because of the presence of horizons, the breakdown of the symmetry does not proceed smoothly.

Regions separated by a horizon distance are causally disconnected, and hence the preferential directions of symmetry breaking are expected to be different in such regions (in this context, the "direction" is not a direction in ordinary space, but in the internal-symmetry space).

Hence, there will be discontinuities, or defects, in the directions of symmetry breakdown ("topological defects").

Among the defects that formed during the breakdown of grand unification are pointlike defects, which have the character of magnetic monopoles, that is, sources or sinks of magnetic field lines.

The number of such defects is expected to be at least one per horizon volume.

Since the horizon volume is quite small and the mass of each monopole is large, about 10^{16} GeV, the mass density of magnetic monopoles is enormous – their mass density would be 10^{14} times as large as the baryonic mass density!

This catastrophic prediction provided the original motivation for the inflationary model of the early universe – a large, inflationary expansion of the universe was supposed to reduce the density of monopoles to a low, unobservable level.

However, theorists later found alternative ways to achieve the spontaneous breakdown of symmetry, without topological defects, so the monopole puzzle resolved itself, and today it is not regarded as a serious problem.

The inflationary model was first proposed by Guth, and it was later modified and improved by Linde and by Albrecht and Steinhardt.

The essential assumption of the inflationary model is that, at about the time of the grand unification symmetry breakdown, the universe goes through an enormous and quick expansion, an expansion so large that a single horizon volume of the pre-expansion universe encompasses the entire observable universe that surrounds us today.

Inflation solves the monopole puzzle – one monopole per horizon volume implies at most one monopole within our observable universe.

Inflation also solves the smoothness puzzle, because before inflation, the matter in the entire observable universe was in causal and thermal contact and it could attain a uniform temperature.

And inflation also solves the flatness problem.

During inflation, the radius of curvature $a(t)$ increases by an enormous factor, perhaps a factor of 10^{43}.

This makes the initially curved geometry almost exactly flat, and it makes the final value of Ω almost exactly 1, regardless of what the initial value of Ω was.

To appreciate the enormous expansion of the universe, note that before inflation, at 10^{-34} s at the beginning of the hadron era, our entire visible universe today was contained in a volume about 10^{-41} cm across.

During inflation, this volume expanded to about 100 cm (and during the following 13 billion years it expanded to about 10^{29} cm).

The inflation is caused by a GUT phase transition of the matter in the universe, from a high-energy, symmetric, grand-unified phase to a low-energy, not symmetric, not unified phase.

At the earliest times and the highest temperatures, the large available thermal energies keep the universe in the grand-unified phase.

But when the temperature of the universe decreases to a critical value of 10^{14} GeV, this high-energy symmetric state cannot be maintained, and the universe settles into the state of less energy, which is not a symmetric state.

That is, the universe makes a phase transition from the high-energy phase to the low-energy phase.

This phase transition is believed to occur gradually, so the universe lingers in what might be called a supercooled condition, in which the symmetric vacuum state still retains its large amount of energy.

As we know from Section 7.3, a vacuum energy shows up in the Einstein equations in the form of a cosmological term, with a nonzero cosmological constant.

A positive energy density corresponds to a positive cosmological constant.

In the lingering, supercooled, condition, the thermal energy density associated with particles is negligible compared with the energy density of the vacuum; consequently, the Einstein equation for the inflationary development of the universe is the de Sitter equation, and the solution for $a(t)$ is Eq. (9.123),

$$a(t) = a(t_I)\exp\sqrt{\frac{\Lambda}{3}}(t - t_I) \tag{10.55}$$

where t_I is the time at the beginning of inflation.

Thus, the inflation of the radius of curvature is exponential.

This expansion lasts until the vacuum finally completes its phase transition.

Of course, the enormous expansion of the universe also leads to an enormous and quick decrease in temperature, to a temperature near zero.

However, the energy released in the phase transition of the vacuum is ultimately changed into thermal energy, and this thermal energy is transferred to the particles in the universe, reheating them to about the same temperature they would have had without inflation.

Figure 10.5 is a sketch of the radius of curvature and the temperature of the universe in the inflationary model as a function of time.

