PART V

RELATIVISTIC STARS

Wherein the reader, armed
with the magic potions and powers
of Geometrodynamics, conquers the stars.

CHAPTER 23

SPHERICAL STARS

§23.1. PROLOG

Beautiful though gravitation theory may be, it is a sterile subject until it touches the real physical world. Only the hard reality of experiments and of astronomical observations can bring gravitation theory to life. And only by building theoretical models of stars (Part V), of the universe (Part VI), of stellar collapse and black holes (Part VII), of gravitational waves and their sources (Part VIII), and of gravitational experiments (Part IX), can one understand clearly the contacts between gravitation theory and reality.

Preview of the rest of this book

The model-building in this book will follow the tradition of theoretical physics. Each Part (stars, universe, collapse, . . .) will begin with the most oversimplified model conceivable, and will subsequently add only those additional touches of realism necessary to make contact with the least complex of actual physical systems. The result will be a tested intellectual framework, ready to support and organize the additional complexities demanded by greater realism. Greater realism will not be attempted in this book. But the reader seeking it could start in no better place than the two-volume treatise on *Relativistic Astrophysics* by Zel'dovich and Novikov (1971, 1974).

Begin, now, with models for relativistic stars. As a major simplification, insist (initially) that all stars studied be static. Thereby exclude not only exploding and pulsating stars, but even quiescent ones with stationary rotational motions. From the static assumption, plus a demand that the star be made of "perfect fluid" (no shear stresses allowed!), plus Einstein's field equations, it probably follows that the star is spherically symmetric. However, nobody has yet given a proof. [For proofs under more restricted assumptions, see Avez (1964) and Kunzle (1971).] In the absence of a proof, assume the result: insist that all stars studied be spherical as well as static.

Static stars must be spherical

§23.2. COORDINATES AND METRIC FOR A STATIC, SPHERICAL SYSTEM

Metric for any static, spherical system:

To deduce the gravitational field for a static spherical star—or for any other static, spherical system—begin with the metric of special relativity (no gravity) in the spherically symmetric form

$$ds^2 = -dt^2 + dr^2 + r^2\, d\Omega^2, \tag{23.1}$$

where

$$d\Omega^2 = d\theta^2 + \sin^2\theta\, d\phi^2. \tag{23.2}$$

(1) generalized from flat spacetime

Try to modify this metric to allow for curvature due to the gravitational influence of the star, while preserving spherical symmetry. The simplest and most obvious guess is to allow those metric components that are already non-zero in equation (23.1) to assume different values:

$$ds^2 = -e^{2\Phi}\, dt^2 + e^{2\Lambda}\, dr^2 + R^2\, d\Omega^2, \tag{23.3}$$

where Φ, Λ, and R are functions of r only. (The static assumption demands $\partial g_{\mu\nu}/\partial t = 0$.) To verify that this guess is good, use it in constructing stellar models, and check that the resulting models have the same generality (same set of quantities freely specifiable) as in Newtonian theory and as expected from general physical considerations. An apparently more general metric

$$ds^2 = -a^2\, dt^2 - 2ab\, dr\, dt + c^2\, dr^2 + R^2\, d\Omega^2 \tag{23.4}$$

actually is not more general in any physical sense. One can perform a coordinate transformation to a new time coordinate t' defined by

$$e^{\Phi}\, dt' = a\, dt + b\, dr. \tag{23.5}$$

By inserting this in equation (23.4), and by defining $e^{2\Lambda} \equiv b^2 + c^2$, one obtains the postulated line element (23.3), apart from a prime on the t.*

The necessity to allow for arbitrary coordinates in general relativity may appear burdensome when one is formulating the theory; but it gives an added flexibility, something one should always try to turn to one's advantage when formulating and solving problems. The $g_{rt} = 0$ simplification (called a *coordinate condition*) in equation (23.3) results from an advantageous choice of the t coordinate. The r coordinate, however, is also at one's disposal (as long as one chooses it in a way that respects spherical symmetry; thus not $r' = r + \cos\theta$). One can turn this freedom to advantage by introducing a new coordinate $r'(r)$ defined by

(2) specialized to "Schwarzschild form"

$$r' = R(r). \tag{23.6}$$

*Of course, equation (23.5) only succeeds in defining a new time coordinate t' if it is integrable as a differential equation for t'. By choosing the integrating factor e^{Φ} to be just $e^{\Phi} = a(r)$, one sees that $t' = t + \int[b(r)/a(r)]\, dr$ is the integral of (23.5); thus the required t' coordinate always exists, no matter what the functions $a(r)$, $b(r)$, $c(r)$, and $R(r)$ in equation (23.4) may be.

With this choice of the radial coordinate, and with the primes dropped, equation (23.3) reduces to

$$ds^2 = -e^{2\Phi}\,dt^2 + e^{2\Lambda}\,dr^2 + r^2\,d\Omega^2, \tag{23.7}$$

a line element with just two unknown functions, $\Phi(r)$ and $\Lambda(r)$. This coordinate system and metric have been used in most theoretical models for relativistic stars since the pioneering work of Schwarzschild (1916b), Tolman (1939), and Oppenheimer and Volkoff (1939). These particular coordinates are sometimes called "curvature coordinates" and sometimes "Schwarzschild coordinates." The central idea of these coordinates, in a nutshell, is (Schwarzschild r-coordinate) = (proper circumference)$/2\pi$.

For a more rigorous proof that in any static spherical system Schwarzschild coordinates can be introduced, bringing the metric into the simple form (23.7), see Box 23.3 at the end of this chapter.

(3) derived more rigorously

EXERCISE

Exercise 23.1. ISOTROPIC COORDINATES AND NEWTONIAN LIMIT

An alternative set of coordinates sometimes used for static, spherical systems is the "isotropic coordinate system" $(t, \bar{r}, \theta, \phi)$. The metric in isotropic coordinates has the form

$$ds^2 = -e^{2\Phi}\,dt^2 + e^{2\mu}[d\bar{r}^2 + \bar{r}^2\,d\Omega^2], \tag{23.8}$$

with Φ and μ being functions of $\bar{r}$.

(a) Exhibit the coordinate transformation connecting the Schwarzschild coordinates (23.7) to the isotropic coordinates (23.8).

(b) From equation (16.2a) [or equivalently (18.15c)], show that, in the Newtonian limit, the metric coefficient Φ of the isotropic line element becomes the Newtonian potential; and μ becomes equal to $-\Phi$. By combining with part (a), discover that $\Lambda = r\,d\Phi/dr$ in the Newtonian limit.

---

§23.3. PHYSICAL INTERPRETATION OF SCHWARZSCHILD COORDINATES

In general relativity, because the use of arbitrary coordinates is permitted, the physical significance of statements about tensor or vector components and other quantities is not always obvious. There are, however, some situations where the interpretation is almost as straightforward as in special relativity. The most obvious example is the center point of a local inertial coordinate system, where the principle of equivalence allows one to treat all local quantities (quantities not involving spacetime curvature) exactly as in special relativity. Schwarzschild coordinates for a spherical system turn out to be a second example.

One's first reaction when meeting a new metric should be to examine it, not in order to learn about the gravitational field, for which the curvature tensor is more

The form of any metric can reveal the nature of the coordinates being used directly informative, but to learn about the coordinates. (Are they, for instance, locally inertial at some point?)

The names given to the coordinates have no intrinsic significance. A coordinate transformation $t' = \theta$, $r' = \phi$, $\theta' = r$, $\phi' = t$ is perfectly permissible, and has no influence on the physics or the mathematics of a relativistic problem. The only thing it affects is easy communication between the investigator who adopts it and his colleagues. Thus the names $tr\theta\phi$ for the Schwarzschild coordinates (23.7) provide a mnemonic device pointing out the geometric content of the coordinates.* In particular, the names θ, ϕ are justified by the fact that on each two-dimensional surface of constant r and t, the distance between two nearby events is given by $ds^2 = r^2\, d\Omega^2$, as befits standard θ, ϕ coordinates on a sphere of radius r. The area of this two-dimensional sphere is clearly

Geometric significance of the Schwarzschild coordinates:

(1) θ, ϕ are angles on sphere

$$A = \int (r\, d\theta)(r \sin\theta\, d\phi) = 4\pi r^2; \tag{23.9}$$

hence, the metric (23.7) tells how to measure the r coordinate that it employs. One can merely measure (in proper length units) the area A of the sphere, composed of all points rotationally equivalent to the point $\mathscr{P}$ for which the value $r(\mathscr{P})$ is desired; and one can then calculate

(2) r measures surface area of sphere

$$r(\mathscr{P}) = \left(\begin{matrix}\text{proper area of sphere}\\ \text{through point } \mathscr{P}\end{matrix} \Big/ 4\pi\right)^{1/2}. \tag{23.9'}$$

The Schwarzschild coordinates have been picked for convenience, and not for the ease with which one could build a coordinate-measuring machine. This makes it more difficult to design a machine to measure t than machines to measure r, θ, ϕ. The geometric properties of t on which a measuring device can be based are: (1) the time-independent distances ($\partial g_{\alpha\beta}/\partial t = 0$) between world lines of constant r, θ, ϕ; (2) the orthogonality ($g_{tr} = g_{t\theta} = g_{t\phi} = 0$) of these world lines to the $t =$ constant hypersurfaces; and (3) a labeling of these hypersurfaces by Minkowski (special relativistic) coordinate time at spatial infinity, where spacetime becomes flat. This labeling produces a constraint

(3) t has 3 special geometric properties

$$\Phi(\infty) = 0 \tag{23.10}$$

in the metric (23.7). [Mathematically, this constraint is imposed by a simple rescaling transformation $t' = e^{\Phi(\infty)}t$, and by then dropping the prime.]

One "machine" design which constructs (mentally) such a t coordinate, and in the process measures it, is the following. Observers using radar sets arrange to move along the coordinate lines $r, \theta, \phi =$ const. They do this by adjusting their velocities until each finds that the radar echos from his neighbors, or from "benchmark" reference points in the asymptotically flat space, require the same round-trip time at each repetition. Equivalently, each returning echo must show zero doppler shift;

(4) description of a "machine" to measure t

*For an example of misleading names, consider those in the equation

$$ds^2 = -e^{2\Phi(\theta')}\, d\phi'^2 + e^{2\Lambda(\theta')}\, d\theta'^2 + \theta'^2\, (dt'^2 + \sin^2 t'\, dr'^2),$$

which is equivalent to equation (23.7), but employs the coordinates $t' = \theta$, $r' = \phi$, $\theta' = r$, $\phi' = t$.

it must return with the same frequency at which it was sent out. Next a master clock is set up near spatial infinity (far from the star). It is constructed to measure proper time—which, for it, is Minkowski time "at infinity"—and to emit a standard one-Hertz signal. Each observer adjusts the rate of his "coordinate clock" to beat in time with the signals he receives from the master clock. To set the zero of his "coordinate clock," now that its rate is correct, he synchronizes with the master clock, taking account of the coordinate time Δt required for radar signals to travel from the master to him. [To compute the transit time, he assumes that for radar signals $(t_{\text{reflection}} - t_{\text{emission}}) = (t_{\text{return}} - t_{\text{reflection}}) = \Delta t$, so that the echo is obtained by time-inversion about the reflection event. This time-reversal invariance distinguishes the time t in the metric (23.7) from the more general t coordinates allowed by equation (23.4).] Each observer moving along a coordinate line ($r, \theta, \phi =$ const.) now has a clock that measures coordinate time t in his neighborhood.

The above discussion identifies the Schwarzschild coordinates of equation (23.7) by their intrinsic geometric properties. Not only are r and t radial and time variables, respectively (in that $\partial/\partial r$ and $\partial/\partial t$ are spacelike and timelike, respectively, and are orthogonal also to the spheres defined by rotational symmetry), but they have particular properties [$4\pi r^2 =$ surface area; $\partial g_{\mu\nu}/\partial t = 0$; $\partial/\partial r \cdot \partial/\partial t = g_{rt} = 0$; $\partial/\partial t \cdot \partial/\partial t = g_{tt} = -1$ at $r = \infty$] that distinguish them from other possible coordinate choices [$r' = f(r)$, $t' = t + F(r)$]. No claim is made that these are the only coordinates that might reasonably be called r and t; for an alternative choice ("isotropic coordinates"), see exercise 23.1. However, they provide a choice that is reasonable, unambiguous, useful, and often used.

Other coordinates are possible, but Schwarzschild are particularly simple

§23.4. DESCRIPTION OF THE MATTER INSIDE A STAR

To high precision, the matter inside any star is a perfect fluid. (Shear stresses are negligible, and energy transport is negligible on a "hydrodynamic time scale.") Thus, it is reasonable in model building to describe the matter by perfect-fluid parameters:

Material inside star to be idealized as perfect fluid

Parameters describing perfect fluid:
(1) ρ, p, n

$$\begin{aligned} \rho &= \rho(r) = \text{density of mass-energy in rest-frame of fluid;} \\ p &= p(r) = \text{isotropic pressure in rest-frame of fluid;} \\ n &= n(r) = \text{number density of baryons in rest-frame of fluid;} \qquad (23.11) \\ u^\mu &= u^\mu(r) = \text{4-velocity of fluid;} \end{aligned}$$

$$T^{\mu\nu} = (\rho + p)u^\mu u^\nu + p g^{\mu\nu} = \text{stress-energy tensor of fluid.} \qquad (23.12)$$

(For Track-1 discussion, see Box 5.1; for greater Track-2 detail, see §§22.2 and 22.3.) In order that the star be static, each element of fluid must remain always at rest in the static coordinate system; i.e., each element must move along a world line of constant r, θ, ϕ; i.e., each element must have 4-velocity components

(2) $\boldsymbol{u}$

$$u^r = dr/d\tau = 0, \qquad u^\theta = d\theta/d\tau = 0, \qquad u^\phi = d\phi/d\tau = 0. \qquad (23.13a)$$

The normalization of 4-velocity,

$$-1 = \boldsymbol{u} \cdot \boldsymbol{u} = g_{\mu\nu} u^\mu u^\nu = g_{tt} u^t u^t = -e^{2\Phi} u^t u^t,$$

then determines u^t,

$$u^t = dt/d\tau = e^{-\Phi}, \qquad \boldsymbol{u} = e^{-\Phi}\,\partial/\partial t; \tag{23.13b}$$

(3) $\boldsymbol{T}$

and this, together with the general form (23.12) of the stress-energy tensor and the form (23.7) of the metric, determines $T^{\mu\nu}$:

$$\begin{gathered} T^{00} = \rho e^{-2\Phi}, \qquad T^{rr} = pe^{-2\Lambda}, \qquad T^{\theta\theta} = pr^{-2}, \qquad T^{\phi\phi} = pr^{-2}\sin^{-2}\theta, \\ T^{\alpha\beta} = 0 \text{ if } \alpha \neq \beta. \end{gathered} \tag{23.14}$$

Although these components of the stress-energy tensor in Schwarzschild coordinates are useful for calculations, the normalization factors $e^{-2\Phi}$, $e^{-2\Lambda}$, r^{-2}, $r^{-2}\sin^{-2}\theta$ make them inconvenient for physical interpretations. More convenient are components on orthonormal tetrads carried by the fluid elements ("proper reference frames"; see §13.6):

Proper reference frame of fluid

$$\boldsymbol{e}_{\hat{t}} \equiv \frac{d}{d\tau} = \frac{1}{e^\Phi}\frac{\partial}{\partial t}, \qquad \boldsymbol{e}_{\hat{r}} = \frac{1}{e^\Lambda}\frac{\partial}{\partial r}, \qquad \boldsymbol{e}_{\hat{\theta}} = \frac{1}{r}\frac{\partial}{\partial \theta}, \qquad \boldsymbol{e}_{\hat{\phi}} = \frac{1}{r\sin\theta}\frac{\partial}{\partial \phi}; \tag{23.15a}$$

$$\boldsymbol{\omega}^{\hat{t}} = e^\Phi\, \boldsymbol{d}t, \qquad \boldsymbol{\omega}^{\hat{r}} = e^\Lambda\, \boldsymbol{d}r, \qquad \boldsymbol{\omega}^{\hat{\theta}} = r\, \boldsymbol{d}\theta, \qquad \boldsymbol{\omega}^{\hat{\phi}} = r\sin\theta\, \boldsymbol{d}\phi; \tag{23.15b}$$

Components of $\boldsymbol{u}$ and $\boldsymbol{T}$ in proper reference frame

$$\boldsymbol{u} = \boldsymbol{e}_{\hat{t}}; \qquad u^{\hat{t}} = 1, \qquad u^{\hat{r}} = u^{\hat{\theta}} = u^{\hat{\phi}} = 0; \tag{23.15c}$$

$$T_{\hat{t}\hat{t}} \equiv T_{\hat{0}\hat{0}} = \rho, \qquad T_{\hat{r}\hat{r}} = T_{\hat{\theta}\hat{\theta}} = T_{\hat{\phi}\hat{\phi}} = p, \qquad T_{\hat{\alpha}\hat{\beta}} = 0 \text{ if } \alpha \neq \beta. \tag{23.15d}$$

See exercise 23.2 below.

The structure of a star—i.e., the set of functions $\Phi(r)$, $\Lambda(r)$, $\rho(r)$, $p(r)$, $n(r)$—is determined in part by the Einstein field equations, $G^{\mu\nu} = 8\pi T^{\mu\nu}$, and in part by the law of local conservation of energy-momentum in the fluid, $T^{\mu\nu}{}_{;\nu} = 0$. However, these are not sufficient to fix the structure uniquely. Also necessary is the functional dependence of pressure p and density ρ on number density of baryons n:

$$p = p(n), \qquad \rho = \rho(n). \tag{23.16}$$

Equation of state:
(1) in general

Normally one cannot deduce p and ρ from a knowledge solely of n. One must know, in addition, the temperature T or the entropy per baryon s; then the laws of thermodynamics plus equations of state will determine all remaining thermodynamic variables:

$$p = p(n, s), \qquad \rho = \rho(n, s), \ldots.$$

(2) idealized to "one-parameter form" $p = p(n)$, $\rho = \rho(n)$

(See §22.2 and Box 22.1 for full Track-2 discussions.) To pass from the given thermodynamic knowledge, $p(n, s)$ and $\rho(n, s)$, to the desired knowledge, $p(n)$ and $\rho(n)$, one needs information about the star's thermal properties, and especially about the way in which energy generation plus heat flow have conspired to distribute the entropy, $s = s(n)$:

$$p(n) = p[n, s(n)], \qquad \rho(n) = \rho[n, s(n)].$$

There exist three important applications of the theory of relativistic stars: neutron stars, white dwarfs, and supermassive stars (stars with $M \gtrsim 10^3 M_\odot$, which may exist according to theory, but the existence of which has never yet been confirmed by observation). In all three cases, happily, the passage from $p = p(n, s)$, $\rho(n, s)$, to $p = p(n)$, $\rho = \rho(n)$, is trivial.

Justification for idealized equation of state:

(1) in neutron stars

Consider first a neutron star. Though hot by ordinary standards, a neutron star is so cold by any nuclear-matter scale of temperatures that essentially all its thermal degrees of freedom are frozen out ("degenerate gas"; "quantum fluid"). It is not important that a detailed treatment of the substance of a neutron star is beyond the capability of present theory (allowance for the interaction between baryon and baryon; production at sufficiently high pressures of hyperons and mesons). The simple fact is that one is dealing with matter at densities comparable to the density of matter in an atomic nucleus (2×10^{14} g/cm^3) and higher. Everything one knows about nuclear matter [see, for example, Bohr and Mottelson (1969)] tells one that it is degenerate, and that one can estimate in order of magnitude its degeneracy temperature by treating it as though it were an ideal Fermi neutron gas. (In a normal atomic nucleus, a little more than 50 per cent of all baryons are neutrons, the rest are protons; in a neutron star, as many as 99 per cent are neutrons.) When approximating the neutron-star matter as an ideal Fermi neutron gas, one considers the neutrons to occupy free-particle quantum states, with two particles of opposite spin in each occupied state, and a sharp drop from 100 per cent occupancy of quantum states to empty states when the particle energy rises to the level of the "Fermi energy" [for more on such an ideal Fermi gas, see Kittel, Section 19 (1958); or at an introductory level, see Sears, Section 16-5 (1953)]. In matter at nuclear density, the Fermi energy is of the order

$$E_{\text{Fermi}} \sim 30 \text{ MeV or } 3 \times 10^{11} \text{ K};$$

and at higher density the temperature required to unfreeze the degeneracy is even greater. In other words, for matter at and above nuclear densities, already at zero temperature the kinetic energy of the particles (governed by the Pauli exclusion principle and by their Fermi energy) is a primary source of pressure. Nuclear forces make a large correction to this pressure, but for $T \lesssim 30 \text{ MeV} = 3 \times 10^{11}$ K, energies of thermal agitation do not.

A star, in collapsing from a normal state to a neutron-star state (see Chapter 24), emits a huge flux of neutrinos at temperatures $\gtrsim 10^{10}$ K, and thereby cools to $T \ll 3 \times 10^{11}$ K within a few seconds after formation. Consequently, in all neutron stars older than a few seconds one can neglect thermal contributions to the pressure and density; i.e., one can set

$$p(n, s) = p(n, s = 0) = p(n), \qquad \rho(n, s) = \rho(n, s = 0) = \rho(n).$$

(2) in white dwarfs

A white dwarf is similar, except that here electrons rather than neutrons are the source of Fermi gas pressure and degeneracy. Typical white-dwarf temperatures satisfy

$$kT \ll E_{\text{Fermi electrons}};$$

the Fermi kinetic energy (Pauli exclusion principle), and not random kT energy, is primarily responsible for the pressure and energy density; and one can set

$$p(n, s) = p(n, s = 0) = p(n), \qquad \rho(n, s) = \rho(n, s = 0) = \rho(n).$$

(3) in supermassive stars

In a supermassive star (see Chapter 24), the situation is quite different. There temperature and entropy are almost the whole story, so far as pressure and energy density are concerned. However, convection keeps the star stirred up and produces a uniform entropy distribution

$$s = \text{const. independent of radius;}$$

so one can write

$$p(n, s) = p_s(n), \qquad \rho(n, s) = \rho_s(n).$$

functions depending on uniform entropy per baryon, s, in the star

In all three cases—neutron stars, white dwarfs, supermassive stars—one regards the relations $p(n)$ and $\rho(n)$ as "equations of state"; and having specified them, one can calculate the star's structure without further reference to its thermal properties.

EXERCISE

Exercise 23.2. PROPER REFERENCE FRAMES OF FLUID ELEMENTS

(a) Verify that equations (23.15a,b) define an orthonormal tetrad and its dual basis of 1-forms, at each event in spacetime.

(b) Verify that the components of the fluid 4-velocity relative to these tetrads are given by equations (23.15c). Why do these components guarantee that the tetrads form "proper reference frames" for the fluid elements?

(c) Verify equations (23.15d) for the components of the stress-energy tensor.

---

§23.5. EQUATIONS OF STRUCTURE

Five equations needed to determine 5 stellar-structure functions: Φ, Λ, p, ρ, n

The structure of a relativistic star is determined by five functions of radius r: the metric functions $\Phi(r)$, $\Lambda(r)$, the pressure $p(r)$, the density of mass-energy $\rho(r)$, and the number density of baryons, $n(r)$. Hence, to determine the structure uniquely, one needs five equations of structure, plus boundary conditions. Two equations of structure, the equations of state $p(n)$ and $\rho(n)$, are already in hand. The remaining three must be the essential content of the Einstein field equations and of the law of local energy-momentum conservation, $T^{\mu\nu}{}_{;\nu} = 0$.

One knows that the law of local energy-momentum conservation for the fluid follows as an identity from the Einstein field equations. Without loss of information, one can therefore impose all ten field equations and ignore local energy-momentum conservation. But that is an inefficient way to proceed. Almost always the equations $T^{\mu\nu}{}_{;\nu} = 0$ can be reduced to usable form more easily than can the field equations. Hence, the most efficient procedure is to: (1) evaluate the four equations $T^{\mu\nu}{}_{;\nu} = 0$; (2) evaluate enough field equations (six) to obtain a complete set $(6 + 4 = 10)$; and (3) evaluate the remaining four field equations as checks of the results of (1) and (2).

The most efficient procedure for solving Einstein equations

The Track-2 reader has learned (§22.3) that the equations $T^{\mu\nu}{}_{;\nu} = 0$ for a perfect fluid take on an especially simple form when projected (1) on the 4-velocity $\boldsymbol{u}$ of the fluid itself, and (2) orthogonal to $\boldsymbol{u}$. Projection along $\boldsymbol{u}$ ($u_\mu T^{\mu\nu}{}_{;\nu} = 0$) gives the local law of energy conservation (22.11a),

$$\frac{d\rho}{d\tau} = -(\rho + p)\boldsymbol{\nabla}\cdot\boldsymbol{u} = \frac{\rho + p}{n}\frac{dn}{d\tau},$$

where $\boldsymbol{u} = d/d\tau$; i.e., τ is proper time along the world line of any chosen element of the fluid. For a static star, or for any other static system, both sides of this equation must vanish identically (no fluid element ever sees any change in its own density).

Projection of $T^{\mu\nu}{}_{;\nu} = 0$ orthogonal to $\boldsymbol{u}$ gives the reasonable equation

Equation of hydrostatic equilibrium derived

$$\begin{pmatrix}\text{inertial mass}\\ \text{per unit volume}\end{pmatrix} \times (\text{4-acceleration}) = -\begin{pmatrix}\text{pressure gradient, projected}\\ \text{perpendicular to } \boldsymbol{u}\end{pmatrix}$$

i.e.,

$$(\rho + p)\boldsymbol{\nabla}_{\boldsymbol{u}}\boldsymbol{u} = -[\boldsymbol{\nabla}p + (\boldsymbol{\nabla}_{\boldsymbol{u}}p)\boldsymbol{u}].$$

[see equation (22.13)]. When applied to a static star, this equation tells how much pressure gradient is needed to prevent a fluid element from falling. Only the radial component of this equation has content, since the pressure depends only on r. The radial component in the Schwarzschild coordinate system says [see the line element (23.7) and the 4-velocity components (23.13)],

$$\begin{aligned}(\rho + p)u_{r;\nu}u^\nu &= -(\rho + p)\Gamma^\alpha{}_{r\nu}u_\alpha u^\nu = -(\rho + p)\Gamma^0{}_{r0}u_0u^0\\ &= (\rho + p)\Phi_{,r} = -p_{,r}.\end{aligned} \tag{23.17}$$

(Track-1 readers can derive this from scratch at the end of the section, exercise 23.3.) In the Newtonian limit, Φ becomes the Newtonian potential (since $g_{00} = -e^{2\Phi} \approx -1 - 2\Phi$), and the pressure becomes much smaller than the mass-energy density; consequently equation (23.17) becomes

$$\rho\Phi_{,r} = -p_{,r}. \tag{23.17N}$$

This is the Newtonian version of the equation describing the balance between gravitational force and pressure gradient.

The pressure gradient that prevents a fluid element from falling appears in Einstein's theory as the source of an acceleration. This acceleration, by keeping the fluid element at a fixed r value, causes it to depart from geodesic motion (from "fiducial world line"; from motion of free fall into the center of the star). Newtonian

Comparison of Newton and Einstein views of hydrostatic equilibrium theory, on the other hand, views as the fiducial world line the one that stays at a fixed r value. It regards the "gravitational force" as trying (without success, because balanced by the pressure gradient) to pull a particle from a fixed-r world line onto a geodesic world line. In the two theories the magnitudes of the acceleration, whether "actually taking place" (Einstein theory) or "trying to take place" (Newtonian theory), are the same to lowest order (but opposite in direction); so it is no surprise that (23.17) and (23.17N) differ only in detail.

Equation for Λ derived

Turn next to the Einstein field equation. Here, as is often the case, the components of the field equation in the fluid's orthonormal frame [equations (23.15a,b)] are simpler than the components in the coordinate basis. One already knows the stress-energy tensor $T_{\hat{\alpha}\hat{\beta}}$ in the orthonormal frame [equation (23.15d)]; and Track-2 readers have already calculated the Einstein tensor $G_{\hat{\alpha}\hat{\beta}}$ (exercise 14.13; Track-1 readers will face the task at the end of this section, exercise 23.4). All that remains is to equate $G_{\hat{\alpha}\hat{\beta}}$ to $8\pi T_{\hat{\alpha}\hat{\beta}}$. Examine first the $\hat{0}\hat{0}$ component of the field equations:

$$\begin{aligned} G_{\hat{0}\hat{0}} &= r^{-2} - r^{-2}e^{-2\Lambda} - r^{-1}(d/dr)(e^{-2\Lambda}) \\ &= r^{-2}(d/dr)[r(1 - e^{-2\Lambda})] = 8\pi T_{\hat{0}\hat{0}} = 8\pi\rho. \end{aligned}$$

This equation becomes easy to solve as soon as one notices that it is a differential equation *linear* in the quantity $e^{-2\Lambda}$; a bit of tidying up then focuses attention on the quantity $r(1 - e^{-2\Lambda})$. Give this quantity the name $2m(r)$ (so far only a name!); thus,

$$2m \equiv r(1 - e^{-2\Lambda}); \qquad e^{2\Lambda} = (1 - 2m/r)^{-1}. \tag{23.18}$$

In this notation the $\hat{0}\hat{0}$ component of the Einstein tensor becomes

$$G_{\hat{0}\hat{0}} = \frac{2}{r^2}\frac{dm(r)}{dr} = 8\pi\rho.$$

Integrate and find

$$m(r) = \int_0^r 4\pi r^2\rho\, dr + m(0). \tag{23.19}$$

For the constant of integration $m(0)$, a zero value means a space geometry smooth at the origin (physically acceptable); a non-zero value means a geometry with a singularity at the origin (physically unacceptable: no local Lorentz frame at $r = 0$):

$$\begin{aligned} ds^2 &= [1 - 2m(0)/r]^{-1}\, dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \\ &\approx -[r/2m(0)]\, dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \qquad \text{at } r \approx 0 \text{ if } m(0) \neq 0; \\ ds^2 &= [1 - (8\pi/3)\rho_c r^2]^{-1}\, dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \\ &\approx dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \qquad \text{at } r \approx 0 \text{ if } m(0) = 0. \end{aligned} \tag{23.20}$$

"Mass-energy inside radius r," $m(r)$, defined

The quantity $m(r)$, defined by equation (23.18) and calculated from equation (23.19) with $m(0) = 0$, is a relativistic analog of the "mass-energy inside radius r." Box 23.1 spells out the analogy in detail.

Box 23.1 MASS-ENERGY INSIDE RADIUS r

The total mass-energy M of an isolated star is well-defined (Chapter 19). But not well-defined, in general, is the distribution of that mass-energy from point to point inside the star and in its gravitational field (no unique "gravitational stress-energy tensor"). This was the crucial message of §20.4 (Track 2).

The message is true in general. But for the case of a spherical star—and only for that case—the message loses its bite. Spherical symmetry allows one to select a distribution of the total mass-energy that is physically reasonable. In Schwarzschild coordinates, it is defined by

$$\text{“total mass-energy inside radius } r\text{”} \equiv m(r) = \int_0^r 4\pi r^2 \rho \, dr. \tag{1}$$

The fully convincing argument for this definition is found only by considering a generalization of it to time-dependent spherically symmetric stars (pulsating, collapsing, or exploding stars; see Chapters 26 and 32, and especially exercise 32.7). For them one finds that the mass-energy m associated with a given ball of matter (fixed baryon number) can change in time only to the extent that locally measurable energy fluxes can be detected at the boundary of the ball. [Such energy fluxes could be the power expended by pressure forces against the moving boundary surface, or heat fluxes, or radiation (photon or neutrino) fluxes. But since spherically symmetric gravitational waves do not exist (Chapters 35 and 36), neither physical intuition nor Einstein's equations require that problems of localizing gravitational-wave energy be faced.] Thus the energy m is localized, not by a mathematical convention, but by the circumstance that transfer of energy (with this definition of m) is detectable by local measurements. [For the mathematical details of $m(r, t)$ in the time-dependent case, see Misner and Sharp (1964), Misner (1965), and exercise 32.7.]

In addition to the critical "local energy flux" property of $m(r)$ described above, there are three further properties that verify its identification as mass-energy. They are: (1) Everywhere outside the star

$$m(r) = M \equiv \begin{pmatrix}\text{total mass-energy of star as measured from} \\ \text{Kepler's third law for distant planets}\end{pmatrix}; \tag{2}$$

see §23.6 for proof. (2) For a Newtonian star, where "mass inside radius r" has a unique meaning, $m(r)$ is that mass. (3) For a relativistic star, $m(r)$ splits nicely into "rest mass-energy" $m_0(r)$ plus "internal energy" $U(r)$ plus "gravitational potential energy" $\Omega(r)$.

To recognize and appreciate the split

$$m(r) = m_0(r) + U(r) + \Omega(r), \tag{3}$$

proceed as follows. First split the total density of mass-energy, ρ, into a part $\mu_0 n$ due to rest mass—where μ_0 is the average rest mass of the baryonic species pres-

Box 23.1 (*continued*)

ent—and a part $\rho - \mu_0 n$ due to internal thermal energy, compressional energy, etc. Next notice that the proper volume of a shell of thickness dr is

$$d\mathcal{V} = 4\pi r^2(e^{\Lambda}\, dr) = 4\pi r^2(1 - 2m/r)^{-1/2}\, dr, \tag{4}$$

not $4\pi r^2\, dr$. Consequently, the total rest mass inside radius r is

$$m_0 = \int_0^r \mu_0 n\, d\mathcal{V} = \int_0^r 4\pi r^2(1 - 2m/r)^{-1/2}\mu_0 n\, dr, \tag{5}$$

and the total internal energy is

$$U = \int_0^r (\rho - \mu_0 n)\, d\mathcal{V} = \int_0^r 4\pi r^2(1 - 2m/r)^{-1/2}(\rho - \mu_0 n)\, dr. \tag{6}$$

Subtract these from the total mass-energy, m; the quantity that is left must be the gravitational potential energy,

$$\begin{aligned} \Omega &= -\int_0^r \rho[(1 - 2m/r)^{-1/2} - 1]4\pi r^2\, dr \\ &\approx -\int_0^r (\rho m/r)4\pi r^2\, dr. \end{aligned} \tag{7}$$

[Newtonian limit, $m/r \ll 1$]

(See exercise 23.7.)

Equation for Φ derived

Turn next to the $\hat{r}\hat{r}$ component of the field equations:

$$\begin{aligned} G_{\hat{r}\hat{r}} &= -r^{-2} + r^{-2}e^{-2\Lambda} + 2r^{-1}e^{-2\Lambda}\, d\Phi/dr \\ &= 8\pi T_{\hat{r}\hat{r}} = 8\pi p. \end{aligned}$$

Solving this equation for the derivative of Φ, and replacing $e^{-2\Lambda}$ by $1 - 2m/r$, one obtains an expression for the gradient of the potential Φ:

$$\frac{d\Phi}{dr} = \frac{m + 4\pi r^3 p}{r(r - 2m)}. \tag{23.21}$$

This expression reduces to the familiar formula

$$d\Phi/dr = m/r^2 \tag{23.21N}$$

in the Newtonian limit.

In most studies of stellar structure, one replaces equation (23.17) by the equivalent equation obtained with the help of (23.21),

Equation of hydrostatic equilibrium rewritten in "OV" form

$$\frac{dp}{dr} = -\frac{(\rho + p)(m + 4\pi r^3 p)}{r(r - 2m)}. \tag{23.22}$$

This is called the Oppenheimer-Volkoff (OV) equation of hydrostatic equilibrium. Its Newtonian limit,

$$dp/dr = -\rho m/r^2, \tag{23.22N}$$

is familiar.

Compare two stellar models, one relativistic and the other Newtonian. Suppose that at a given radius r [determined in both cases by (proper area) $= 4\pi r^2$], the two configurations have the same values of ρ, p, and m. Then in the relativistic model the pressure gradient is

$$\begin{aligned}\frac{dp}{d(\text{proper radial distance})} &= \frac{dp}{e^{\Lambda}\,dr}\\ &= -\frac{(\rho + p)(m + 4\pi r^3 p)}{r^2(1 - 2m/r)^{1/2}}.\end{aligned} \tag{23.23}$$

In contrast, Newtonian theory gives for the pressure gradient

$$\frac{dp}{d(\text{proper radial distance})} = \frac{dp}{dr} = -\frac{\rho m}{r^2}. \tag{23.23N}$$

Comparison of pressure gradients in Newtonian and relativistic stars

The relativistic expression for the gradient is larger than the Newtonian expression (1) because the numerator is larger (added pressure term in both factors) and (2) because the denominator is smaller [shrinkage factor $(1 - 2m/r)^{1/2}$]. Therefore, as one proceeds deeper into the star, one finds pressure rising faster than Newtonian gravitation theory would predict. Moreover, this rise in pressure is in a certain sense "self-regenerative." The more the pressure goes up, the larger the pressure-correction terms become in the numerator of (23.23); and the larger these terms become, the faster is the further rise of the pressure as one probes still deeper into the star. The geometric factor $[1 - 2m(r)/r]^{1/2}$ in the denominator of (23.23) further augments this regenerative rise of pressure towards the center. It is appropriate to summarize the situation in short-hand terms by saying that general relativity predicts stronger gravitational forces in a stationary body than does Newtonian theory. These forces, among their other important effects, can pull certain white-dwarf stars and supermassive stars into gravitational collapse under circumstances (see Chapter 24) where Newtonian theory would have predicted stable hydrostatic equilibrium. As the most elementary indication that a new factor has surfaced in the analysis of stability, note that no star in hydrostatic equilibrium can ever have $2m(r)/r \geq 1$ (see Box 23.2 for one illustration and §23.8 for discussion), a phenomenon alien to Newtonian theory.

Equations of stellar structure summarized

Now in hand are five equations of structure [two equations of state (23.16); equation (23.19), expressing $m(r) = \frac{1}{2}r(1 - e^{-2\Lambda})$ as a volume integral of ρ; the source equation (23.21) for Φ; and the OV equation of hydrostatic equilibrium (23.22)] for the five structure functions $\rho, p, n, \Phi, \Lambda$. If the theory of relativistic stars as outlined above is well posed, then each of the remaining eight Einstein field equations $G_{\hat{\alpha}\hat{\beta}} = 8\pi T_{\hat{\alpha}\hat{\beta}}$ must be either vacuous ("0 = 0"), or must be a consequence of the five equations of structure. This is, indeed, the case, as one can verify by straightforward but tedious computations.

To construct a stellar model, one needs boundary conditions as well as structure equations. To facilitate the presentation of boundary conditions, the next section will examine the star's external gravitational field.

EXERCISES

Exercise 23.3. LAW OF LOCAL ENERGY-MOMENTUM CONSERVATION (for readers who have not studied Chapter 22)

Evaluate the four components of the equation $T^{\alpha\beta}{}_{;\beta} = 0$ for the stress-energy tensor (23.14) in the Schwarzschild coordinate system of equation (23.7). [Answer: only $T^{r\beta}{}_{;\beta} = 0$ gives a nonvacuous result; it gives equation (23.17).]

Exercise 23.4. EINSTEIN CURVATURE TENSOR (for readers who have not studied Chapter 14)

Calculate the components of the Einstein curvature tensor, $G_{\alpha\beta}$, in Schwarzschild coordinates. Then perform a transformation to obtain $G_{\hat{\alpha}\hat{\beta}}$, the components in the orthonormal frame of equations (23.15a,b). [See Box 8.6, or Box 14.2 and equation (14.7).]

Exercise 23.5. TOTAL NUMBER OF BARYONS IN A STAR

Show that, if $r = R$ is the location of the surface of a static star, then the total number of baryons inside the star is

$$A = \int_0^R 4\pi r^2 n e^{\Lambda}\, dr. \tag{23.24}$$

[Hint: See the discussion of m_0 in Box 23.1.]

Exercise 23.6. BUOYANT FORCE IN A STAR

An observer at rest at some point inside a relativistic star measures the radial pressure-buoyant force, F_{buoy}, on a small fluid element of volume V. Let him use the usual laboratory techniques. Do not confuse him by telling him he is in a relativistic star. What value will he find for F_{buoy}, in terms of ρ, p, m, V, and dp/dr? If he equates this buoyant force to an equal and opposite gravitational force, F_{grav}, what will F_{grav} be in terms of ρ, p, m, V, and r? (Use equation 23.22.) How do these results differ from the corresponding Newtonian results?