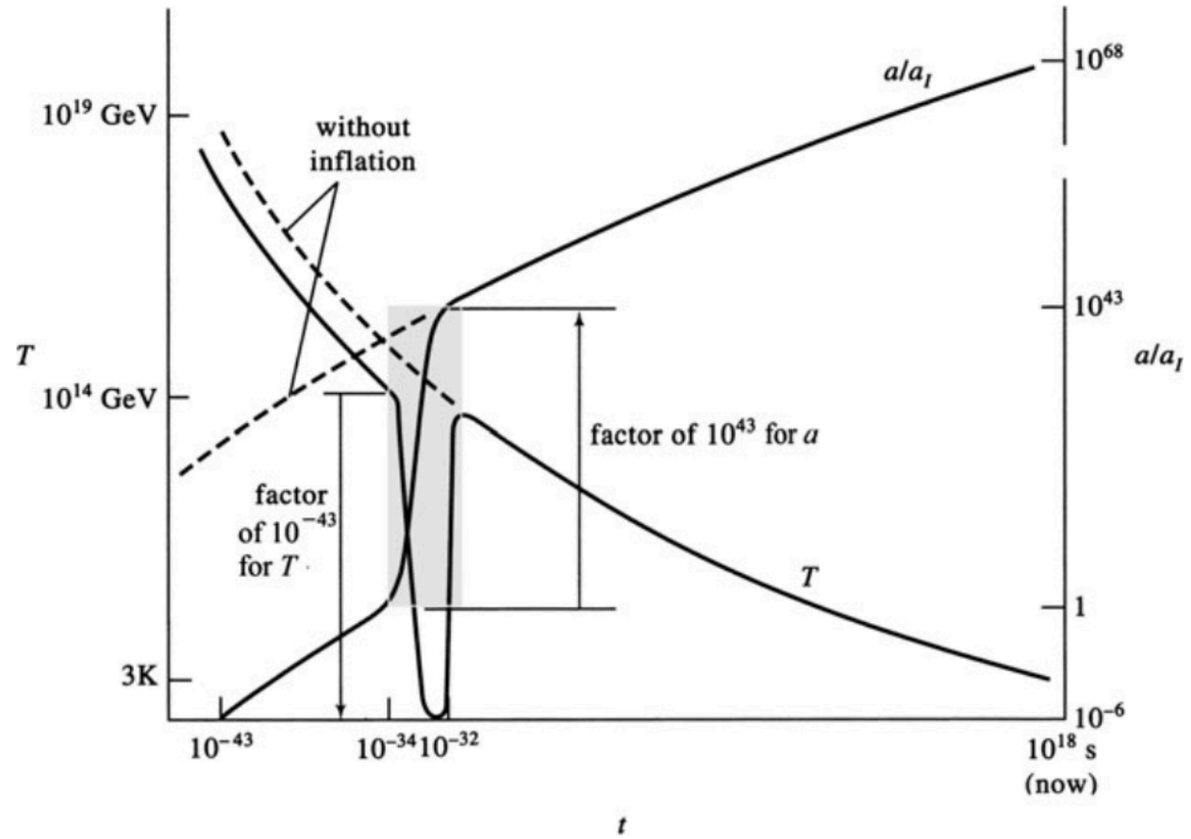


Fig. 10.5 *a* and *T* as a function of time in the inflationary model. The gray band indicates the inflation interval.

The GUT phase transition of the universe can be thought of as analogous to the phase transition of water to ice.

Water is an isotropic substance, whereas ice has preferential directions associated with its crystalline structure – in a geometric sense, it has less symmetry than water.

Thus, the freezing of water into ice is a phase transition with a loss of symmetry, analogous to the GUT phase transition with its "freezing out" of the full symmetry of all interactions.

The phase transition of water, like the GUT phase transition, releases energy, in the form of heat of transformation.

When we freeze water, we usually remove this heat, to allow the phase transition to proceed.

But we could imagine freezing strongly supercooled water, in which case the heat of transformation would simply be redeposited in the ice, raising its temperature to some extent.