Exercise 23.7. GRAVITATIONAL ENERGY OF A NEWTONIAN STAR

Calculate in Newtonian theory the energy one would gain from gravity if one were to construct a star by adding one spherical shell of matter on top of another, working from the inside outward. Use Laplace's equation $(r^2\Phi_{,r})_{,r} = 4\pi r^2\rho$ and the equation of hydrostatic equilibrium $p_{,r} = -\rho\Phi_{,r}$ to put the answer in the following equivalent forms:

$$
\begin{aligned}
\text{(energy gained from gravity)} &\equiv -(\text{gravitational potential energy}) \\
&= \int_0^R (\rho r \Phi_{,r}) 4\pi r^2\, dr = \int_0^R (\rho m/r) 4\pi r^2\, dr \\
&= -\frac{1}{2}\int_0^R (\rho\Phi) 4\pi r^2\, dr = \frac{1}{8\pi}\int_0^\infty (\Phi_{,r})^2 4\pi r^2\, dr \\
&= 3\int_0^R 4\pi r^2 p\, dr.
\end{aligned}
$$

---

§23.6. EXTERNAL GRAVITATIONAL FIELD

Outside a star the density and pressure vanish, so only the metric parameters Φ and $\Lambda = -\frac{1}{2}\ln(1 - 2m/r)$ need be considered. From equation (23.19) one sees that "the mass inside radius r," $m(r)$, stays constant for values of r greater than R (outside the star). Its constant value is denoted by M:

$$m(r) = M \qquad \text{for } r > R \text{ (i.e., outside the star).} \tag{23.25}$$

By integrating equation (23.21) with $p = 0$ and $m = M$, and by imposing the boundary condition (23.10) on Φ at $r = \infty$ ("normalization of scale of time at $r = \infty$"), one finds

$$\Phi(r) = \frac{1}{2}\ln(1 - 2M/r) \qquad \text{for } r > R. \tag{23.26}$$

Consequently, outside the star the spacetime geometry (23.7) becomes

Spacetime outside star possesses "Schwarzschild" geometry

$$ds^2 = -\left(1 - \frac{2M}{r}\right)dt^2 + \frac{dr^2}{(1 - 2M/r)} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{23.27}$$

This is called the "Schwarzschild geometry" or "Schwarzschild gratitational field" or "Schwarzschild line element," because Karl Schwarzschild (1916a) discovered it as an exact solution to Einstein's field equations a few months after Einstein formulated general relativity theory.

In that region of spacetime, $r \gg 2M$, where the geometry is nearly flat, Newton's theory of gravity is valid, and the Newtonian potential is

$$\Phi = -M/r \qquad \text{for } r > R,\ r \gg 2M. \tag{23.26N}$$

Total mass-energy of star

Consequently, M is the mass that governs the Keplerian motions of planets in the distant, Newtonian gravitational field—i.e., it is the star's "total mass-energy" (see Chapters 19 and 20). Since the metric (23.27) far outside the star is precisely diagonal ($g_{tj} \equiv 0$), the star's total angular momentum must vanish. This result accords with the absence of internal fluid motions.

§23.7. HOW TO CONSTRUCT A STELLAR MODEL

Equations of stellar structure collected together

The equations of stellar structure (23.16), (23.19), (23.21), (23.22), and associated boundary conditions (to be discussed below), all gathered together along with the line element, read as follows.

Line Element

$$ds^2 = -e^{2\Phi}\,dt^2 + \frac{dr^2}{1 - 2m/r} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{23.27$'$}$$

$$= -\left(1 - \frac{2M}{r}\right)dt^2 + \frac{dr^2}{1 - 2M/r} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \quad \text{for } r > R.$$

Mass Equation

$$m = \int_0^r 4\pi r^2 \rho\, dr, \text{ with } m(r = 0) = 0. \tag{23.28a}$$

OV Equation of Hydrostatic Equilibrium

$$\frac{dp}{dr} = -\frac{(\rho + p)(m + 4\pi r^3 p)}{r(r - 2m)}, \text{ with } p(r = 0) = p_c = \text{central pressure.} \tag{23.28b}$$

Equations of State

$$p = p(n), \tag{23.28c}$$

$$\rho = \rho(n). \tag{23.28d}$$

Source Equation for Φ

$$\frac{d\Phi}{dr} = \frac{(m + 4\pi r^3 p)}{r(r - 2m)}, \qquad \text{with } \Phi(r = R) = \frac{1}{2}\ln(1 - 2M/R). \tag{23.28e}$$

How to solve the equations of stellar structure

To construct a stellar model one can proceed as follows. First specify the equations of state (23.28c,d) and a value of the central pressure, p_c. Also specify an arbitrary (later to be renormalized) value, Φ_0, for $\Phi(r = 0)$. The boundary conditions $p(r = 0) = p_c$, $\Phi(r = 0) = \Phi_0$, $m(r = 0) = 0$ are sufficient to determine uniquely the solution to the coupled equations (23.28). Integrate these coupled equations outward from $r = 0$ until the pressure vanishes. [The OV equation, (23.28b), guarantees that the pressure will decrease monotonically so long as the equations of state obey the reasonable restriction $\rho \geq 0$ for all $p \geq 0$.] The point at which the pressure reaches zero is the star's surface; the value of r there is the star's radius, R; and the value of m there is the star's total mass-energy, M. Having reached the surface, renormalize Φ by adding a constant to it everywhere, so that it obeys the boundary condition (23.28e). The result is a relativistic stellar model whose structure functions Φ, m, ρ, p, n satisfy the equations of structure.

Notice that for any fixed choice of the equations of state $p = p(n)$, $\rho = \rho(n)$, the stellar models form a one-parameter sequence (parameter p_c). Once the central pressure has been specified, the model is determined uniquely.

The next chapter describes a variety of realistic stellar models constructed numerically by the above prescription. For an idealized stellar model constructed analytically, see Box 23.2.

Exercise 23.8. NEWTONIAN STARS OF UNIFORM DENSITY **EXERCISE**

Calculate the structures of uniform-density configurations in Newtonian theory. Show that the relativistic configurations of Box 23.2 become identical to the Newtonian configurations in the weak-gravity limit. Also show that there are no mass or radius limits in Newtonian theory.

(*continued on page 612*)

Box 23.2 RELATIVISTIC MODEL STAR OF UNIFORM DENSITY

For realistic equations of state (see next chapter), the equations of stellar structure (23.28) cannot be integrated analytically; numerical integration is necessary. However, analytic solutions exist for various idealized and *ad hoc* equations of state. One of the most useful analytic solutions [Karl Schwarzschild (1916b)] describes a star of uniform density,

$$\rho = \rho_0 = \text{constant for all } p. \tag{1}$$

It is not necessary to indulge in the fiction of "an incompressible fluid" to accept this model as interesting. Incompressibility would imply a speed of sound, $v = (dp/d\rho)^{1/2}$, of unlimited magnitude, therefore in excess of the speed of light, and therefore in contradiction with a central principle of special relativity ("principle of causality") that no physical effect can be propagated at a speed $v > 1$. (If a source could cause an effect so quickly in one local Lorentz frame, then there would exist another local Lorentz frame in which the effect would occur before the source had acted!) However, that the part of the fluid in the region of high pressure has the same density as the part of the fluid in the region of low pressure is an idea easy to admit, if only one thinks of the fluid having a composition that varies from one

r value to another ("hand-tailored"). Whether one thinks along this line, or simply has in mind a globe of water limited in size to a small fraction of the dimensions of the earth, one has in Schwarzschild's model an instructive example of hydrostatics done in the framework of Einstein's theory.

The mass equation (23.28a) gives immediately

$$m = \begin{Bmatrix} (4\pi/3)\rho_0 r^3 & \text{for } r < R \\ M = (4\pi/3)\rho_0 R^3 & \text{for } r > R \end{Bmatrix}. \tag{2}$$

from which follows the length-correction factor in the metric

$$\frac{d(\text{proper distance})}{dr} = e^{\Lambda} = [1 - 2m(r)/r]^{-1/2}. \tag{3}$$

When for ease of visualization the space geometry (r, ϕ) of an equatorial slice through the star is viewed as embedded in a Euclidean 3-geometry (z, r, ϕ) [see §23.8], the "lift" out of the plane $z = 0$ is

$$z(r) = \begin{cases} (R^3/2M)^{1/2}[1 - (1 - 2Mr^2/R^3)^{1/2}] & \text{for } r \le R, \\ (R^3/2M)^{1/2}[1 - (1 - 2M/R)^{1/2}] + [8M(r - 2M)]^{1/2} - [8M(R - 2M)]^{1/2} & \\ \qquad \text{for } r \ge R. \end{cases} \tag{4}$$

The knowledge of $m(r)$ from (2) allows the equation of hydrostatic equilibrium (23.28b) to be integrated to give the pressure:

$$p = \rho_0 \left\{ \frac{(1 - 2Mr^2/R^3)^{1/2} - (1 - 2M/R)^{1/2}}{3(1 - 2M/R)^{1/2} - (1 - 2Mr^2/R^3)^{1/2}} \right\} \text{ for } r < R. \tag{5}$$

The pressure in turn leads via (23.28e) to the time-correction factor in the metric.

$$\frac{d(\text{proper time})}{dt} = e^{\Phi} = \begin{Bmatrix} \frac{3}{2}\left(1 - \frac{2M}{R}\right)^{1/2} - \frac{1}{2}\left(1 - \frac{2Mr^2}{R^3}\right)^{1/2} & \text{for } r < R \\ (1 - 2M/r)^{1/2} & \text{for } r > R \end{Bmatrix}. \tag{6}$$

Several features of these uniform-density configurations are noteworthy. (1) For fixed energy density, ρ_0, the central pressure

$$p_c = \rho_0 \left\{ \frac{1 - (1 - 2M/R)^{1/2}}{3(1 - 2M/R)^{1/2} - 1} \right\}, \tag{7}$$

increases monotonically as the radius, R, increases—and, hence, also as the mass, $M = (4\pi/3)\rho_0 R^3$, and the ratio ("strength of gravity")

$$2M/R = (8\pi/3)\rho_0 R^2 \tag{8}$$

increase. This is natural, since, as more and more matter is added to the star, a greater and greater pressure is required to support it. (2) The central pressure becomes infinite when M, R, and $2M/R$ reach the limiting values

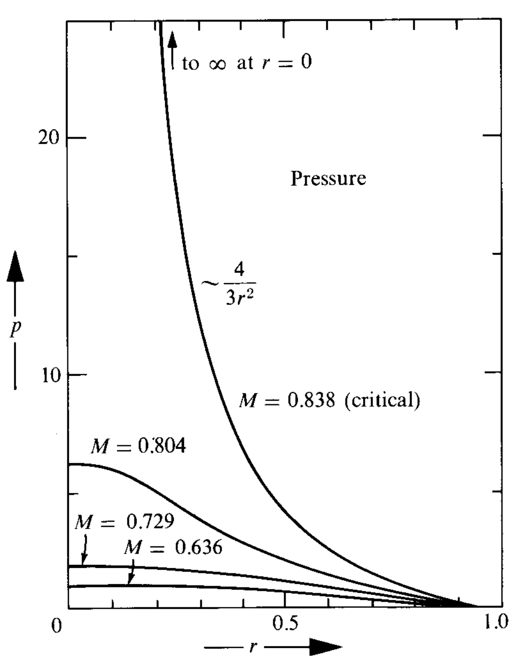
$$R_{\rm lim} = (9/4)M_{\rm lim} = (3\pi\rho_0)^{-1/2}, \tag{9}$$

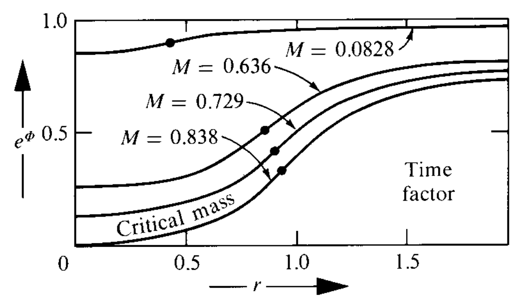
$$(2M/R)_{\rm lim} = 8/9. \tag{10}$$

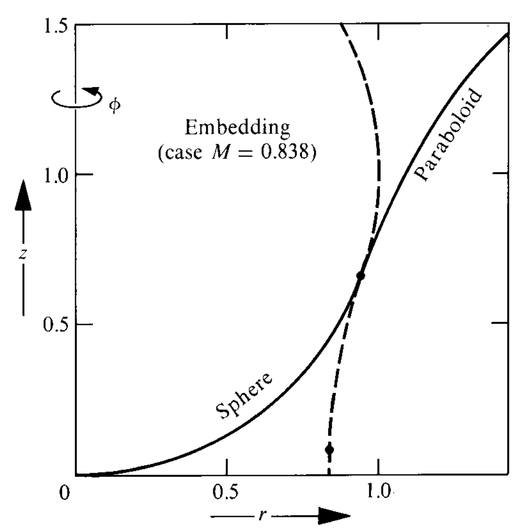
No star of uniform density can have a mass and radius exceeding these limits. These limits are purely relativistic phenomena; no such limits occur in Newtonian theory. (3) Inside the star the space geometry (geometry of a hypersurface $t = constant$) is that of a three-dimensional spherical surface with radius of curvature

$$a = (3/8\pi\rho_0)^{1/2}. \tag{11}$$

[See equation (4), above.] Outside the star the (Schwarzschild) space geometry is that of a three-dimensional paraboloid of revolution. The interior and exterior geometries join together smoothly. All these details are shown in the following three diagrams. There all quantities are given in the following geometric units (to convert mass in g or density in g/cm^3 into mass in cm or density in cm^{-2}, multiply by 0.742×10^{-28} cm/g): lengths, in units $(3/8\pi\rho_0)^{1/2}$; pressure, in units ρ_0; mass, in units $(3/32\pi\rho_0)^{1/2}$.







Box 23.2 (*continued*)

The mass "after assembly" is what is called M. The mass of the same fluid, dispersed in droplets at infinite separation, is called M_{before} in the following table.

M_{before}	small	0.0882	0.894	1.0913	1.374
M	small	0.0828	0.636	0.729	0.838 (critical)
Difference (binding):	$\frac{3}{10}M^{5/3}$	0.0054	0.258	0.362	0.536

§23.8. THE SPACETIME GEOMETRY FOR A STATIC STAR

Surface area of spheres, $4\pi r^2$:

(1) increases monotonically from center of star outward

For a highly relativistic star, the spacetime geometry departs strongly from Euclid-Lorentz flatness. Consequently, there is no *a priori* reason to expect that the surface area $4\pi r^2$, and hence also the radial coordinate r, will increase monotonically as one moves from the center of the star outward. Fortunately, *the equations of stellar structure guarantee that* r *will increase monotonically from* 0 *at the star's center to* ∞ *at an infinite distance away from the star,* so long as $\rho \geq 0$ and so long as the star is static (equilibrium).

The monotonicity of r can be seen as follows. Introduce as a new radial coordinate proper distance, ℓ, from the center of the star. By virtue of expression (23.27′) for the metric, ℓ and r are related by

$$dr = \pm(1 - 2m/r)^{1/2}\, d\ell. \tag{23.29}$$

Note that r is zero at the center of the star (where $m \propto r^3$), and note that r is always nonnegative by definition. Therefore r must at first increase with ℓ as one moves outward from $\ell = 0$; $r(\ell)$ can later reach a maximum and start decreasing only at a point where $2m/r$ becomes unity [see equation (23.29)]. Such a behavior can and does happen in a closed model universe, a 3-sphere of uniform density and radius a, where

$$r(\ell) = a \sin(\ell/a)$$

[see Chapter 27; especially the embedding diagram of Box 27.2(A)]. However, the field equations demand that such a system be dynamic. Here, on the contrary, attention is limited to a system where conditions are static. In such a system, the condition of hydrostatic equilibrium (23.28b) applies. Then the pressure gradient is given by an expression with the factor $[1 - 2m(r)/r]$ in its denominator. If $2m/r$ approaches unity with increasing ℓ in some region of the star, the pressure gradient there becomes so large that one comes to the point $p = 0$ (surface of the star) before one comes to any point where $2m(r)/r$ might attain unit value. Moreover, after the surface of the star is passed, m remains constant, $m(r) = M$, and $2m(r)/r$ decreases. Consequently, *$2m/r$ is always less than unity;* and $r(\ell)$ cannot have a maximum, Q.E.D. (Details of the proof are left to the reader as exercise 23.9.)

Although the radii of curvature, r, and corresponding spherical surface areas, $4\pi r^2$, increase monotonically from the center of a star outward, they do not increase at the same rate as they would in flat spacetime. In flat spacetime the rate of increase is given by dr/d(proper radial distance) $= dr/d\ell = 1$. In a star it is given by $dr/d\ell = (1 - 2m/r)^{1/2} < 1$. Consequently, if one were to climb a long ladder outward from the center of a relativistic star, measuring for each successive spherical shell its Schwarzschild r-value ("proper circumference"/2π), one would find these r-values to increase surprisingly slowly.

(2) but increases more slowly than in flat spacetime

This strange behavior is most easily visualized by means of an "embedding diagram." It would be too much for any easy visualization if one were to attempt to embed the whole curved four-dimensional manifold in some higher-dimensional flat space. [See, however, Fronsdal (1959) and Clarke (1970) for a global embedding in 5 + 1 dimensions, and Kasner (1921b) for a local embedding in 4 + 2 dimensions. One can never embed a non-flat, vacuum metric ($G_{\mu\nu} = 0$) in a flat space of 5 dimensions (Kasner, 1921c).] Therefore seek a simpler picture (Flamm 1916). Space at one time in the context of a static system has the same 3-geometry as space at another time. Therefore, depict 3-space only as it is at one time, $t = constant$. Moreover, at any one time the space itself has spherical symmetry. Consequently, one slice through the center, $r = 0$, that divides the space symmetrically into two halves (for example, the equatorial slice, $\theta = \pi/2$) has the same 2-geometry as any other such slice (any selected angle of tilt, at any azimuth) through the center. Therefore limit attention to the 2-geometry of the equatorial slice. The geometry on this slice is described by the line element

Embedding of spacetime in a flat space of higher dimensionality

$$ds^2 = [1 - 2m(r)/r]^{-1}\, dr^2 + r^2\, d\phi^2. \tag{23.30}$$

Now one may embed this two-dimensional curved-space geometry in the flat geometry of a Euclidean three-dimensional manifold.

If the curvature of the two-dimensional slice is zero or negligible, the embedding is trivial. In this event, identify the 2-geometry with the slice $z = 0$ of the Euclidean 3-space. Moreover, introduce into that 3-space the familiar cylindrical coordinates z, r, ϕ, that one employs for any problem with axial symmetry (see Fig. 23.1 and Box 23.2 for more detail). Then one recognizes the flat two-dimensional slice as the set of points of the Euclidean space with $z = 0$, with ϕ running from 0 to 2π, and r from 0 to ∞. One has identified the r and ϕ of the slice with the r and ϕ of the Euclidean 3-space.

Construction of "embedding diagram" for equatorial slice through star

If the 2-geometry is curved, as it is when the equatorial section is taken through a real star, then maintain the identification between the r, ϕ, of the slice and the r, ϕ, of the Euclidean 3-geometry, but bend up the slice out of the plane $z = 0$ (except at the origin, $r = 0$). At the same time, insist that the bending be axially symmetric. In other words, require that the amount of the "lift" above the plane $z = 0$ shall

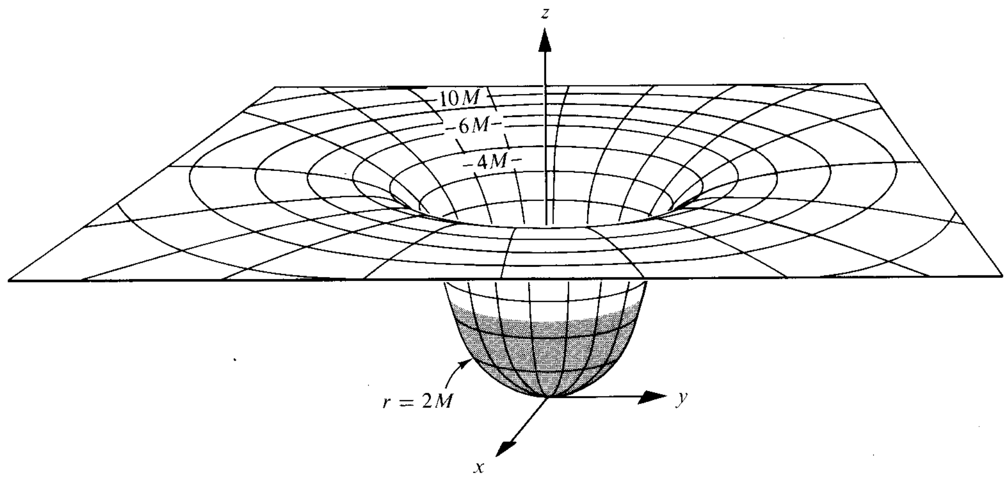


Figure 23.1.
Geometry within (grey) and around (white) a star of radius $R = 2.66M$, schematically displayed. The star is in hydrostatic equilibrium and has zero angular momentum (spherical symmetry). The two-dimensional geometry

$$ds^2 = [1 - 2m(r)/r]^{-1}\, dr^2 + r^2\, d\phi^2$$

of an equatorial slice through the star ($\theta = \pi/2$, $t =$ constant) is represented as embedded in Euclidean 3-space, in such a way that distances between any two nearby points (r, ϕ) and $(r + dr, \phi + d\phi)$ are correctly reproduced. Distances measured off the curved surface have no physical meaning; points off that surface have no physical meaning; and the Euclidean 3-space itself has no physical meaning. Only the curved 2-geometry has meaning. A circle of Schwarzschild coordinate radius r has proper circumference $2\pi r$ (attention limited to equatorial plane of star, $\theta = \pi/2$). Replace this circle by a sphere of proper area $4\pi r^2$, similarly for all the other circles, in order to visualize the entire 3-geometry in and around the star at any chosen moment of Schwarzschild coordinate time t. The factor $[1 - 2m(r)/r]^{-1}$ develops no singularity as r decreases within $r = 2M$, because $m(r)$ decreases sufficiently fast with decreasing r.

be independent of ϕ, whatever may be its dependence on r. Thus the whole story of the embedding is summarized by the single function, the lift,

$$z = z(r) \text{ ("embedding formula").}$$

The geometry on this curved two-dimensional locus in Euclidean space (a made-up 3-space; it has nothing whatever to do with the real world) is to be identical with the geometry of the two-dimensional equatorial slice through the actual star; in other words, the line elements in the two cases are to be identical. To work out this requirement in mathematical terms, write the line element in three-dimensional Euclidean space in the form

$$ds^2 = dz^2 + dr^2 + r^2\, d\phi^2. \tag{23.31}$$

Restrict to the chosen locus ("lifted surface") by writing $z = z(r)$ or $dz = (dz/dr)\, dr$. Thus have

$$ds^2 = \left[1 + \left(\frac{dz(r)}{dr}\right)^2\right] dr^2 + r^2\, d\phi^2 \tag{23.32}$$

on the two-dimensional locus in the 3-geometry, to be identified with

$$ds^2 = [1 - 2m(r)/r]^{-1}\, dr^2 + r^2\, d\phi^2$$

in the actual star. Compare and conclude

$$\left(\frac{dz(r)}{dr}\right)^2 + 1 = [1 - 2m(r)/r]^{-1}. \tag{23.33}$$

This equation is information enough to find the lift as a function of r; thus,

$$z(r) = \int_0^r \frac{dr}{\left[\dfrac{r}{2m(r)} - 1\right]^{1/2}} \quad \text{everywhere,} \tag{23.34a}$$

$$z(r) = [8M(r - 2M)]^{1/2} + \text{constant} \qquad \text{outside the star.} \tag{23.34b}$$

Description of embedded surface

Outside the star this embedded surface is a segment of a paraboloid of revolution. Its form inside the star depends on how the mass, m, varies as a function of r. Recall that $m(r)$ varies as $(4\pi/3)\rho_c r^3$ near the center of the star. Conclude that the embedded surface there looks like a segment of a sphere of radius $a = (3/8\pi\rho_c)^{1/2}$; thus,

$$[a - z(r)]^2 + r^2 = a^2 \qquad \text{for } r \ll a = (3/8\pi\rho_c)^{1/2}. \tag{23.34c}$$

In the special case of a star with uniform density (Box 23.2), the entire interior is of the spherical form (23.34c); in the general case it is not. In all cases, because $r > 2m(r)$, equation (23.34a) produces a surface with z and r as monotonically increasing functions of each other. This means that the embedded surface always opens upward and outward like a bowl; it always looks qualitatively like Figure 23.1; it never has a neck, and it never flattens out except asymptotically at $r = \infty$. At the star's surface, even though the density may drop discontinuously to zero (ρ finite inside when $p = 0$; ρ zero outside), the interior and exterior geometries will join together smoothly [dz/dr, as given by equation (23.33), is continuous].

It must be emphasized that only points lying on the embedded 2-surface have physical significance so far as the stellar geometry is concerned: the three-dimensional regions inside and outside the bowl of Figure 23.1 are physically meaningless. So is the Euclidean embedding space. It merely permits one to visualize the geometry of space around the star in a convenient manner.

EXERCISES

Exercise 23.9. GOOD BEHAVIOR OF r

Carry out explicitly the full details of the proof, at the beginning of this section, that $2m/r$ is always less than unity and r is a monotonic function of ℓ.

Exercise 23.10. CENTER OF STAR OCCUPIED BY IDEAL FERMI GAS AT EXTREME RELATIVISTIC LIMIT

Opposite to the idealization of a star built from an incompressible fluid is the idealization in which it is built from an ideal Fermi gas [ideal neutron star; see Oppenheimer and Volkoff (1939)] at zero temperature, so highly compressed that the particles have relativistic energies, in comparison with which any rest mass they possess is negligible. In this limit, with two particles per occupied cell of volume h^3 in phase space, one has

$$\begin{pmatrix}\text{number density}\\ \text{of fermions}\end{pmatrix} = n = (2/h^3)4\pi \int_0^{p_F} p^2\,dp = 8\pi {p_F}^3/3h^3,$$

$$\begin{pmatrix}\text{density of}\\ \text{mass-energy}\end{pmatrix} = \rho = (2/h^3)4\pi \int_0^{p_F} cp\cdot p^2\,dp = 2\pi c{p_F}^4/h^3,$$

and finally

$$p = -\frac{d\begin{pmatrix}\text{energy}\\ \text{per particle}\end{pmatrix}}{d\begin{pmatrix}\text{volume}\\ \text{per particle}\end{pmatrix}} = -\frac{d(\rho/n)}{d(1/n)} = 2\pi c{p_F}^4/3h^3 = \rho/3,$$

as if one were dealing with radiation instead of particles (p_F = Fermi momentum; momentum of highest occupied state).

Box 23.3 RIGOROUS DERIVATION OF THE SPHERICALLY SYMMETRIC LINE ELEMENT

Section 23.2 gave a heuristic derivation of the general spherically symmetric line element (23.7). This box attempts a more rigorous derivation, applicable to nonstatic systems, as well as static ones.

Begin with a manifold M^4 on which a metric ds^2 of Lorentz signature is defined. Assume M^4 to be spherically symmetric in the sense that to any 3×3 rotation matrix A there corresponds a mapping (rotation) of M^4, also called A ($A: M^4 \longrightarrow M^4: \mathscr{P} \longrightarrow A\mathscr{P}$), that preserves the lengths of all curves. Further assumptions and constructions will be numbered (i), (ii), etc., so one can see what specializations are needed to get to the line element (23.7). Daggers (†) indicate assumptions that are found inapplicable to some other physically interesting situations.

For any point $\mathscr{P}$, form the set $s = \mathcal{S}(\mathscr{P}) = \{A\mathscr{P} \in M^4 | A \in SO(3)\}$ of all points equivalent to $\mathscr{P}$ under rotations. Assume (i)† that s is a two-dimensional surface (except for center points, where s is zero-dimensional), and (ii) that the metric on s is that of a standard 2-sphere. Then on s one will have

$$(ds^2)_s = R^2(s)\,d\Omega^2, \tag{1}$$

where $d\Omega^2$ is the standard metric of a unit sphere ($d\Omega^2 = d\theta^2 + \sin^2\theta\,d\phi^2$ for some θ, ϕ, defined on s), and where $2\pi R$ is the circumference of s. If M^2 is the set of all such surfaces s, then $\mathcal{S}: M^4 \longrightarrow M^2: \mathscr{P} \longrightarrow s = \mathcal{S}(\mathscr{P})$ allows one to obtain, from $R: M^2 \longrightarrow \mathscr{R}: s \longrightarrow R(s)$ [the "circumference" function on M^2 as defined by equation (1)], a corresponding function $R: M^4 \longrightarrow \mathscr{R}: \mathscr{P} \longrightarrow R(\mathcal{S}(\mathscr{P}))$ on M^4 which in some cases can eventually be used as a coordinate on M^4. (Note: $\mathscr{R}$ denotes here the real numbers.)

Now assume (iii)† there is a spherically symmetric 4-velocity field $\boldsymbol{u}$, defined so that if $\mathscr{P} = \mathcal{C}(\tau)$ is one trajectory of $\boldsymbol{u}$ with $\boldsymbol{u} = d/d\tau$, then each curve $\mathscr{P} = A\mathcal{C}(\tau)$ obtained by a rotation must also be a trajectory of $\boldsymbol{u}$. The orthogonal projection of $\boldsymbol{u}$ onto any sphere s must then vanish, as there are no rotation invariant non-zero vector fields on 2-spheres. Thus $\boldsymbol{u}$ is orthogonal to each s. Also, if two trajectories of $\boldsymbol{u}$ start on some same sphere s, so $\mathcal{C}_1(0) = A\mathcal{C}_2(0)$, then the same rotation A will always relate them, $\mathcal{C}_1(\tau) = A\mathcal{C}_2(\tau)$, since trajectories are uniquely defined by any one point on them. Then $\mathcal{S}(\mathcal{C}_1(\tau))$ and $\mathcal{S}(\mathcal{C}_2(\tau))$ are both the same curve in M^2, whose tangent $d/d\tau$ one can call also $\boldsymbol{u}$; in this way one obtains a vector field $\boldsymbol{u}$ on M^2. Give each trajectory of $\boldsymbol{u}$ on M^2 a different label r to define a function $r(s)$ on M^2. Denote by $r = r(\mathcal{S}(\mathscr{P}))$ a corresponding function r on M^4 with $dr/d\tau = 0$. Since functions and their gradients on M^4 define corresponding quantities on M^2, inner products such as $\boldsymbol{d}f\cdot\boldsymbol{d}g$ can be defined on M^2 by their values on M^4; thus, from the metric on M^4 one obtains a metric on M^2. Then by equa-

(a) Write out the relativistic equation of hydrostatic equilibrium for a substance satisfying the equation of state $p = \rho/3$.

(b) Show that there exists a well-defined analytic solution for the limiting case of infinite central density, in which $m(r)/r$ has the value $3/14$.

(c) Find $\rho(r)$, $p(r)$, and $n(r)$.

(d) Show that the number of particles out to any finite r-value is finite, despite the fact that $n(r)$ is infinite at the origin.

(e) Show that the 3-geometry has a "conical singularity" at $r = 0$.

(f) Make an "embedding diagram" for this 3-geometry ["lift" $z(r)$ as a function of r from (23.34)]. (Note that the conical singularity at $r = 0$, otherwise physically unreasonable, arises because the density of mass-energy goes to infinity at that point. Note also that the calculated mass of the system diverges to infinity as $r \longrightarrow \infty$. In actuality with decreasing density the Fermi momentum falls from relativistic to nonrelativistic values, the equation of state changes its mathematical form, and the total mass M converges to a finite value).

---

tion (23.5) or equivalently by drawing curves in M^2 orthogonal to the $r =$ const. lines, and giving each a different label t, one obtains coordinates with $g^{rt} = \boldsymbol{dr} \cdot \boldsymbol{dt} = 0$. Both r and t labels were assigned arbitrarily on the corresponding curves, so it is clear that transformations $t' = t'(t)$ and $r' = r'(r)$ are not excluded.

On one 2-sphere s in M^4, on the $t = 0$ hypersurface, choose a set of θ, ϕ coordinates by picking the pole ($\theta = 0$) and the prime meridian ($\phi = 0$) arbitrarily. Then extend the definition of θ, ϕ, over the $t = 0$ hypersurface by requiring θ and ϕ to be constant on curves orthogonal to each 2-sphere s, i.e., by demanding that $(\partial/\partial r)_{\theta\phi}$ be orthogonal to each s at $t = 0$. Extend the definition of θ and ϕ to $t \neq 0$ by requiring them to be constant on curves with tangent $\boldsymbol{u}$, so $(\partial/\partial t)_{r\theta\phi} \propto \boldsymbol{u}$. But each s is a surface of constant r and t; so $(\partial/\partial\theta)_{rt\phi}$ and $(\partial/\partial\phi)_{rt\theta}$ are tangent to s, while $\boldsymbol{u} \propto (\partial/\partial t)$ is orthogonal to each s. Consequently,

$$g_{t\theta} = (\partial/\partial t) \cdot (\partial/\partial\theta) = 0 \tag{2}$$

and

$$g_{t\phi} = (\partial/\partial t) \cdot (\partial/\partial\phi) = 0 \tag{3}$$

in the $tr\theta\phi$ coordinate system just constructed. The vector $(\partial/\partial r)_{t\theta\phi}$ does not depend on the arbitrary directions introduced in the original choice of θ, ϕ coordinates on one sphere s; it is invariant under transformations $\theta = \theta(\theta', \phi')$, $\phi = \phi(\theta', \phi')$. But nothing except θ and ϕ introduced nonrotationally invariant elements into the discussion; so $(\partial/\partial r)_{t\theta\phi}$ must be a rotationally invariant vector field (unlike, say, $\partial/\partial\phi$); so it is, like $\boldsymbol{u}$, orthogonal to each 2-sphere s. This invariance then gives

$$g_{r\theta} = (\partial/\partial r) \cdot (\partial/\partial\theta) = 0, \tag{4}$$

$$g_{r\phi} = (\partial/\partial r) \cdot (\partial/\partial\phi) = 0, \tag{5}$$

which, with $g^{tr} = 0$ as previously established, gives $g_{tr} = 0$. The result is a line element of the form (23.3). Further specialization, a change of radial and time coordinates to R and T, where R is defined by (1) above and

$$\boldsymbol{dT} = e^{\psi}\left[\frac{1}{g_{rr}}\frac{\partial R}{\partial r}\,\boldsymbol{dt} - \frac{1}{g_{tt}}\frac{\partial R}{\partial t}\,\boldsymbol{dr}\right],$$

$$e^{\psi} = \begin{pmatrix}\text{integrating} \\ \text{factor}\end{pmatrix},$$

followed by a change of notation, leads to Schwarzschild coordinates and the line element (23.7)—though such a transformation is possible (i.e., nonsingular) only where $\boldsymbol{dR} \wedge \boldsymbol{dT} \neq 0$:

$$(\boldsymbol{\nabla} R)^2 = \frac{(\partial R/\partial t)^2}{g_{tt}} + \frac{(\partial R/\partial r)^2}{g_{rr}} \neq 0.$$

If (iv)† spacetime is asymptotically flat, so $r \longrightarrow \infty$ is a region where the metric can take on its special relativity values, then the arbitrariness in the t coordinate, $t' = t'(t)$, can be eliminated by requiring $g_{tt} = -1$ as $r \longrightarrow \infty$. Then $(\partial/\partial t)_{r\theta\phi}$ is uniquely determined by natural requirements (independent of the arbitrary θ, ϕ, choices), and whenever it is desired to make the further physical assumption (v)† of a time-independent geometry, this can be appropriately restated as $\partial g_{\mu\nu}/\partial t = 0$.

CHAPTER 24

PULSARS AND NEUTRON STARS; QUASARS AND SUPERMASSIVE STARS

Go, wond'rous creature, mount where Science guides,
Go, measure earth, weigh air, and state the tides;
Instruct the planets in what orbs to run,
Correct old time, and regulate the sun.

ALEXANDER POPE (1733)

§24.1. OVERVIEW

Types of stellar configurations where relativity should be important

Five kinds of stellar configurations are recognized in which relativistic effects should be significant: white dwarfs, neutron stars, black holes, supermassive stars, and relativistic star clusters. The key facts about each type of configuration are summarized in Box 24.1; and the most important details are described in the text of this chapter (white dwarfs in §24.2; neutron stars and their connection to pulsars in §§24.2 and 24.3; supermassive stars and their possible connection to quasars and galactic nuclei in §§24.4 and 24.5; and relativistic star clusters in §24.6; a detailed discussion of black holes is delayed until Chapter 33).

The book *Stars and Relativity* by Zel'dovich and Novikov (1971) presents a clear and very complete treatment of all these astrophysical applications of relativistic stellar theory. In a sense, that book can be regarded as a companion volume to this one; it picks up, with astrophysical emphasis, all the topics that this book treats with gravitational emphasis. This chapter is meant only to give the reader a brief survey of the material to be found in *Stars and Relativity*.

(continued on page 621)

Box 24.1. STELLAR CONFIGURATIONS WHERE RELATIVISTIC EFFECTS ARE IMPORTANT

[For detailed analyses and references on all these topics, see Zel'dovich and Novikov (1971).]

A. White Dwarf Stars

Are stars of about one solar mass, with radii about 5,000 kilometers and densities about 10^6 g/cm^3 $\sim$ 1 ton/cm^3; support themselves against gravity by the pressure of degenerate electrons; have stopped burning nuclear fuel, and are gradually cooling as they radiate away their remaining store of thermal energy.

Were observed and studied astronomically long before they were understood theoretically.

Key points in history:

August 1926, Dirac (1926) formulated Fermi-Dirac statistics, following Fermi (February).

December 1926, R. H. Fowler (1926) used Fermi-Dirac statistics to explain the nature of white dwarfs; he invoked electron degeneracy pressure to hold the star out against the inward pull of gravity.

1930, S. Chandrasekhar (1931a,b) calculated white-dwarf models taking account of special relativistic effects in the electron-degeneracy equation of state; he discovered that *no white dwarf can be more massive than* $\sim$*1.2 solar masses* (*"Chandrasekhar Limit"*).

1932, L. D. Landau (1932) gave an elementary explanation of the Chandrasekhar limit.

1949, S. A. Kaplan (1949) derived the effects of general relativity on the mass-radius curve for massive white dwarfs, and deduced that general relativity probably induces an instability when the radius becomes smaller than 1.1×10^3 km.

Role of general relativity in white dwarfs:

negligible influence on structure;

significant influence on stability, on pulsation frequencies, and on form of mass-radius curve near the Chandrasekhar limit (i.e., in massive white dwarfs). Electron capture also significant. See, e.g., Zel'dovich and Novikov (1971); Faulkner and Gribbin (1968).

B. Neutron Stars

Are stars of about one solar mass, with radii about 10 km and densities about 10^{14} g/cm^3 (same as density of an atomic nucleus); are supported against gravity by the pressure of degenerate neutrons and by nucleon-nucleon strong-interaction forces; are not burning nuclear fuel; the energy being radiated is the energy of rotation and the remaining store of internal thermal energy.

Theoretical calculations predicted their existence in 1934, but they were not verified to exist observationally until 1968.

Key points in history:

1932, neutron discovered by Chadwick (1932).

1933–34, Baade and Zwicky (1934a,b,c) (1) invented the concept of neutron star; (2) identified a new class of astronomical objects which they called "supernovae"; (3) suggested that supernovae might be created by the collapse of a normal star to form a neutron star. (See Figure 24.1.)

1939, Oppenheimer and Volkoff (1939) performed the first detailed calculations of the structures of neutron stars; in the process, they laid the foundations of the general relativistic theory of stellar structure as presented in Chapter 23. (See Figure 24.1.)

1942, Duyvendak (1942) and Mayall and Oort (1942) deduced that the Crab nebula is a remnant of the supernova observed by Chi-

Box 24.1 *(continued)*

nese astronomers in A.D. 1054. Baade (1942) and Minkowskii (1942) identified the "south preceding star," near the center of the Crab Nebula, as probably the (collapsed) remnant of the star that exploded in 1054 (see frontispiece).

1967, Pulsars were discovered by Hewish *et al.* (1968).

1968, Gold (1968) advanced the idea that pulsars are rotating neutron stars; and subsequent observations confirmed this suggestion.

1969, Cocke, Disney, and Taylor (1969) discovered that the "south preceding star" of the Crab nebula is a pulsar, thereby clinching the connection between supernovae, neutron stars, and pulsars.

Role of general relativity in neutron stars:

significant effects (as much as a factor of 2) on structure and vibration periods;

gravitational radiation reaction may be the dominant force that damps nonradial vibrations.

C. Black Holes

Are objects created when a star collapses to a size smaller than twice its geometrized mass ($R < 2M \sim (M/M_\odot) \times 3$ km), thereby creating such strong spacetime curvatures that it can no longer communicate with the external universe (detailed analysis of black holes in Chapters 33 and 34).

No one who accepts general relativity has found any way to escape the prediction that black holes must exist in our galaxy. This prediction depends in no way on the complexity of the collapse that forms the black holes, or on unknown properties of matter at high density. However, the existence of black holes has not yet been verified observationally.

Key points in history:

1795, Laplace (1795) noted that, according to Newtonian gravity and Newton's corpuscular theory of light, light cannot escape from a sufficiently massive object (Figure 24.1).

1939, Oppenheimer and Snyder (1939) calculated the collapse of a homogeneous sphere of pressure-free fluid, using general relativity, and discovered that the sphere cuts itself off from communication with the rest of the universe. This was the first calculation of how a black hole can form (Figure 24.1).

1965, Beginning of an era of intensive theoretical investigation of black-hole physics.

Role of general relativity in black-hole physics:

No sensible account of black holes possible in Newtonian theory. The physics of black holes calls on Einstein's description of gravity from beginning to end.