This is essentially what happens at the end of GUT phase transition, when the energy released in the freezing out of the symmetry is redeposited in the universe, so as to reheat it to what might be regarded its normal temperature.

The inflationary model makes the definite assertion that Ω_0 is now almost exactly 1.

The high-z supernova data provide an observational test of this assertion, and confirm it.

Another observational test of the model emerges from the study of the primordial density fluctuations that serve as seeds for the formation of the clumps of matter that ultimately grow into galaxies and clusters of galaxies.

In the inflationary model, such density fluctuations in the post-inflationary era can be traced to fluctuations of whatever quantum fields existed in the deSitter spacetime.

Because the initial spectrum of these fluctuations is known from quantum theory, the post-inflationary spectrum can be calculated.

The prediction of the inflationary model is that the resulting spectrum of primordial fluctuations in the gravitational potential is approximately uniform, so $\delta\Phi$ is a constant, independent of the length scale of the fluctuations.

This kind of scale-invariant spectrum is called the Harrison-Zel'dovich spectrum.

The corresponding spectrum of fluctuations in the mass distribution is $\delta M/M \propto M^{-2/3}$.

At the end of inflation, these primordial fluctuations in the gravitational potential and the mass distribution cause perturbations in the baryon-electron-photon plasma, which take the form of sound waves, that is, density and temperature oscillations whose wavelengths and frequencies are related by the characteristic speed of sound waves in an ultrarelativistic gas, $\lambda\nu = 1/\sqrt{3}$, or $\lambda\nu = c/\sqrt{3}$ in cgs units.

Theses sound waves have a broad spectrum of wavelengths, corresponding to the broad spectrum of the primordial quantum fluctuations that created them.

The sound waves continue to oscillate until the decoupling time $t_{dec} \cong 10^{13}$ s, at which point the plasma suddenly becomes transparent and the photons become independent of the baryon-electron plasma.

The photons then cease to participate in the sound-wave oscillations, and they begin to stream freely through the universe at their normal speed *c*.

Some are ultimately captured by our detectors and give us a "snapshot" of the conditions of the universe at t_{dec}.

Thus, the map of small-scale anisotropies detected in the temperature distribution of the cosmic background radiation over the sky (Fig. 9.7 b) is a snapshot of the superposed amplitudes of all the sound waves that existed at t_{dec}.

By a Fourier analysis of the angular distribution of these temperature anisotropies over the sky, we can discover which sound waves (of what wavelength) had large amplitudes at t_{dec} and which had small amplitudes.

Figure 10.6 shows the power spectrum of the observed anisotropies expressed as a function of the multipole index *l* of the spherical harmonics used in the Fourier analysis.