D. Supermassive Stars

Are stars of mass between 10^3 and 10^9 solar masses, constructed from a hot plasma of density typically less than that in normal stars; are supported primarily by the pressure of photons, which are trapped in the plasma and are in thermal equilibrium with it; burn nuclear fuel (hydrogen) at some stages in their evolution.

Theoretical calculations suggest (but *not* with complete confidence) that supermassive stars exist in the centers of galaxies and quasars, and perhaps elsewhere. Supermassive stars conceivably could be the energy sources for some quasars and galactic nuclei. However, astronomical observations have not yet yielded definitive evidence about their existence or their roles in the universe if they do exist.

Key points in history:

1963, Hoyle and Fowler (1963a,b) conceived the idea of supermassive stars, calculated their properties, and suggested that they might be associated with galactic nuclei and quasars.

1963–64, Chandrasekhar (1964a,b) and Feynman (1964) developed the general relativistic theory of stellar pulsations; and Feynman used it to show that supermassive stars, although Newtonian in structure, are subject to a general-relativistic instability.

1964 and after, calculations by many workers have elaborated on and extended the ideas of Hoyle and Fowler, but have not produced any spectacular breakthrough.

Role of general relativity in supermassive stars:

negligible influence on structure, except in the extreme case of a compact, rapidly rotating, disc-like configuration [see Bardeen and Wagoner (1971); Salpeter and Wagoner (1971)].

significant influence on stability.

E. Relativistic Star Clusters

Are clusters of stars so dense that relativistic corrections to Newtonian theory modify their structure.

Theoretical calculations suggest that relativistic star clusters might, but quite possibly do not, form in the nuclei of some galaxies and quasars; if they do try to form, they might be destroyed during formation by star-star collisions, which convert the cluster into supermassive stars or into a dense conglomerate of stars and gas. Astronomical observations have yielded no definitive evidence, as yet, about the existence of relativistic clusters.

Key points in history:

1965, Zel'dovich and Podurets (1965) conceived the idea of relativistic star clusters, developed the theory of their structure using general relativity and kinetic theory (cf. §25.7), and speculated about their stability.

1968, Ipser (1969) developed the theory of star-cluster stability and showed (in agreement with the Zel'dovich-Podurets speculations) that, when it becomes too dense, a cluster begins to collapse to form a black hole.

Role of general relativity in star clusters:

significant effect on structure when gravitational redshift from center to infinity exceeds $z_c \equiv \Delta\lambda/\lambda \sim 0.05$.

induces collapse of cluster to form black hole when central redshift reaches $z_c \approx 0.50$.

§24.2. THE ENDPOINT OF STELLAR EVOLUTION

After the normal stages of evolution, stars "die" by a variety of processes. Some stars explode, scattering themselves into the interstellar medium; others contract into a white-dwarf state; and others—according to current theory—collapse to a neutron-star state, or beyond, into a black hole. Although one knows little at present about a star's dynamic evolution into its final state, much is known about the final states themselves. The final states include dispersed nebulae, which are of no interest here; cold stellar configurations, the subject of this section; and "black holes," the subject of Part VII.

(continued on page 624)

JANUARY 15, 1934 PHYSICAL REVIEW VOLUME 45

Proceedings of the American Physical Society

Minutes of the Stanford Meeting, December 15–16, 1933

38. Supernovae and Cosmic Rays. W. Baade, *Mt. Wilson Observatory*, and F. Zwicky, *California Institute of Technology*.—Supernovae flare up in every stellar system (nebula) once in several centuries. The lifetime of a supernova is about twenty days and its absolute brightness at maximum may be as high as $M_{vis} = -14^M$. The visible radiation L_ν of a supernova is about 10^8 times the radiation of our sun, that is, $L_\nu = 3.78 \times 10^{41}$ ergs/sec. Calculations indicate that the total radiation, visible and invisible, is of the order $L_\tau = 10^7 L_\nu = 3.78 \times 10^{48}$ ergs/sec. The supernova therefore emits during its life a total energy $E_\tau \geq 10^5 L_\tau = 3.78 \times 10^{53}$ ergs. If supernovae initially are quite ordinary stars of mass $M < 10^{34}$ g, E_τ / c^2 is of the same order as M itself. In the *supernova* process *mass in bulk is annihilated*. In addition the hypothesis suggests itself that *cosmic rays are produced by supernovae*. Assuming that in every nebula one supernova occurs every thousand years, the intensity of the cosmic rays to be observed on the earth should be of the order $\sigma = 2 \times 10^{-3}$ erg/cm² sec. The observational values are about $\sigma = 3 \times 10^{-3}$ erg/cm² sec. (Millikan, Regener). With all reserve we advance the view that supernovae represent the transitions from ordinary stars into *neutron stars*, which in their final stages consist of extremely closely packed neutrons.

FEBRUARY 15, 1939 PHYSICAL REVIEW VOLUME 55

On Massive Neutron Cores

J. R. Oppenheimer and G. M. Volkoff
Department of Physics, University of California, Berkeley, California

(Received January 3, 1939)

It has been suggested that, when the pressure within stellar matter becomes high enough, a new phase consisting of neutrons will be formed. In this paper we study the gravitational equilibrium of masses of neutrons, using the equation of state for a cold Fermi gas, and general relativity. For masses under $\frac{1}{3}\odot$ only one equilibrium solution exists, which is approximately described by the nonrelativistic Fermi equation of state and Newtonian gravitational theory. For masses $\frac{1}{3}\odot < m < \frac{3}{4}\odot$ two solutions exist, one stable and quasi-Newtonian, one more condensed, and unstable. For masses greater than $\frac{3}{4}\odot$ there are no static equilibrium solutions. These results are qualitatively confirmed by comparison with suitably chosen special cases of the analytic solutions recently discovered by Tolman. A discussion of the probable effect of deviations from the Fermi equation of state suggests that actual stellar matter after the exhaustion of thermonuclear sources of energy will, if massive enough, contract indefinitely, although more and more slowly, never reaching true equilibrium.

Figure 24.1.
Two important arrivals on the astrophysical scene:
the neutron star (1933) and the black hole (1795, 1939).
No proper account of either can forego general relativity.

EXPOSITION

DU SYSTEME

DU MONDE,

PAR PIERRE-SIMON LAPLACE,
de l'Institut National de France, et
du Bureau des Longitudes.

TOME SECOND.

A PARIS,

De l'Imprimerie du CERCLE-SOCIAL, rue du
Théâtre Français, N°. 4.

L'AN IV DE LA RÉPUBLIQUE FRANÇAISE.

aussi sensibles à la distance qui nous en sépare ; et combien ils doivent surpasser ceux que nous observons à la surface du soleil ? Tous ces corps devenus invisibles, sont à la même place où ils ont été observés, puisqu'ils n'en ont point changé, durant leur apparition ; il existe donc dans les espaces célestes, des corps obscurs aussi considérables, et peut être en aussi grand nombre, que les étoiles. Un astre lumineux de même densité que la terre, et dont le diamètre serait deux cents cinquante fois plus grand que celui du soleil, ne laisserait en vertu de son attraction, parvenir aucun de ses rayons jusqu'à nous ; il est donc possible que les plus grands corps lumineux de l'univers, soient par cela même, invisibles. Une étoile qui, sans être de cette grandeur, surpasserait considérablement le soleil ; affaiblirait sensiblement la vitesse de la lumière, et augmenterait ainsi l'étendue de son aberration. Cette différence dans l'aberration des étoiles ; un catalogue de celles qui ne font que paraître, et leur position observée au moment de leur éclat passager ; la détermination de toutes les étoiles changeantes,

SEPTEMBER 1, 1939 PHYSICAL REVIEW VOLUME 56

On Continued Gravitational Contraction

J. R. OPPENHEIMER AND H. SNYDER
University of California, Berkeley, California

(Received July 10, 1939)

When all thermonuclear sources of energy are exhausted a sufficiently heavy star will collapse. Unless fission due to rotation, the radiation of mass, or the blowing off of mass by radiation, reduce the star's mass to the order of that of the sun, this contraction will continue indefinitely. In the present paper we study the solutions of the gravitational field equations which describe this process. In I, general and qualitative arguments are given on the behavior of the metrical tensor as the contraction progresses: the radius of the star approaches asymptotically its gravitational radius; light from the surface of the star is progressively reddened, and can escape over a progressively narrower range of angles. In II, an analytic solution of the field equations confirming these general arguments is obtained for the case that the pressure within the star can be neglected. The total time of collapse for an observer comoving with the stellar matter is finite, and for this idealized case and typical stellar masses, of the order of a day; an external observer sees the star asymptotically shrinking to its gravitational radius.

"Final state of stellar evolution," and "cold, catalyzed matter" defined

What does one mean in principle by the term "the final state of stellar evolution"? Start with a star containing a given number, A, of baryons and let it evolve to the absolute, burned-out end point of thermonuclear combustion (minimum mass-energy possible for the A-baryon system). If the normal course of thermonuclear combustion is too slow, speed it up by catalysis. If an explosion occurs, collect the outgoing matter, extract its kinetic energy, and let it fall back onto the system. Repeat this operation as many times as needed to arrive at burnout (cold Fe^{56} for the part of the system under modest pressure; other nuclear species in the region closer to the center; "cold matter catalyzed to the end point of thermonuclear combustion" throughout). End up finally with the system in its absolutely lowest energy state, with all angular momentum removed and all heat extracted, so that it sits at the absolute zero of temperature and has zero angular velocity. Such a "dead" system, depending upon its mass and prior history (two distinct energy minima for certain A-values), ends up as a cold stellar configuration (neutron star, or "white" dwarf), or as a "dead" black hole.

Equation of state for cold, catalyzed matter

The analysis of a cold stellar configuration demands an equation of state. The temperature is fixed at zero; the nuclear composition in principle is specified uniquely by the density; and therefore the pressure is also fixed uniquely once the density has been specified [equation of state $p(\rho)$ for "cold catalyzed matter"].

The white dwarfs and neutron stars observed by astronomers are not really built of cold catalyzed matter. However, the matter in them is sufficiently near the end point of thermonuclear evolution and sufficiently cold that it can be idealized with fair accuracy as cold and catalyzed (see §23.4).

The equation of state, $\rho(p)$, for cold catalyzed matter is shown graphically in Figure 24.2. This version of the equation of state was constructed by Harrison and Wheeler in 1958. Other versions constructed more recently [see Cameron (1970) and Baym, Bethe, and Pethick (1971) for references] are almost identical to the Harrison-Wheeler version at densities well below nuclear densities, $\rho < 3 \times 10^{13}$ g/cm^3. At nuclear and supernuclear densities, all versions differ because of differing assumptions about nucleon-nucleon interactions. Along with the equation of state, in Figure 24.2 are shown properties of the models of cold stars constructed from this equation of state by integrating numerically the equations of structure (23.28).

The equation of state can be understood by following the transformations that occur as a sample of cold catalyzed matter is compressed to higher and higher densities. At each stage in the compression, each possible thermonuclear reaction is to be catalyzed to its endpoint and the resultant thermal energy is to be removed.

When the sample is at zero pressure, it is a ball of pure, cold Fe^{56}, since Fe^{56} is the most tightly bound of all nuclei. It has the density 7.86 g/cm^3. As the sample is compressed, its internal pressure is provided at first by normal solid-state forces; but the atoms are soon squeezed so closely together that the electrons become quite oblivious of their nuclei, and begin to form a degenerate Fermi gas. By the time a density of $\rho = 10^5$ g/cm^3 has been reached, valence forces are completely negligible, the degenerate electron pressure dominates, and the compressibility index, γ (see legend for Figure 24.2), is 5/3, the value for a nonrelativistically degenerate Fermi gas. Between 10^5 and 10^7 g/cm^3, the pressure-providing electrons gradually

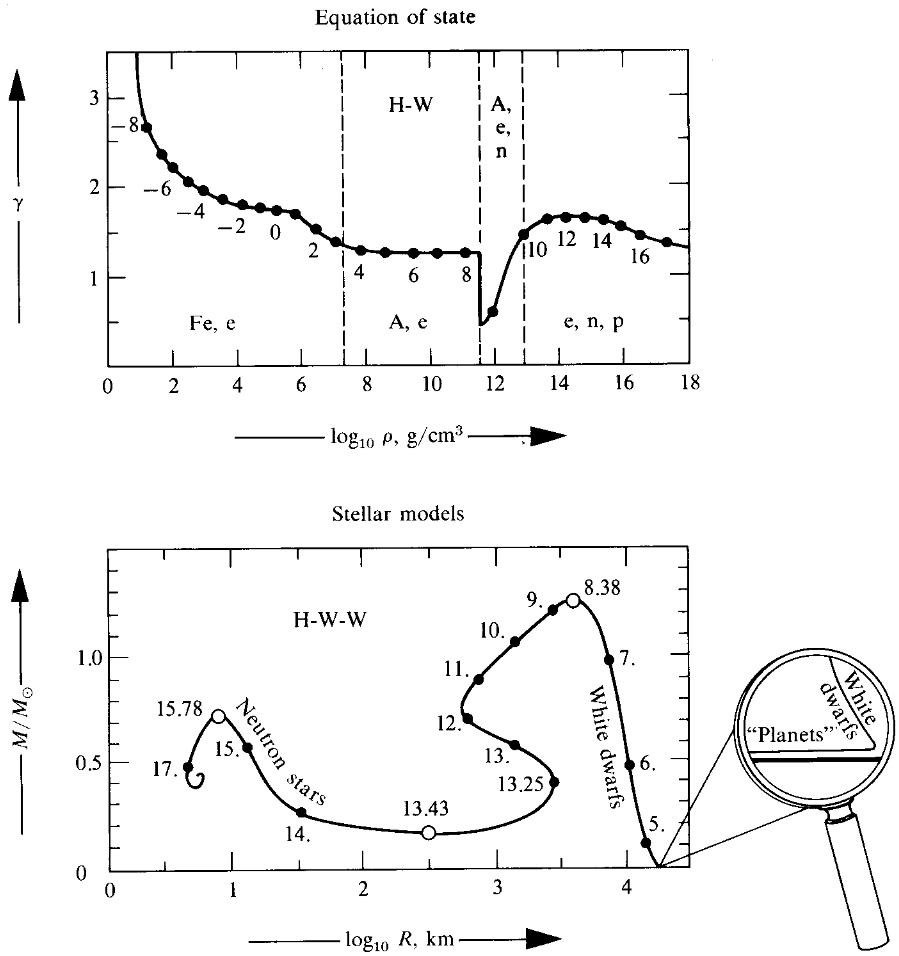


Figure 24.2.
The Harrison-Wheeler equation of state for cold matter at the absolute end point of thermonuclear evolution, and the corresponding Harrison-Wakano-Wheeler stellar models. The equation of state is exhibited in the form of a plot of "compressibility index,"

$$\gamma = \frac{\rho + p}{p}\frac{dp}{d\rho},$$

as a function of density of mass-energy, ρ. (Small γ corresponds to easy compressibility.) The curve is parameterized by the logarithm of the pressure, $\log_{10}p$, in units of g/cm^3 [same units as ρ; note that $p(\text{g/cm}^3) = (1/c^2) \times p(\text{dyne/cm}^2)$]. The chemical composition of the matter as a function of density is indicated as follows: Fe, Fe^{56} nuclei; A, nuclei more neutron rich than Fe^{56}; e, electrons; n, free neutrons; p, free protons.

The first law of thermodynamics [equation (22.6)], when applied to cold matter (zero entropy) says $d\rho/(\rho + p) = dn/n$; i.e.,

$$n = \frac{\rho + p}{\mu_{\text{Fe}}/56} \exp\left(-\int_0^p \frac{dp}{\rho + p}\right).$$

Here μ_{Fe}, the rest mass of an Fe^{56} atom, is the ratio between $\rho + p \approx \rho$ and $n/56$ in the limit of zero density. From this equation and a knowledge of $\rho(p)$—(see Figure)—one can calculate $n(p)$.

The equilibrium configurations are represented by curves of total mass-energy, M, versus radius, R. (R is defined such that $4\pi R^2$ is the star's surface area.) The $M(R)$ curve is parameterized by the logarithm of the central density, $\log_{10}\rho_c$, measured in g/cm^3. Only configurations along two branches of the curve are stable against small perturbations and can therefore exist in nature: the white dwarfs, with $\log_{10}\rho_c < 8.38$, and the neutron stars, with $13.43 < \log_{10}\rho_c < 15.78$ (see Box 26.1).

For greater detail on both the equation of state and the equilibrium configurations, see Harrison, Thorne, Wakano, and Wheeler (1965); also, for an updated table of the equation of state, see Hartle and Thorne (1968).

become relativistically degenerate, and γ approaches 4/3. Above $\rho = 1.4 \times 10^7$ g/cm^3, the rest mass of 62 Fe^{56}_{26} nuclei, plus the rest mass of 44 electrons, plus the rather large Fermi kinetic energy of 44 electrons at the top of the Fermi sea, exceeds the rest mass of 56 Ni^{62}_{28} nuclei. Consequently, as the catalyzed sample of matter is compressed past $\rho = 1.4 \times 10^7$ g/cm^3, the nuclear reaction

$$\begin{aligned} &62\ \mathrm{Fe}^{56}_{26}\begin{pmatrix}\text{highly compressed}\\ \text{neutral atoms}\end{pmatrix} \longrightarrow \\ &56\ \mathrm{Ni}^{62}_{28}\begin{pmatrix}\text{highly compressed}\\ \text{neutral atoms}\end{pmatrix}\end{aligned} \tag{24.1}$$

goes to its end point, with a release of energy. As the compression continues beyond this point, the rising Fermi energy of the electrons induces new nuclear reactions similar to (24.1), but involving different nuclei. In these reactions more and more electrons are swallowed up to form new nuclei, which are more and more neutron-rich. When the density reaches $\rho = 3 \times 10^{11}$ g/cm^3, the nuclei are so highly neutron-rich (Y^{122}_{39}) that neutrons begin to drip off them. The matter now becomes highly compressible for a short time ($3 \times 10^{11} \lesssim \rho \lesssim 4 \times 10^{11}$), since most of the remaining electrons are swallowed up very rapidly by the dripping nuclei. Above $\rho \sim 4 \times 10^{11}$ g/cm^3 free neutrons become plentiful and their degeneracy pressure exceeds that of the electrons. Further compression to $\rho \sim 10^{13}$ g/cm^3 completely disintegrates the remaining nuclei, leaving the sample almost pure neutrons with $\gamma = 5/3$, the value for a nonrelativistically degenerate Fermi gas. Intermixed with the neutrons are just enough degenerate electrons to prevent the neutrons from decaying, and just enough protons to maintain charge neutrality. Compression beyond $\rho \sim 10^{13}$ g/cm^3 pushes the sample into the domain of nuclear densities where the physics of matter is only poorly understood. This Harrison-Wheeler version of the equation of state ignores all nucleon-nucleon interactions at and above nuclear densities; it idealizes matter as a noninteracting mixture of neutrons, protons, and electrons with neutrons dominating; and it shows a compressibility index of 5/3 while the neutrons are nonrelativistic, but 4/3 after they attain relativistic Fermi energies. Other versions of the equation of state attempt to take into account the nucleon-nucleon interactions in a variety of ways [see Cameron (1970), Baym, Bethe, and Pethick (1971), and many references cited therein].

Equilibrium configurations for cold, catalyzed matter:

(1) forms and stability

Corresponding to each value of the central density, ρ_c, there is one stellar equilibrium configuration. Equilibrium, yes; but is the equilibrium stable? Stability studies (Chapter 26, especially Box 26.1) show that many of the models are unstable against small radial perturbations, which lead to gravitational collapse. Only white-dwarf stars in the range $\log_{10}\rho_c < 8.4$ and neutron stars in the range $13.4 \lesssim \log_{10}\rho_c \lesssim 15.8$ are stable. Instability for the region of $\log_{10}\rho_c$ values between 8.4 and 13.4 is caused by a combination of (1) relativistic strengthening of the gravitational forces, and (2) high compressibility of the matter due to electron capture and neutron drip by the atomic nuclei. Neutron stars are stable for a simple reason. Neutron-dominated matter is so difficult to compress that even the relativistically strengthened gravitational forces cannot overcome it. Above $\log_{10}\rho_c \sim 15.8$, the gravitational forces become strong enough to win out over the pressure of the nuclear matter, and the stars are all unstable. [See Gerlach (1968) for the possibility—which, however, he rates as unlikely—that there might exist a third family of stable equilibrium configurations, additional to white dwarfs and neutron stars.]

(2) white-dwarf stars

The white-dwarf stars have masses below 1.2 $M_\odot$ and radii between ~3000 and ~20,000 km. They are supported almost entirely by the pressure of the degenerate electron gas. Relativistic deviations from Newtonian structure are only a fraction of a per cent, but relativistic effects on stability and pulsations are important from $\rho_c \approx 10^8$ g/cm^3 to the upper limit of the white-dwarf family at $\rho_c = 10^{8.4}$ g/cm^3 [see, e.g., Faulkner and Gribbin (1968)]. The properties of white-dwarf models are fairly independent of whose version of the equation of state is used in the calculations.

(3) neutron stars

The properties of neutron stars are moderately dependent on the equation of state used. However, all versions lead to upper and lower limits on the mass and central density. The correct lower limits probably lie in the range

$$\begin{aligned} 13.4 \lesssim \log_{10}\rho_{c\,\mathrm{min}} \lesssim 14.0, \\ 0.05\ M_\odot \lesssim M_{\mathrm{min}} \lesssim 0.2\ M_\odot; \end{aligned} \tag{24.2}$$

the correct upper limits are probably in the range

$$\begin{aligned} 15.0 \lesssim \log_{10}\rho_{c\,\mathrm{max}} \lesssim 16.0, \\ 0.5\ M_\odot \lesssim M_{\mathrm{max}} \lesssim 3\ M_\odot \end{aligned} \tag{24.3}$$

[see Rhoades (1971)]. Neutron stars typically have radii between ~6 km and ~100 km. Relativistic deviations from Newtonian structure are great, sometimes more than 50 per cent.

It appears certain that no cold stellar configuration can have a mass exceeding ~5 $M_\odot$ [Rhoades (1971)] (1.2 $M_\odot$ according to the Harrison-Wheeler equation of state, Figure 24.2). Any star more massive than this must reduce its mass below this limit if it is to fade away into quiet obscurity, otherwise relativistic gravitational forces will eventually pull it into catastrophic gravitational collapse past white-dwarf radii, past neutron-star radii, and into a black hole a few kilometers in size (see Part VII).

(4) black holes

§24.3. PULSARS

Theory predicts that, when a star more massive than the Chandrasekhar limit of 1.2 $M_\odot$ has exhausted the nuclear fuel in its core and has compressed its core to white-dwarf densities, an instability pushes the star into catastrophic collapse. The

Birth of a neutron star by stellar collapse

core implodes upon itself until nucleon-nucleon repulsion halts the implosion. The result is a neutron star, unless the core's mass is so great that gravity overcomes the nucleon-nucleon repulsion and pulls the star on in to form a black hole. Not all the star's mass should become part of the neutron star or black hole. Much of it, perhaps most, can be ejected into interstellar space by the violence that accompanies the collapse—violence due to flash nuclear burning, shock waves, and energy transport by neutrinos ("stick of dynamite in center of star, ignited by collapse").

Dynamics of a newborn neutron star

The collapsed core holds more interest for gravitation theory than the ejected envelope. That core, granted a mass small enough to avoid the black-hole fate, will initially be a hot, wildly pulsating, rapidly rotating glob of nuclear matter with a strong, embedded magnetic field (see Figure 24.3). The pulsations must die out quickly. They emit a huge flux of gravitational radiation, and radiation reaction damps them in a characteristic time of ~ 1 second [see Wheeler (1966); Thorne (1969a)]. Moreover, the pulsations push and pull elementary particle reactions back and forth by raising and lowering the Fermi energies in the core's interior; these particle reactions can convert pulsation energy into heat at about the same rate as the pulsation energy is radiated by gravity. [See Langer and Cameron (1969); also §11.5 of Zel'dovich and Novikov (1971) for details and references.]

The result, after a few seconds, is a rapidly rotating centrifugally flattened neutron star with a strong (perhaps 10^{12} gauss) magnetic field; all the pulsations are gone. If the star is deformed from axial symmetry (e.g., by centrifugal forces or by a nonsymmetric magnetic field), its rotation produces a steady outgoing stream of gravitational waves, which act back on the star to remove rotational energy. Whether or not this occurs, the rotating magnetic field itself radiates electromagnetic waves. They slow the rotation and transport energy into the surrounding, exploding gas cloud (nebula). [See Pacini (1968), Goldreich and Julian (1968), and Ostriker and Gunn (1969) for basic considerations.]

Neutron star as a pulsar

Somehow, but nobody understands in detail how, the rotating neutron star beams coherent radio waves and light out into space. Each time the beam sweeps past the Earth optical and radio telescopes see a pulse of radiation. The light is emitted synchronously with the radio waves, but the light pulses reach Earth earlier (~ 1 second for the pulsar in the crab nebula) because of the retardation of the radio waves by the plasma along the way. This is the essence of the 1973 theory of pulsars, accepted by most astrophysicists.

Pulsar radiation as a tool for studying neutron stars

Although the mechanism of coherent emission is not understood, the pulsar radiation can nevertheless be a powerful tool in the experimental study of neutron stars. Anything that affects the stellar rotation rate, even minutely (fractional changes as small as 10^{-9}) will produce measurable irregularities in the timing of the pulses at Earth. If the star's crust and mantle are crystalline, as 1973 theory predicts, they may be subject to cracking, faulting, or slippage ("starquake") that changes the moment of inertia, and thence the rotation rate. Debris falling into the star will also change its rotation. Whichever the cause, after such a disturbance the star may rotate differentially for awhile; and how it returns to rigid rotation may depend on such phenomena as superfluidity in its deep interior. Thus, pulsar-timing data may eventually give information about the interior and crust of the neutron star, and

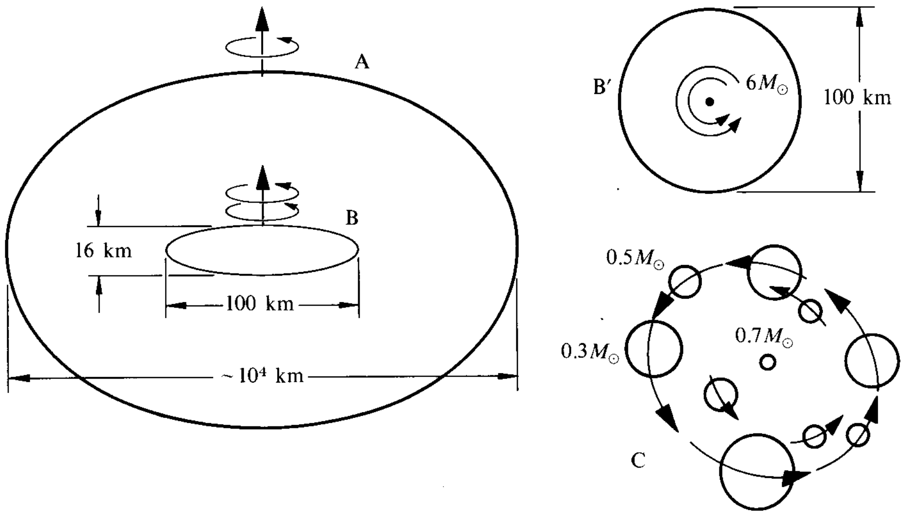


Figure 24.3.
"Collapse, pursuit, and plunge scenario" [schematic from Ruffini and Wheeler (1971b)].

- A star with white-dwarf core (A), slowly rotating,
  - evolves by straightforward astrophysics,
  - arrives at the point of gravitational instability,
  - collapses, and
  - ends up as a rapidly spinning neutron-star pancake (B,B′).
- It then fragments (C) because it has too much angular momentum to collapse into a single stable object. If the substance of the neutron-star pancake were an incompressible fluid, the fragmentation would have a close tie to well-known and often observed phenomena ("drop formation"). However, the more massive a neutron star is, the smaller it is, so one's insight into this and subsequent stages of the scenario are of necessity subject to correction or amendment. One can not today guarantee that fragmentation takes place at all; nevertheless, fragmentation will be assumed in what follows.
- The fragments dissipate energy and angular momentum via gravitational radiation.
- One by one as they revolve they coalesce ("pursuit and plunge scenario").
- In each such plunge a pulse of gravitational radiation emerges.
- Fragments of debris fall onto the coalesced objects (neutron stars or black holes, as the case may be), changing their angular momenta.
- Eventually the distinct neutron stars or black holes or both unite into one such collapsed object with a final pulse of gravitational radiation.
- The details of the complete scenario differ completely from one evolving star to another, depending on
  - the mass of its core, and
  - the angular momentum of this core.
- An entirely different kind of picture therefore has to be drawn for altered values of these two parameters.
- Even for the values of these parameters adopted in the drawing, the present picture can at best possess only qualitative validity.
- Detailed computer analysis would seem essential for any firm prediction about the course of any selected scenario.

thence (by combination with theory) about its mass and radius. These issues are discussed in detail in a review article by Ruderman (1972) as well as in Zel'dovich and Novikov (1971).

§24.4. SUPERMASSIVE STARS AND STELLAR INSTABILITIES

Theory of the stability of Newtonian stars

When a Newtonian star of mass M oscillates adiabatically in its fundamental mode, the change in its radius, δR, obeys a harmonic-oscillator equation,

$$M\,\delta\ddot{R} = -k\,\delta R, \tag{24.4}$$

with a "spring constant" k that depends on the star's mean adiabatic index $\bar{\Gamma}_1$ [recall: $\Gamma_1 \equiv (n/p)(\partial p/\partial n)_{\text{const. entropy}}$], on its gravitational potential energy Ω, on the trace $I = \int \rho r^2\, d\mathcal{V}$ of the second moment of its mass distribution, and on its mass M,

$$k = 3M(\bar{\Gamma}_1 - 4/3)|\Omega|/I \tag{24.5}$$

(See Box 24.2). If $\bar{\Gamma}_1 > 4/3$ the Newtonian star is stable and oscillates; if $\bar{\Gamma}_1 < 4/3$ the star is unstable and either collapses or explodes, depending on its initial conditions and overall energetics. This result is a famous theorem in Newtonian stellar theory—but it is relevant only for adiabatic oscillations.

Box 24.2 OSCILLATION OF A NEWTONIAN STAR

The following is a volume-averaged analysis of the lowest mode of radial oscillation. Such analyses are useful in understanding the qualitative behavior and stability of a star. [See Zel'dovich and Novikov (1971) for an extensive exploitation of them.] However, for precise quantitative results, one must perform a more detailed analysis [see, e.g., Ledoux and Walraven (1958); also Chapter 26 of this book].

1. Let M = star's total mass
   R = star's radius
   $\bar{\rho}$ = mean density = $(3/4\pi)M/R^3$
   $\bar{p}$ = mean pressure
   $\bar{\Gamma}_1$ = mean adiabatic index = $(\bar{n}/\bar{p})(\partial\bar{p}/\partial\bar{n})_{\text{adiabatic}}$
   $= (\bar{\rho}/\bar{p})(\partial\bar{p}/\partial\bar{\rho})_{\text{adiabatic}}$ in Newtonian limit, where $\rho = \text{const.} \times n$.
2. Then the mean pressure-buoyancy force $\bar{F}_{\text{buoy}}$ and the counterbalancing gravitational force $\bar{F}_{\text{grav}}$ in the equilibrium star are

$$\begin{aligned}\bar{F}_{\text{buoy}} &= \bar{p}/R \\ &= \bar{F}_{\text{grav}} = \bar{\rho}M/R^2 = (4\pi/3)\bar{\rho}^2 R.\end{aligned}$$

3. When the oscillating star has expanded or contracted so its radius is $R + \delta R$, then its mean density will have changed to

$$\bar{\rho} + \delta\bar{\rho} = (3/4\pi)M[R^{-3} + \delta(R^{-3})] = \bar{\rho} - 3(\bar{\rho}/R)\,\delta R,$$

and its mean pressure will be

$$\bar{p} + \delta\bar{p} = \bar{p} + (\bar{p}/\bar{\rho})\bar{\Gamma}_1\,\delta\bar{\rho} = \bar{p} - 3(\bar{\Gamma}_1\bar{p}/R)\,\delta R.$$

The corresponding changes in the forces will be

$$\delta\bar{F}_{\text{buoy}} = \frac{\delta\bar{p}}{R} - \frac{\bar{p}}{R^2}\,\delta R = -(3\bar{\Gamma}_1 + 1)\frac{\bar{p}}{R}\frac{\delta R}{R} = -(3\bar{\Gamma}_1 + 1)\bar{F}_{\text{buoy}}\left(\frac{\delta R}{R}\right),$$

$$\delta\bar{F}_{\text{grav}} = \left(\frac{4\pi}{3}\right)(2\bar{\rho}R\,\delta\bar{\rho} + \bar{\rho}^2\,\delta R) = \left(\frac{4\pi}{3}\bar{\rho}^2R\right)\left(-5\,\frac{\delta R}{R}\right) = -5\bar{F}_{\text{grav}}\left(\frac{\delta R}{R}\right).$$

Consequently, the restoring force will be (recall: $\bar{F}_{\text{buoy}} = \bar{F}_{\text{grav}}$)

$$\delta\bar{F}_{\text{grav}} - \delta\bar{F}_{\text{buoy}} = 3\left(\bar{\Gamma}_1 - \frac{4}{3}\right)\bar{F}_{\text{grav}}\,\frac{\delta R}{R}.$$

4. This restoring force produces an acceleration,

$$\delta\bar{F}_{\text{grav}} - \delta\bar{F}_{\text{buoy}} = -\bar{\rho}\,\delta\ddot{R}.$$

Hence, the equation of motion for the oscillations is

$$\delta\ddot{R} = -3(\bar{\Gamma}_1 - 4/3)(4\pi/3)\bar{\rho}\,\delta R,$$

corresponding to a "spring constant" k and angular frequency of oscillation ω, given by $\omega^2 = 4\pi(\bar{\Gamma}_1 - 4/3)\bar{\rho}$, and $k = M\omega^2$.

5. A more nearly exact analysis (see exercise 39.7 for details, or Box 26.2 for an alternative derivation) yields the improved formula

$$\omega^2 = 3(\bar{\Gamma}_1 - 4/3)|\Omega|/I,$$

$$\Omega = \begin{pmatrix}\text{star's self-gravitational}\\ \text{energy}\end{pmatrix} = \frac{1}{2}\int \rho\Phi\, d\mathcal{V} = -\frac{1}{2}\int \frac{\rho\rho'}{|\boldsymbol{x} - \boldsymbol{x}'|}\, d\mathcal{V}\, d\mathcal{V}',$$

$$I = \begin{pmatrix}\text{trace of second moment of}\\ \text{star's mass distribution}\end{pmatrix} = \int \rho r^2\, d\mathcal{V},$$

for the square of the oscillation frequency.

6. Note that $\bar{\Gamma}_1 > 4/3$ corresponds to stable oscillations; $\bar{\Gamma}_1 < 4/3$ corresponds to exponentially developing collapse or explosion.

In a real star no oscillation is precisely adiabatic. The oscillations in temperature cause corresponding oscillations in the stellar opacity and in nuclear burning rates. These insert energy into or extract energy from the gas vibrations.

All main-sequence stars thus far observed and studied have masses below 60 $M_\odot$. For such small masses, theory predicts low enough temperatures that gas pressure dominates over radiation pressure, and the adiabatic index is nearly that of nonrelativistic gas, $\bar{\Gamma}_1 \approx 5/3$. Such stars vibrate stably. The net effect of the oscillating opacity and burning rate is usually to extract energy from the vibrations. Thus, they damp. (The vibrations of Cepheid variable stars are a notable exception.)

Stability theory predicts "engine-driven oscillations" and quick death for stars of $M > 60 M_\odot$

No one has yet seen a main-sequence star with mass above about 60 $M_\odot$. This is explained as follows. For masses above 60 $M_\odot$, the temperature should be so high that radiation pressure dominates over gas pressure, and the adiabatic index $\bar{\Gamma}_1$ is only slightly above the value 4/3 for pure radiation. Consequently the "spring constant" of the star, although positive, is very small. On the inward stroke of an oscillation, the central temperature rises, and nuclear burning speeds up. (The nuclear burning rate goes as a very high power of the central temperature; for example, in a massive star HCNO burning releases energy at a rate $\varepsilon_{\mathrm{HCNO}} \propto T_c^{11}$.) Because the spring constant is so small, the inward stroke lasts for a long time, and the enhanced nuclear burning produces a significant excess of thermal energy and pressure. Hence, on the outward stroke the star expands more vigorously than it contracted ("engine"). Successive vibrations are driven to higher and higher amplitudes. Eventually, calculations suggest, the star either explodes, or it ejects enough mass by its vigorous vibrations to drop below the critical limit of $M \sim 60\ M_\odot$. Hence, stars of mass above 60 $M_\odot$ should not live long enough that astronomers could have a reasonable probability of discovering them.

Possible existence of supermassive stars

Of course, this "engine action" does not prevent massive stars from forming, living a short time, and then disrupting themselves. Such a possibility is particularly intriguing for *supermassive stars* [M between $10^3\ M_\odot$ and $10^9\ M_\odot \sim 0.01 \times$ (mass of a galaxy)]. Although such stars may be exceedingly rare, by their huge masses and huge release of explosive energy they might play an important role in the universe. Moreover, it is conceivable that the oscillations of such stars, like those of Cepheid variables, might be sustained at large amplitudes for long times (a million years?), with nonlinear damping processes preventing their further growth.

Relativistic instabilities in a supermassive star

Theory predicts that general relativistic effects should strongly influence the oscillations of a supermassive star. The increase in "gravitational force," δF_{grav}, acting on a shell of matter on the inward stroke is greater in general relativity than in Newtonian theory, and the decrease on the outward stroke is also greater. Consequently the "effective index" $\Gamma_{1\,\mathrm{crit}}$ of gravitational forces is increased above the Newtonian value of 4/3; thus,

$$\left(\begin{array}{l}\text{fractional increase in}\\ \text{“pressure-like force of}\\ \text{gravity” per unit fractional}\\ \text{change in baryon-number}\\ \text{density}\end{array}\right) \equiv \Gamma_{1\,\mathrm{crit}} = (4/3) + \alpha(M/R) + O(M^2/R^2), \quad (24.6)$$

where α is a constant of the order of unity that depends on the structure of the star (see Box 26.2). To resist gravity, one has only the elasticity of the relativistic material of the star:

$$\begin{pmatrix}\text{fractional increase in} \\ \text{“pressure-like resisting} \\ \text{force” per unit fractional} \\ \text{change in baryon number} \\ \text{density}\end{pmatrix} = \bar{\Gamma}_1 = \left\langle \frac{p}{n}\left(\frac{\partial p}{\partial n}\right)_s \right\rangle_{\substack{\text{effective average} \\ \text{over star}}}. \tag{24.7}$$

The effective spring constant for the vibrations of the star is governed by the delicate margin between these two indices:

$$k = \begin{pmatrix}\text{effective} \\ \text{spring constant}\end{pmatrix} = \begin{pmatrix}\text{contribution} \\ \text{of “elastic} \\ \text{forces”}\end{pmatrix} - \begin{pmatrix}\text{contribution} \\ \text{of gravity}\end{pmatrix}$$

$$= 3M(\bar{\Gamma}_1 - \Gamma_{1\,\mathrm{crit}})\frac{|\Omega|}{I}. \tag{24.8}$$

(derivation in Chapter 26). The relativistic rise in the effective index of gravity above 4/3 [equation (24.6)] brings on the transition from stability (positive k; vibration) to instability (negative k; explosion or collapse) under conditions when one otherwise would have expected stability. For supermassive stars, Fowler and Hoyle (1964) show that

$$\bar{\Gamma}_1 = 4/3 + \zeta(M/M_\odot)^{-1/2},$$

where ζ is a constant of order unity. As a newly formed supermassive star contracts inward, heating up, but not yet hot enough to ignite its nuclear fuel, it approaches nearer and nearer to instability against collapse. Unless burning halts the contraction, collapse sets in at a radius R_{crit} given by

$$\bar{\Gamma}_1 = 4/3 + \zeta(M/M_\odot)^{-1/2} = \Gamma_{1\,\mathrm{crit}} = 4/3 + \alpha M/R;$$

i.e.,

$$\begin{aligned} R &= (\alpha/2\zeta)(M/M_\odot)^{1/2} \times (\text{Schwarzschild Radius}) \\ &\sim 10^4 \times (\text{Schwarzschild Radius}) \text{ if } M = 10^8\, M_\odot. \end{aligned}$$

The relativistic instability occurs far outside the Schwarzschild radius when the star is very massive. Relativity hardly modifies the star's structure at all; but because of the delicate balance between $\delta\bar{F}_{\mathrm{grav}}$ and $\delta\bar{F}_{\mathrm{buoy}}$ in the Newtonian oscillations (Box 24.2), tiny relativistic corrections to these forces can completely change the stability.

In practice, the story of a supermassive star is far more complicated than has been indicated here. Rotation can stabilize it against relativistic collapse for a while. However, after the star has lost all angular momentum in excess of the critical value

Temporary stabilization by rotation

Possible scenarios for evolution and death of a supermassive star

$J_{\text{crit}} = M^2$ ("extreme Kerr limit"; see Chapter 33), and after it has contracted to near the Schwarzschild radius, rotation is helpless to stave off implosion. Depending on its mass and angular momentum, the star may ignite its fuel before or after relativistic collapse begins, and before or after implosion through the Schwarzschild radius. When the fuel is ignited, it can wreak havoc, because even if the star is not then imploding, its adiabatic index will be very near the critical one, and the burning may drive oscillations to higher and higher amplitudes. These processes are so complex that in 1973 one is far from having satisfactory analyses of them, but for reviews of what is known and has been done, the reader can consult Fowler (1966), Thorne (1967), and Zel'dovich and Novikov (1971).

The theory of stellar pulsations in general relativity is presented for Track-2 readers in Chapter 26 of this book.

§24.5. QUASARS AND EXPLOSIONS IN GALACTIC NUCLEI

Supermassive stars as possible energy sources for quasars and galactic nuclei

Supermassive stars were first conceived by Hoyle and Fowler (1963a,b) as an explanation for explosions in the nuclei of galaxies. Shortly thereafter, when quasars were discovered, Hoyle and Fowler quite naturally appealed to their supermassive stars for an explanation of these puzzles as well. Whether galactic explosions or quasars are driven by supermassive stars remains a subject of debate in astronomical circles even as this book is being finished, in 1973. Hence, this book will avoid the issue except for the following remark.

Other possible energy sources: dense star clusters; black holes

Whatever is responsible for quasars and galactic explosions must be a machine of great mass ($M \sim 10^6$ to $10^{10}\, M_\odot$) and small radius (light-travel time across the machine, as deduced from light variations, is sometimes less than a day). The machine might be a coherent object, i.e., a supermassive star; or it might be a dense mixture of ordinary stars and much gas. Actually these two possibilities may not be distinct. Star-star collisions in a dense cluster can lead to stellar coalescence and the gradual building up of one or more supermassive stars [Sanders (1970); Spitzer (1971); Colgate (1967)]. Thus, at one stage in its life, a galactic nucleus or quasar might be driven by collisions in a dense star cluster; and at a later stage it might be driven by a supermassive star; and at a still later stage that star might collapse to leave behind a massive black hole (10^6–$10^9\, M_\odot$), but a black hole that is still "live" and active (Chapter 33).

§24.6. RELATIVISTIC STAR CLUSTERS

Relativistic star clusters

The normal astrophysical evolution of a galactic nucleus is estimated [Sanders (1970); Spitzer (1971)] to lead under some circumstances to a star cluster so dense that general relativity influences its structure and evolution. The theory of relativistic star clusters is closely related to that of relativistic stars, as developed in Chapter 23. A star is a swarm of gas molecules that collide frequently; a star cluster is a swarm of stars that collide rarely. But the frequency of collisions is relatively unimportant in a steady state. For the theory of relativistic star clusters, see: §25.7 of this book; Zel'dovich and Podurets (1965); Fackerell, Ipser, and Thorne (1969); Chapter 12 of Zel'dovich and Novikov (1971); and references cited there. A relativistic star cluster is a latent volcano. No future is evident for it except to evolve with enormous energy release to a massive black hole, either by direct collapse (possibly a star at a time) or by first coalescing into a supermassive star that later collapses.

CHAPTER **25**

THE "PIT IN THE POTENTIAL" AS THE CENTRAL NEW FEATURE OF MOTION IN SCHWARZSCHILD GEOMETRY

"Eccentric, intervolved, yet regular
Then most, when most irregular they seem;
And in their motions harmony divine"
MILTON, 1665

This chapter is entirely Track 2, *except* for Figures 25.2 and 25.6, and Boxes 25.6 and 25.7 (pp. 639, 660, 674, and 677), which Track-1 readers should peruse for insight and flavor. No earlier Track-2 material is needed as preparation for it.
§25.2 (symmetries) is needed as preparation for Box 30.2 (Mixmaster cosmology). The rest of the chapter is not essential for any later chapter, but it will be helpful in understanding
(1) Chapters 31–34 (gravitational collapse and black holes), and
(2) Chapter 40 (solar-system experiments).

§25.1. FROM KEPLER'S LAWS TO THE EFFECTIVE POTENTIAL FOR MOTION IN SCHWARZSCHILD GEOMETRY

No greater glory crowns Newton's theory of gravitation than the account it gives of the principal features of the solar system: a planet in its motion sweeps out equal areas in equal times; its orbit is an ellipse, with one focus at the sun; and the cube of the semimajor axis, a, of the ellipse, multiplied by the square of the average angular velocity of the planet in its orbit ($\omega = 2\pi/\text{period}$) gives a number with the dimensions of a length, the same number for all the planets (Box 25.1), equal to the mass of the sun:

$$M = \omega^2 a^3.$$

Overview of this chapter

Exactly the same is true for the satellites of Jupiter (Figure 25.1), and of the Earth (Box 25.1), and true throughout the heavens. What more can one possibly expect of Einstein's theory of gravity when it in its turn grapples with this centuries-old theme of a test object moving under the influence of a spherically symmetric center of attraction? The principal new result can be stated in a single sentence: The particle is governed by an "effective potential" (Figure 25.2 and §§25.5, 25.6) that possesses not only (1) the long distance $-M/r$ attractive behavior and (2) the shorter distance (angular momentum)$^2/r^2$ repulsive behavior of Newtonian gravitational theory, but also (3) at still shorter distances *a pit in the potential,* which (1) captures a particle that comes too close; (2) establishes a critical distance of closest approach for this black-hole capture process; (3) for a particle that approaches this critical point without crossing it, lengthens the turn-around time as compared to Newtonian expectations; and thereby (4) makes the period for a radial excursion longer than the period of a revolution; (5) causes an otherwise Keplerian orbit to precess; and (6) deflects a fast particle and a photon through larger angles than Newtonian theory would predict.

The *pit in the potential* being thus the central new feature of motion in Schwarzschild geometry and the source of major predictions (Box 25.2), it is appropriate to look for the most direct road into the concept of effective potential and its meaning and application. In this search no guide is closer to hand than Newtonian mechanics.

Analytic mechanics offers several ways to deal with the problem of motion in a central field of force, and among them are two of central relevance here: (1) the world-line method, which includes second-order differential equations of motion, Lagrange's equations, search for constants of integration, reduction to first-order equations, and further integration in rather different ways according as one wants the shape of the orbit, $\theta = \theta(r)$, or the time to get to a given point on the world line, $t = t(r)$; and (2) the wave-crest method, otherwise known as the "eikonal method" or "Hamilton-Jacobi method," which gives the motion by the condition of "constructive interference of wave crests," thus making a single leap from the Hamilton-Jacobi equation to the motion of the test object. Both methods are em-

(*continued on page 641*)

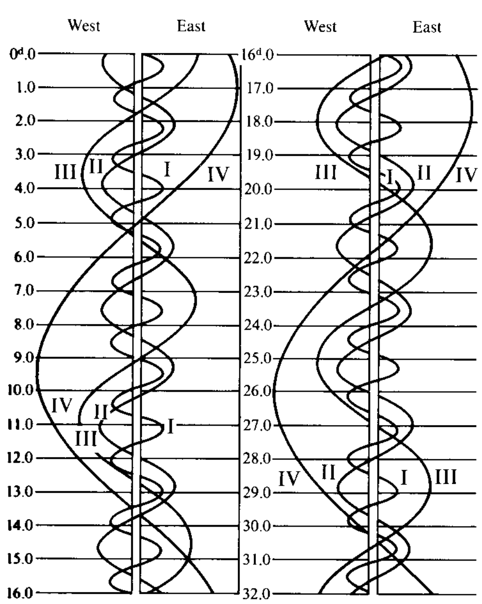


Figure 25.1.
Jupiter's satellites, as followed from night to night with field glasses or telescope, provide an opportunity to check for oneself the central ideas of gravitation physics in the Newtonian approximation (distances large compared to Schwarzschild radius). For the practically circular orbits of these satellites, Kepler's law becomes $M^1 = \omega^2 r^3$ ("1-2-3 principle") and the velocity in orbit is $\beta = \omega r$. Out of observations on any two of the quantities β, M, ω, r, one can find the other two. (In the opposite limiting case of two objects, each of mass M, going around their common center of gravity with separation r, one has $M = \omega^2 r^3/2$, $\beta = \omega r/2$). The configurations of satellites I–IV of Jupiter as given here for December 1964 (days 0.0, 1.0, 2.0, etc. in "universal time," for which see any good dictionary or encyclopedia) are taken from *The American Ephemeris and Nautical Almanac for 1964* [U.S. Government Printing Office (1962)].

Box 25.1 MASS FROM MEAN ANGULAR FREQUENCY AND SEMIMAJOR AXIS: $M = \omega^2 a^3$

Appropriateness of Newtonian analysis shown by smallness of mass (or "half-Schwarzschild radius" or "extension of the pit in the potential") as listed in last column compared to the semimajor axis a in the next-to-last column. Basic data from compilation of Allen (1963).

Object	*Period*[a] *(days)*	$\omega(cm^{-1})$	$a(cm)$	$\omega^2 a^3(cm)$
Planets				
Mercury	87.9686	275.8×10^{-19}	0.5791×10^{13}	1.477×10^5
Venus	224.700	107.95	1.0821	1.477
Earth	365.257	66.41	1.4960	1.477
Mars	686.980	35.31	2.2794	1.477
Jupiter	4332.587	5.599	7.783	1.478
Saturn	10759.20	2.255	14.27	1.477
Uranus	30685	0.7905	28.69	1.476
Neptune	60188	0.4030	44.98	1.478
Pluto	90700	0.2674×10^{-19}	59.00×10^{13}	1.469×10^5
Major satellites of Jupiter				
Io	1.769 138	13.711×10^{-16}	0.422×10^{11}	141.3
Europa	3.551 181	6.831	0.671	141.0
Ganymede	7.154 553	3.391	1.070	140.8
Callisto	16.689 018	1.454×10^{-16}	1.883×10^{11}	141.1
Two satellites of Earth				
OSO 5[b]	95.6 min.	3.65×10^{-14}	6.916×10^8	0.442
Moon	27.32	0.888×10^{-16}	3.84×10^{10}	0.446

[a] Sidereal period: time to make one revolution relative to fixed stars.
[b] Orbiting scientific observatory launched Jan. 22, 1969, to observe x-rays and ultraviolet radiation from the sun. Perigee 531 km, apogee 560 km, above earth.

SOME TYPICAL MASSES AND TIMES IN CONVENTIONAL AND GEOMETRIC UNITS. Conversion factor for mass, $G/c^2 = 0.742 \times 10^{-28}$ cm/g

Mass	*Galaxy*	*Sun*	*Jupiter*	*Earth*
M_{conv} (g)	2.2×10^{44}	1.989×10^{33}	1.899×10^{30}	5.977×10^{27}
M (cm)	1.6×10^{16}	1.47×10^5	112	0.444

Conversion factor for time, $c = 2.998 \times 10^{10}$ cm/sec. One sidereal year = 365.256 days or 3.1558×10^7 sec.

Period	*1 sec*	*1 min*	*1 hr*	*1 day*
ω_{conv} (sec^{-1})	6.28	1.046×10^{-1}	1.75×10^{-3}	7.28×10^{-5}
ω (cm^{-1})	2.09×10^{-10}	3.48×10^{-12}	5.80×10^{-14}	2.42×10^{-15}

	1 week	*1 month*	*1 year*
ω_{conv} (sec^{-1})	1.04×10^{-5}	2.39×10^{-6}	1.99×10^{-7}
ω (cm^{-1})	3.46×10^{-16}	7.95×10^{-17}	6.63×10^{-18}

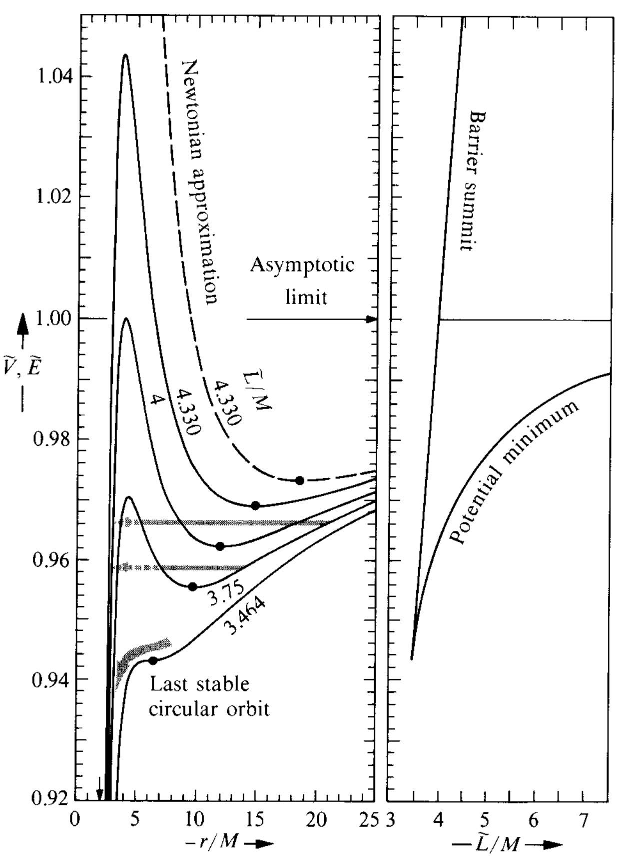


Figure 25.2.
Effective potential for motion of a test particle in the Schwarzschild geometry of a concentrated mass M. Energy, in units of the rest mass μ of the particle, is denoted $\tilde{E} = E/\mu$; angular momentum, $\tilde{L} = L/\mu$. The quantity r denotes the Schwarzschild r coordinate. The effective potential (also in units of μ) is defined by equation (25.16) or, equivalently, by the equation

$$\left(\frac{dr}{d\tau}\right)^2 + \tilde{V}^2(r) = \tilde{E}^2$$

(see also §25.5) and has the value

$$\tilde{V} = [(1 - 2M/r)(1 + \tilde{L}^2/r^2)]^{1/2}.$$

It represents that value of $\tilde{E}$ at which the radial kinetic energy of the particle, at r, reduces to zero ($\tilde{E}$-value that makes r into a "turning point": $\tilde{V}(r) = \tilde{E}$. Note that one could equally well regard $\tilde{V}^2(r)$ as the effective potential, and define a turning point by the condition $\tilde{V}^2 = \tilde{E}^2$. Which definition one chooses depends on convenience, on the intended application, on the tie to the archetypal differential equation $\frac{1}{2}\dot{x}^2 + V(x) = E$, and on the stress one wishes to put on correspondence with the effective potential of Newtonian theory). Stable circular orbits are possible (representative point sitting at minimum of effective potential) only for $\tilde{L}$ values in excess of $2\sqrt{3}\,M$. For any such fixed $\tilde{L}$ value, the motion departs slightly from circularity as the energy is raised above the potential minimum (see the two heavy horizontal lines for $\tilde{L} = 3.75\,M$). In classical physics, the motion is limited to the region of positive kinetic energy. In quantum physics, the particle can tunnel through the region where the kinetic energy, as calculated classically, is negative (dashed prolongations of heavy horizontal lines) and head for the "pit in the potential" (capture by black hole). Such tunneling is absolutely negligible when the center of attraction has any macroscopic dimension, but in principle becomes important for a black hole of mass 10^{17} g (or 10^{-11} cm) if such an object can in principle exist.

The diagram at the right gives values of the minimum and maximum of the potential as they depend on the angular momentum of the test particle. The roots of $\partial\tilde{V}/\partial r$ are given in terms of the "reduced angular momentum parameter" $L^\dagger = \tilde{L}/M = L/M\mu$ by

$$r = \frac{6M}{1 + (1 - 12/L^{\dagger 2})^{1/2}},$$

$$\tilde{E}^2 = \frac{(L^{\dagger 2} + 36) + (L^{\dagger 2} - 12)(1 - 12/L^{\dagger 2})^{1/2}}{54}$$

$[= 8/9$ for $L^\dagger = (12)^{1/2}$; 1 for $L^\dagger = 4$; $(L^{\dagger 2}/27) + (1/3) + (1/L^{\dagger 2}) + \cdots$ for $L^\dagger \longrightarrow \infty]$

(plus root for maximum of the effective potential; minus root for minimum; see exercise 25.18).

Box 25.2 MOTION IN SCHWARZSCHILD GEOMETRY REGARDED AS A CENTRAL POINT OF DEPARTURE FOR MAJOR APPLICATIONS OF EINSTEIN'S GEOMETRODYNAMICS

1. Newtonian effect of sun on planets and of earth on moon and man.
2. Bending of light by sun.
3. Red shift of light from sun.
4. Precession of the perihelion of Mercury around the sun.
5. Capture of a test object by a black hole as simple exemplar of gravitational collapse.
6. Dynamics of Friedmann universe derived from model of Schwarzschild "lattice universe." Lattice universe is constructed from 120 or some other magic number of concentrations of mass, each mass in an otherwise empty lattice cell of its own. Each lattice cell, though actually polygonal, is idealized (see Wigner-Seitz approximation of solid-state physics) as spherical. A test object at the interface between two cells falls toward the center of each [standard radial motion in Schwarzschild geometry; see discussion following equation (25.27). Therefore the two masses fall toward each other at a calculable rate. From this simple argument follows the entire dynamics of the closed 3-sphere lattice universe, in close concord with the predictions of the Friedmann model [see Lindquist and Wheeler (1957)].
7. Perturbations of Schwarzschild geometry, I. Gravitational waves are incident on, scattered by, and captured into a black hole. Waves with wavelength short compared to the Schwarzschild radius can be analyzed to good approximation by the methods of geometric optics (exercises 35.15 and 35.16), as employed in this chapter to treat the motions of particles and photons. For longer wavelengths, there occur important physical-optics corrections to this geometric-optics idealization (see §35.8 and exercises 32.10, 32.11). Similar considerations apply to electromagnetic and de Broglie waves.
8. Lepton number for an electron in its lowest quantum state in the geometry ("gravitational field of force") of a black hole is calculated to be transcended (capture of the electron!) or not according as the mass of this black hole is large or small compared to a certain critical mass $M_{*e} = M^{*2}/m_e$ ($\sim 10^{17}$ g or 10^{-11} cm) [Hartle (1971, 1972); Wheeler (1971b,c); Teitelboim (1972b,c)]. Similarly (with another value for the critical mass) for conservation of baryon number [Bekenstein (1972a,b), Teitelboim (1972a)]. To analyze "transcendence or not" one must solve quantum-mechanical wave equations, of which the Hamilton-Jacobi equation for particle and photon orbits is a classical limit. These quantum wave equations contain effective potentials identical—aside from spin-dependent and wavelength-dependent corrections—to the effective potentials for particle and photon motion.
9. Perturbations of Schwarzschild geometry, II. Those small changes in standard Schwarzschild black-hole geometry which remain stationary in time describe the alterations in a "dead" black hole that make it into a "live" black hole, one endowed with angular momentum as well as mass (see Chapter 33). To analyze such changes in black-hole geometry, one must again solve wave equations, but wave equations which are now classical. Once more the wave equations are closely related to the Hamilton-Jacobi equation, and their effective potentials are close kin to those for particle and photon motion.

ployed here in turn because each gives special insights. The Hamilton-Jacobi method (Box 25.3) leads quickly to the major results of interest (Box 25.4), and it has a close tie to the quantum principle. The world-line method (§§25.2, 25.3, 25.4) starts with the geodesic equations of motion themselves. It provides a more familiar way into the subject for a reader not acquainted with the Hamilton-Jacobi approach. Moreover, in attempting to solve the geodesic equations of motion, one must analyze symmetry properties of the geometry, an enterprise that continues to pay dividends when one moves from Schwarzschild geometry to Kerr-Newman geometry (Chapter 33), and from Friedmann cosmology (Chapter 27) to more general cosmologies (Chapter 30).

(continued on page 650)

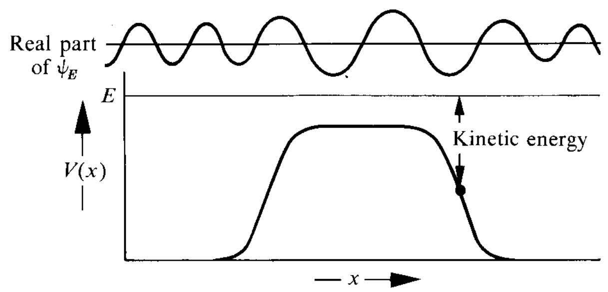
Box 25.3 THE HAMILTON-JACOBI DESCRIPTION OF MOTION: NATURAL BECAUSE RATIFIED BY THE QUANTUM PRINCIPLE

1. Purely classical (nonquantum).
2. Originated with William Rowan Hamilton out of conviction that mechanics is similar in its character to optics; that the "particle world line" of mechanics is an idealization analogous to the "light ray" of geometric optics. Localization of energy of light ray is approximate only. Its spread is governed by wavelength of light ("geometric optics"). Hamilton had glimmerings of same idea for particles: "quantum physics before quantum physics." The way that he and Jacobi developed to analyze motion through the Hamilton-Jacobi function $S(x, t)$—to take the example of a dynamic system with only one degree of freedom, x—makes the leap from classical ideas to quantum ideas as short as one knows how to make it. Moreover, the real world is a quantum world. Classical mechanics is not born out of a vacuum. It is an idealization of and approximation to quantum mechanics.
3. Key idea is idealization to a particle wavelength so short that quantum-mechanical spread or uncertainty in location of particle (or spread of configuration coordinates of more complex system) is negligible. No better way was ever discovered to unite the spirit of quantum mechanics and the precision of location of classical mechanics.
4. Call Hamiltonian $H(p, x) = p^2/2m + V(x)$. Call energy of particle E. Then there is no way whatever consistent with the quantum principle to describe the motion of the particle in space and time. The uncertainty principle forbids (sharply defined energy $\Delta E \longrightarrow 0$, in $\Delta E\, \Delta t \geq \hbar/2$, implies uncertainty $\Delta t \longrightarrow \infty$; also $\Delta p \longrightarrow 0$ in $\Delta p\, \Delta x \geq \hbar/2$ implies $\Delta x \longrightarrow \infty$). The quantum-mechanical wave function is spread out over all space. This spread shows in the so-called semi-

Box 25.3 *(continued)*

classical or Wentzel-Kramers-Brillouin ["WKB"; see, for example, Kemble (1937)] approximation for the probability amplitude function,

$$\psi_E(x, t) = \begin{pmatrix}\text{slowly varying} \\ \text{amplitude function}\end{pmatrix} e^{(i/\hbar)S_E(x,t)}. \tag{1}$$



5. It is of no help in localizing the probability distribution that $\hbar = 1.054 \times 10^{-27}$ g cm^2/s [or $\hbar = (1.6 \times 10^{-33}\text{ cm})^2$ in geometric units] is very small compared to the "quantities of action" or "magnitudes of the Hamilton-Jacobi function, S" or "dynamic phase, S" encountered in most everyday applications.

6. It is of no help in localizing the probability distribution that this dynamic phase obeys the simple Hamilton-Jacobi law of propagation,

$$-\frac{\partial S}{\partial t} = H\left(\frac{\partial S}{\partial x}, x\right) = \frac{1}{2m}\left(\frac{\partial S}{\partial x}\right)^2 + V(x). \tag{2}$$

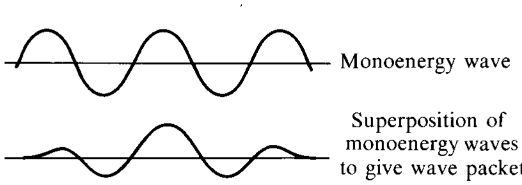
7. It is of no help in localizing the probability distribution that the solution of this equation for a particle of energy E is extraordinarily simple,

$$S(x, t) = -Et + \int_{x_0}^{x} \{2m[E - V(x)]\}^{1/2}\, dx + \delta_E \tag{3}$$

(with δ_E an arbitrary additive phase constant). The probability amplitude is still spread all over everywhere. There is not the slightest trace of anything like a localized world line, $x = x(t)$.

8. To localize the particle, build a probability-amplitude wave packet by superposing monofrequency (monoenergy) terms, according to a prescription qualitatively of the form

$$\psi(x, t) = \psi_E(x, t) + \psi_{E+\Delta E}(x, t) + \cdots \tag{4}$$

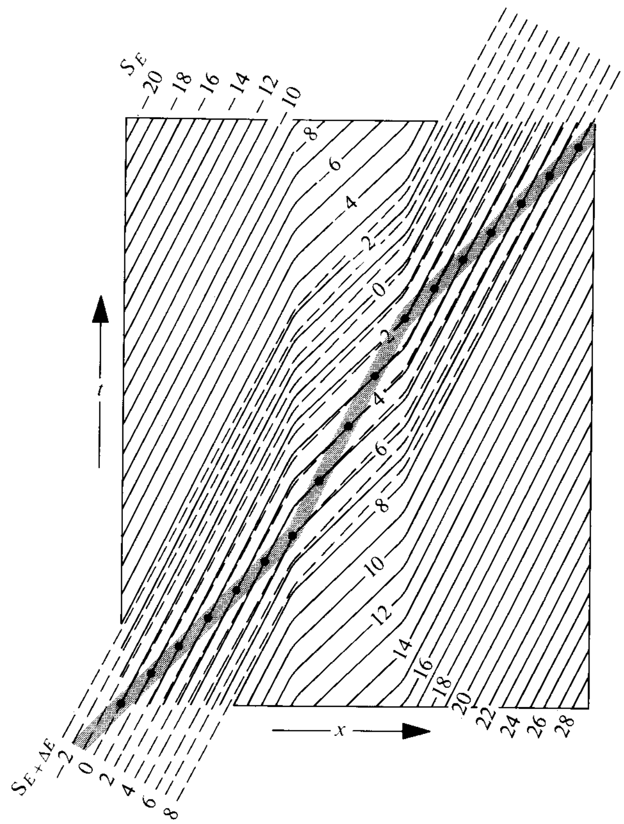


Destructive interference takes place almost everywhere. The wave packet is concentrated in the region of constructive interference. There the phases of the various waves agree; thus

$$S_E(x, t) = S_{E+\Delta E}(x, t). \tag{5}$$

At last one has moved from a wave spread everywhere to a localized wave and thence, in the limit of indefinitely small wavelength, to a classical world line. This one equation of constructive interference ties together x and t (locus of world line in x, t, diagram). Smooth lines -20, -19, -18, etc. are wave crests of ψ_E; dashed lines, wave crests for $\psi_{E+\Delta E}$. Shaded area is region of constructive interference (wave packet). Black dots mark locus of classical world line,

$$\lim_{\Delta E \to 0} \frac{S_{E+\Delta E}(x, t) - S_E(x, t)}{\Delta E} = 0.$$



9. The Newtonian course of the world line through spacetime follows at once from this condition of constructive interference when one goes to the classical limit ($\hbar$ negligible compared to amounts of action involved; hence wavelength negligibly short; hence spread of energies ΔE required to build well-localized wave packet also negligible); thus

$$\frac{S_{E+\Delta E}(x, t) - S_E(x, t)}{\Delta E} = 0$$

reduces to

$$\frac{\partial S_E(x, t)}{\partial E} = 0.$$

Box 25.3 (*continued*)

This condition in turn, applied to expression (3), gives the time required to travel to the point x; thus,

$$-t + \int_{x_0}^{x} \frac{dx}{\{(2/m)[E - V(x)]\}^{1/2}} + t_0 = 0,$$

where t_0 is an abbreviation for the quantity

$$t_0 = d\,\delta_E/dE$$

("difference in base value of dynamic phase per unit difference of energy").

10. Not one trace of the quantum of action comes into this final Newtonian result, for a simple reason: $\hbar$ has been treated as negligible and the wavelength has been treated as negligible. In this limit the location of the wave "packet" reduces to the location of the wave crest. The location of the wave crest is precisely what is governed by $S_E(x, t)$; and the condition of "constructive interference" $\partial S_E(x, t)/\partial E = 0$ gives without approximation the location of the sharply defined Newtonian world line $x = x(t)$.
11. Relevance in the context of motion in a central field of force? Quickest known route to the concept of effective potential (Box 25.4).

Box 25.4 MOTION UNDER GRAVITATIONAL ATTRACTION OF A CENTRAL MASS ANALYZED BY HAMILTON-JACOBI METHOD

A. Newtonian Theory of Gravitation

Hamiltonian
$$\widetilde{H} = \frac{\widetilde{p}_r^{\,2}}{2} + \frac{\widetilde{p}_\theta^{\,2}}{2r^2} + \frac{\widetilde{p}_\phi^{\,2}}{2r^2 \sin^2\theta} - \frac{M}{r} \tag{1}$$

(tildes over energy, momentum, etc., refer to test object of unit mass; test particle of mass μ follows same motion with energy $E = \mu\widetilde{E}$, momentum $\boldsymbol{p} = \mu\widetilde{\boldsymbol{p}}$, etc.).

Equation of Hamilton-Jacobi for propagation of wave crests:

$$-\frac{\partial \widetilde{S}}{\partial t} = \frac{1}{2}\left(\frac{\partial \widetilde{S}}{\partial r}\right)^2 + \frac{1}{2r^2}\left(\frac{\partial \widetilde{S}}{\partial \theta}\right)^2 + \frac{1}{2r^2 \sin^2\theta}\left(\frac{\partial \widetilde{S}}{\partial \phi}\right)^2 - \frac{M}{r}. \tag{2}$$

Box 25.4 ***(continued)***

Solve by "method of separation of variables" with convention that $\sqrt{a^2} \equiv \pm a$,

$$\begin{aligned}\tilde{S} = -\tilde{E}t + \tilde{p}_\phi \phi + \int^\theta \left(\tilde{L}^2 - \frac{\tilde{p}_\phi{}^2}{\sin^2\theta}\right)^{1/2} d\theta \\ + \int^r \left[2\left(\tilde{E} + \frac{M}{r} - \frac{\tilde{L}^2}{2r^2}\right)\right]^{1/2} dr + \delta_{\tilde{p}_\phi, \tilde{L}, \tilde{E}}.\end{aligned} \tag{3}$$

(Check by substituting into Hamilton-Jacobi equation. Solution as *sum* of four terms corresponding to the four independent variables goes hand in hand with expression of probability amplitude in quantum mechanics as *product* of four factors, because $iS/\hbar = i\mu\tilde{S}/\hbar$ is exponent in approximate expression for the probability amplitude.)

Constructive interference of waves:

(1) with slightly different $\tilde{E}$ values (impose "condition of constructive interference" $\partial \tilde{S}_{\tilde{p}_\phi, \tilde{L}, \tilde{E}}(t, r, \theta, \phi)/\partial \tilde{E} = 0$) tells when the particle arrives at a given r (that is, gives relation between t and r);

(2) with slightly different values of the "parameter of total angular momentum per unit mass," $\tilde{L}$ (impose condition of constructive interference $\partial \tilde{S}_{\tilde{p}_\phi, \tilde{L}, \tilde{E}}(t, r, \theta, \phi)/\partial \tilde{L} = 0$) tells correlation between θ and r (a major feature of the shape of the orbit);

(3) with slightly different values of the "parameter of azimuthal angular momentum per unit mass," $\tilde{p}_\phi$ (impose condition $\partial \tilde{S}/\partial \tilde{p}_\phi = 0$) gives correlation between θ and ϕ,

$$0 = \frac{\partial \tilde{S}}{\partial \tilde{p}_\phi} = \phi - \int^\theta \frac{(\tilde{p}_\phi/\tilde{L})\, d\theta}{\sin\theta\, (\sin^2\theta - \tilde{p}_\phi{}^2/\tilde{L}^2)^{1/2}} \tag{4}$$

Planar character of the orbit.

Puzzle out the value of this last integral with the help of a table of integrals? It is quicker and clearer to capture the content without calculation: the particle moves in a plane. The vector associated with the angular momentum $\tilde{L}$ stands perpendicular to this plane. The projection of this angular momentum along the z-axis is $\tilde{p}_\phi = \tilde{L} \cos\alpha$ (definition of orbital inclination, α). Straight line connecting origin with particle cuts unit sphere in a point $\mathscr{P}$. As time runs on, $\mathscr{P}$ traces out a great circle on the unit sphere. The plane of this great circle cuts the equatorial plane in a "line of nodes," at which "hinge-line" the two planes are separated by a dihedral angle, α. The orbit of the point $\mathscr{P}$ is described by $\hat{x} = r\cos\psi$, $\hat{y} = r\sin\psi$, $\hat{z} = 0$ in a Cartesian system of coordinates in which $\hat{y}$ runs along the line of nodes and in which $\hat{x}$ lies in the plane of the *orbit*.

Box 25.4 (*continued*)

In a coordinate system in which y runs along the line of nodes and x lies in the plane of the *equator,* one has:

$$r\cos\theta = z = \hat{z}\cos\alpha + \hat{x}\sin\alpha = r\cos\psi\sin\alpha;$$

$$r\sin\theta\cos\phi = x = -\hat{z}\sin\alpha + \hat{x}\cos\alpha = r\cos\psi\cos\alpha;$$

$$r\sin\theta\sin\phi = y = \hat{y} = r\sin\psi.$$

Eliminate reference to the Cartesian coordinates and, by taking ratios, also eliminate reference to r. Thus find the equation of the great circle route in parametric form,

$$\tan\phi = \tan\psi/\cos\alpha$$

and

$$\cos\theta = \cos\psi\sin\alpha.$$

Here increasing values of ψ spell out successive points on the great circle. Eliminate ψ via the relation

$$\sec^2\psi - \tan^2\psi = 1$$

to find

$$\frac{\sin^2\alpha}{\cos^2\theta} - \tan^2\phi\cos^2\alpha = 1$$

or, more briefly,

$$\sec\phi = \tan\alpha\tan\theta. \tag{5}$$

One verifies that ϕ as calculated from (5) provides an integral of (4), thus confirming the physical argument just traced out. Moreover, the arbitrary constant of integration that comes from (4), left out for the sake of simplicity from (5), is easily inserted by replacing ϕ there by $\phi - \phi_0$ (rotation of line of nodes to a new azimuth). The kind of physics just done in tracing out the relation between θ and ϕ is evidently elementary solid geometry and nothing more. The same geometric relationships also show up, with no relativistic corrections whatsoever (how could there be any?!) for motion in Schwarzschild geometry. Therefore it is appropriate to drop this complication from attention here and hereafter. Let the particle move entirely in the direction of increasing θ, not at all in the direction of increasing ϕ; that is, let it move in an orbit of zero angular momentum $\tilde{p}_\phi$ (total angular momentum vector $\tilde{L}$ inclined at angle $\alpha = \pi/2$ to z-axis). Consequently the dynamic phase S (to be divided by $\hbar$ to obtain phase of Schrödinger wave function when one turns from classical to quantum mechanics) becomes

$$\tilde{S} = -\tilde{E}t + \tilde{L}\theta + \int^r \left[2\left(\tilde{E} + \frac{M}{r} - \frac{\tilde{L}^2}{2r^2}\right)\right]^{1/2} dr + \delta_{\tilde{L},\tilde{E}}. \tag{6}$$

Shape of orbit:

$$0 = \frac{\partial \widetilde{S}}{\partial \widetilde{L}} = \theta - \int^r \frac{\widetilde{L}\, dr/r^2}{[2(\widetilde{E} + M/r - \widetilde{L}^2/2r^2)]^{1/2}}, \tag{7}$$

whence

$$r = \frac{\widetilde{L}^2/M}{1 + e\cos\theta}. \tag{8}$$

Here e is an abbreviation for the eccentricity of the orbit,

$$e = (1 + 2\widetilde{E}\widetilde{L}^2/M^2)^{1/2} \tag{9}$$

(greater than 1 for positive $\widetilde{E}$, hyperbolic orbit; equal to 1 for zero $\widetilde{E}$, parabolic orbit; less than 1 for negative $\widetilde{E}$, elliptic orbit). A constant of integration has been omitted from (8) for simplicity. To reinstall it, replace θ by $\theta - \theta_0$ (rotation of direction of principal axis in the plane of the orbit). Other features of the orbit:

(semimajor axis of orbit when elliptic)
$$a = \frac{\widetilde{L}^2/M}{1 - e^2} = \frac{M}{(-2\widetilde{E})}; \tag{10}$$

(semiminor axis of orbit when elliptic)
$$b = \frac{\widetilde{L}^2/M}{(1 - e^2)^{1/2}} = \frac{\widetilde{L}}{(-2\widetilde{E})^{1/2}}; \tag{11}$$

("impact parameter" for hyperbolic orbit, or "distance of closest approach in absence of deflection")
$$\begin{aligned} b &= \frac{(\text{angular momentum per unit mass})}{(\text{linear momentum per unit mass})} \\ &= \frac{\widetilde{L}}{(2\widetilde{E})^{1/2}}; \end{aligned} \tag{12}$$

(actual distance of closest approach)
$$r_{\min} = \frac{\widetilde{L}^2/M}{(1 + 2\widetilde{E}\widetilde{L}^2/M^2)^{1/2} + 1}; \tag{13}$$

(angle of deflection in hyperbolic orbit)
$$\begin{aligned} \Theta &= \pi - 2 \arccos(1/e) \\ &= 2 \arctan[M/(2\widetilde{E})^{1/2}\widetilde{L}] \\ &= 2 \arctan[M/2\widetilde{E}b]; \end{aligned} \tag{14}$$

(differential scattering cross section)
$$\begin{aligned} \frac{d\sigma}{d\Omega} &= \frac{2\pi b\, db}{2\pi \sin\Theta\, d\Theta} \\ &= \frac{M^2}{(4\widetilde{E}\sin^2\Theta/2)^2} \text{ (Rutherford).} \end{aligned} \tag{15}$$

Box 25.4 (*continued*)

Time as correlated with position:

$$0 = \frac{\partial \widetilde{S}}{\partial \widetilde{E}} = -t + \int^r \frac{dr}{\left[2\left(\widetilde{E} + \dfrac{M}{r} - \dfrac{\widetilde{L}^2}{2r^2}\right)\right]^{1/2}}. \tag{16}$$

Write

$$r = \frac{M}{(-2\widetilde{E})}(1 - e\cos u) \tag{17}$$

to simplify the integration. Get

$$t = \frac{M}{(-2\widetilde{E})^{3/2}}(u - e\sin u), \tag{18}$$

$$\begin{pmatrix}\text{mean circular}\\ \text{frequency}\end{pmatrix} = \frac{2\pi}{(\text{period})} = \omega = \frac{(-2\widetilde{E})^{3/2}}{M} = \left(\frac{M}{a^3}\right)^{1/2}. \tag{19}$$

Here u is the so-called "mean eccentric anomaly" (Bessel's time parameter). In terms of this quantity, one has also:

$$\sin u = \frac{(1 - e^2)^{1/2}\sin\theta}{1 + e\cos\theta};$$

$$\cos u = \frac{\cos\theta + e}{1 + e\cos\theta};$$

$$\cos\theta = \frac{\cos u - e}{1 - e\cos u};$$

$$\sin\theta = \frac{(1 - e^2)^{1/2}\sin u}{1 - e\cos u};$$

$$x = r\cos\theta = \frac{M}{(-2\widetilde{E})}(\cos u - e); \tag{20}$$

$$y = r\sin\theta = \frac{\widetilde{L}}{(-2\widetilde{E})^{1/2}}\sin u. \tag{21}$$

These expressions lend themselves to Fourier analysis into harmonic functions of the time, with coefficients that are standard Bessel functions:

$$J_n(z) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(z\sin u - nu)}\,du; \tag{22}$$

$$x/a = -\frac{3}{2}e + \sum_{\substack{k=-\infty\\ k\neq 0}}^{+\infty} k^{-1}J_{k-1}(ke)\cos k\omega t; \tag{23}$$

$$y/a = (1 - e^2)^{1/2}\sum_{\substack{k=-\infty\\ k\neq 0}}^{+\infty} k^{-1}J_{k-1}(ke)\sin k\omega t \tag{24}$$

[for these and further formulas of this type, see, for example, Wintner (1941), Siegel (1956), and Siegel and Moser (1971)]. Via such Fourier analysis one is in a position to calculate the intensity of gravitational radiation emitted at the fundamental circular frequency ω and at the overtone frequencies (see Chapter 36).