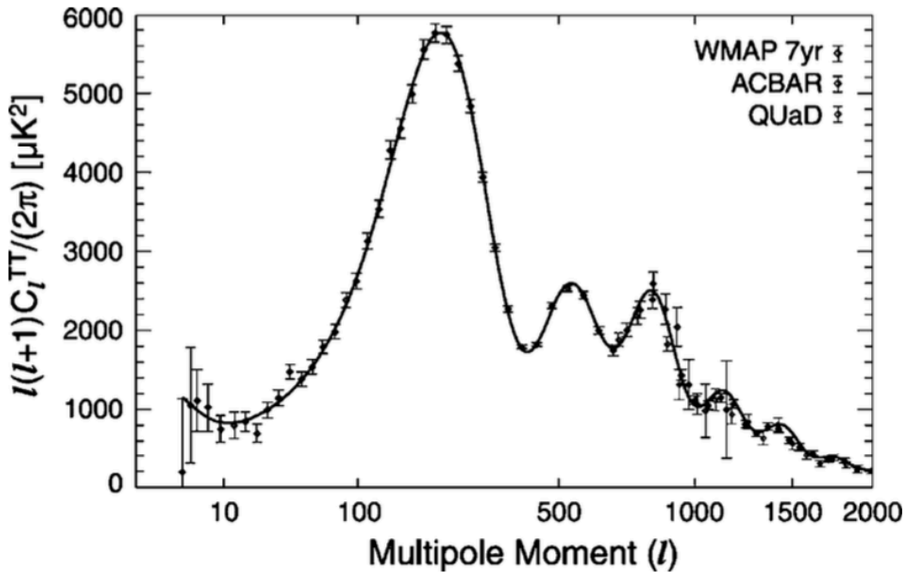


Fig. 10.6 Spectrum of small-scale anisotropies in the cosmic background radiation. In a Fourier analysis by spherical harmonics, the angular size of the anisotropy is characterized by a multipole index *l*. Roughly, a multipole of index *l* corresponds to an anisotropy of an angular size $180^\circ / l$, and the wavelength of the sound wave responsible for this anisotropy is proportional to this angular size [see Eq. (10.56)]. (From Komatsu et al., 2011.)

The figure reveals a series of "acoustic peaks" in the power spectrum, at progressively shorter angular scales.

The strong peak at $l \cong 200$ to 300 indicates a large amplitude of sound waves with an angular scale of about 1°.

The next peaks indicate prominent but somewhat weaker amplitudes for sound waves at smaller angular scales.

We can understand the mechanism that produces this series of acoustic peaks by considering the oscillations of sound waves of various wavelengths and frequencies in the baryon-electron-photon plasma within the time interval from the end of inflation to t_{dec}.

The initial condition set by inflation is that all these Fourier modes start their oscillations at the end of inflation with equal initial phases.

The waves all start at maximum amplitudes, with zero initial time derivatives; this characterizes these waves as standing waves.

The waves then oscillate through successive minima and maxima of amplitude.

A wave whose half-period coincides with t_{dec} will attain maximum negative amplitude at that time; in the power spectrum, this shows up as a peak, the first acoustic peak in Fig. 10.6.

A wave whose full period coincides with t_{dec} will return to maximum positive amplitude at that time; in the power spectrum, this shows up as the second acoustic peak, and so on.

Note that all the acoustic peaks in Fig. 10.6 involve waves of quite long periods.

The first peak is associated with a wave period of $2 \times t_{dec} = 7 \times 10^5$ y, the second peak with a period of $1 \times t_{dec} = 3.5 \times 10^5$ y, and so on.

Correspondingly, the wavelengths are quite large, 7×10^5 y $\times$ c/ $\sqrt{3}$ $= 4 \times 10^5$ ly for the first peak, 2×10^5 ly for the second, and so on.

However the corresponding angular sizes observed on the sky are fairly small.

To convert the wavelength 4×10^5 ly into an observed angular size, we need the angular-diameter distance to the surface of decoupling.

Since decoupling happened early in the history of the universe, the proper distance to this surface now is almost as large as the horizon distance of our universe, $l_0 = 4.5 \times 10^{10}$ ly.

But at the time of emission of the light that we observe now, this surface was closer to us by a factor a/a_0 = 1,100; that is, it was at a distance of 4.5×10^{10} ly/1100 = 4.1×10^7 ly.

This initial distance determines the angular geometry of the light rays emitted at that time, so this initial distance is the angular-size distance d_A.

Hence the linear size 4×10^5 ly corresponds to an observed angular size:

$$\theta \cong \frac{4 \times 10^5 \text{ly}}{d_A} \cong \frac{4 \times 10^5 \text{ly}}{4.1 \times 10^7 \text{ly}} \cong 0.01 \text{ rad} \cong 0.6^\circ \tag{10.56}$$

Accordingly, we expect that the acoustic oscillations of the post-inflation era should produce anisotropies on a scale of 0.6° or smaller in the observed cosmic background radiation.

In the Fourier analysis, an anisotropy of angular scale of 0.6° or smaller corresponds to a multipole index $l = 180^\circ/0.6^\circ \cong 300$ or larger.

The observed spectrum of anisotropies (see Fig. 10.6) shows strong peaks at $l \cong 300$ and at several larger values, in agreement with this estimate.

The detailed features of the observed spectrum of anisotropies depend on various cosmological parameters and can therefore be used to determine these parameters.

For instance, the positions of the acoustic peaks reveal the curvature of the universe and establish that the radius of curvature is much larger than the radius of the visible universe, that is, the universe is nearly flat.

In a positively curved universe, the observed angular scale of the largest acoustically caused anisotropies would be larger than in a flat universe, whereas in a negatively curved universe it would be smaller [in a positively curved universe the angular-size distance d_A is smaller, whereas in a negatively curved universe it is larger; this changes the angle given by Eq. (10.56)].

The 0.6° scale found in the spectrum of multipoles is consistent with the spatially flat universe of the inflationary scenario, but it is not consistent with strongly curved universes.

The anisotropy data tell us that $\Omega_0 = 1$, to within better than $\pm 1\%$.

The relative heights of the acoustic peaks reveal the baryon density and the density of cold dark mass.

Furthermore, the enhancement of baryon density at the first acoustic peak triggers the formation of galaxy clusters at the length scale associated with that peak.

From this absolute dimension of such clusters and the observed angular size, we can determine the distance to the clusters, and the measured redshift then gives us the Hubble constant.

Here is a (partial) list of various cosmological parameters that have been determined by a detailed analysis of the spectrum of anisotropies:

$$\begin{aligned} &\Omega_B = 0.046 \pm 0.003, \quad \Omega_{m,0} = 0.27 \pm 0.01, \quad \Omega_{\Lambda,0} = 0.727 \pm 0.003 \\ &H_0 = 70 \pm 3\,\mathrm{km/(s \cdot Mpc)}, \quad t_0 = (1.38 \pm 0.01) \times 10^{10}\mathrm{y} \end{aligned} \tag{10.57}$$

These results are impressive – the error bars are drastically smaller than for determinations available from supernova Ia data.

Figure 10.7 compares the values of $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ obtained from analysis of the spectrum of anisotropies with the values obtained from recent, improved supernova Ia data.

Although this comparison suggests that the anisotropy analysis is far superior, the limitations of this approach for evaluating the cosmological parameters must be kept in mind.

The values listed in Eq. (10.57) were obtained in the context of a specific model of the universe: inflation, a cosmological term, and cold dark matter.

If a different model had been adopted for the analysis of the observed anisotropy spectrum, the results would have been different.

This model dependence of the analysis raises some questions about the reliability of the results.