B. Einstein's Geometric Theory of Gravitation

Connection between energy and momentum for a test particle of rest mass μ traveling in curved space,

$$g^{\alpha\beta}p_\alpha p_\beta + \mu^2 = 0. \tag{25}$$

Gravitation shows up in no way other than in curvature of the geometry, in which the particle moves as free of all "real" force. Refer all quantities to basis of a test object of unit rest mass by dealing throughout with $\tilde{\boldsymbol{p}} = \boldsymbol{p}/\mu$. Also write $\tilde{p}_\alpha = \partial\tilde{S}/\partial x^\alpha$. Thus Hamilton-Jacobi equation for propagation of wave crests in Schwarzschild geometry (external field of a star; §23.6) becomes

$$-\frac{1}{(1-2M/r)}\left(\frac{\partial\tilde{S}}{\partial t}\right)^2 + (1-2M/r)\left(\frac{\partial\tilde{S}}{\partial r}\right)^2 + \frac{1}{r^2}\left(\frac{\partial\tilde{S}}{\partial\theta}\right)^2 + \frac{1}{r^2\sin^2\theta}\left(\frac{\partial\tilde{S}}{\partial\phi}\right)^2 + 1 = 0. \tag{26}$$

Solve Hamilton-Jacobi equation. As in Newtonian problem, simplify by eliminating all motion in direction of increasing ϕ. Thus set $0 = \tilde{p}_\phi = \partial\tilde{S}/\partial\phi$ (dynamic phase independent of ϕ) and have

$$\tilde{S} = -\tilde{E}t + \tilde{L}\theta + \int^r [\tilde{E}^2 - (1-2M/r)(1+\tilde{L}^2/r^2)]^{1/2}\frac{dr}{(1-2M/r)}. \tag{27}$$

Find shape of orbit by "principle of constructive interference"; thus,

$$0 = \frac{\partial\tilde{S}}{\partial\tilde{L}} = \theta - \int^r \frac{\tilde{L}\,dr/r^2}{[\tilde{E}^2 - (1-2M/r)(1+\tilde{L}^2/r^2)]^{1/2}}. \tag{28}$$

[See equation (25.41) and associated discussion in text; also Figure 25.6.]

Find time to get to given r by considering "interference of wave crests" belonging to slightly different $\tilde{E}$ values:

$$0 = \frac{\partial\tilde{S}}{\partial\tilde{E}} = -t + \int^r \frac{\tilde{E}}{[\tilde{E}^2 - (1-2M/r)(1+\tilde{L}^2/r^2)]^{1/2}}\frac{dr}{(1-2M/r)}. \tag{29}$$

[See equation (25.32) and associated discussion in text; also Figure 25.5 and exercise 25.15.]

§25.2. SYMMETRIES AND CONSERVATION LAWS

From symmetries to conservation laws by:
(1) Lagrangian or Hamiltonian approach
(2) Killing-vector approach

In analytic mechanics, one knows that symmetries of a Lagrangian or Hamiltonian result in conservation laws. Exercises 25.1 to 25.4 describe how these general principles are used to deduce, from the symmetries of Schwarzschild spacetime, constants of motion for the trajectories (geodesics) of freely falling particles in the gravitational field outside a star. The same constants of motion are obtained in a different language in differential geometry, where a "Killing vector" is the standard tool for the description of symmetry. This section considers the general question of metric symmetries before proceeding to Schwarzschild spacetime.

Killing vector, $\boldsymbol{\xi}$, defined

Let the metric components $g_{\mu\nu}$ relative to some coordinate basis $\boldsymbol{d}x^\alpha$ be independent of one of the coordinates x^K, so

$$\partial g_{\mu\nu}/\partial x^\alpha = 0 \text{ for } \alpha = K. \tag{25.1}$$

This relation possesses a geometric interpretation. Any curve $x^\alpha = c^\alpha(\lambda)$ can be translated in the x^K direction by the coordinate shift $\Delta x^K = \varepsilon$ to form a "congruent" (equivalent) curve:

$$x^\alpha = c^\alpha(\lambda) \text{ for } \alpha \neq K \text{ and } x^K = c^K(\lambda) + \varepsilon.$$

Let the original curve run from $\lambda = \lambda_1$ to $\lambda = \lambda_2$ and have length

$$L = \int_{\lambda_1}^{\lambda_2} [g_{\mu\nu}(x(\lambda))(dx^\mu/d\lambda)(dx^\nu/d\lambda)]^{1/2}\, d\lambda.$$

Then the displaced curve has length

$$L(\varepsilon) = \int_{\lambda_1}^{\lambda_2} \left[\left\{g_{\mu\nu}(x(\lambda)) + \varepsilon\frac{\partial g_{\mu\nu}}{\partial x^K}\right\}(dx^\mu/d\lambda)(dx^\nu/d\lambda)\right]^{1/2} d\lambda.$$

But the coefficient of ε in the integrand is zero. Therefore the length of the new curve is identical to the length of the original curve: $dL/d\varepsilon = 0$.

The vector

$$\boldsymbol{\xi} \equiv d/d\varepsilon = (\partial/\partial x^K) \tag{25.2}$$

provides an infinitesimal description of these length-preserving "translations." It is called a "Killing vector." It satisfies Killing's equation*

Killing's equation derived

$$\xi_{\mu;\nu} + \xi_{\nu;\mu} = 0 \tag{25.3}$$

(condition on the vector field $\boldsymbol{\xi}$ necessary and sufficient to ensure that all lengths are preserved by the displacement $\varepsilon\boldsymbol{\xi}$). This condition is expressed in covariant form.

*Historical note: Wilhelm K. J. Killing, born May 10, 1847, in Burbach, Westphalia, died February 11, 1923 in Münster, Westphalia; Professor of Mathematics at the University of Münster, 1892–1920. The key article that gives the name "Killing vector" to the kind of isometries considered here appeared almost a century ago [Killing (1892)].

Therefore it is enough to establish it in the preferred coordinate system of (25.1) in order to have it hold in every coordinate system. In that preferred coordinate system, the vector field, according to (25.2), has components

$$\xi^\mu = \delta^\mu{}_K.$$

Therefore the covariant derivative of this vector field has components

$$\begin{aligned}\xi_{\mu;\nu} = g_{\mu\alpha}\xi^\alpha{}_{;\nu} &= g_{\mu\alpha}\left(\frac{\partial\xi^\alpha}{\partial x^\nu} + \Gamma^\alpha_{\nu\sigma}\xi^\sigma\right)\\ &= g_{\mu\alpha}\Gamma^\alpha_{\nu K} = \Gamma_{\mu\nu K} = \frac{1}{2}\left(\frac{\partial g_{\mu K}}{\partial x^\nu} + \frac{\partial g_{\mu\nu}}{\partial x^K} - \frac{\partial g_{\nu K}}{\partial x^\mu}\right) \qquad (25.4)\\ &= \frac{1}{2}(g_{\mu K,\nu} - g_{\nu K,\mu}).\end{aligned}$$

One sees that $\xi_{\mu;\nu}$ is antisymmetric in the labels μ and ν, as stated in Killing's equation (25.3).

The geometric significance of a Killing vector is spelled out in Box 25.5.

Conservation of $\boldsymbol{p}\cdot\boldsymbol{\xi}$ for geodesic motion

From Killing's equation, $\xi_{(\mu;\nu)} = 0$, and from the geodesic equation $\nabla_{\boldsymbol{p}}\boldsymbol{p} = 0$ for the tangent vector $\boldsymbol{p} = d/d\lambda$ to any geodesic, follows an important theorem: *In any geometry endowed with a symmetry described by a Killing vector field $\boldsymbol{\xi}$, motion along any geodesic whatsoever leaves constant the scalar product of the tangent vector with the Killing vector:*

$$p_K = \boldsymbol{p}\cdot\boldsymbol{\xi} = \text{constant}. \qquad (25.5)$$

In verification of this result, evaluate the rate of change of the constant p_K along the course of the typical geodesic (affine parameter λ; result therefore as applicable to light rays—with zero lapse of proper time—as to particles); thus,

$$dp_K/d\lambda = (p^\mu\xi_\mu)_{;\nu}p^\nu = (p^\mu{}_{;\nu}p^\nu)\xi_\mu + p^{(\mu}p^{\nu)}\xi_{[\mu;\nu]} = 0. \qquad (25.6)$$

Turn back from a general coordinate system to the coordinates of (25.1), where the Killing vector field of the symmetry lets itself be written $\boldsymbol{\xi} = \partial/\partial x^K$. Then the scalar product of (25.5) becomes constant $\equiv p_\alpha\xi^\alpha = p_\alpha\delta^\alpha{}_K = p_K$. *The symmetry of the geometry guarantees the conservation of the K-th covariant coordinate-based component of the momentum.*

On a timelike geodesic in spacetime, the momentum of a test particle of mass μ is

$$\boldsymbol{p} \equiv d/d\lambda = \mu\boldsymbol{u} = \mu d/d\tau. \qquad (25.7)$$

Thus the affine parameter λ most usefully employed in the above analysis, when it is concerned with a particle, is not proper time τ but rather the ratio $\lambda = \tau/\mu$.

Terminology: "cyclic coordinate," "conjugate momentum"

When the metric $g_{\mu\nu}$ is independent of a coordinate x^K, that coordinate is called cyclic, and the corresponding conserved quantity, p_K, is called the "momentum conjugate to that cyclic coordinate" in a terminology borrowed from nonrelativistic mechanics.

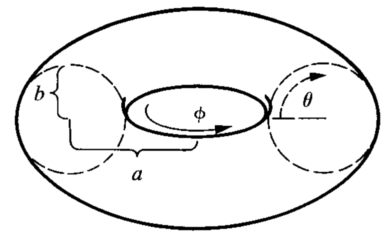
(continued on page 654)

Box 25.5 KILLING VECTORS AND ISOMETRIES (Illustrated by a Donut)

A. On a given manifold (e.g., spacetime, or the donut pictured here), in a given coordinate system, the metric components are independent of a particular coordinate x^K. Example of donut:

$$ds^2 = b^2\, d\theta^2 + (a - b\cos\theta)^2\, d\phi^2;$$

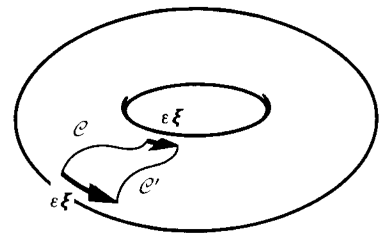
$$g_{\mu\nu} \text{ independent of } x^K = \phi.$$



B. Translate an arbitrary curve $\mathscr{C}$ through the infinitesimal displacement

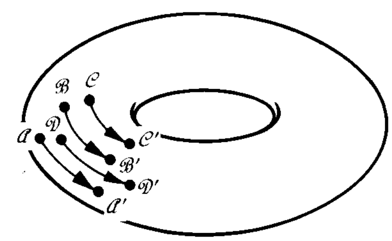
$$\varepsilon\boldsymbol{\xi} \equiv \varepsilon(\partial/\partial x^K) = \varepsilon(\partial/\partial\phi), \qquad \varepsilon \ll 1$$

to form a new curve $\mathscr{C}'$. In coordinate language $\mathscr{C}$ is $\theta = \theta(\lambda)$, $\phi = \phi(\lambda)$; while $\mathscr{C}'$ is $\theta = \theta(\lambda)$, $\phi = \phi(\lambda) + \varepsilon$. (Translation of all points through $\Delta\phi = \varepsilon$.) Because $\partial g_{\mu\nu}/\partial\phi = 0$, the curves $\mathscr{C}$ and $\mathscr{C}'$ have the same length (see text).



C. Pick a set of neighboring points $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$, $\mathscr{D}$; and translate each of them through $\varepsilon\boldsymbol{\xi}$ to obtain points $\mathscr{A}'$, $\mathscr{B}'$, $\mathscr{C}'$, $\mathscr{D}'$. Since the length of every curve is preserved by this translation, the distances between neighboring points are also preserved:

(distance between $\mathscr{A}'$ and $\mathscr{B}'$) =
(distance between $\mathscr{A}$ and $\mathscr{B}$).



But geometry is equivalent to a table of all infinitesimal distances (see Box 13.1). *Thus the geometry of the manifold is left completely unchanged by a translation of all points through* $\varepsilon\boldsymbol{\xi}$. [This is the coordinate-free version of the statement $\partial g_{\mu\nu}/\partial\phi = 0$.] One says that $\boldsymbol{\xi} = \partial/\partial\phi$ *is the generator of an "isometry"* (*or "group of motions"*) *on the manifold.*

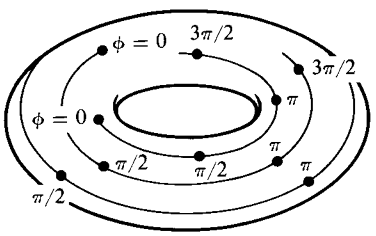
D. In general (see text), a vector field $\boldsymbol{\xi}(\mathscr{P})$ generates an isometry if and only if it satisfies Killing's equation $\xi_{(\alpha;\beta)} = 0$.

E. If $\boldsymbol{\xi}(\mathscr{P})$ generates an isometry (i.e. if $\boldsymbol{\xi}$ is a "Killing vector"), then the curves

$$\mathscr{P}(\underset{\substack{\uparrow\\ \left[\substack{\text{parameter}\\ \text{on curve}}\right]}}{x^K}, \underbrace{\alpha_1, \ldots, \alpha_n}_{\substack{\uparrow\\ \left[\substack{\text{labels to tell}\\ \text{"which" curve}}\right]}})$$

to which $\boldsymbol{\xi}$ is tangent [$\boldsymbol{\xi} = (\partial\mathscr{P}/\partial x^K)_{\alpha_1,\ldots,\alpha_n}$] are called "trajectories of the isometry."

Three different trajectories on a donut



Parameter on trajectories is $x^K = \phi$

F. The geometry is invariant under a translation of all points of the manifold through the same Δx^K along these trajectories [$\mathscr{P}(x^K, \alpha_1, \ldots, \alpha_n) \longrightarrow \mathscr{P}(x^K + \Delta x^K, \alpha_1, \ldots, \alpha_n)$; "finite motion" built up from many "infinitesimal motions" $\varepsilon\boldsymbol{\xi}$.]

G. This isometry is described in physical terms as follows. Station a family of observers throughout the manifold. Have each observer report to a central computer (1) all aspects of the manifold's geometry near him, and (2) the distances and directions to all neighboring observers (directions relative to "preferred" directions that are determined by anisotropies in the local geometry). Through each observer's position passes a unique trajectory of the isometry. Move each observer through the same fixed Δx^K (e.g., $\Delta x^K = 17$) along his trajectory, leaving the manifold itself unchanged. Then have each observer report to the central computer the same geometric information as before his motion. The information received by the computer after the motion will be identical to that received before the motion. There is no way whatsoever, by geometric measurements, to discover that the motion has occurred! This is the significance of an isometry.

EXERCISES

Exercise 25.1. CONSTANT OF MOTION OBTAINED FROM HAMILTON'S PRINCIPLE

Prove the above theorem of conservation of $p_K \equiv \boldsymbol{p} \cdot \boldsymbol{\xi}$ from Hamilton's principle (Box 13.3)

$$\delta \int \frac{1}{2} g_{\mu\nu}(x)(dx^\mu/d\lambda)(dx^\nu/d\lambda)\, d\lambda = 0 \tag{25.8}$$

as applied to geodesic paths. Recall: In this action principle, $g_{\mu\nu}$ is to be regarded as a known function of position, x, along the path; and the path itself, $x^\mu(\lambda)$, is to be varied.

Exercise 25.2. SUPER-HAMILTONIAN FORMALISM FOR GEODESIC MOTION

Show that a set of differential equations in Hamiltonian form results from varying p_μ and x^μ independently in the variational principle $\delta I = 0$, where

$$I = \int (p_\mu\, dx^\mu - \mathcal{H}\, d\lambda) \tag{25.9}$$

and

$$\mathcal{H} \equiv \frac{1}{2} g^{\mu\nu}(x) p_\mu p_\nu. \tag{25.10}$$

Show that the "super-Hamiltonian" $\mathcal{H}$ is a constant of motion, and that the solutions of these equations are geodesics. What do the choices $\mathcal{H} = +\frac{1}{2}$, $\mathcal{H} = 0$, $\mathcal{H} = -\frac{1}{2}\mu^2$, or $\mathcal{H} = -\frac{1}{2}$ mean for the geodesic and its parametrization?

Exercise 25.3. KILLING VECTORS IN FLAT SPACETIME

Find ten Killing vectors in flat Minkowski spacetime that are linearly independent. (Restrict attention to linear relationships with constant coefficients).

Exercise 25.4. POISSON BRACKET AS KEY TO CONSTANTS OF MOTION

If $\boldsymbol{\xi}$ is a Killing vector, show that $p_K \equiv \xi^\mu p_\mu$ commutes (has vanishing Poisson bracket) with the Hamiltonian $\mathcal{H}$ of the previous problem, $[\mathcal{H}, p_K] = 0$, so $dp_K/d\lambda = 0$. (Hint: Use a convenient coordinate system.)

Exercise 25.5. COMMUTATOR OF KILLING VECTORS IS A KILLING VECTOR

Consider two Killing vectors, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, which happen not to commute [as differential operators; i.e., the commutator of equations (8.13) does not vanish; consider rotations about perpendicular directions as a case in point]; thus,

$$[\boldsymbol{\xi}, \boldsymbol{\eta}] \equiv \boldsymbol{\zeta} \neq 0.$$

(a) Show that no single coordinate system can be simultaneously adapted, in the sense of equation (25.1), to both the $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ symmetries (see exercise 9.9).

(b) Let $p_\xi = p_\mu \xi^\mu$, $p_\eta = p_\mu \eta^\mu$, and $p_\zeta = p_\mu \zeta^\mu$, and derive the Poisson-bracket relationship $[p_\xi, p_\eta] = -p_\zeta$. In a geometry, the symmetries of which are related in this way, show that p_ζ is also a constant of motion.

(c) In a coordinate system where $\boldsymbol{\zeta} = (\partial/\partial x^K)$, define $\mathcal{H}$ as in (25.10) and show from $[\mathcal{H}, p_\zeta] = 0$ that $\boldsymbol{\zeta}$ is a Killing vector.

Thus the commutator of two Killing vectors is itself a Killing vector.

Exercise 25.6. EIGENVALUE PROBLEM FOR KILLING VECTORS

Show that any Killing vector satisfies $\xi^{\mu}{}_{;\mu} = 0$, and is an eigenvector with eigenvalue $\kappa = 0$ of the equation

$$\xi^{\mu;\nu}{}_{;\nu} + R^{\mu}{}_{\nu}\xi^{\nu} = \kappa\xi^{\mu}. \tag{25.11}$$

Find a variational principle (Raleigh-Ritz type) for this eigenvalue equation.

---

§25.3. CONSERVED QUANTITIES FOR MOTION IN SCHWARZSCHILD GEOMETRY

Consider a test particle moving in the Schwarzschild geometry, described by the line element

$$ds^2 = -(1 - 2M/r)\,dt^2 + \frac{dr^2}{(1 - 2M/r)} + r^2(d\theta^2 + \sin^2\theta\,d\phi^2). \tag{25.12}$$

Why attention focuses on particle orbits around a black hole

This expression for the geometry applies outside any spherically symmetric center of attraction of total mass-energy M. It makes no difference, for the motion of the particle outside, what the geometry is inside, because the particle never gets there; before it can, it collides with the surface of the star—if the center of attraction is a star, that is to say, a fluid mass in hydrostatic equilibrium. At each point throughout such an equilibrium configuration, the Schwarzschild coordinate r exceeds the local value of the quantity $2m(r)$; see §23.8. Therefore the Schwarzschild coordinate R of the surface exceeds $2M$. Consequently, expression (25.12) applies outside any equilibrium configuration, no matter how compact ($r > R > 2M$ implies that one need not face the issue of the "singularity" at $r = 2M$). The more compact the configuration, however, the more of the Schwarzschild geometry the test particle can explore. The ideal limit is not a star in hydrostatic equilibrium. It is a star that has undergone complete gravitational collapse to a black hole. Then (25.12) applies arbitrarily close to $r = 2M$. This idealization is assumed here ("black hole"), because the analysis can then cover the maximum range of possible situations.

Wherever the test particle lies, and however fast it moves, project that point and project that 3-velocity radially onto a sphere of some fixed r value, say, the unit sphere $r = 1$. The point and the vector together define a point and a vector on the surface of the unit sphere; and they in turn mark the beginning and define the totality of a great circle. As the particle continues on its way, the radial projection of its position will continue to lie on that great circle. To depart from the great circle on one side or the other would be to give preference to the one hemisphere or the other of the unit sphere, contrary to the symmetry of the situation.

Choice of coordinates to make particle orbit lie in "equator," $\theta = \pi/2$

Orient the coordinate system so that the radial projection of the orbit coincides with the equator, $\theta = \pi/2$, of the polar coordinates (see Box 25.4 for the spherical trigonometry of a more general orientation of the orbit, and for eventual specialization to a polar orbit, in contrast to the equatorial orbit considered here). In polar coordinates as so oriented, the particle has at the start, and continues to have, zero momentum in the θ direction; thus,

$$p^\theta = d\theta/d\lambda = 0.$$

Conserved quantities for particle motion:

(1) E
(2) L

The expression (25.12) for the line element shows that the geometry is unaffected by the translations $t \longrightarrow t + \Delta t$, $\phi \longrightarrow \phi + \Delta\phi$. Thus the coordinates t and ϕ are "cyclic." The conjugate momenta $p_0 \equiv -E$ and $p_\phi \equiv \pm L\ (L \geq 0)$ are therefore conserved. This circumstance allows one immediately to deduce the major features of the motion, as follows.

The magnitude of the 4-vector of energy-momentum is given by the rest mass of the particle,

(3) μ

$$g_{\alpha\beta}p^\alpha p^\beta + \mu^2 = g^{\alpha\beta}p_\alpha p_\beta + \mu^2 = 0 \tag{25.13}$$

or

$$-\frac{E^2}{(1-2M/r)} + \frac{1}{(1-2M/r)}\left(\frac{dr}{d\lambda}\right)^2 + \frac{L^2}{r^2} + \mu^2 = 0. \tag{25.14}$$

Moreover, one knows from the equivalence principle that test particles follow the same world line regardless of mass. Therefore what is relevant for the motion of particles is not the energy and angular momentum themselves, but the ratios

(4) $\tilde{E} \equiv E/\mu$

(5) $\tilde{L} \equiv L/\mu$

$$\tilde{E} = E/\mu = \begin{pmatrix}\text{energy per unit}\\ \text{rest mass}\end{pmatrix},$$

$$\tilde{L} = L/\mu = \begin{pmatrix}\text{angular momentum}\\ \text{per unit rest mass}\end{pmatrix}. \tag{25.15}$$

Recall also

$$\lambda = \tau/\mu = \begin{pmatrix}\text{proper time per}\\ \text{unit rest mass}\end{pmatrix}.$$

Then (25.14) becomes an equation for the change of r-coordinate with proper time in which the rest mass makes no appearance:

Effective potential $\tilde{V}$, and equations for orbit when $\mu \neq 0$

$$\begin{aligned}\left(\frac{dr}{d\tau}\right)^2 &= \tilde{E}^2 - (1-2M/r)(1+\tilde{L}^2/r^2)\\ &= \tilde{E}^2 - \tilde{V}^2(r).\end{aligned} \tag{25.16a}$$

Here

$$\tilde{V}(r) = [(1-2M/r)(1+\tilde{L}^2/r^2)]^{1/2} \tag{25.16b}$$

is the "effective potential" mentioned in §25.1 and Figure 25.2 and to be discussed in §25.5. For the rate of change of the other two relevant coordinates with proper time, one has, assuming a "direct" orbit ($d\phi/d\tau > 0$; $p_\phi = +L$ rather than $-L$),

$$\frac{d\phi}{d\tau} = \frac{1}{\mu}\frac{d\phi}{d\lambda} = \frac{p^\phi}{\mu} = \frac{g^{\phi\phi}L}{\mu} = \frac{\tilde{L}}{r^2} \tag{25.17}$$

and

$$\frac{dt}{d\tau} = \frac{1}{\mu}\frac{dt}{d\lambda} = \frac{p^0}{\mu} = -\frac{g^{00}E}{\mu} = \frac{\tilde{E}}{1 - 2M/r}. \tag{25.18}$$

Knowing r as a function of τ from (25.16), one can find ϕ and t in their dependence on τ from (25.17) and (25.18). Symmetry considerations have in effect reduced the four coupled second-order differential equations $p^\mu{}_{;\nu}p^\nu = 0$ of geodesic motion to the single first-order equation (25.16).

For objects of zero rest mass, it makes no sense to refer to proper time, and a slightly different treatment is appropriate (§25.6).

Before looking, in §25.5, at the motions predicted by equations (25.16) to (25.18), it is useful to analyze the physical significance of the constants p_0 and p_ϕ, and to identify other physically significant quantities whose values will be of interest in studying these orbits. One calls $E = -p_0$ the "energy at infinity"; and $L = |p_\phi|$, for equatorial orbits, the "total angular momentum." To justify these names, compare them with standard quantities measured by an observer at rest on the equator of the Schwarzschild coordinate system as the test particle flies past him in its orbit. Let

Interpretation of E as "energy at infinity" and L as "angular momentum"

$$\begin{aligned} E_{\text{local}} &\equiv p^{\hat{0}} \equiv \langle \boldsymbol{\omega}^{\hat{0}}, \boldsymbol{p}\rangle \equiv \langle |g_{00}|^{1/2}\, \boldsymbol{d}t, \boldsymbol{p}\rangle = |g_{00}|^{1/2}p^0 \\ &= |g_{00}|^{1/2}\, dt/d\lambda = (1 - 2M/r)^{1/2}\, dt/d\lambda \end{aligned} \tag{25.19}$$

be the energy he measures in his proper reference frame, and let

$$\begin{aligned} v_{\hat{\phi}} &\equiv \frac{p^{\hat{\phi}}}{p^{\hat{0}}} \equiv \frac{\langle \boldsymbol{\omega}^{\hat{\phi}}, \boldsymbol{p}\rangle}{E_{\text{local}}} = \frac{\langle |g_{\phi\phi}|^{1/2}\, \boldsymbol{d}\phi, d/d\lambda\rangle}{E_{\text{local}}} \\ &= \frac{r(d\phi/d\lambda)}{E_{\text{local}}} = \frac{p_\phi}{rE_{\text{local}}} \end{aligned} \tag{25.20}$$

be the tangential component of the ordinary velocity he measures. [Note: $\boldsymbol{\omega}^{\hat{\alpha}}$ are the basis one-forms of the observer's proper reference frame; see equations (23.15a,b).] In terms of these locally measured quantities, the energy-at-infinity is

$$E = -p_0 = -g_{00}p^0 = |g_{00}|^{1/2}E_{\text{local}} = (1 - 2M/r)^{1/2}E_{\text{local}} = \text{constant}. \tag{25.21}$$

It therefore represents the locally measured energy E_{local}, corrected by a factor $|g_{00}|^{1/2}$. For any particle that flies freely (geodesic motion) from this observer to $r = \infty$, the correction factor reduces to unity, and E_{local} (as measured by a second observer, this time at infinity) becomes identical with E. Similarly the angular momentum from (25.20) is

$$p_\phi = E_{\text{local}}v_{\hat{\phi}}r = \text{constant}. \tag{25.22}$$

This, like $E = -p_0$, represents a quantity that is conserved, and whose interpretation for $r \longrightarrow \infty$ on any orbit is familiar. Finally, recall that the total 4-momentum of two colliding particles $\boldsymbol{p}_1 + \boldsymbol{p}_2$ or $(p_\mu)_1 + (p_\mu)_2$ is conserved in a point collision (at any r). Therefore the totals $(E)_1 + (E)_2 = (-p_0)_1 + (-p_0)_2$ and $(p_\phi)_1 + (p_\phi)_2$ are also conserved. One of the colliding particles may be on an orbit that could never reach out to $r = \infty$, but this makes no difference. This conservation principle allows and forces one to take over the terms $E =$ "energy at infinity" and $L =$ "angular momentum," valid for orbits that do reach to infinity, for an orbit that does not reach to infinity.

EXERCISES

Exercise 25.7. RADIAL VELOCITY OF A TEST PARTICLE

Obtain a formula for the radial component of velocity $v_{\hat{r}}$ that an observer at r would measure [see (25.20) for $v_{\hat{\phi}}$]. Express E_{local}, $v_{\hat{r}}$, and $v_{\hat{\phi}}$ in terms of r and the constants E, p_ϕ.

Exercise 25.8. ROTATIONAL KILLING VECTORS FOR SCHWARZSCHILD GEOMETRY

(a) Show that in the isotropic coordinates of exercise 23.1, the metric for the Schwarzschild geometry takes the form

$$ds^2 = -(1 - M/2\bar{r})^2(1 + M/2\bar{r})^{-2}\,dt^2 + (1 + M/2\bar{r})^4(d\bar{r}^2 + \bar{r}^2\,d\Omega^2). \tag{25.23}$$

(b) Exhibit a coordinate transformation that brings this into the form

$$ds^2 = -(1 - M/2\bar{r})^2(1 + M/2\bar{r})^{-2}\,dt^2 + (1 + M/2\bar{r})^4(dx^2 + dy^2 + dz^2), \tag{25.24}$$

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

(c) Show that $\boldsymbol{\xi}_x = y(\partial/\partial z) - z(\partial/\partial y)$ and similar vectors $\boldsymbol{\xi}_y$ and $\boldsymbol{\xi}_z$ are each Killing vectors, by verifying (see exercise 25.5c) that the Poisson brackets $[\mathcal{H}, L_K]$ vanish for each $L_K = \boldsymbol{p}\cdot\boldsymbol{\xi}_K$, $K = x, y, z$.

(d) Show that $\boldsymbol{\xi}_z = (\partial/\partial\phi)_{t,r,\theta}$; and show that for orbits in the equatorial plane $L_z = p_\phi$, $L_x = L_y = 0$.

Exercise 25.9. CONSERVATION OF TOTAL ANGULAR MOMENTUM OF A TEST PARTICLE

Prove by a Poisson-bracket calculation that the total angular momentum squared, $L^2 = p_\theta{}^2 + (\sin\theta)^{-2}p_\phi{}^2$ is a constant of motion for any Schwarzschild geodesic.

Exercise 25.10. SELECTING EQUATION BY SELECTING WHAT IS VARIED

Write out the integral I that is varied in (25.8) for the special case of the Schwarzschild metric (25.12). What equation results from the demand $\delta I = 0$ if only $\phi(\lambda)$ is varied? If only $t(\lambda)$?

Exercise 25.11. MOTION DERIVED FROM SUPER-HAMILTONIAN

Write out the super-Hamiltonian (25.10) for the special case of the Schwarzschild metric. Deduce from its form that p_0 and p_ϕ are constants of motion. Derive (25.14), (25.17), and (25.18) from this super-Hamiltonian formalism.

2

§25.4. GRAVITATIONAL REDSHIFT

Law of "energy redshift" ("gravitational redshift")

The conservation law $|g_{00}|^{1/2}E_{\text{local}} = \text{constant}$ (equation 25.21), which is valid in this form for any time-independent metric with $g_{0j} = 0$ and for particles with both zero and non-zero rest mass, is sometimes called the "law of energy red-shift." It describes how the locally measured energy of any particle or photon changes (is "red-shifted" or "blue-shifted") as it climbs out of or falls into a static gravitational field. For a particle of zero rest mass (photon or neutrino), the locally measured energy E_{local}, and wavelength λ_{local} (not to be confused with affine parameter!), are related by $E_{\text{local}} = h/\lambda_{\text{local}}$, where h is Planck's constant. Consequently, the law of energy red-shift can be rewritten as

$$\lambda_{\text{local}}|g_{00}|^{-1/2} = \text{constant}. \tag{25.25}$$

A photon emitted by an atom at rest in the gravitational field at radius r, and received by an astronomer at rest at infinity is red-shifted by the amount

$$z \equiv \Delta\lambda/\lambda = (\lambda_{\text{received}} - \lambda_{\text{emitted}})/\lambda_{\text{emitted}} = |g_{00}(r)|^{-1/2} - 1,$$

$$z = (1 - 2M/r)^{-1/2} - 1, \tag{25.26}$$

$$z \approx M/r \text{ in Newtonian limit.} \tag{25.26N}$$

Note that these expressions are valid whether the photon travels along a radial path or not.

EXERCISE

Exercise 25.12. REDSHIFT BY TIMED PULSES

Derive expression (25.26) for the photon redshift by considering two pulses of light emitted successively by an atom at rest at radius r. [*Hint:* If $\Delta\tau_{\text{em}}$ is the proper time between pulses as measured by the emitting atom, and $\Delta\tau_{\text{rec}}$ is the proper time separation as measured by the observer at $r = \infty$, then one can idealize λ_{em} as $\Delta\tau_{\text{em}}$ and λ_{rec} as $\Delta\tau_{\text{rec}}$.]

---

§25.5. ORBITS OF PARTICLES

Turn attention now from energy red-shift to the orbit of a particle in the Schwarzschild geometry. The position as a function of proper time follows upon solving (25.16a),

$$\left(\frac{dr}{d\tau}\right)^2 + \tilde{V}^2(r) = \tilde{E}^2, \tag{25.16a}$$

where $\tilde{V}$ is the "effective potential" defined by

$$\tilde{V}^2(r) = (1 - 2M/r)(1 + \tilde{L}^2/r^2) \tag{25.16b}$$

and illustrated in Figure 25.2 and Box 25.6. The first diagram in Box 25.6 gives $\tilde{V}^2(r)$ as a function of r. It is relevant even in the "domain inside the black hole" ($r < 2M$), where $\tilde{V}^2$ is negative (see Chapter 31). It serves as a model for, and is closely related to, the "effective potential" $B^{-2}(r)$ used in §25.6 to analyze photon orbits. The final diagram in Box 25.6 gives $\tilde{V}(r)$ itself as a function of r. Energy levels in this diagram or in Figure 25.2 can be interpreted as in any conventional energy-level diagram. The difference in energy between two levels represents energy, as measured at infinity, of the photon given off in the transition from the one level to the other. Whether one plots $\tilde{V}(r)$ or $\tilde{V}^2(r)$ as a function of r is largely a matter of convenience. The important point is this: a value of r where $\tilde{V}(r)$ becomes equal to the available energy $\tilde{E}$, or $\tilde{V}^2(r)$ becomes equal to $\tilde{E}^2$, is a *turning point*. A particle that was moving to larger r values, once arrived at a turning point, turns around and moves to smaller r values. Or when a particle moving to smaller r values comes to a turning point, it reverses its motion and proceeds to larger r values. A turning point is not a point of equilibrium. A stone thrown straight up does not sit at a point of equilibrium at the top of its flight. However, when $\tilde{E} - \tilde{V}(r)$, or $\tilde{E}^2 - \tilde{V}^2(r)$, instead of having a single root, has a double root, then one does deal with a point of equilibrium (only possible because of "centrifugal force" fighting against gravity). When this equilibrium occurs at a minimum of $\tilde{V}(r)$, it is a stable equilibrium; at a maximum, an unstable equilibrium. Thus all the major features of the motion in the r direction can be read from a plot of the effective potential as a function of r (plot depends on value of $\tilde{L}$) and from a knowledge of the $\tilde{E}$ value (Figure 25.2, with further details in Box 25.6).

Qualitative features of orbits diagnosed from effective-potential diagram

Box 25.6 QUALITATIVE FEATURES OF ORBITS OF A PARTICLE MOVING IN SCHWARZSCHILD GEOMETRY

A. Equations Governing Orbit (see text for derivation)

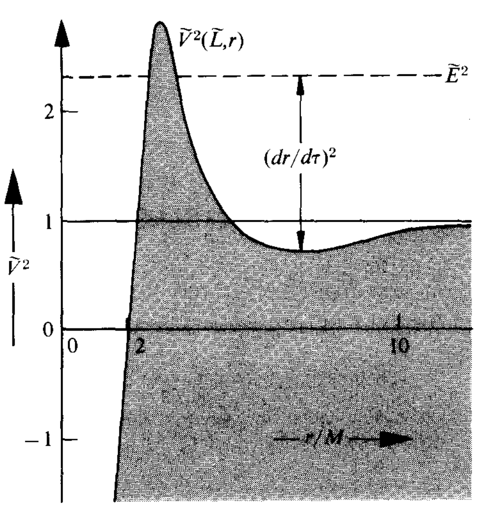
1. Effective-potential equation for radial part of motion:

$$(dr/d\tau)^2 + \tilde{V}^2(\tilde{L}, r) = \tilde{E}^2,$$
$$\tilde{V}^2(\tilde{L}, r) = (1 - 2M/r)(1 + \tilde{L}^2/r^2),$$

$\tilde{E} =$ (energy at infinity per unit rest mass),
$\tilde{L} =$ (angular momentum per unit rest mass).

2. Supplementary equations for angular and time motion for "direct" orbit, $d\phi/d\tau > 0$:

$$d\phi/d\tau = \tilde{L}/r^2,$$
$$\frac{dt}{d\tau} = \frac{\tilde{E}}{1 - 2M/r}.$$



"Turning points" of orbit occur where horizontal line of height $\tilde{E}^2$ crosses $\tilde{V}^2$

B. Newtonian Limit, $|\tilde{E} - 1| \ll 1$, $M/r \ll 1$, $\tilde{L}/r \ll 1$

1. Speak not about "energy-at-infinity per unit rest mass," $\tilde{E} = E/\mu = (1 - v_\infty^2)^{-1/2}$, but instead about the "nonrelativistic energy per unit rest mass,"

$$\varepsilon \equiv \frac{1}{2}(\tilde{E}^2 - 1) \approx \tilde{E} - 1 \approx \frac{1}{2} v_\infty{}^2.$$

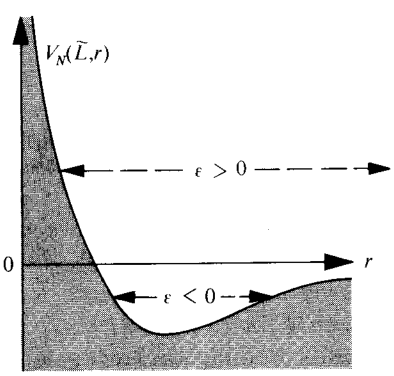
2. Speak not about $\tilde{V}^2(\tilde{L}, r)$ but instead about the Newtonian effective potential,

$$V_N(\tilde{L}, r) \equiv \frac{1}{2}(\tilde{V}^2 - 1) \approx -\frac{M}{r} + \frac{\tilde{L}^2}{2r^2}.$$

3. Rewrite effective-potential equation in the form

$$\frac{1}{2}\left(\frac{dr}{d\tau}\right)^2 + V_N(\tilde{L}, r) = \varepsilon.$$

4. From the effective-potential diagram and the subsidiary equation $d\phi/d\tau = \tilde{L}/r^2$, conclude that:
   a. Particles with $\varepsilon \geq 0$ ($\tilde{E} \geq 1$) come in from $r = \infty$ along hyperbolic or parabolic orbits, are reflected off the effective potential at $\varepsilon = V_N$[$\tilde{E}^2 = \tilde{V}^2$; "turning point"; $(dr/d\tau)^2 = 0$], and return to $r = \infty$.
   b. Particles with $\varepsilon < 0$ ($\tilde{E} < 1$) move back and forth in an effective potential well between periastron (inner turning point of elliptic orbit) and apastron (outer turning point).

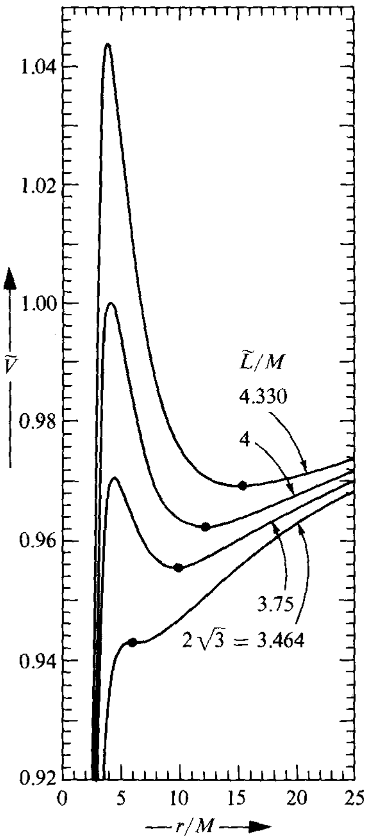


C. Relativistic Orbits

Use the effective-potential diagram of part A (reproduced here for various $\tilde{L}$), in the same way one uses the Newtonian diagram of part B, to deduce the qualitative features of the orbits. The main conclusions are these.