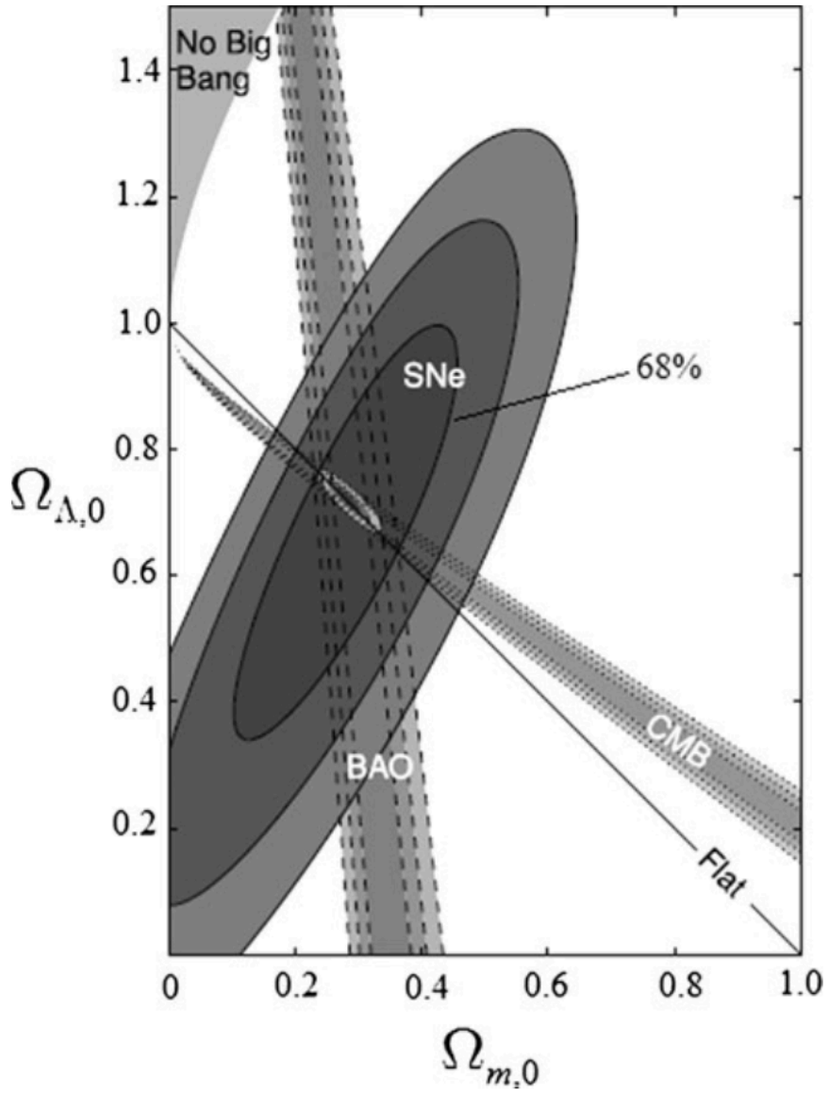


Fig. 10.7 The elliptical contours show the best-fit determination for $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ according to recent, improved supernova Ia data (confidence limits of 68%, 95%, and 99.7%, as in Fig. 9.4). The bands show determinations according to the WMAP data on anisotropies in the cosmic background radiation. The BAO band relies on anisotropies arising from the coupling of the radiation to the acoustic baryon oscillations, and the CMB band relies on anisotropies arising independently from the evolution of the primordial anisotropies of the radiation on its own. Although the new supernova contours give significantly smaller error bars than the contours in Figure 9.4 (note that the scale of Fig. 9.4 is different), the intersection of the BAO and CMB bands gives by far the smallest errors bars. (From Suzuki et al., 2012).

Optimistically, we might regard the excellent agreement between the values obtained here and those obtained in Chapter 9 as evidence in favor of the model.

In particular, the existence of acoustic oscillations hinges on the inflationary scenario, and the observational evidence for acoustic peaks therefore speaks in favor of this scenario.

Besides the small-scale anisotropies arising from acoustic oscillations, there are also larger scale anisotropies that arise from fluctuations of large size that enter their horizons at or after the time of decoupling.

They then evolve only gravitationally, and they can affect the cosmic background radiation only by the perturbation they produce in the gravitational potential.

Wherever the gravitational potential is deeper or shallower, the background photons suffer a relative redshift or a blueshift, and correspondingly, δT will be negative or positive (the Sachs-Wolfe effect, already mentioned in Section 9.5).

Thus, the inflationary model makes definite predictions for the spectrum of the temperature fluctuations in the cosmic background radiation on angular scales of 1° or more.

The inflationary model has proved very successful in explaining the isotropy of the cosmic background radiation as well as the observed small-scale deviations from this isotropy.

But this success is achieved at a cost: The model is carefully contrived to give the answers we want, that is, the model is largely ad hoc, and many of its features lack any fundamental justification.

The model answers many questions about the early universe, but also raises many others and leaves us with a medley of unanswered questions.

Was there a singularity in the primordial universe?

What came before the Big Bang?

Was the universe before the Big Bang a contracting universe?

What were the quantum-gravity effects in the universe before the Planck time?

How does the quantum-mechanical superposition of states of the primordial universe achieve reduction ("collapse") into a definite state?

What determines the amount of energy released in the GUT transition and the magnitude of the cosmological constant that drove the inflation?

What determines the final value of the cosmological constant that is now accelerating our universe?

What determines the amplitude of the quantum fluctuations that survive inflation into later times to trigger growing density perturbations that lead to galaxy formation?

Are there alternatives to inflation?

What determines the prevalence of matter over antimatter in the universe?

What does the dark mass consist of and what determines its abundance?

How much of this is cold dark mass and how much hot?

Although these questions concern the universe on both large and small scales, their answers will have to be found by exploring the realm of physics on a small scale – we urgently need a deeper understanding of the earliest stages of the universe, before and after the Planck time, where we are likely to find a wealth of unanticipated new physics.