Box 25.6 (*continued*)

1. Orbits with periastrons at $r \gg M$ are Keplerian in form, except for the periastron shift (exercise 25.16; §40.5) familiar for Mercury.
2. Orbits with periastrons at $r \lesssim 10M$ differ markedly from Keplerian orbits.
3. For $\tilde{L}/M \leq 2\sqrt{3}$ there is no periastron; any incoming particle is necessarily pulled into $r = 2M$.
4. For $2\sqrt{3} < \tilde{L}/M < 4$ there are bound orbits in which the particle moves in and out between periastron and apastron; but any particle coming in from $r = \infty$ (unbound; $\tilde{E}^2 \geq 1$) necessarily gets pulled into $r = 2M$.
5. For $L^\dagger = \tilde{L}/M > 4$, there are bound orbits; particles coming in from $r = \infty$ with
$$\tilde{E}^2 < \tilde{V}_{\max}{}^2 = (1 - 2u_m)(1 + L^{\dagger 2}u_m{}^2),$$
$$u_m \equiv \frac{1 + \sqrt{1 - 12/L^{\dagger 2}}}{6}$$
reach periastrons and then return to $r = \infty$; but particles from $r = \infty$ with $\tilde{E}^2 > \tilde{V}_{\max}{}^2$ get pulled into $r = 2M$.
6. There are stable circular orbits at the minimum of the effective potential; the minimum moves inward from $r = \infty$ for $\tilde{L} = \infty$ to $r = 6M$ for $L^\dagger = \tilde{L}/M = 2\sqrt{3}$. The most tightly bound, stable circular orbit ($\tilde{L}/M = 2\sqrt{3}$, $r = 6M$) has a fractional binding energy of
$$\frac{\mu - E}{\mu} = 1 - \tilde{E} = 1 - \sqrt{8/9} = 0.0572.$$
7. There are unstable circular orbits at the maximum of the effective potential; the maximum moves outward from $r = 3M$ for $\tilde{L} = \infty$ to $r = 6M$ for $\tilde{L}/M = 2\sqrt{3}$. A particle in such a circular orbit, if perturbed inward, will spiral into $r = 2M$. If perturbed outward, and if it has $\tilde{E}^2 > 1$, it will escape to $r = \infty$. If perturbed outward, and if it has $\tilde{E}^2 < 1$, it will either reach an apastron and then enter a spiraling orbit that eventually falls into the star (e.g., if $\delta\tilde{E} > 0$, with unchanged angular momentum); or it will move out and in between apastron and periastron, in a stable bound orbit (e.g., if $\delta\tilde{E} < 0$, again with unchanged angular momentum).



When one turns from qualitative features to quantitative results, one finds it appropriate to write down explicitly the proper time $\Delta\tau$ required for the particle to augment its Schwarzschild coordinate by the amount Δr; thus (with the convention that square roots may be negative or positive, $\sqrt{a^2} \equiv \pm a$)

$$\tau = \int d\tau = \int \frac{dr}{[\tilde{E}^2 - (1 - 2M/r)(1 + \tilde{L}^2/r^2)]^{1/2}}. \tag{25.27}$$

Radial orbits:

The integration is especially simple for a particle falling straight in, or climbing straight out, for then the angular momentum vanishes and the integral can be written in an elementary form that applies (with the change $\tau \longrightarrow t$) even in Newtonian mechanics,

$$\tau = \int d\tau = \int \frac{dr}{[2M/r - 2M/R]^{1/2}}. \tag{25.27$'$}$$

Here $R \equiv 2M/(1 - \tilde{E}^2)$ is the radius at which the particle has zero velocity ("apastron"). The motion follows the same "cycloid principle" that is so useful in nonrelativistic mechanics (Figure 25.3). Thus, in parametric form, one has

(1) "cycloidal" form of $r(\tau)$ for radial bound orbits

$$r = \frac{R}{2}(1 + \cos\eta),$$
$$\tau = \frac{R}{2}\left(\frac{R}{2M}\right)^{1/2}(\eta + \sin\eta), \tag{25.28}$$

with the total proper time to fall from rest at $r = R$ into $r = 0$ given by the expression

$$\tau = \frac{\pi}{2} R \left(\frac{R}{2M}\right)^{1/2} \tag{25.29}$$

(shorter by a factor $1/\sqrt{2}$ than the time for fall under pull of the same mass, distributed over a sphere of radius R; see dotted curve in Figure 25.3).

What about the Schwarzschild-coordinate time taken for a given motion? Take equation (25.16a) for general motion (radial or nonradial), and where $dr/d\tau$ appears, replace it by

$$\frac{dr}{d\tau} = \frac{dr}{dt}\frac{dt}{d\tau} = \frac{dr}{dt}\frac{\tilde{E}}{1 - 2M/r} = \tilde{E}\frac{dr^*}{dt}. \tag{25.30}$$

Here r^* is an abbreviation for a new "tortoise coordinate,"

$$r^* = \int dr^* = \int \frac{dr}{1 - 2M/r} = r + 2M \ln\left(\frac{r}{2M} - 1\right), \tag{25.31}$$

which was introduced by Wheeler (1955) and popularized by Regge and Wheeler (1957). Thus find the equation

(2) "tortoise" radial coordinate as function of coordinate time, $r^*(t)$

$$\left(\tilde{E}\frac{dr^*}{dt}\right)^2 + \tilde{V}^2 = \tilde{E}^2. \tag{25.32}$$

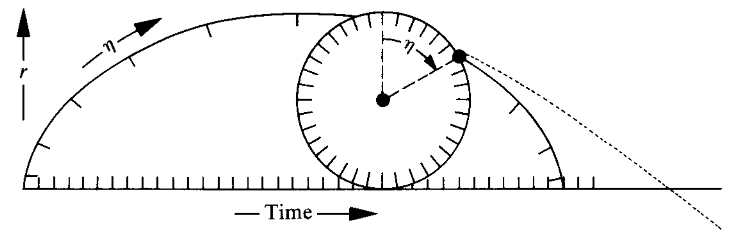


Figure 25.3.
A cycloid gives the relation between proper time and Schwarzschild r coordinate for a test particle falling straight in toward center of gravitational attraction of negligible dimensions. The angle of turn of the wheel as it rolls on the base line and generates the cycloid is denoted by η. In terms of this parameter, one has

$$r = \frac{R}{2}(1 + \cos\eta) \qquad \text{(Schwarzschild } r\text{-coordinate)}$$

$$\tau = \frac{R}{2}\left(\frac{R}{2M}\right)^{1/2}(\eta + \sin\eta) \qquad \text{(proper time)}$$

(note difference in scale factors in expressions for r and for τ). The total lapse of proper time to fall from $r = R$ to $r = 0$ is $\tau = (\pi/2)(R^3/2M)^{1/2}$. The same cycloid relation and the same expression for time to fall holds in Newton's nonrelativistic theory of gravitation, except that there the symbol τ is to be replaced by the symbol t (ordinary time). Were one dealing in Newtonian theory with the same attracting mass M spread uniformly over a sphere of radius R, with a pipe thrust through it to make a channel for the motion of the test particle, then that particle would execute simple harmonic oscillations (dotted curve above). The angular frequency ω of these vibrations would be identical with the angular frequency of revolution of the test particle in a circle just grazing the surface of the planet, a frequency given by Kepler's law $M = \omega^2 R^3$. In this case, the time to fall to the center would be $(\pi/2)(R^3/M)^{1/2}$, longer by a factor $2^{1/2}$ than for a concentrated center of attraction (concentrated mass: stronger acceleration and higher velocity in the later phases of the fall). The expression for the Schwarzschild-coordinate time t required to reach any point r in the fall under the influence of a concentrated center of attraction is complicated and is not shown here (see equation 25.37 and Figure 25.5).

The same cycloidal relation that connects r with time for free fall of a particle also connects the radius of the "Friedmann dust-filled universe" with time (see Box 27.1), except that there the cycloid diagram applies directly, without any difference in scale between the two key variables:

$$\begin{pmatrix}\text{radius of}\\ \text{3-sphere}\end{pmatrix} = \frac{a}{2}(1 - \cos\eta) \simeq \frac{a}{4}\eta^2 \text{ (for small } \eta),$$

$$\begin{pmatrix}\text{coordinate time}\\ \text{identical with}\\ \text{proper time as}\\ \text{measured on dust}\\ \text{particle}\end{pmatrix} = \frac{a}{2}(\eta - \sin\eta) \simeq \frac{a}{12}\eta^3 \text{ (for small } \eta).$$

The starting point of η is renormalized to time of start of expansion; see Lindquist and Wheeler (1957) for more on correlation between fall of particle and expansion of universe.

Here the effective potential is the same effective potential that one dealt with before,

$$\tilde{V} = [(1 - 2M/r)(1 + \tilde{L}^2/r^2)]^{1/2}. \tag{25.33}$$

Moreover, the $\tilde{E}$ on the righthand side is the same $\tilde{E}$ that appeared in the earlier equation for $(dr/d\tau)^2$. Therefore the turning points and the qualitative description of the motion are both the same as before. "A turning point is a turning point is a turning point." Right? Right about turning points; wrong about the conclusion.

The story has it that Achilles never could pass the tortoise. Whenever he caught up with where it had been, it had moved ahead to a new location; and when he got there, it was still further ahead; and so on *ad infinitum*. Imagine the race between Achilles and the tortoise as running to the left and the expected point of passing as lying at $r = 2M$. The r-coordinate has no inhibition about passing through the value $r = 2M$. Not so r^*, the "tortoise coordinate." It can go arbitrarily far in the direction of minus infinity (corresponding to the infinitely many times when Achilles catches up with where the tortoise was) and still r remains outside $r = 2M$:

(3) details of the approach to the Schwarzschild radius ($r = 2M$)

$r/2M$	1.000001	1.0001	1.01	1.278465	2	5	10	10,000
$r^*/2M$	−12.8155	−8.2102	−3.5952	0	2	6.386	12.303	10,009.210

It follows that there is a great difference between the description of the motion in terms of the proper time τ of a clock on the falling particle (r goes all the way from $r = R$ down to $r = 0$ in the finite proper time of 25.29) and a description of the motion in terms of the Schwarzschild-coordinate time t appropriate for the faraway observer (r^* goes all the way from $r^* = R^*$ down to $r^* = -\infty$; infinite t required for this; but even in infinite time, as r^* goes down to $-\infty$, r is only brought asymptotically down to $r \sim 2M$). Thus the second description of the motion leaves out, and has no alternative but to leave out, the whole range of r values from $r = 2M$ down to zero: perfectly good physics, and physics that the falling particle is going to see and explore, but physics that the faraway observer never will see and never can see. If the tortoise coordinate did not exist, it would have to be invented. It invests each factor ten of closer approach to $r = 2M$ with the same interest as the last factor ten and the next to come. It proportions itself in accord with the amount of Schwarzschild-coordinate time available to the faraway observer to study these more and more microscopic amounts of motion in more and more detail.

Figure 25.4 shows the effective potential $\widetilde{V}$ of (25.33) and of Figure 25.2 replotted as a function of the tortoise coordinate. The approach of $\widetilde{V}$ to zero at $r = 2M$ shows up as an exponential approach of $\widetilde{V}$ to zero as r^* goes to minus infinity. Thus in moving "towards the black hole" ($r = 2M$, $r^* = -\infty$), the particle, as described in coordinate time t, soon casts off any effective influence of any potential, and moves essentially freely toward decreasing r^*, in accordance with the equation

$$\left(\widetilde{E}\frac{dr^*}{dt}\right)^2 \simeq \widetilde{E}^2; \tag{25.34}$$

that is, "with the speed of light" ($dr^*/dt \simeq -1$). This dependence of r^* on t implies at once an asymptotic dependence of r itself on Schwarzschild-coordinate time t, of the form

$$r = 2M + (\text{constant} \times e^{-t/2M}). \tag{25.35}$$

This result is independent of the angular momentum of the particle and independent also of the energy, provided only that the energy-per-unit-mass $\widetilde{E}$ is enough to

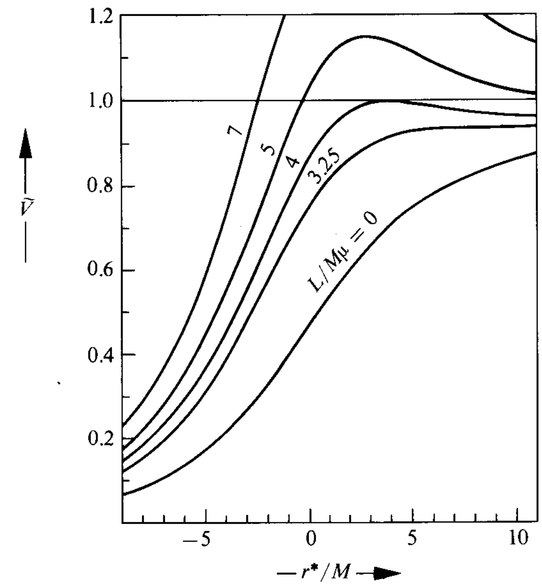


Figure 25.4.
Effective potential for motion in Schwarzschild geometry, expressed as a function of the tortoise coordinate, for selected values of the angular momentum of the test particle. The angular momentum L is expressed in units $M\mu$, where M is the mass of the black hole and μ the mass of the test particle. The effective potential (including rest mass) is expressed in units μ; thus, $\widetilde{V} = V/\mu$. The tortoise coordinate $r^* = r + 2M \ln (r/2M - 1)$ is given in units M.

surmount the barrier (Figure 25.4) of the effective potential-per-unit-mass $\widetilde{V}$. (More will be said on the approach to $r = 2M$ in Chapter 32, on gravitational collapse.)

To replace the asymptotic formula (25.35) by a complete formula requires one to integrate (25.32); thus,

$$\begin{aligned} t = \int dt &= \int \frac{\widetilde{E}\, dr^*}{[\widetilde{E}^2 - \widetilde{V}^2]^{1/2}} \\ &= \int \frac{\widetilde{E}}{[\widetilde{E}^2 - (1 - 2M/r)(1 + \widetilde{L}^2/r^2)]^{1/2}} \frac{dr}{(1 - 2M/r)}. \end{aligned} \tag{25.36}$$

The integration here is not easy, even for pure radial motion ($\widetilde{L} = 0$), as is seen in the complication of the resulting expression (Khuri 1957):

$$\begin{aligned} t = &\left[\left(\frac{R}{2} + 2M\right)\left(\frac{R}{2M} - 1\right)^{1/2}\right]\eta + \frac{R}{2}\left(\frac{R}{2M} - 1\right)^{1/2} \sin \eta \\ &+ 2M \ln \left| \frac{(R/2M - 1)^{1/2} + \tan (\eta/2)}{(R/2M - 1)^{1/2} - \tan (\eta/2)} \right|. \end{aligned} \tag{25.37}$$

Here η is the same cycloid parameter that appears in equation (25.28) and Figure 25.3 (see the detailed plot in Figure 25.5 of the correlation between r and t, illustrat-

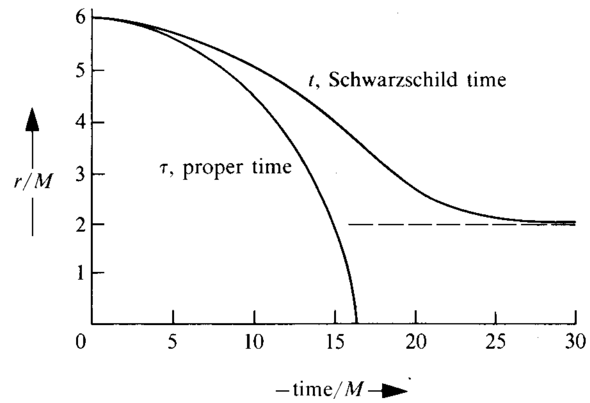


Figure 25.5.
Fall toward a Schwarzschild black hole as described (a) by a comoving observer (proper time τ) and (b) by a faraway observer (Schwarzschild-coordinate time t). In the one description, the point $r = 0$ is attained, and quickly [see equation (25.28)]. In the other description, $r = 0$ is never reached and even $r = 2M$ is attained only asymptotically [equations (25.35) and (25.37)]. The qualitative features of the motion in both cases are most easily deduced by inspection of the "effective potential-per-unit-mass" $\widetilde{V}$ in its dependence on r (Figure 25.2) when one is interested in proper time; or the same effective potential $\widetilde{V}$ in its dependence on the "tortoise coordinate" r^* [Figure 25.4 and equation (25.31)] when one is interested in Schwarzschild-coordinate time t.

ing the asymptotic approach to $r = 2M$). The difficulty in the integration for t, as compared to the ease of the integration for τ (25.28), has a simple origin. Only two r-values appear in (25.27a) as special points when $\widetilde{L}$ is zero: the starting point, $r = R$, where the velocity vanishes, and the point $r = 0$, where $dr/d\tau$ becomes infinite. In contrast (25.36), rewritten as

$$t = \int dt = \int \frac{[1 - 2M/R]^{1/2}}{[2M/r - 2M/R]^{1/2}} \frac{dr}{(1 - 2M/r)}, \tag{25.36$'$}$$

contains three special points: $r = R$, $r = 0$, and the added point with all the new physics, $r = 2M$. To admit angular momentum is to increase the number of special points still further, and to make the integral unmanageable except numerically or qualitatively (via the potential diagram of Figure 25.4), or in terms of elliptic functions [Hagihara (1931)].

(4) free-fall from $r = \infty$

It is often convenient to abstract away from the precise value $r = R$ at the start of the collapse. In this event, one deals with the limit $R \longrightarrow \infty$. Then it is convenient to displace the zero of proper time to the instant of final catastrophe. In this limit, one has

$$\tau/2M = -(2/3)(r/2M)^{3/2},$$

$$t/2M = -(2/3)(r/2M)^{3/2} - 2(r/2M)^{1/2} + \ln \frac{(r/2M)^{1/2} + 1}{(r/2M)^{1/2} - 1}. \tag{25.38}$$

At very large negative time, the particle is far away and approaching only very slowly. Then one can write

$$r = (9M\tau^2/2)^{1/3} \simeq (9Mt^2/2)^{1/3} \tag{25.39a}$$

whether one refers to coordinate time or to proper time. However, the final stages of infall are again very different, when expressed in terms of proper time ($\tau \longrightarrow 0$, $r \longrightarrow 0$), from what they are as expressed in terms of Schwarzschild-coordinate time,

$$r/2M = 1 + 4e^{-8/3}e^{-t/2M}. \tag{25.39b}$$

Nonradial orbits:
(1) Fourier analysis

Turning from pure radial motion to motion endowed with angular momentum, one has a situation where one would like to express the principal quantities of the motion (components of displacement, velocity, and acceleration) in Fourier series (in Schwarzschild-coordinate time), these being so convenient in the Newtonian limit in analyzing radiation and perturbations of one orbit by another and tidal perturbations of the moving particle itself by the tide-producing action of the center of attraction. Any exact evaluation of these coefficients would appear difficult. For the time being, the values of the Fourier amplitudes would seem best developed by successive approximations starting from the Newtonian analysis (see Box 25.4 and references cited there).

In connection with any such Fourier analysis, it is appropriate to recall that the fundamental frequency alone appears, and all higher harmonics have zero amplitude, when the motion takes place in an exactly circular orbit (opposite extreme from the pure radial motion of $\tilde{L} = 0$). Therefore it is of interest to note (exercise 25.19) that the circular frequency ω of this motion, as measured by a faraway observer, is correlated with the Schwarzschild r-value of the orbit by exactly the Keplerian formula of non-relativistic physics:

$$\omega^2 r^3 = M \qquad \text{(exact; general relativity).} \tag{25.40}$$

(2) details of angular motion

Turn now from the correlation between r and time to the correlation between r and angle of revolution (ϕ in the analysis here; θ in the Hamilton-Jacobi analysis of Box 25.4; this difference in name is irrelevant in what follows). Return to equation (25.16),

$$\left(\frac{dr}{d\tau}\right)^2 + \tilde{V}^2(r) = \tilde{E}^2,$$

and recall also equation (25.17)

$$\frac{d\phi}{d\tau} = \frac{\tilde{L}}{r^2}.$$

Solve the second equation for $d\tau$, and substitute into the first to find

$$\left(\frac{\tilde{L}}{r^2}\frac{dr}{d\phi}\right)^2 + \tilde{V}^2(r) = \tilde{E}^2, \tag{25.41}$$

or equivalently, with $u = M/r$ and $L^\dagger = \tilde{L}/M = L/M\mu$,

$$\left(\frac{du}{d\phi}\right)^2 = \frac{\tilde{E}^2 - (1-2u)(1+L^{\dagger 2}u^2)}{L^{\dagger 2}}. \tag{25.42}$$

Exercise 25.16 presents an alternative differential equation derived from this formula, and uses it to obtain the following expression for the angle swept out by the particle or planet, moving in a nearly circular orbit, between two successive points of closest approach:

(3) nearly circular orbits: periastron shift

$$\Delta\phi = \frac{2\pi}{(1 - 6M/r_0)^{1/2}}. \tag{25.43}$$

The radial motion turns around from ingoing to outgoing, or from outgoing to ingoing, whenever the quantity $\tilde{E}^2 - \tilde{V}^2(r)$, or $\tilde{E} - \tilde{V}(r)$, plotted as a function of r, undergoes a change of sign, and this as clearly here in the correlation between r and ϕ as in the earlier correlation between r and time. Recall again the curves of Figure 25.2 for $\tilde{V}(r)$ as a function of r for selected $\tilde{L}$ values. From them one can read out, without any calculation at all, the principal features of typical orbits (Box 25.6) obtained by detailed numerical calculation. Characteristic features are

(4) qualitative features of angular motion

(1) circular orbit when $\tilde{E}$ coincides with a minimum of the effective potential $\tilde{V}(r)$,

(2) precession when $\tilde{E}$ is a little more than $\tilde{V}_{\text{min}}$,

(3) temporary "orbiting" (many turns around the center of attraction) when $\tilde{E}$ is close to a maximum $\tilde{V}_{\text{max}}$ of the effective potential,

(4) "capture into the black hole" when $\tilde{E}$ exceeds $\tilde{V}_{\text{max}}$.

A more detailed analysis appears in Box 25.6. [For explicit analytic calculation of orbits in the Schwarzschild geometry, see Hagihara (1931), Darwin (1959 and 1961), and Mielnik and Plebanski (1962).]

For orbits of positive energy, no feature of the inverse-square force is better known than the Rutherford scattering formula. It gives the "effective amount of target area" presented by the center of attraction for throwing particles into a faraway receptor that picks up everything coming off into a unit solid angle at a specified angle of deflection Θ:

Scattering of incoming particles:

$$\frac{d\sigma}{d\Omega} = \frac{M^2}{[4(\tilde{E} - 1)\sin^2\Theta/2]^2} \text{ (Rutherford; nonrelativistic)} \tag{25.44}$$

(1) Rutherford (nonrelativistic) cross section

(derivation in equations 8 to 15 of Box 25.4). When one turns from the Newtonian analysis to the general-relativity treatment, one finds two striking new features of the scattering associated with the phenomenon of orbiting. (1) The particles that come off at a given angle of deflection Θ now include not only those that have really been deflected by Θ (the only contribution in Rutherford scattering), but also those that have been deflected by $\Theta + 2\pi$, $\Theta + 4\pi$, . . . etc. (an infinite series of contributions). (2) These supplementary contributions, while finite in amount, and even finite in amount "per unit range of Θ," are not finite in amount when expressed "per unit of solid angle $d\Omega = 2\pi \sin\Theta\, d\Theta$" in either the forward direction ($\Theta = 0$) or the backward direction ($\Theta = \pi$). This circumstance produces no spectacular change in the forward scattering, for that is already infinite in the nonrelativistic approximation (infinity in Rutherford value of $d\sigma/d\Omega$ as $\Theta = 0$ is approached, arising from

(2) new features due to relativistic gravity

particles flying past with large impact parameters and experiencing small deflections; see exercise 25.21). In contrast, the backward scattering, which was perfectly finite in the Rutherford analysis, acquires also an infinity:

$$\left(\frac{d\sigma}{d\Omega}\right)_{\Theta\sim\pi} \sim \frac{\text{constant}}{\sin\Theta}. \tag{25.45}$$

This concentration of scattering in the backward direction is known as a "glory." The effect is most readily seen by looking at the brilliant illumination that surrounds the shadow of one's plane on clouds far below (180° scattering of light ray within waterdrop). It is also clearly seen in observations on the scattering of atoms by atoms near $\Theta = 180°$. No dwarf star, not even any neutron star, is sufficiently compact to be out of the way of a high-speed particle trying to make such a 180° turn. Only a black hole is compact enough to produce this effect.

Further interesting features of motion in Schwarzschild geometry appear in the exercises below.

EXERCISES

Exercise 25.13. QUALITATIVE FORMS OF PARTICLE ORBITS

Verify the statements about particle orbits made in part C of Box 25.6.

Exercise 25.14. IMPACT PARAMETER

For a scattering orbit (i.e., unbound orbit), show that $\tilde{L} = \tilde{E}v_\infty b$, where b is the impact parameter and v_∞ the asymptotic ordinary velocity; also show that

$$b = \tilde{L}/(\tilde{E}^2 - 1)^{1/2}. \tag{25.46}$$

Draw a picture illustrating the physical significance of the impact parameter.

Exercise 25.15. TIME TO FALL TO $r = 2M$

Show from equation (25.16) and the first picture in Box 25.6 that orbits (general $\tilde{L}$ value!) which approach $r = 2M$ do so in a finite proper time, but (equation 25.32) an infinite coordinate time t. For equilibrium stars, which must have radii $R > 2M$, the coordinate time t to fall to the surface is finite, of course.

Exercise 25.16. PERIASTRON SHIFT FOR NEARLY CIRCULAR ORBITS

Rewrite equation (25.42) in the form

$$(du/d\phi)^2 + (1 - 6u_0)(u - u_0)^2 - 2(u - u_0)^3 = (\tilde{E}^2 - \tilde{E}_0^2)/L^{\dagger 2}. \tag{25.47}$$

Express the constant $u_0 \equiv M/r_0$ in terms of $\tilde{L}/M$, and express $\tilde{E}_0$ in terms of u_0. Show for a nearly circular orbit of radius r_0 that the angle swept out between two successive periastra (points of closest approach to the star) is

$$\Delta\phi = 2\pi(1 - 6M/r_0)^{-1/2}. \tag{25.48}$$

Sketch the shape of the orbit for $r_0 = 8M$.

Exercise 25.17. ANGULAR MOTION DURING INFALL

From equation (25.42), deduce that the total angle $\Delta\phi$ swept out on a trajectory falling into $r = 0$ is finite. The computation is straightforward; but the interpretation, in view of the behavior of $t(\lambda)$ on the same trajectory (equation 25.32 and exercise 25.15), is not. The interpretation will be elucidated in Chapter 31.

Exercise 25.18. MAXIMUM AND MINIMUM OF EFFECTIVE POTENTIAL

Derive the expressions given in the caption of Figure 25.2 for the locations of the maximum and the minimum of the effective potential as a function of angular momentum. Determine also the limiting form of the dependence of barrier height on angular momentum in the limit in which $\widetilde{L}$ is very large compared to M.

Exercise 25.19. KEPLER LAW VALID FOR CIRCULAR ORBITS

From $d\phi/d\tau$ of (25.17) and $dt/d\tau$ of (25.18), deduce an expression for the circular frequency of revolution as seen by a faraway observer; and from the results of exercise 25.18 (or otherwise) show that it fulfills exactly the Kepler relation

$$\omega^2 r^3 = M$$

for any circular orbit of Schwarzschild r-value equal to r, whether stable (potential minimum) or unstable (potential maximum).

Exercise 25.20. HAMILTON-JACOBI FUNCTION

Construct the locus in the r, θ diagram of points of constant dynamic phase $\widetilde{S}(t, r, \theta) = 0$ for $t = 0$ and for values $\widetilde{L} = 4M$, $\widetilde{E} = 1$ (or for $\widetilde{L} = 2\sqrt{3}M$, $\widetilde{E} = (8/9)^{1/2}$, or for some other equally simple set of values for these two parameters). Show that the whole set of surfaces of constant $\widetilde{S}$ can be obtained by rotating the foregoing locus through one angle, then another and another, and recopying or retracing. Interpret physically the principal features of the resulting pattern of curves.

Exercise 25.21. DEFLECTION BY GRAVITY CONTRASTED WITH DEFLECTION BY ELECTRIC FORCE

A test particle of arbitrary velocity β flies past a mass M at an impact parameter b so great that the deflection is small. Show that the deflection is

$$\theta = \frac{2M}{b\beta^2}(1 + \beta^2). \tag{25.49}$$

Derive the deflection according to Newtonian mechanics for a particle moving with the speed of light. Show that (25.49) in the limit $\beta \longrightarrow 1$ is twice the Newtonian deflection. Derive also (flat-space analysis) the contrasting formula for the deflection of a fast particle of rest mass μ and charge e by a nucleus of charge Ze,

$$\theta = \frac{2Ze^2}{\mu b\beta^2}(1 - \beta^2)^{1/2}. \tag{25.50}$$

How feasible is it to rule out a "vector" theory of gravitation [see, for example, Brillouin (1970)], patterned after electromagnetism, by observations on the bending of light by the sun? [Hint: To simplify the mathematical analysis, go back to (25.42). Differentiate once with respect to ϕ to convert into a second-order equation. Rearrange to put on the left all those terms that would be there in the absence of gravity, and on the right all those that originate from the $-2u$ term (gravitation) in the factor $(1 - 2u)$. Neglect the right-hand side of the equation and solve exactly (straight-line motion). Evaluate the perturbing term on the right as a function of ϕ by inserting in it the unperturbed expression for $u(\phi)$. Solve again and get the deflection.]

Exercise 25.22. CAPTURE BY A BLACK HOLE

Over and above any scattering of particles by a black hole, there is direct capture into the black hole. Show that the cross section for capture is πb_{crit}^2, with the critical impact parameter b_{crit} given by $L_{\text{crit}}/(E^2 - \mu^2)^{1/2}$. From the formulas in the caption of Fig. 25.2 or otherwise, show that for high-energy particles this cross section varies with energy as

$$\sigma_{\text{capt}} = 27\pi M^2\left(1 + \frac{2}{3\widetilde{E}^2} + \cdots\right) \tag{25.51}$$

(photon limit for $\widetilde{E} \longrightarrow \infty$) and for low energies as

$$\sigma_{\text{capt}} = 16\pi M^2/\beta^2, \tag{25.52}$$

where β is the velocity relative to the velocity of light [Bogorodsky (1962)].

---

§25.6. ORBIT OF A PHOTON, NEUTRINO, OR GRAVITON IN SCHWARZSCHILD GEOMETRY

Orbits for particles of zero rest mass:

The concepts of "energy per unit of rest mass" and "angular momentum per unit of rest mass" make no sense for an object of zero rest mass (photon, neutrino, even the graviton of exercise 35.16). However, there is nothing about the motion of such an entity that cannot be discovered by considering the motion of a particle of finite rest mass μ and going to the limit $\mu \longrightarrow 0$. In this limit the quantities

$$\widetilde{E} = E/\mu$$

and

$$\widetilde{L} = L/\mu$$

individually go to infinity; but the ratio

(1) impact parameter defined

$$\begin{pmatrix}\text{impact para-}\\\text{meter}\end{pmatrix} = b = \frac{\begin{pmatrix}\text{angular}\\\text{momentum}\end{pmatrix}}{\begin{pmatrix}\text{linear}\\\text{momentum}\end{pmatrix}} = \frac{L}{(E^2-\mu^2)^{1/2}} = \frac{\widetilde{L}}{(\widetilde{E}^2-1)^{1/2}} \tag{25.53}$$

goes to the finite value

$$\lim_{\mu\to 0} \frac{\widetilde{L}}{\widetilde{E}} = b. \tag{25.54}$$

(2) shape of orbit

In this limit, equation (25.41) for the shape of the orbit reduces at once to the simple form

$$\left(\frac{1}{r^2}\frac{dr}{d\phi}\right)^2 + \frac{1 - 2M/r}{r^2} = \frac{1}{b^2}, \tag{25.55}$$

or

$$\left(\frac{1}{r^2}\frac{dr}{d\phi}\right)^2 + B^{-2}(r) = b^{-2}, \tag{25.56}$$

or

$$\left(\frac{du}{d\phi}\right)^2 + u^2(1 - 2u) = \left(\frac{M}{b}\right)^2 \equiv \frac{1}{\bar{b}^2}. \tag{25.57}$$

Whichever way the differential equation for the orbit is written, one term in it depends on the choice of orbit (the term $1/b^2$) the other on the properties of the Schwarzschild geometry, but not on the choice of orbit. This second term defines a kind of effective potential,

(3) effective potential

$$\begin{pmatrix}\text{"effective}\\ \text{potential for}\\ \text{photon"}\end{pmatrix} \equiv B^{-2}(r) \equiv \frac{1 - 2M/r}{r^2}. \tag{25.58}$$

No attempt is made here to take the square root, as was done for a particle of finite rest mass. There one took the root in order to have a quantity that reduced to the Newtonian effective potential (plus the rest mass) in the nonrelativistic limit; but for light ($v = 1$) there is no nonrelativistic limit. Therefore the effective potential (25.58) is plotted directly in Box 25.7, and used there to analyze some of the principal features of the orbits of a photon in Schwarzschild geometry.

On occasion it has proved useful to plot as a function of r, not the "effective potential" of (25.58), but the "potential impact parameter $B(r)$" calculated from that formula [see, for example, Power and Wheeler (1957), Zel'dovich and Novikov (1971)]. This potential impact parameter has the following interpretation: A ray, in order to reach the point r, must have an impact parameter b that is equal to or less than $B(r)$:

$$b \leq B(r) \text{ ("condition of accessibility").} \tag{25.59}$$

(4) critical impact parameter

A ray with zero impact parameter (head-on impact), or any impact parameter less than $b_{\text{crit}} = \min[B(r)] = 3\sqrt{3}M$, can get to any and all r values.

(5) affine parameter

The beautifully simple "effective potential" defined by (25.58) is used in (25.56) to determine the shape of an orbit; that is, the azimuth ϕ that the photon has when it gets to a given r-value. In other connections, it can be equally interesting to know when, or at what Schwarzschild coordinate time, the photon gets to a given r value. More broadly, the geodesic of a photon, for which proper time has no meaning, admits of analysis from first principles by way of an affine parameter λ, as contrasted with the device of first considering a particle and then going to the limit $\mu \longrightarrow 0$.

(*continued on page 676*)

Box 25.7 QUALITATIVE ANALYSIS OF ORBITS OF A PHOTON IN SCHWARZSCHILD GEOMETRY

A. Equations Governing Orbit

1. Effective-potential equation for radial part of motion:

$$\left(\frac{dr}{d\lambda}\right)^2 + B^{-2}(r) = b^{-2};$$

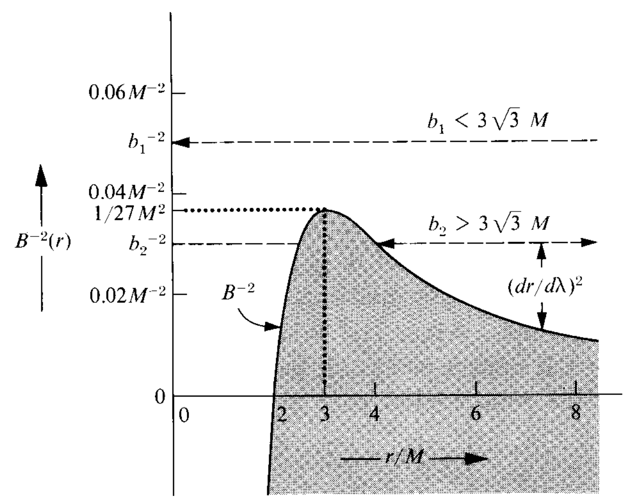
$$B^{-2}(r) = r^{-2}(1 - 2M/r);$$

$$b = \text{(impact parameter)}.$$

2. Supplementary equations to determine angular and time motion:

$$d\phi/d\lambda = 1/r^2;$$

$$dt/d\lambda = b^{-1}(1 - 2M/r)^{-1}.$$



B. Qualitative Features of Orbits (deduced from effective-potential diagram)

1. A zero-mass particle with $b > 3\sqrt{3}M$, which falls in from $r = \infty$, is "reflected off the potential barrier" (periastron; $b = B$; $dr/d\lambda = 0$) and returns to infinity.
   a. For $b \gg 3\sqrt{3}M$, the orbit is a straight line, except for a slight deflection of angle $4M/b$ (exercise 25.21; §40.3).
   b. For $0 < b - 3\sqrt{3}M \ll M$, the particle circles the star many times ("unstable circular orbit) at $r \approx 3M$ before flying back to $r = \infty$.

2. A zero-mass particle with $b < 3\sqrt{3}M$, which falls in from $r = \infty$, falls into $r = 2M$ (no periastron).

3. A zero-mass particle emitted from near $r = 2M$ escapes to infinity only if it has $b < 3\sqrt{3}M$; otherwise it reaches an apastron and then gets pulled back into $r = 2M$.

C. Escape Versus Capture as a Function of Propagation Direction

An observer at rest in the Schwarzschild gravitational field measures the ordinary velocity of a zero-mass particle relative to his orthonormal frame [equations (23.15)]:

$$v_{\hat{r}} = \frac{|g_{rr}|^{1/2}\, dr/d\lambda}{|g_{00}|^{1/2}\, dt/d\lambda} = \pm(1 - b^2/B^2)^{1/2};$$

$$v_{\hat{\phi}} = \frac{|g_{\phi\phi}|^{1/2}\, d\phi/d\lambda}{|g_{00}|^{1/2}\, dt/d\lambda} = b/B;$$

$$(v_{\hat{r}})^2 + (v_{\hat{\phi}})^2 = 1;$$

$$\begin{aligned}\delta &\equiv (\text{angle between propagation direction and radial direction})\\ &= \cos^{-1} v_{\hat{r}} = \sin^{-1} v_{\hat{\phi}}.\end{aligned}$$

To be able to cross over the potential barrier, the particle must have $b < 3\sqrt{3}M$, or $v_{\hat{\phi}}{}^2B^2 < 27M^2$, or $\sin^2\delta < 27M^2/B^2$. This result, restated:

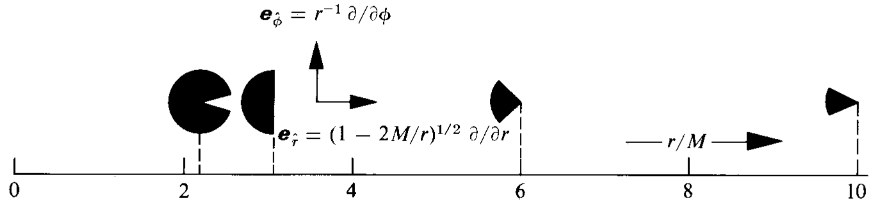
1. *A particle of zero rest mass at $r < 3M$ will eventually escape to infinity, rather than be captured by a black hole at $r = 2M$ if and only if v_r is positive and*

$$\sin\delta < 3\sqrt{3}MB^{-1}(r).$$

2. *A particle of zero rest mass at $r > 3M$ will eventually escape to infinity if and only if:* (1) *v_r is positive, or* (2) *v_r is negative and*

$$\sin\delta > 3\sqrt{3}MB^{-1}(r).$$

White, escape; black, to black hole; directions in proper reference frame



Return to the statement of the conservation laws (25.17) and (25.18) in the form that makes reference to the affine parameter λ but no reference to the rest mass μ; thus

$$\frac{d\phi}{d\lambda} = \frac{L}{r^2} \tag{25.60}$$

and

$$\frac{dt}{d\lambda} = \frac{E}{1 - 2M/r}. \tag{25.61}$$

Recall that the course of a photon in a gravitational field is governed by its direction but not by its energy. Therefore neither E nor L individually are relevant but only their ratio, the impact parameter $b = L/E$ of (25.54) and exercise 25.14. This circumstance leads one to replace the affine parameter λ by a new affine parameter,

$$\lambda_{\text{new}} = L\lambda, \tag{25.62}$$

(6) equations for orbit

that is equally constant along the world line of the photon. In this notation (drop the subscript "new" hereafter), the conservation laws take the form

$$\frac{d\phi}{d\lambda} = \frac{1}{r^2}, \tag{25.63}$$

$$\frac{dt}{d\lambda} = \frac{1}{b(1 - 2M/r)}. \tag{25.64}$$

The statement that the world line of the photon is a line of zero lapse of proper time,

$$g_{\alpha\beta}\frac{dx^\alpha}{d\lambda}\frac{dx^\beta}{d\lambda} = 0 \tag{25.65}$$

leads to the "radial equation"

$$\left(\frac{dr}{d\lambda}\right)^2 + B^{-2}(r) = b^{-2}. \tag{25.66}$$

Here one encounters again the "effective potential" $B^{-2}(r)$ of (25.58). The present fuller set of equations for the geodesic of a photon have the advantage that they reach beyond space to a description of the world line in spacetime.

(7) scattering cross section

Return to space! Figure 25.6 shows typical orbits for a photon in Schwarzschild geometry. Figure 25.7 shows angle of deflection as a function of impact parameter. From the information contained in this curve, one can evaluate the contributions to the differential scattering cross section

$$\frac{d\sigma}{d\Omega} = \sum_{\text{"branches"}} \left|\frac{2\pi b\, db}{2\pi \sin\Theta\, d\Theta}\right| \tag{25.67}$$

from the various "branches" of the scattering curve of Figure 25.7 [one turn around the center of attraction, two turns, etc.; for more on these branches and the central

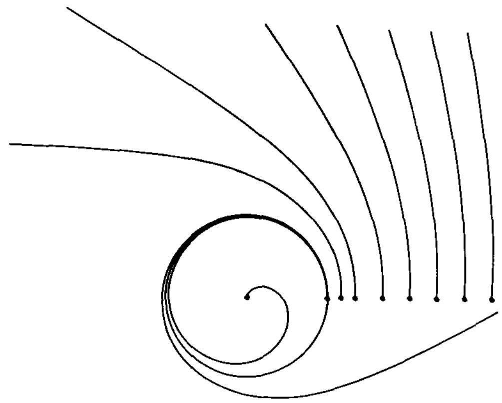


Figure 25.6.
The orbit of a photon in the "equatorial plane" of a black hole, plotted in terms of the Schwarzschild coordinates r and ϕ, for selected values of the turning point of the orbit, $r_{TP}/M = 2.99$, 3.00 (unstable circular orbit), 3.01, 3.5, 4, 5, 6, 7, 8, 9. The impact parameter is given by the formula $b = r_{TP}(1 - 2M/r_{TP})^{-1/2}$. In none of the cases shown, even for the inward plunging spiral, is the impact parameter less than $b_{crit} = (27)^{1/2} M$, nor are any of these orbits able to cross the circle $r = 3\,M$. That only happens for orbits with b less than b_{crit}. For such orbits there is no turning point; the photon comes in from infinity and ends up at $r = 0$: straight in for $b = 0$ (head-on impact); only after many loops near $r = 3M$, when $b/M = (27)^{1/2} - \varepsilon$, where ε is a very small quantity. Appreciation is expressed to Prof. R. H. Dicke for permission to publish these curves, which he had a digital calculator compute and plot out directly from the formula $d^2u/d\phi^2 = 3u^2 - u$, where $u = M/r$.

role of the deflection function $\Theta = \Theta(b)$ in the analysis of scattering, see, for example, Ford and Wheeler (1959a,b)]. For small angles the "Rutherford" part of the scattering predominates. The major part of the small-angle scattering, and in the limit $\Theta \longrightarrow 0$ all of it, comes from large impact parameters, for which one has

$$\Theta = \frac{4M}{b} \tag{25.68}$$

(see exercises 25.21 and 25.24). It follows that the limiting form of the cross section is

$$\frac{d\sigma}{d\Omega} = \left(\frac{4M}{\Theta^2}\right)^2 \qquad \text{(small } \Theta\text{)}. \tag{25.69}$$

Also, at $\Theta = \pi$ one has a singularity in the differential scattering cross section, with the character of a glory [see discussion following equation (25.44)]. Writing down the contributions of the several branches of the scattering function to the differential cross section, and summing them, one has, near $\Theta = \pi$,

$$\frac{d\sigma}{d\Omega} = \frac{M^2}{\pi - \Theta}(1.75 + 0.0029 + 0.0000055 + \cdots) = 1.75\frac{M^2}{\pi - \Theta}. \tag{25.70}$$

Thus, in principle, if one shines a powerful source of light onto a black hole, one gets a direct return of a few photons from it. Equation (25.70) provides a means to calculate the strength of this return. See exercise 25.26.

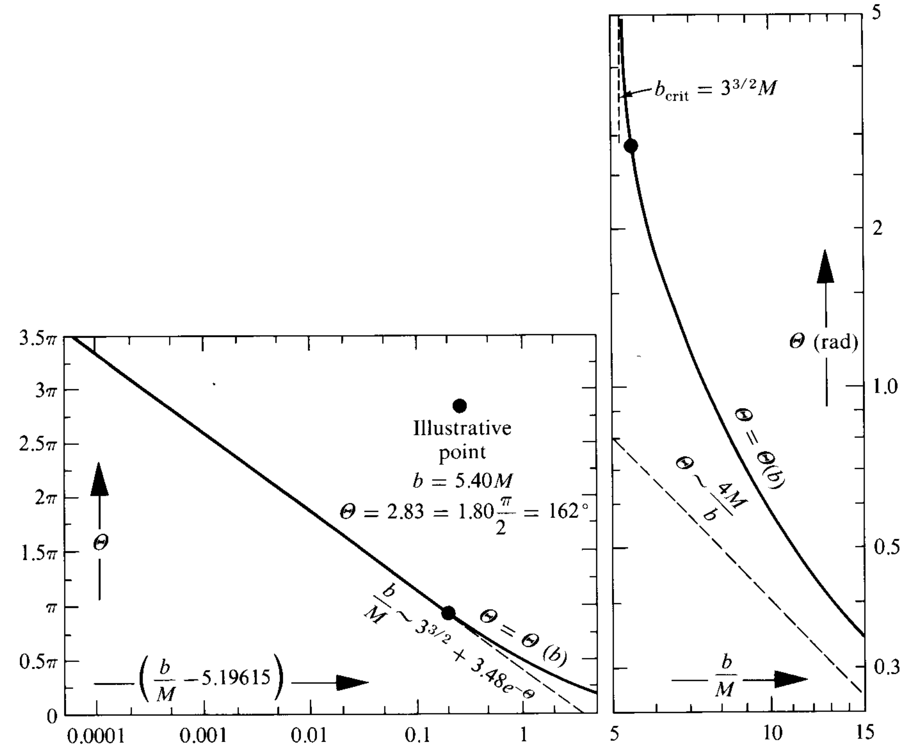


Figure 25.7.
Deflection of a photon by a Schwarzschild black hole, or by any spherically symmetric center of attraction small enough not to block the trajectory of the photon. The accurate calculations (smooth curves) are compared with formulas (dashed curves) valid asymptotically in the two limiting cases of an impact parameter, b: (1) very close to $b_{crit} = 3^{3/2}M$ (many turns around the center of attraction); and (2) very large compared to b_{crit} (small deflection). The algorithm for the accurate calculation of the deflection proceeds as follows (all distances being given, for simplicity, in units of the mass value, M). (1) Choose a value, $r = R$, for the Schwarzschild coordinate of the point of closest approach. (2) Calculate the impact parameter, b, from $b^2 = R^3/(R-2)$. (3) Calculate Q from $Q^2 = (R-2)(R+6)$. (4) Determine the modulus, k, of an "elliptic integral of the first kind" from $\sin^2\theta = k^2 = (Q - R + 6)/2Q$. (5) Determine the so-called amplitude $\phi = \phi_{min}$ of the same elliptic function from $\mathrm{sn}^2 u_{min} = \sin^2\phi_{min} = (2 + Q - R)/(6 + Q - R)$. (6) Then the total deflection is

$$\Theta = 4(R/Q)^{1/2}[F(\pi/2, \theta) - F(\phi_{min}, \theta)] - \pi.$$

The values plotted here were kindly calculated by James A. Isenberg on the basis of the work of C. G. Darwin (1959, 1961).

(8) gravitational lens effect

When the source of illumination, instead of being on the observer's side of the black hole, is on the opposite side, then in addition to the "lens effect" experienced by photons flying by with large impact parameter [literature too vast to summarize here, but see, e.g., Refsdal (1964)], and subsumed in equation (25.68), there is a glory type of illumination (intensity $\sim 1/\sin\Theta$, with now, however, Θ close to zero) received from photons that have experienced deflections $\Theta = 2\pi, 4\pi, \ldots$. This illumination comes from "rings of brightness" located at impact parameters given by $b/M - 3^{3/2} = 0.0065, 0.000012, \ldots$. Interesting though all these optical effects are as matters of principle, they are, among all the ways to observe a black hole, the worst; see part VI, C, of Box 33.3 for a detailed discussion.

EXERCISES

Exercise 25.23. QUALITATIVE FEATURES OF PHOTON ORBITS

Verify all the statements about orbits for particles of zero rest mass made in Box 25.7.

Exercise 25.24. LIGHT DEFLECTION

Using the dimensionless variable $u = M/r$ in place of r itself, and $u_b = M/b$ in place of the impact parameter, transform (25.55) into the first-order equation

$$\left(\frac{du}{d\phi}\right)^2 + (1 - 2u)u^2 = u_b^2 \tag{25.71}$$

and thence, by differentiation, into

$$\frac{d^2u}{d\phi^2} + u = 3u^2. \tag{25.72}$$

(a) In the large-impact-parameter or small-u approximation, in which the term on the right is neglected, show that the solution of (25.72) yields elementary rectilinear motion (zero deflection).

(b) Insert this zero-order solution into the perturbation term $3u^2$ on the righthand side of (25.72), and solve anew for u ("rectilinear motion plus first-order correction"). In this way, verify the formula for the bending of light by the sun given by putting $\beta = 1$ in equation (25.49).

Exercise 25.25. CAPTURE OF LIGHT BY A BLACK HOLE

Show that a Schwarzschild black hole presents a cross section $\sigma_{\text{capt}} = 27\pi M^2$ for capture of light.

Exercise 25.26. RETURN OF LIGHT FROM A BLACK HOLE

Show that flashing a powerful pulse of light onto a black hole leads in principle to a return from rings of brightness located at $b/M - 3^{3/2} = 0.151, 0.00028, \ldots$. How can one evaluate the difference in time delays of these distinct returns? Show that the intensity I of the return (erg/cm^2) as a function of the energy E_0(erg/steradian) of the original pulse, the mass M(cm) of the black hole, the distance R to it, and the lateral distance r from the "flashlight" to the receptor of returned radiation is

$$I = \frac{E_0}{R^3 r} \sum_{\Theta = (2N+1)\pi} \left|\frac{2b\,db}{d\Theta}\right| = \frac{E_0 M^2}{R^3 r}(1.75 + 0.0029 + 0.0000055 + \cdots)$$

under conditions where diffraction can be neglected.

§25.7. SPHERICAL STAR CLUSTERS

By combining orbit theory, as developed in this chapter, with kinetic theory in curved spacetime as developed in §22.6, one can formulate the theory of relativistic star clusters.

Static, spherical star clusters:

Consider, for simplicity, a spherically symmetric cluster of stars (e.g., a globular cluster, but one so dense that relativistic gravitational effects might be important).

Demand that the cluster be static, in the sense that the number density in phase space $\mathcal{N}$ is independent of time. (New stars, flying along geodesic orbits, enter a fixed region in phase space at the same rate as "old" stars leave it.) Ignore collisions and close encounters between stars; i.e., treat each star's orbit as a geodesic in the spherically symmetric spacetime of the cluster as a whole.

(1) foundations for analysis

With these idealizations accepted, one can write down a manageable set of equations for the structure of the cluster.* Since the cluster is static and spherical, so must be its gravitational field. Consequently, one can introduce the same kind of coordinate system ("Schwarzschild coordinates") as was used for a static spherical star in Chapter 23:

$$ds^2 = -e^{2\Phi}\, dt^2 + e^{2\Lambda}\, dr^2 + r^2\, d\Omega^2; \qquad \Phi = \Phi(r), \qquad \Lambda = \Lambda(r). \tag{25.73}$$

In the tangent space at each event in spacetime reside the momentum vectors of the swarming stars. For coordinates in this tangent space ("momentum space"), it is convenient to use the physical components of 4-momentum, $p^{\hat{\alpha}}$—i.e., components on the orthonormal frame

$$\boldsymbol{\omega}^{\hat{t}} = e^{\Phi}\, \boldsymbol{d}t, \qquad \boldsymbol{\omega}^{\hat{r}} = e^{\Lambda}\, \boldsymbol{d}r, \qquad \boldsymbol{\omega}^{\hat{\theta}} = r\, \boldsymbol{d}\theta, \qquad \boldsymbol{\omega}^{\hat{\phi}} = r \sin\theta\, \boldsymbol{d}\phi. \tag{25.74}$$

Then the number density of stars in phase space is a spherically symmetric, static function

$$\mathcal{N} = \mathcal{N}[r, p^{\hat{0}}, p^{\hat{r}}, (p^{\hat{\theta}2} + p^{\hat{\phi}2})^{1/2}]. \tag{25.75}$$

[$\mathcal{N}$ is independent of t because the cluster is static; and independent of θ, ϕ, and angle $\Theta = \tan^{-1}(p^{\hat{\phi}}/p^{\hat{\theta}})$ because of spherical symmetry.]

The functions describing the structure of the cluster, Φ, Λ, and $\mathcal{N}$, are determined by the kinetic (also, in this context, called the Vlasoff) equation (§22.6)

$$d\mathcal{N}/d\lambda = 0, \text{ i.e., } \mathcal{N} \text{ conserved along orbit of each star in phase space;} \tag{25.76a}$$

and by the Einstein field equations

$$G^{\hat{\alpha}\hat{\beta}} = 8\pi T^{\hat{\alpha}\hat{\beta}} = 8\pi \int (\mathcal{N} p^{\hat{\alpha}} p^{\hat{\beta}}) \mu^{-1}\, dp^{\hat{0}}\, dp^{\hat{r}}\, dp^{\hat{\theta}}\, dp^{\hat{\phi}}. \tag{25.76b}$$

[The Vlasoff equation for Newtonian star clusters is treated by Ogorodnikov (1965). The above expression for the stress-energy tensor of a swarm of particles (stars) was derived in exercise 22.18. Here, as in exercise 22.18, the particles (stars) are assumed *not* all to have the same rest mass. Note that rest mass is here denoted μ, but in Chapter 22 it was denoted m.]

(2) solution of Vlasoff equation

To solve the Vlasoff equation, one need only note that $\mathcal{N}$ is conserved along stellar orbits and therefore must be a function of the constants of the orbital motion. There is a constant of motion corresponding to each Killing vector in the cluster's static, spherical spacetime (see exercise 25.8):

*These equations were first derived and explored by Zel'dovich and Podurets (1965).

$$E = \text{“energy at infinity”} = -\boldsymbol{p} \cdot (\partial/\partial t) = -p_0,$$
$$L_z = \text{“}z\text{-component of angular momentum”} = p \cdot \boldsymbol{\xi}_z = p \cdot (\partial/\partial\phi) = p_\phi,$$
$$L_y = \text{“}y\text{-component of angular momentum”} = p \cdot \boldsymbol{\xi}_y, \tag{25.77a}$$
$$L_x = \text{“}x\text{-component of angular momentum”} = p \cdot \boldsymbol{\xi}_x.$$

In addition, each star's rest mass

$$\mu = (p^{\hat{0}2} - p^{\hat{r}2} - p^{\hat{\theta}2} - p^{\hat{\phi}2})^{1/2} \tag{25.77b}$$

is a constant of its motion. The general solution of the Vlasoff equation, then, has the form

$$\mathcal{N} = H(E, L_x, L_y, L_z, \mu).$$

But this general solution is not spherically symmetric. For example, the distribution function

$$\mathcal{N} = H(E, \mu, L_z)\, \delta(L_y)\, \delta(L_x),$$

corresponds to a cluster of stars with orbits all in the equatorial plane $\theta = \pi/2$ ($L_y = L_x = 0$ for all stars in cluster). To be spherical the cluster's distribution function must depend only on the magnitude

$$L = (L_x{}^2 + L_y{}^2 + L_z{}^2)^{1/2}$$

of the angular momentum, and not on its direction (not on the orientation of a star's orbital plane). Thus, the general spherical solution to the Vlasoff equation in a static, spherical spacetime must have the form

$$\mathcal{N} = F(E, L, \mu). \tag{25.78}$$

To use this general solution, one must reexpress the constants of the motion E, L, μ, in terms of the agreed-on phase-space coordinates $(t, r, \theta, \phi, p^{\hat{0}}, p^{\hat{r}}, p^{\hat{\theta}}, p^{\hat{\phi}})$. The rest mass of a star is given by (25.77b). The energy-at-infinity is obtained by red-shifting the locally measured energy

$$E = -p_0 = e^{\Phi} p^{\hat{0}}. \tag{25.79a}$$

For an orbit in the equatorial plane ($p_\theta = p^\theta = p^{\hat{\theta}} = 0$; $L_x = L_y = 0$), the total angular momentum has the form

$$L = |L_z| = |p_\phi| = |rp^{\hat{\phi}}| = r \times (\text{“tangential” component of 4-momentum}).$$

By symmetry, the equation $L = r \times$ (“tangential” component of $\boldsymbol{p}$) must hold true also for orbits in other planes; it must be perfectly general:

$$L = rp^{\hat{T}}, \tag{25.79b}$$

$$p^{\hat{T}} \equiv (\text{tangential component of 4-momentum}) = [(p^{\hat{\theta}})^2 + (p^{\hat{\phi}})^2]^{1/2} \tag{25.80}$$

(see exercise 25.9).

2

(3) "smeared-out" stress-energy tensor due to stars

Before solving the Einstein field equations, one finds it useful to reduce the stress-energy tensor to a more explicit form than (25.76b). The off-diagonal components $T^{\hat{0}\hat{j}}$ and $T^{\hat{j}\hat{k}}$ ($j \neq k$) all vanish because their integrands are odd functions of $p^{\hat{j}}$. The integrands for the diagonal components $T^{\hat{0}\hat{0}}$, $T^{\hat{r}\hat{r}}$, and $\frac{1}{2}(T^{\hat{\theta}\hat{\theta}} + T^{\hat{\phi}\hat{\phi}})$ are independent of angle $\Theta \equiv \tan^{-1}(p^{\hat{\phi}}/p^{\hat{\theta}})$ in the tangential momentum plane; so the momentum volume element can be rewritten as

$$dp^{\hat{0}}\, dp^{\hat{r}}\, dp^{\hat{\theta}}\, dp^{\hat{\phi}} \longrightarrow 2\pi p^{\hat{T}}\, dp^{\hat{T}}\, dp^{\hat{r}}\, dp^{\hat{0}}.$$

Changing variables from $(p^{\hat{T}}, p^{\hat{r}}, p^{\hat{0}})$ to $(p^{\hat{T}}, \mu, p^{\hat{0}})$ where

$$\mu = [(p^{\hat{0}})^2 - (p^{\hat{r}})^2 - (p^{\hat{T}})^2]^{1/2},$$

and recognizing that two values of $p^{\hat{r}}$ ($\pm p^{\hat{r}}$) correspond to each value of μ, one brings the volume element into the form

$$2\pi p^{\hat{T}}\, dp^{\hat{T}}\, dp^{\hat{r}}\, dp^{\hat{0}} \longrightarrow 4\pi (p^{\hat{T}}\mu/p^{\hat{r}})\, dp^{\hat{T}}\, dp^{\hat{0}}\, d\mu.$$

The diagonal components of $\boldsymbol{T}$ [equation (25.76b)] then read

$$\begin{aligned} \rho \equiv T^{\hat{0}\hat{0}} &= \text{(total density of mass-energy)} \\ &= 4\pi \int F(e^{\Phi}p^{\hat{0}}, rp^{\hat{T}}, \mu)(p^{\hat{0}2}p^{\hat{T}}/p^{\hat{r}})\, dp^{\hat{T}}\, dp^{\hat{0}}\, d\mu, \end{aligned} \tag{25.81a}$$

$$\begin{aligned} P_T \equiv \frac{1}{2}(T^{\hat{\theta}\hat{\theta}} + T^{\hat{\phi}\hat{\phi}}) &= T^{\hat{\theta}\hat{\theta}} = T^{\hat{\phi}\hat{\phi}} = \text{(tangential pressure)} \\ & \text{[by spherical symmetry]} \\ &= 2\pi \int F(e^{\Phi}p^{\hat{0}}, rp^{\hat{T}}, \mu)[(p^{\hat{T}})^3/p^{\hat{r}}]\, dp^{\hat{T}}\, dp^{\hat{0}}\, d\mu, \end{aligned} \tag{25.81b}$$

$$\begin{aligned} P_r \equiv T^{\hat{r}\hat{r}} &= \text{(radial pressure)} \\ &= 4\pi \int F(e^{\Phi}p^{\hat{0}}, rp^{\hat{T}}, \mu)(p^{\hat{r}}p^{\hat{T}})\, dp^{\hat{T}}\, dp^{\hat{0}}\, d\mu. \end{aligned} \tag{25.81c}$$

When performing these integrals, one must express $p^{\hat{r}}$ in terms of the variables of integration,

$$p^{\hat{r}} = [(p^{\hat{0}})^2 - (p^{\hat{T}})^2 - \mu^2]^{1/2}. \tag{25.81d}$$

(4) solution of field equations

The Einstein field equations for this stress-energy tensor and the metric (25.73), after use of expressions (14.43) for $G^{\hat{\alpha}\hat{\beta}}$ and after manipulations analogous to those for a spherical star (§23.5), reduce to

$$e^{2\Lambda} = (1 - 2m/r)^{-1}, \qquad m = \int_0^r 4\pi r^2 \rho\, dr; \tag{25.82a}$$

$$\frac{d\Phi}{dr} = \frac{m + 4\pi r^3 P_r}{r(r - 2m)}. \tag{25.82b}$$

These equations, together with the assumed form $F(E, L, \mu)$ of the distribution function and the integrals (25.81) for ρ, P_r, and P_T, determine the structure of the cluster. Box 25.8 gives an overview of these structure equations, and specializes them for an isotropic velocity distribution. Box 25.9 presents and discusses the solution to the equations for an isothermal star cluster (truncated Maxwellian velocity distribution).

EXERCISES

Exercise 25.27. ISOTROPIC STAR CLUSTER

For a cluster with distribution function independent of angular momentum, derive properties B.1 to B.6 of Box 25.8.

Exercise 25.28. SELF-SIMILAR CLUSTER [See Bisnovatyi-Kogan and Zel'dovich (1969), Bisnovatyi-Kogan and Thorne (1970).]

(a) Find a solution to the equations of structure for a spherical star of infinite central density, with the equation of state $P = \gamma\rho$, where γ is a constant $(0 < \gamma < 1/3)$.

(b) Find an isotropic distribution function $F(E, \mu)$ that leads to a star cluster with the same distributions of ρ, P, m, and Φ as in the gas sphere of part (a). (See Box 25.8.) [*Answer:*

$$P = \gamma\rho = \frac{\gamma^2}{1 + 6\gamma + \gamma^2}\frac{1}{2\pi r^2},$$

$$e^{2\Lambda} = (1 - 2m/r)^{-1} = (1 + 6\gamma + \gamma^2)/(1 + \gamma)^2,$$

$$e^{2\Phi} = Br^{4\gamma/(1+\gamma)}, \qquad B = \text{const};$$

$$F = A(E/B^{1/2})^{-(1+\gamma)/\gamma}\,\delta(\mu - \mu_0) = Ar^{-2}(E_{\text{local}})^{-(1+\gamma)/\gamma}, \qquad A = \text{const.}]$$

Exercise 25.29. CLUSTER WITH CIRCULAR ORBITS

What must be the form of the distribution function to guarantee that all stars move in circular orbits? Specialize the equations of structure to this case. Analyze the stability of the orbits of individual stars in the cluster, using an effective-potential diagram. What conditions must the distribution function satisfy if all orbits are to be stable? [See Einstein (1939), Zapolsky (1968).]

Box 25.8 EQUATIONS OF STRUCTURE FOR A SPHERICAL STAR CLUSTER

A. To Build a Model for a Star Cluster, Proceed as Follows

1. Specify the distribution function $\mathfrak{N} = F(E, L, \mu)$, where

$$E = \text{energy-at-infinity of a star},$$
$$L = \text{angular momentum of a star},$$
$$\mu = \text{rest mass of a star}.$$

2. Solve the following two integro-differential equations for the metric functions $m = \frac{1}{2}r(1 - e^{-2\Lambda})$ and Φ of the line element

2

Box 25.8 (*continued*)

$$ds^2 = -e^{2\Phi}\, dt^2 + e^{2\Lambda}\, dr^2 + r^2\, d\Omega^2:$$

$$m = \int_0^r 4\pi r^2 \rho\, dr,$$

$$\frac{d\Phi}{dr} = \frac{m + 4\pi r^3 P_r}{r(r - 2m)},$$

where

$$\rho = 4\pi \int F(e^{\Phi} p^{\hat{0}}, rp^{\hat{T}}, \mu)[(p^{\hat{0}})^2 p^{\hat{T}}/p^{\hat{r}}]\, dp^{\hat{T}}\, dp^{\hat{0}}\, d\mu,$$

$$P_T = 2\pi \int F(e^{\Phi} p^{\hat{0}}, rp^{\hat{T}}, \mu)[(p^{\hat{T}})^3/p^{\hat{r}}]\, dp^{\hat{T}}\, dp^{\hat{0}}\, d\mu,$$

$$P_r = 4\pi \int F(e^{\Phi} p^{\hat{0}}, rp^{\hat{T}}, \mu)(p^{\hat{r}} p^{\hat{T}})\, dp^{\hat{T}}\, dp^{\hat{0}}\, d\mu,$$

$$p^{\hat{r}} = [(p^{\hat{0}})^2 - (p^{\hat{T}})^2 - \mu^2]^{1/2}.$$

The integrations for ρ, P_T, and P_r go over all positive $p^{\hat{T}}$, $p^{\hat{0}}$, μ for which $(p^{\hat{0}})^2 - (p^{\hat{T}})^2 - \mu^2 \geq 0$.

B. If the Distribution Function is Independent of Angular Momentum, Then

1. $F = F(E, \mu)$.
2. The distribution of stellar velocities at each point in the cluster is isotropic.
3. $\rho = 4\pi \int F(e^{\Phi} p^{\hat{0}}, \mu)[(p^{\hat{0}})^2 - \mu^2]^{1/2} (p^{\hat{0}})^2\, dp^{\hat{0}}\, d\mu$.
4. The pressure is isotropic:

$$P_r = P_T \equiv P \equiv \frac{4\pi}{3} \int F(e^{\Phi} p^{\hat{0}}, \mu)(p^{\hat{0}2} - \mu^2)^{3/2}\, dp^{\hat{0}}\, d\mu.$$

5. The total density of mass-energy ρ, the pressure P, and the metric functions Φ and $m = \frac{1}{2}r(1 - e^{-2\Lambda})$ satisfy the equations of structure for a gas sphere ("star"),

$$m = \int 4\pi r^2 \rho\, dr,$$

$$\frac{d\Phi}{dr} = \frac{m + 4\pi r^3 P}{r(r - 2m)},$$

$$\frac{dP}{dr} = -\frac{(\rho + P)(m + 4\pi r^3 P)}{r(r - 2m)}.$$

6. Thus, to every static, spherical star cluster with isotropic velocity distribution, there corresponds a unique gas sphere that has the same distributions of ρ, P, m, and Φ.
7. Conversely [see Fackerell (1968)], given a gas sphere (solution to equations of stellar structure for ρ, P, m, and Φ), one can always find a distribution function $F(E, \mu)$ that describes a cluster with the same ρ, P, m, and Φ. But for some gas spheres F is necessarily negative in part of phase space, and is thus unphysical.

Box 25.9 ISOTHERMAL STAR CLUSTERS

A. Distribution Function

1. In any relativistic star cluster, one might expect that occasional close encounters between stars would "thermalize" the stellar distribution function. This suggests that one study isotropic, spherical clusters with the Boltzmann distribution function (tacitly assumed zero for $p^{\hat{0}} = Ee^{-\Phi} < \mu_0$)

$$\mathcal{N} = F(E, L, \mu) = Ke^{-E/T}\,\delta(\mu - \mu_0). \tag{1}$$

Here K is a normalization constant, T is a constant "temperature," and for simplicity the stars are all assumed to have the same rest mass μ_0.

2. In such a cluster, an observer at radius r sees a star of energy-at-infinity E to have locally measured energy

$$p^{\hat{0}} = (\text{rest mass-energy}) + (\text{kinetic energy}) = \frac{\mu_0}{(1 - v^2)^{1/2}} = Ee^{-\Phi(r)}. \tag{2}$$

Consequently, the stars in his neighborhood have a Boltzmann distribution

$$\frac{dN}{d^3\hat{p}\, d^3\hat{x}\, d\mu} = \mathcal{N} = K \exp(-p^{\hat{0}}/T_{\text{loc}})\, \delta(\mu - \mu_0) \tag{3}$$

with locally measured temperature

$$T_{\text{loc}}(r) = Te^{-\Phi(r)}. \tag{4}$$

Thus, the temperature of the cluster is subject to identically the same redshift-blueshift effects as photons, particles, and stars that move about in the cluster. (For a derivation of this same temperature-redshift law for a gas in thermal equilibrium, see part (e) of exercise 22.7.)

3. Actually, the Boltzmann distribution (1) can never be achieved. Stars with $E > \mu_0$ are gravitationally unbound from the cluster and will escape. The Boltzmann distribution presumes that, as such stars go zooming off toward $r = \infty$, an equal number of stars with the same energies come zooming in from $r = \infty$ to maintain an unchanged distribution function. Such a situation is clearly unrealistic. Instead, one expects the escape of stars to truncate the distribution at some energy E_{max} slightly less than μ_0. The result, in idealized form, is the "truncated Boltzmann distribution"

$$\mathcal{N} = F(E, L, \mu) = \begin{cases} Ke^{-E/T}\,\delta(\mu - \mu_0), & E < E_{\text{max}}, \\ 0, & E > E_{\text{max}}. \end{cases} \tag{5}$$

Box 25.9 (*continued*)

B. Structure and Stability of Cluster Models

1. Models for star clusters with truncated Boltzmann distributions have been constructed by Zel'dovich and Podurets (1965), by Fackerell (1966), and by Ipser (1969), using the procedure of Box 25.8. Ipser has analyzed the collisionless radial vibrations of such clusters.

2. In general, these clusters form a 4-parameter family (K, T, μ_0, E_{max}). Replace the parameter K by the total rest mass of the cluster, $M_0 = \mu_0 N$, where N is the total number of stars. Replace T by the temperature per unit rest mass, $\widetilde{T} = T/\mu_0$. Replace E_{max} by the maximum energy per unit rest mass, $\widetilde{E}_{\text{max}} = E_{\text{max}}/\mu_0$. Then the clusters are parametrized by (M_0, $\widetilde{T}$, μ_0, $\widetilde{E}_{\text{max}}$). When one now doubles μ_0, holding M_0, $\widetilde{T}$, $\widetilde{E}_{\text{max}}$ fixed (and thus halving the total number of stars), all macroscopic features of the cluster remain unchanged. In this sense μ_0 is a "trivial parameter" and can henceforth be ignored or changed at will. The total rest mass of the cluster M_0 can be regarded as a "scaling factor"; all dimensionless features of the cluster are independent of it. For example, if ρ_c is the central density of mass-energy [equation (25.81a), evaluated at $r = 0$], then $\rho_c M_0{}^2$ is dimensionless and is thus independent of M_0, which means that $\rho_c \propto M_0{}^{-2}$. Only two nontrivial parameters remain: $\widetilde{T}$ and $\widetilde{E}_{\text{max}}$.

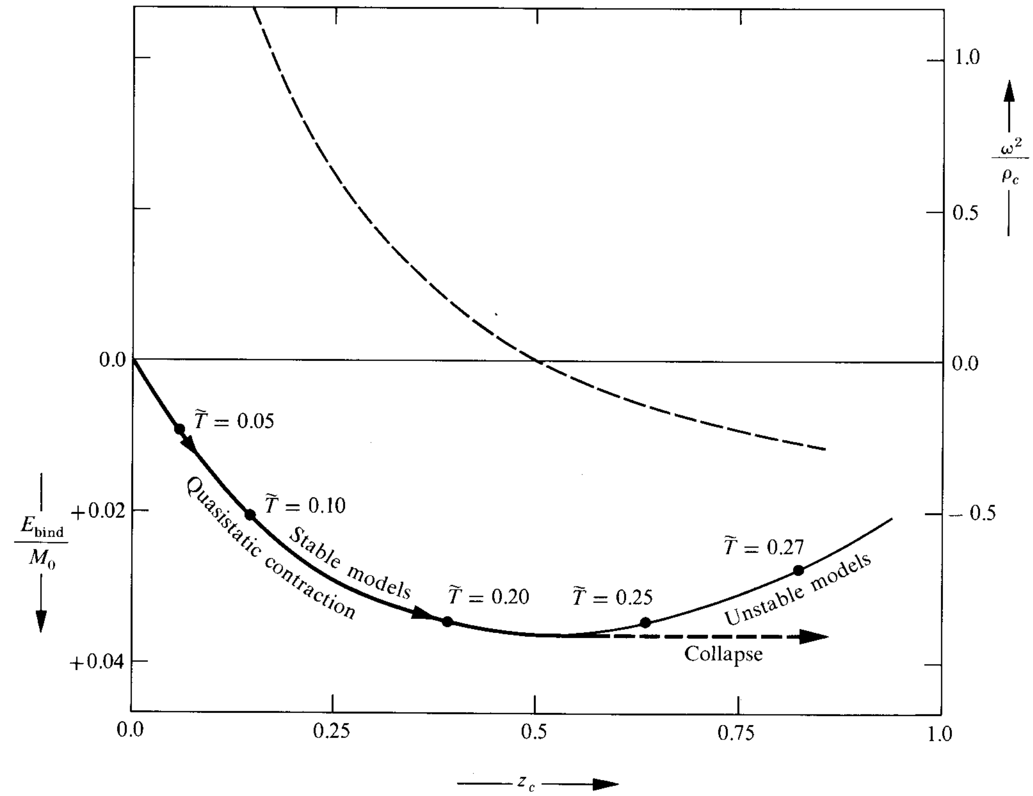
3. Consider as an instructive special case [Zel'dovich and Podurets (1965)] the one-parameter sequence with $\widetilde{E}_{\text{max}} = 1 - \frac{1}{2}\widetilde{T}$. The following figure, computed by Ipser (1969), plots for this sequence the fractional binding energy,

$$E_{\text{bind}}/M_0 \equiv (M_0 - M)/M_0 \tag{6}$$

(here M is total mass-energy); the square of the angular frequency for collisionless vibrations (vibration amplitude $\propto e^{-i\omega t}$) divided by central density of mass-energy, ω^2/ρ_c; and the redshift, z_c, of photons emitted from the center of the cluster and received at infinity. All these quantities are dimensionless, and thus depend only on the choice of $\widetilde{T} = T/\mu_0$.

4. Notice that all models beyond the point of maximum binding energy ($z_c \gtrsim 0.5$) are unstable against collisionless radial perturbations (ω imaginary; amplitude of perturbation $\propto e^{|\omega|t}$). When perturbed slightly, such clusters must collapse to form black holes. (See Chapter 26 for an analysis of the analogous instability in stars).

5. These results suggest an idealized story of the evolution of a spherical cluster [Zel'dovich and Podurets (1965); Fackerell, Ipser, and Thorne (1969)]. The



cluster would evolve quasistatically along a sequence of spherical equilibrium configurations such as those of the figure. The evolution would be driven by stellar collisions and by the evaporation of stars. When two stars collide and coalesce, they increase the cluster's rest mass and hence its fractional binding energy. When a star gains enough energy from such encounters to escape from the cluster, it carries away excess kinetic energy, leaving the cluster more tightly bound. Thus, both collisions and evaporation should drive the cluster toward states of tighter and tighter binding. When the cluster reaches the point, along its sequence, of maximum fractional binding energy, it can no longer evolve quasistatically. Relativistic gravitational collapse sets in: the stars spiral inward through the gravitational radius of the cluster toward its center, leaving behind a black hole with, perhaps, some remaining stars orbiting it.

It is tempting to speculate that violent events in the nuclei of some galaxies and in quasars might be associated with the onset of such a collapse, or with encounters between an already collapsed cluster (black hole) and surrounding stars.

CHAPTER 26

STELLAR PULSATIONS

§26.1. MOTIVATION

This chapter is entirely Track 2, but it neither depends on nor prepares for any other chapter.

The *raison d'etre* of this chapter

In relativistic astrophysics, as elsewhere in physics, most problems of deep physical interest are too difficult and too complex to be solved exactly. They can be solved only by use of approximation techniques. And of all approximation techniques, the one that has the widest range of application is perturbation theory.

Perturbation calculations are typically long, tedious, and filled with complicated mathematical expressions. Therefore, they are not appropriate for a textbook such as this. Nevertheless, because it is so important that aspiring astrophysicists know how to set up and carry out perturbation calculations in general relativity, the authors have chosen to present one example in detail.

The example chosen is an analysis of adiabatic, radial pulsations of a nonrotating, relativistic star. Two features of this example are noteworthy: (1) it is sufficiently complex to be instructive, but sufficiently simple to be presentable; (2) in the results of the calculation one can discern and quantify the relativistic instability that is so important for modern astrophysics (see Chapter 24).

The calculation presented here is patterned after that of Chandrasekhar (1964a,b), which first revealed the existence of the relativistic instability. For an alternative calculation, based on the concept of "extremal energy," see Appendix B of Harrison, Thorne, Wakano, and Wheeler (1965); and for a calculation based on extremal entropy, see Cocke (1965).

The authors are deeply indebted to Mr. Carlton M. Caves, who found and corrected many errors in the equations of this chapter and of a dozen other chapters.

§26.2. SETTING UP THE PROBLEM

The system to be analyzed is a sphere of perfect fluid, pulsating radially with very small amplitude. To analyze the pulsations one needs (a) the exact equations governing the equilibrium configuration about which the sphere pulsates; (b) a coordinate system for the vibrating sphere that reduces, for zero pulsation amplitude, to the standard Schwarzschild coordinates of the equilibrium sphere; (c) a set of small functions describing the pulsation (radial displacement and velocity, pressure and density perturbations, first-order changes in metric coefficients), in which to linearize; and (d) a set of equations governing the evolution of these "perturbation functions."

Setting up the analysis of stellar pulsations

a. Equilibrium Configuration

The equations of structure for the equilibrium sphere are those summarized in §23.7. It will be useful to rewrite them here, with a few changes of notation (use of subscript "o" to denote "unperturbed configuration"; use of $\Lambda = -\frac{1}{2}\ln(1 - 2m/r)$ in place of m in all equations; use of a prime to denote derivatives with respect to r):

Equilibrium configuration of star

$$ds^2 = -e^{2\Phi_o}\,dt^2 + e^{2\Lambda_o}\,dr^2 + r^2(d\theta^2 + \sin^2\theta\,d\phi^2), \tag{26.1a}$$

$$\Lambda_o' = \frac{1}{2r}(1 - e^{2\Lambda_o}) + 4\pi r\rho_o e^{2\Lambda_o}, \tag{26.1b}$$

$$p_o' = -(\rho_o + p_o)\Phi_o', \tag{26.1c}$$

$$\Phi_o' = -\frac{1}{2r}(1 - e^{2\Lambda_o}) + 4\pi r p_o e^{2\Lambda_o}. \tag{26.1d}$$

b. Coordinates for Perturbed Configuration

The gas sphere pulsates in a radial, i.e., spherically symmetric, manner. Consequently, its spacetime geometry must be spherical. In Box 23.3 it is shown that for any spherical spacetime, whether dynamic or static, one can introduce Schwarzschild coordinates with a line element

$$ds^2 = -e^{2\Phi}\,dt^2 + e^{2\Lambda}\,dr^2 + r^2(d\theta^2 + \sin^2\theta\,d\phi^2), \tag{26.2}$$
$$\Phi = \Phi(t, r), \qquad \Lambda = \Lambda(t, r).$$

Coordinates for perturbed configuration

One uses these coordinates for the pulsating sphere because they reduce to the unperturbed coordinates when the pulsations have zero amplitude.

c. Perturbation Functions

When the pulsations have very small amplitude, the metric coefficients, Φ and Λ, and the thermodynamic variables p, ρ, and n as measured in the fluid's rest frame have very nearly their unperturbed values. Denote by $\delta\Phi$, $\delta\Lambda$, δp, $\delta\rho$, and δn the perturbations at fixed coordinate locations:

Perturbation functions

$$\begin{aligned} \Phi(t,r) &= \Phi_o(r) + \delta\Phi(t,r), \qquad & \Lambda(t,r) &= \Lambda_o(r) + \delta\Lambda(t,r), \\ p(t,r) &= p_o(r) + \delta p(t,r), \qquad & \rho(t,r) &= \rho_o(r) + \delta\rho(t,r), \\ & & n(t,r) &= n_o(r) + \delta n(t,r). \end{aligned} \tag{26.3a}$$

Besides $\delta\Phi$, $\delta\Lambda$, δp, $\delta\rho$, and δn, one more perturbation function is needed to describe the pulsations: the radial displacement ξ of the fluid from its equilibrium position:

> A fluid element located at coordinate radius r in the unperturbed configuration is displaced to coordinate radius $r + \xi(r, t)$ at coordinate time t in the vibrating configuration. (26.3b)

To make the analysis of the pulsations tractable, all equations will be linearized in the perturbation functions ξ, $\delta\Phi$, $\delta\Lambda$, δp, $\delta\rho$, and δn.

d. Equations of Evolution

How to derive equations governing the perturbation functions

The evolution of the perturbation functions with time will be governed by the Einstein field equations, the local law of conservation of energy-momentum $\boldsymbol{\nabla}\cdot\boldsymbol{T} = 0$, and the laws of thermodynamics—all appropriately linearized. The analysis from here on is nothing but a reduction of those equations to "manageable form." Of course, the reduction will proceed most efficiently if one knows in advance what form one seeks. The goal in this calculation and in most similar calculations is simple: (1) obtain a set of *dynamic equations* for the true dynamic degrees of freedom (only the fluid displacement ξ in this case; the fluid displacement and the amplitudes of the gravitational waves in a nonspherical case, where waves are possible); and (2) obtain a set of *initial-value equations* expressing the remaining perturbation functions ($\delta\Phi$, $\delta\Lambda$, δp, $\delta\rho$, and δn in this case) in terms of the dynamic degrees of freedom (ξ).

§26.3. EULERIAN VERSUS LAGRANGIAN PERTURBATIONS

Eulerian perturbations defined

Lagrangian perturbations defined

Before deriving the dynamic and initial-value equations, it is useful to introduce a new concept: the "Lagrangian perturbation" in a thermodynamic variable. The perturbations δp, $\delta\rho$, and δn of equations (26.3) are *Eulerian perturbations* in p, ρ, and n; i.e., they are changes measured by an observer who sits forever at a fixed point (t, r, θ, ϕ) in the coordinate grid. By contrast, the *Lagrangian perturbations*—denoted Δp, $\Delta\rho$, and Δn—are changes measured by an observer who moves with the fluid; i.e., by an observer who would sit at radius r in the unperturbed configuration, but sits at $r + \xi(t, r)$ in the perturbed configuration:

Relation between Eulerian and Lagrangian perturbations

$$\begin{aligned} \Delta p(t, r) &= p[t, r + \xi(t, r)] - p_o(r) \\ &\approx \delta p + p_o{}'\xi; \end{aligned} \tag{26.4a}$$

$$\begin{aligned} \Delta\rho(t, r) &= \rho[t, r + \xi(t, r)] - \rho_o(r) \\ &\approx \delta\rho + \rho_o{}'\xi; \end{aligned} \tag{26.4b}$$

$$\begin{aligned} \Delta n(t, r) &= n[t, r + \xi(t, r)] - n_o(r) \\ &\approx \delta n + n_o{}'\xi. \end{aligned} \tag{26.4c}$$

§26.4. INITIAL-VALUE EQUATIONS

a. Baryon Conservation

The law of baryon conservation, $\boldsymbol{\nabla} \cdot (n\boldsymbol{u}) = 0$ (§22.2), governs the evolution of perturbations Δn and δn in baryon number. By applying the chain rule to the divergence and using the relation $\boldsymbol{u} \cdot \boldsymbol{\nabla} n = \boldsymbol{\nabla}_{\boldsymbol{u}} n = dn/d\tau$, one can rewrite the conservation law as

Derivation of initial value equations:
(1) for baryon perturbations Δn and δn

$$\underbrace{dn/d\tau}_{\text{[derivative of } n \text{ along fluid world line]}} = -n(\boldsymbol{\nabla} \cdot \boldsymbol{u}).$$

In terms of Δn, the perturbation measured by an observer moving with the fluid, this equation can be rewritten as

$$\frac{d\,\Delta n}{d\tau} = -n(\boldsymbol{\nabla} \cdot \boldsymbol{u}). \tag{26.5}$$

To reduce this equation further, one needs an expression for the fluid's 4-velocity. It is readily derived from

$$\frac{u^r}{u^t} = \left(\frac{dr/d\tau}{dt/d\tau}\right) = \left(\frac{dr}{dt}\right)_{\text{along world line}} = \frac{\partial\xi}{\partial t} \equiv \dot{\xi},$$

$$(u^t)^2 e^{2\Phi} - (u^r)^2 e^{2\Lambda} = 1.$$

The result to first order in ξ, $\delta\Lambda$, and $\delta\Phi$ is

$$u^t = e^{-\Phi} = e^{-\Phi_o}(1 - \delta\Phi), \qquad u^r = \dot{\xi} e^{-\Phi_o}. \tag{26.6}$$

Using these components in equation (26.5), and using the relations

$$\frac{d}{d\tau} = \boldsymbol{u} = u^\alpha \frac{\partial}{\partial x^\alpha}, \qquad \boldsymbol{\nabla} \cdot \boldsymbol{u} = \frac{1}{\sqrt{-g}}(\sqrt{-g}\,u^\alpha)_{,\alpha},$$

together with the vibrating metric (26.2), one reduces equation (26.5) to a relation whose time integral is

$$\Delta n = -n_o[r^{-2}e^{-\Lambda_o}(r^2 e^{\Lambda_o}\xi)' + \delta\Lambda]. \tag{26.7}$$

This is the initial value equation for Δn in terms of the dynamic variable ξ. The initial-value equation for δn, which will not be needed later, one obtains by combining with equation (26.4c).

b. Adiabaticity

(2) for pressure perturbations Δp and δp

For adiabatic vibrations (negligible heat transfer between neighboring fluid elements), the Lagrangian changes in number density and pressure are related by

$$\left(\frac{\partial \ln p}{\partial \ln n}\right)_s \equiv \Gamma_1 = \frac{n}{p}\frac{\Delta p}{\Delta n}. \tag{26.8}$$

[definition of adiabatic index, Γ_1]

Combining this adiabatic relation with equation (26.7) for Δn, and equation (26.4a) for δp in terms of Δp, one obtains the following *initial-value equation for* δp:

$$\delta p = -\Gamma_1 p_o[r^{-2}e^{-\Lambda_o}(r^2 e^{\Lambda_o}\xi)' + \delta\Lambda] - \xi p_o'. \tag{26.9}$$

c. Energy Conservation

(3) for density perturbations $\Delta\rho$ and $\delta\rho$

The local law of energy conservation [first law of thermodynamics; $\boldsymbol{u}\cdot(\boldsymbol{\nabla}\cdot\boldsymbol{T}) = 0$; see §§22.2 and 22.3] says that

$$\frac{d\rho}{d\tau} = \frac{(\rho + p)}{n}\frac{dn}{d\tau}.$$

Rewritten in terms of Lagrangian perturbations (recall: $d/d\tau$ is a time derivative as measured by an observer moving with the fluid), this reads

$$\frac{d\,\Delta\rho}{d\tau} = \frac{\rho + p}{n}\frac{d\,\Delta n}{d\tau},$$

which has as its time integral (first-order analysis!)

$$\Delta\rho = \frac{\rho_o + p_o}{n_o}\Delta n. \tag{26.10}$$

(The constant of integration is zero, because, when $\Delta n = 0$, $\Delta\rho$ must also vanish.) Combining this with equation (26.7) for Δn and equation (26.4b) for $\delta\rho$ in terms of $\Delta\rho$, one obtains the following *initial-value equation for* $\delta\rho$:

$$\delta\rho = -(\rho_o + p_o)[r^{-2}e^{-\Lambda_o}(r^2 e^{\Lambda_o}\xi)' + \delta\Lambda] - \xi\rho_o'. \tag{26.11}$$

d. Einstein Field Equations

Two of the Einstein field equations, when linearized, reduce to initial-value equations for the metric perturbations $\delta\Lambda$ and $\delta\Phi$. The equations needed, expressed in an orthonormal frame

(4) for metric perturbations $\delta\Lambda$ and $\delta\Phi$

$$\boldsymbol{\omega}^{\hat{t}} = e^{\Phi}\,\boldsymbol{d}t, \qquad \boldsymbol{\omega}^{\hat{r}} = e^{\Lambda}\,\boldsymbol{d}r, \qquad \boldsymbol{\omega}^{\hat{\theta}} = r\,\boldsymbol{d}\theta, \qquad \boldsymbol{\omega}^{\hat{\phi}} = r\sin\theta\,\boldsymbol{d}\phi, \tag{26.12}$$

are $G_{\hat{r}\hat{t}} = 8\pi T_{\hat{r}\hat{t}}$, and $G_{\hat{r}\hat{r}} = 8\pi T_{\hat{r}\hat{r}}$. The components of the Einstein tensor in this orthonormal frame were evaluated in exercise 14.16:

$$G_{\hat{r}\hat{t}} = 2(\dot{\Lambda}/r)e^{-(\Lambda+\Phi)} \overset{\text{[linearized]}}{=} 2r^{-1}e^{-(\Lambda_o+\Phi_o)}\,\delta\dot{\Lambda}; \tag{26.13a}$$

$$\begin{aligned} G_{\hat{r}\hat{r}} &= 2(\Phi'/r)e^{-2\Lambda} + r^{-2}(e^{-2\Lambda} - 1) \\ &\underset{\text{[linearized]}}{=} (G_{\hat{r}\hat{r}})_o + 2r^{-1}e^{-2\Lambda_o}\,\delta\Phi' - 2e^{-2\Lambda_o}(2r^{-1}\Phi_o{}' + r^{-2})\,\delta\Lambda. \end{aligned} \tag{26.13b}$$

The components of the stress-energy tensor, $T_{\hat{\alpha}\hat{\beta}} = (\rho + p)u_{\hat{\alpha}}u_{\hat{\beta}} + p\eta_{\alpha\beta}$, as calculated using the 4-velocity (26.6) [transformed into the form $u_{\hat{0}} = -1$, $u_{\hat{r}} = \dot{\xi}e^{\Lambda_o - \Phi_o}$] and using expressions (26.3a) for ρ and p, reduce to

$$T_{\hat{r}\hat{t}} = -(\rho_o + p_o)e^{\Lambda_o - \Phi_o}\dot{\xi}, \qquad T_{\hat{r}\hat{r}} = p_o + \delta p. \tag{26.14}$$

Consequently, the field equation $G_{\hat{r}\hat{t}} = 8\pi T_{\hat{r}\hat{t}}$—after integration with respect to time and choice of the constant of integration, so that $\delta\Lambda = 0$ when $\xi = 0$—reduces to

$$\delta\Lambda = -4\pi(\rho_o + p_o)re^{2\Lambda_o}\xi = -(\Lambda_o' + \Phi_o')\xi. \tag{26.15}$$

This is the initial-value equation for $\delta\Lambda$. The field equation $G_{\hat{r}\hat{r}} = 8\pi T_{\hat{r}\hat{r}}$, after using (26.15) to remove $\delta\Lambda$ and (26.9) to remove δp, and (26.1c) to remove $\Phi_o{}'$, reduces to

$$\begin{aligned} \delta\Phi' = &-4\pi\Gamma_1 p_o r^{-1}e^{2\Lambda_o+\Phi_o}(r^2e^{-\Phi_o}\xi)' \\ &+ [4\pi p_o{}'r - 4\pi(\rho_o + p_o)]e^{2\Lambda_o}\xi. \end{aligned} \tag{26.16}$$

This is the initial-value equation for $\delta\Phi$.

§26.5. DYNAMIC EQUATION AND BOUNDARY CONDITIONS

The dynamic evolution of the fluid displacement $\xi(t, r)$ is governed by the Euler equation (22.13):

Derivation of equation of motion for fluid displacement ξ

$$(\rho + p) \times (\text{4-acceleration}) = -(\text{projection of } \boldsymbol{\nabla}p \text{ orthogonal to } \boldsymbol{u}). \tag{26.17}$$

The 4-acceleration $\boldsymbol{a} = \boldsymbol{\nabla}_{\boldsymbol{u}}\boldsymbol{u}$ corresponding to the 4-velocity (26.6) in the metric (26.2) has as its only non-zero, linearized, covariant component:

$$a_r = \Phi_o{}' + \delta\Phi' + e^{2(\Lambda_o - \Phi_o)}\ddot{\xi}.$$

[The component a_t is trivial in the sense that it leads to an Euler equation that duplicates (26.1c).] Combining this with $\rho + p = \rho_o + p_o + \delta\rho + \delta p$, with the radial component $p_o' + \delta p'$ for the projection of ∇p, and with the zero-order equation of hydrostatic equilibrium (26.1c), one obtains for the Euler equation

$$(\rho_o + p_o)e^{2(\Lambda_o - \Phi_o)}\ddot{\xi} = -\delta p' - (\delta\rho + \delta p)\Phi_o' - (\rho_o + p_o)\,\delta\Phi'. \qquad (26.18)$$

This equation of motion is put into its most useful form by using the initial-value equations (26.9), (26.11), and (26.16) to reexpress δp, $\delta\rho$, and $\delta\Phi'$ in terms of ξ, and by then manipulating terms extensively with the aid of the zero-order equations of structure (26.1). The result is

$$W\ddot{\zeta} = (P\zeta')' + Q\zeta, \qquad (26.19)$$

where ζ is a "renormalized displacement function," and W, P, Q are functions of radius determined by the structure of the equilibrium star:

$$\zeta \equiv r^2 e^{-\Phi_o}\xi; \qquad (26.20)$$

$$W \equiv (\rho_o + p_o)r^{-2}e^{3\Lambda_o + \Phi_o}; \qquad (26.21a)$$

$$P \equiv \Gamma_1 p_o r^{-2} e^{\Lambda_o + 3\Phi_o}; \qquad (26.21b)$$

$$Q \equiv e^{\Lambda_o + 3\Phi_o}\left[\frac{(p_o')^2}{\rho_o + p_o}r^{-2} - 4p_o' r^{-3} - 8\pi(\rho_o + p_o)p_o r^{-2}e^{2\Lambda_o}\right]. \qquad (26.21c)$$

Equation (26.19) is the dynamic equation governing the stellar pulsations. [This equation could be written in other forms; for instance, it could be multiplied by W^{-1} or any other non-zero factor, and terms could be regrouped. The form given in equation (26.19) is preferred because it leads to a self-adjoint eigenvalue problem for the oscillation frequencies, as indicated in Box 26.1.]

Boundary conditions on fluid displacement

Not all solutions of the dynamic equation are acceptable. To be physically acceptable, the displacement function must produce noninfinite density and pressure perturbations ($\delta\rho$ and δp) at the center of the sphere, which means

$$(\xi/r) \text{ finite or zero in limit as } r \longrightarrow 0 \qquad (26.22a)$$

[see (26.9) and (26.11)]; also, it must leave the pressure equal to zero at the star's surface, which means

$$\Delta p = -\Gamma_1 p_o r^{-2} e^{\Phi_o}(r^2 e^{-\Phi_o}\xi)' \longrightarrow 0 \text{ as } r \longrightarrow \underbrace{R}_{\text{[surface radius]}} \qquad (26.22b)$$

[see (26.8), (26.7), and (26.15)].

§26.6. SUMMARY OF RESULTS

Summary of theory of stellar pulsations

If an initial displacement of the fluid, $\xi(t = 0, r)$, is specified subject to the boundary conditions (26.22), then its subsequent evolution $\xi(t, r)$ can be calculated by integrating the dynamic equation (26.19); and the form of the pressure, density, and metric perturbations can be calculated from $\xi(t, r)$ using the initial-value equations (26.9), (26.11), (26.15), and (26.16).

Several important consequences of these results are explored in Boxes 26.1 and 26.2.

(*continued on page 699*)

Box 26.1 EIGENVALUE PROBLEM AND VARIATIONAL PRINCIPLE FOR NORMAL-MODE PULSATIONS OF A STAR

Assume that the renormalized displacement function (26.20) has a sinusoidal time dependence:

$$\zeta = \zeta(r)e^{-i\omega t}.$$

Then the dynamic equation (26.19) and boundary conditions (26.22) reduce to an eigenvalue problem for the angular frequency ω and amplitude $\zeta(r)$:

$$(P\zeta')' + Q\zeta + \omega^2 W\zeta = 0, \tag{1}$$

$$\zeta/r^3 \text{ finite or zero as } r \longrightarrow 0, \tag{2a}$$

$$\Gamma_1 p_0 r^{-2} e^{\Phi_0}\zeta' \longrightarrow 0 \text{ as } r \longrightarrow R. \tag{2b}$$

Methods for solving this eigenvalue problem are catalogued and discussed by Bardeen, Thorne, and Meltzer (1966). One method (but *not* the best for numerical calculations) is the variational principle:

$$\omega^2 = \text{extremal value of} \left\{ \frac{\int_0^R (P\zeta'^2 - Q\zeta^2)\, dr}{\int_0^R W\zeta^2\, dr} \right\}, \tag{3}$$

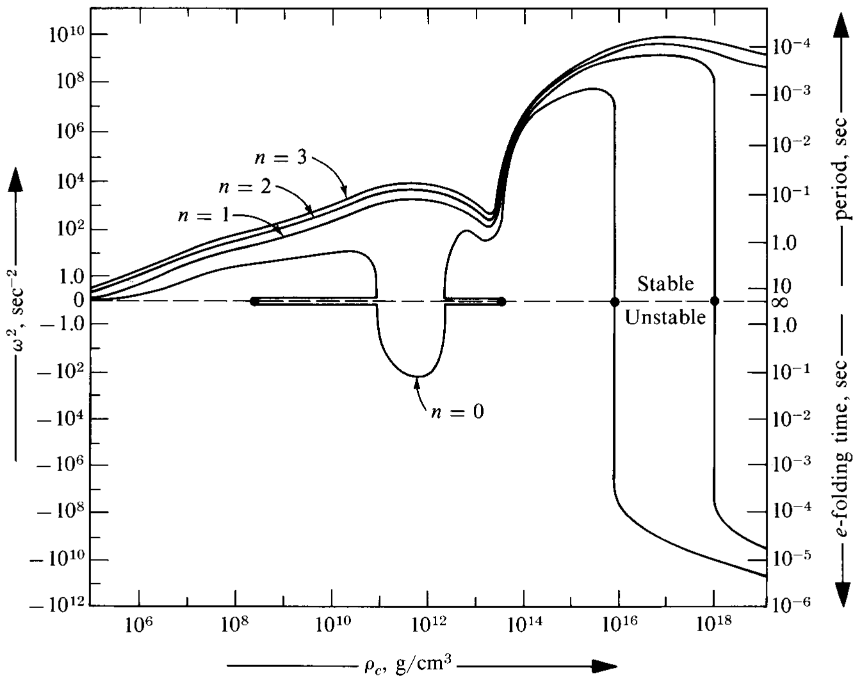
where ζ is varied over all functions satisfying the boundary conditions (2). [See e.g., §12.3 of Mathews and Walker (1965) for discussion of the equivalence between this variational principle and the original eigenvalue problem.]

The absolute minimum value of expression (3) is the squared frequency of the fundamental mode of pulsation. If it is negative, the star is unstable ($e^{-i\omega t}$ grows exponentially in time). If it is positive, the star is stable against adiabatic, radial perturbations. Therefore, since the denominator of expression (3) is positive definite,

$$\begin{bmatrix}\text{stability against}\\ \text{adiabatic radial}\\ \text{perturbations}\end{bmatrix} \Longleftrightarrow \begin{bmatrix}\int_0^R (P\zeta'^2 - Q\zeta^2)\, dr > 0 \text{ for all functions}\\ \zeta \text{ satisfying (2)}\end{bmatrix}. \tag{4}$$

Box 26.1 ***(continued)***

By numerical solution of the eigenvalue equation (1), the pulsation frequencies have been calculated for a wide variety of models of neutron stars and supermassive stars. *Example:* The figure gives a plot of pulsation frequency as a function of central density for the lowest four normal modes of the Harrison-Wakano-Wheeler models at the endpoint of stellar evolution. (Make a detailed comparison with Figure 24.2.) These curves are based on calculations by Meltzer and Thorne (1966), with corrections for the fundamental mode of massive white dwarfs by Faulkner and Gribbin (1968).



Box 26.2 THE CRITICAL ADIABATIC INDEX FOR NEARLY NEWTONIAN STARS

A. Fully Newtonian Stars

1. For a Newtonian star that pulsates sinusoidally, $\xi = \xi(r)e^{-i\omega t}$, the dynamic equation (26.19) reduces to

$$[\Gamma_1 p_o r(\xi/r)']' + 3(\Gamma_1 p_o \xi/r)' - 4p_o'\xi/r + \omega^2 \rho_o \xi = 0. \tag{1}$$

2. If $\Gamma_1 = 4/3$ throughout the star, the physically acceptable solution [solution satisfying boundary conditions (26.22)] for the fundamental mode of vibration (mode with lowest value of ω^2) is

$$\omega^2 = 0, \qquad \xi = \epsilon r, \qquad \epsilon = \text{const.} \tag{2}$$

Thus, for $\Gamma_1 = 4/3$ the fundamental mode is "neutrally stable" and has a "homologous" displacement function—independent of the star's equation of state or structure.

3. If Γ_1 is allowed to differ slightly from 4/3 in an r-dependent way, then $\xi(r)$ will differ slightly from the homologous form:

$$\xi = \epsilon r[1 + r\text{-dependent corrections of magnitude } (\Gamma_1 - 4/3)].$$

Consequently, if one uses the homologous expression $\xi = \epsilon r$ as a trial function in the variational principle of Box 26.1, one will obtain ω^2 accurate to $O[(\Gamma_1 - 4/3)^2]$. (Recall: first-order errors in trial function produce second-order errors in value of variational expression.) The Newtonian limit of the variational expression [equation (3) of Box 26.1] becomes, with the homologous choice of trial function,

$$\omega^2 = (3\bar{\Gamma}_1 - 4)\frac{\int_0^R 3p_o r^2\, dr}{\int_0^R \rho_o r^4\, dr} + O[(3\bar{\Gamma}_1 - 4)^2], \tag{3}$$

where $\bar{\Gamma}_1$ is the pressure-averaged adiabatic index

$$\bar{\Gamma}_1 = \frac{\int_0^R \Gamma_1 p_o 4\pi r^2\, dr}{\int_0^R p_o 4\pi r^2\, dr}. \tag{4}$$

Box 26.2 (*continued*)

By use of the Newtonian virial theorem for the nonpulsating star [equation (39.21b) or exercise 23.7], one can convert equation (3) into the form

$$\omega^2 = (3\bar{\Gamma}_1 - 4)|\Omega|/I, \tag{5}$$

where Ω is the star's self-gravitational energy and $I = \int(\rho_o r^2)4\pi r^2\, dr$ is the trace of the second moment of its mass distribution (see Box 24.2 and exercise 39.6).

B. Nearly Newtonian Stars

1. When one takes into account first-order relativistic corrections (corrections of order M/R), but ignores higher-order corrections, one can rewrite the variational expression [equation (3) of Box 26.1] in the form

$$\omega^2 = \frac{\int_0^R p_o[\Gamma_1 r^4\eta'^2 + (3\Gamma_1 - 4)(r^3\eta^2)'](1 + \Lambda_o + 3\Phi_o)\, dr - \int_0^R F_o\eta^2\, dr}{\int_0^R \rho_o r^4(1 + 3\Lambda_o + \Phi_o + p_o/\rho_o)\eta^2\, dr}, \tag{6}$$

where

$$F_o \equiv 8\pi r^4 p_o\rho_o + 8rm_o p_o + \rho_o m_o{}^2, \qquad \eta = \zeta/r^3 = (\xi/r)(1 - \Phi_o), \tag{7}$$

and $m_o(r)$ is the equilibrium mass inside radius r.

2. For a relativistic star with $\Gamma_1 - 4/3$ of order M/R and with $M/R \ll 1$, the homologous trial function $\xi = \epsilon r$ will still be highly accurate. Equally accurate, and easier to work with, will be $\xi = \epsilon r e^{\Phi_o} \approx \epsilon r(1 + \Phi_o)$, which corresponds to $\eta = \epsilon =$ constant. Its fractional errors will be of order M/r; and the errors which it produces in ω^2 will be of order $(M/R)^2$. By inserting this trial function into the variational principle (6) and keeping only relativistic corrections of order M/R, one obtains

$$\omega^2 = 3(\bar{\Gamma}_1 - \Gamma_{1\,\mathrm{crit}})|\Omega|/I. \tag{8}$$

Here $\bar{\Gamma}_1$ is the pressure-averaged adiabatic index, and the critical value of the adiabatic index $\Gamma_{1\,\mathrm{crit}}$ is

$$\Gamma_{1\,\mathrm{crit}} = \frac{4}{3} + \alpha M/R, \tag{9}$$

with α a positive constant of order unity given by

$$\alpha = \frac{1}{3}\frac{R}{M|\Omega|}\int_0^R \left(3\rho_o \frac{m_o^{\,2}}{r^2} + 4p_o \frac{m_o}{r}\right) 4\pi r^2\, dr. \tag{10}$$

Expressions (8) and (9) for the pulsation frequency and the adiabatic index play an important role in the theory of supermassive stars (§24.4).

3. For alternative derivations of the above result, see Chandrasekhar (1964a,b; 1965c), Fowler (1964, 1965), Wright (1964).

EXERCISE

Exercise 26.1. DRAGGING OF INERTIAL FRAMES BY A SLOWLY ROTATING STAR

A fluid sphere rotates very slowly. Analyze its rotation using perturbation theory; keep only effects and terms *linear* in the angular velocity of rotation. [Hints: (1) Centrifugal forces are second-order in angular velocity. Therefore, to first order the star is undeformed; its density and pressure distributions remain spherical and unperturbed. (2) Show, by symmetry and time-reversal arguments, that one can introduce coordinates in which

$$\begin{aligned} ds^2 = {} & -e^{2\Phi}\, dt^2 + e^{2\Lambda}\, dr^2 + r^2[d\theta^2 + \sin^2\theta\, d\phi^2] \\ & - 2(r^2 \sin^2\theta)\omega\, d\phi\, dt, \end{aligned} \tag{26.23}$$

where

$$\Phi = \Phi(r),\ \Lambda = \Lambda(r), \text{ and } \omega = \omega(r, \theta). \tag{26.24}$$

Show that $\Phi = \Phi_0$ and $\Lambda = \Lambda_0$ (no perturbations!) to first-order in angular velocity. (3) Adopt the following precise definition of the angular velocity $\Omega(r, \theta)$:

$$\Omega \equiv u^\phi/u^t = (d\phi/dt)_{\text{moving with the fluid}}. \tag{26.25}$$

Assuming $u^r = u^\theta = 0$ (i.e., rotation in the ϕ direction), calculate the 4-velocity of the fluid. (4) Use the Einstein field equations to derive a differential equation for the metric perturbation ω in terms of the angular velocity Ω. (5) Solve that differential equation outside the star in terms of elementary functions, and express the solution for $\omega(r, \theta)$ in terms of the star's total angular momentum $\boldsymbol{S}$, as measured using distant gyroscopes (see Chapter 19).] For the original analyses of this problem and of related topics, see Gurovich (1965), Doroshkevich, Zel'dovich, and Novikov (1965), Hartle and Sharp (1965), Brill and Cohen (1966), Hartle (1967), Krefetz (1967), Cohen and Brill (1968), Cohen (1968).

PART VI

THE UNIVERSE

Wherein the reader, flushed with joy at his conquest of the stars, seeks to control the entire universe, and is foiled by an unfathomed mystery: the Initial Singularity.

CHAPTER 27

IDEALIZED COSMOLOGIES

From my point of view one cannot arrive, by way of theory, at any at least somewhat reliable results in the field of cosmology, if one makes no use of the principle of general relativity.

ALBERT EINSTEIN (1949b, p. 684)

§27.1. THE HOMOGENEITY AND ISOTROPY OF THE UNIVERSE

The universe: fine-scale condensations contrasted with large-scale homogeneity

Astronomical observations reveal that the universe is homogeneous and isotropic on scales of $\sim 10^8$ light years and larger. Taking a "fine-scale" point of view, one sees the agglomeration of matter into stars, galaxies, and clusters of galaxies in regions of size ~ 1 light year, $\sim 10^6$ light years, and $\sim 3 \times 10^7$ light years, respectively. But taking instead a "large-scale" viewpoint, one sees little difference between an elementary volume of the universe of the order of 10^8 light years on a side centered on the Earth and other elementary volumes of the same size located elsewhere.

Cosmology, summarized in its simplest form in Box 27.1, takes the large-scale viewpoint as its first approximation; and as its second approximation, it treats the fine-scale structure as a perturbation on the smooth, large-scale background. This chapter (27) treats in detail the large-scale, homogeneous approximation. Chapter 28 considers such small-scale phenomena as the primordial formation of the elements, and the condensation of galaxies out of the primeval plasma during the expansion of the universe. Chapter 29 discusses observational cosmology.

Evidence for large-scale homogeneity and isotropy

Evidence for the large-scale homogeneity and isotropy of the universe comes from several sources. (1) There is evidence in the distribution of galaxies on the sky and in the distribution of their apparent magnitudes and redshifts [see, e.g., Hubble (1934b, 1936); Sandage (1972a); Sandage, Tamman, and Hardy (1972); but note the papers claiming "hierarchic" deviations from homogeneity, which Sandage cites and attacks]. (2) There is evidence in the isotropy of the distribution of radio sources on the sky [see, e.g., Holden (1966), and Hughes and Longair (1967)]. (3) There is evidence in the remarkable isotropy of the cosmic microwave radiation [see, e.g., Boughn, Fram, and Partridge (1971)]. For a review of most of the evidence, see Chapter 2 of Peebles (1971).

(continued on page 711)

Box 27.1 COSMOLOGY IN BRIEF

Uniform density. Idealize the stars and atoms as scattered like dust through the heavens with an effective average density ρ of mass-energy everywhere the same.

Geometry homogeneous and isotropic. Idealize the curvature of space to be everywhere the same.

Closure. Accept the term, "Einstein's geometric theory of gravity" as including not only his field equation $\boldsymbol{G} = 8\pi\boldsymbol{T}$, but also his boundary condition of closure imposed on any solution of this equation.*

A three-sphere satisfies the three requirements of homogeneity, isotropy, and closure, and is the natural generalization of the metric on a circle and a 2-sphere:

Spheres of selected dimensionality	*Visualized as embedded in a Euclidean space of one higher dimension*[a]	*Transformation from Cartesian to polar coordinates*	*Metric on S^n expressed in terms of these polar coordinates*
S^1	$x^2 + y^2 = a^2$	$x = a\cos\phi$ $y = a\sin\phi$	$ds^2 = a^2\,d\phi^2$
S^2	$x^2 + y^2 + z^2 = a^2$	$x = a\sin\theta\cos\phi$ $y = a\sin\theta\sin\phi$ $z = a\cos\theta$	$ds^2 = a^2(d\theta^2 + \sin^2\theta\,d\phi^2)$
S^3	$x^2 + y^2 + z^2 + w^2 = a^2$	$x = a\sin\chi\sin\theta\cos\phi$ $y = a\sin\chi\sin\theta\sin\phi$ $z = a\sin\chi\cos\theta$ $w = a\cos\theta$	$ds^2 = a^2[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\,d\phi^2)]$

[a] Excursion off the sphere is physically meaningless and is forbidden. The superfluous dimension is added to help the reason in reasoning, not to help the traveler in traveling. Least of all does it have anything whatsoever to do with time.

The spacetime geometry is described by the metric

$$ds^2 = -dt^2 + a^2(t)[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\,d\phi^2)]. \tag{1}$$

The dynamics of the geometry is known in full when one knows the radius a as a function of the time t.

---

*"Thus we may present the following arguments against the conception of a space-infinite, and for the conception of a space-bounded, universe:

"1. From the standpoint of the theory of relativity, the condition for a closed surface is very much simpler than the corresponding boundary condition at infinity of the quasi-Euclidean structure of the universe.

"2. The idea that Mach expressed, that inertia depends upon the mutual action of bodies, is contained, to a first approximation, in the equations of the theory of relativity; . . . But this idea of Mach's corresponds only to a finite universe, bounded in space, and not to a quasi-Euclidean, infinite universe" [Einstein (1950), pp. 107–108].

Many workers in cosmology are skeptical of Einstein's boundary condition of closure of the universe, and will remain so until astronomical observations confirm it.

Einstein's field equation (doubled, for convenience), $2\boldsymbol{G} = 16\pi\boldsymbol{T}$, has its whole force concentrated in its $\hat{0}\hat{0}$ (or $\hat{t}\hat{t}$) component,

$$\frac{6}{a^2}\left(\frac{da}{dt}\right)^2 + \frac{6}{a^2} = 16\pi\rho \tag{2}$$

[equation (5a) of Box 14.5]. This component of Einstein's equation is as central as the component $\nabla \cdot \boldsymbol{E} = 4\pi\rho$ of Maxwell's equations. It is described in the Track-2 Chapter 21 as the "initial-value equation" of geometrodynamics. There the two terms on the left receive separate names: the "second invariant" of the "extrinsic curvature" of a "spacelike slice" through the 4-geometry (tells how rapidly all linear dimensions are being stretched from instant to instant); and the "intrinsic curvature" or three-dimensional scalar curvature invariant ${}^{(3)}R$ of the "spacelike slice" (here a 3-sphere) at the given instant itself.

The amount of mass-energy in the universe changes from instant to instant in accordance with the work done by pressure during the expansion,

$$d\left[\begin{pmatrix}\text{density of}\\ \text{mass-energy}\end{pmatrix} \times (\text{volume})\right] = -(\text{pressure})\, d(\text{volume}). \tag{3}$$

Today the pressure of radiation is negligible compared to the density of mass-energy, and the righthand side of this equation ("work done") can be neglected. The same was true in the past, one estimates, back to a time when linear dimensions were about a thousand times smaller than they are today. During this "matter-dominated phase" of the expansion of the universe, the product

$$\begin{pmatrix}\text{density of}\\ \text{mass-energy}\end{pmatrix} \times (\text{volume})$$

remained a constant,

$$\rho \cdot 2\pi^2 a^3 = M. \tag{4}$$

Here the symbol M can look like mass in the form of matter, and can even be called mass; but one has to recall again (see §19.4) that the concept of total mass-energy of a closed universe has absolutely no well-defined meaning whatsoever, not least because there is no "platform" outside the universe on which to stand to measure its attraction via periods of Keplerian orbits or in any other way. More convenient than M, because more significant in what follows, is the quantity a_{max} ("radius of universe at phase of maximum expansion") defined by

$$a_{\text{max}} = 4M/3\pi. \tag{5}$$

Box 27.1 (*continued*)

The decisive component of the Einstein field equation, in the terms of this notation, becomes

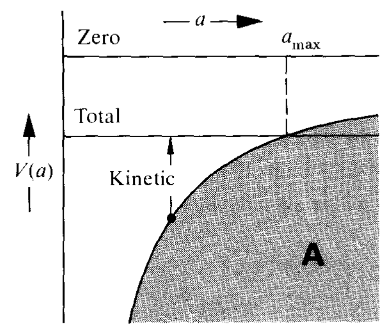
$$\frac{6}{a^2}\left(\frac{da}{dt}\right)^2 + \frac{6}{a^2} = \frac{6a_{max}}{a^3}$$

or

$$\left(\frac{da}{dt}\right)^2 - \frac{a_{max}}{a} = -1. \tag{6}$$

The first term in (6) has the qualitative character of "kinetic energy" in an elementary problem in Newtonian mechanics. The second term has the qualitative character of a "potential energy,"

$$V(a) = -\frac{a_{max}}{a}$$



(see diagram **A**), resulting from an inverse-square Newtonian force. Pursuing the analogy, one identifies the "-1" on the righthand side with the total energy in the Newtonian problem. The qualitative character of the dynamics shows up upon an inspection of diagram **A**. Values of the radius of the universe, a, greater than a_{max} are not possible. If a were to become greater than a_{max}, the "potential energy" would exceed the total "energy" and the "kinetic energy" of expansion would have to become negative, which is impossible. Consequently the geometrodynamic system can never be in a state more expanded than $a = a_{max}$. Starting in a state of small a, ($a \ll a_{max}$) and expanding, the universe has for each a value a perfectly definite da/dt value. This velocity of expansion decreases as the expansion proceeds. It falls to zero at the turning point $a = a_{max}$. Thereafter the system recontracts.

Lack of option is the striking feature of the dynamics. Granted a specific amount of matter [specific M value in (5)], one has at his disposal no free parameter whatsoever. The value of $a_{\rm max}$ is uniquely specified by the amount of matter present, and by nothing more. There is no such thing as an "adjustable constant of energy," such as there would have been in a traditional problem of Newtonian dynamics. Where such an adjustable parameter might have appeared in equation (6), there appears instead the fixed number "-1." This fixity is the decisive feature of a system bound up into closure. Were one dealing with a collection of rocks out in space, one would have a choice about the amount of dynamite one placed at their center. With a low charge of explosive, one would find the rocks flying out for only a limited distance before gravity halted their flight and brought them to collapse together again. With more propellant, they would fly out with escape velocity and never return. But no such options present themselves here, exactly because Einstein's condition of closure has been imposed; and once closed, always closed. Collapse of the universe is universal. This is simple cosmology in brief.

Einstein's unhappiness at this result was great. At the time he developed general relativity, the permanence of the universe was a fixed item of belief in Western philosophy: "The heavens endure from everlasting to everlasting." Yet the reasoning that led to the fixed equation left open no natural way to change that equation or its fantastic prediction. Therefore Einstein (1917), much against his will, introduced the least unnatural change he could imagine, a so-called cosmological term (§27.11), the whole purpose of which was to avoid the expansion of the universe. A decade later, Hubble (1929) verified the predicted expansion. Thereupon Einstein abandoned the cosmological term, calling it "the biggest blunder of my life" [Einstein (1970)]. Thus ended the first great cycle of apparent contradiction to general relativity, test, and dramatic vindication. Will one ever penetrate the mystery of creation? There is no more inspiring evidence that the answer will someday be "yes" than man's power to predict, and predict correctly, and predict against all expectations, so fantastic a phenomenon as the expansion of the universe.

"*Newtonian cosmology*" provides an "equation of energy" similar to that of Einstein cosmology, but fails to provide any clean or decisive argument for closure or for the unique constant "-1." It considers the mass in any elementary spherical region of space of momentary radius r, and the gravitational acceleration of a test particle at the boundary of this sphere toward the center of the sphere; thus,

$$\frac{d^2r}{dt^2} = -\frac{(\text{mass})}{(\text{distance})^2} = -\frac{(4\pi/3)\rho r^3}{r^2} = -\frac{4\pi\rho}{3}r. \tag{7}$$

Consider such imaginary spheres of varied radii drawn in the cosmological medium with the same center. Note that doubling the radius doubles the acceleration. This proportionality between acceleration and distance is compatible with a homogeneous

Box 27.1 (*continued*)

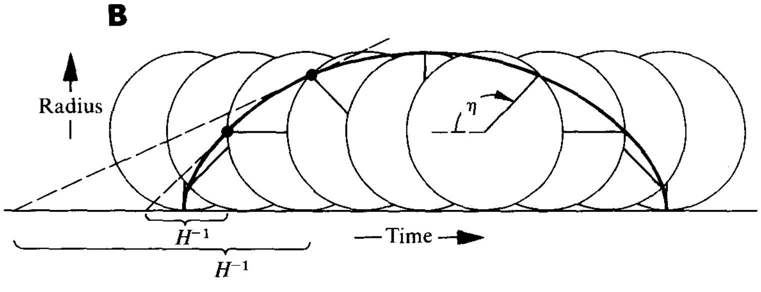
deceleration of the expansion of the universe. Therefore define an expansion parameter a^* as the ratio between the radius of any one of these spheres now and the radius of the same sphere at some fiducial instant; thus, $a^* = r/r_0$ is to be considered as independent of the particular sphere under consideration. Write $\rho = \rho_0 r_0^3/r^3$, where ρ_0 is the density at the fiducial instant. Insert this expression for ρ into the deceleration equation (7), multiply both sides of the equation through by dr/dt, integrate, and translate the result from an equation for dr/dt to an equation for da^*/dt, finding

$$\left(\frac{da^*}{dt}\right)^2 - \frac{(8\pi\rho_0/3)}{a^*} = \text{constant}, \tag{8}$$

in agreement with equation (6), except for (1) the trivial differences that arise because a^* is a dimensionless expansion ratio, whereas a is an absolute radius with the dimensions of cm, and (2) the all-important difference that here the constant is disposable, whereas in standard Einstein geometrodynamics it has the unique canonical value "-1." For more on Newtonian insights into cosmology, see especially Bondi (1961).

Free fall of a particle towards a Newtonian center of attraction according to Newtonian mechanics gives an equation of energy of the same form as (6), except that the "radius of the universe," a, is replaced by distance, r, from the center of

B



attraction. The solution of this problem of free fall is described by a cycloid (diagram **B**; see also Figure 25.3 and Box 25.4), generated by rolling a circle of diameter a_{max} on a line through an ever increasing angle η; thus,

$$\begin{aligned} a &= \frac{1}{2} a_{\text{max}}(1 - \cos\eta), \\ t &= \frac{1}{2} a_{\text{max}}(\eta - \sin\eta). \end{aligned} \tag{9}$$

Immediately observable today is the present rate of expansion of the universe, with every distance increasing at a rate directly proportional to the magnitude of that distance:*

$$\frac{\begin{pmatrix}\text{velocity of recession}\\ \text{of a galaxy}\end{pmatrix}}{(\text{distance to a galaxy})} = (\text{Hubble "constant,"}\ H_0) \sim 55\ \text{km/sec megaparsec}$$
$$= \frac{1}{18 \times 10^9\ \text{yr}}\ \text{or}\ \frac{1}{1.7 \times 10^{28}\ \text{cm}}$$
$$= \frac{\begin{pmatrix}\text{rate of increase of the}\\ \text{radius of the universe itself}\end{pmatrix}}{(\text{radius of the universe})} = \frac{da/dt}{a}. \tag{10}$$

The Hubble time, $H_0^{-1} \sim 18 \times 10^9$ yr (linearly extrapolated back to zero separation on the basis of the expansion rate observed today, as illustrated in the diagram) is predicted to be greater by a factor 1.5 or more (Box 27.3) than the actual time back to the start of the expansion as deduced from the rate of the development of stars ($\sim 10 \times 10^9$ yr). No such satisfactory concord between prediction and observation on this inequality existed in the 1940's. The scale of distances between galaxy and galaxy in use at that time was short by a factor more than five. The error arose from misidentifications of Cepheid variable stars and of HII regions, which are used as standards of intensity to judge the distance of remote galaxies. The linearly extrapolated time,

$$(\text{Hubble time}) = \frac{(\text{distance today})}{(\text{recession velocity today})},$$

back to the start of the expansion was correspondingly short by a factor more than five. The Hubble time came out to be only of the order of 3×10^9 yr. This number obviously violates the inequality

$$\begin{pmatrix}\sim 3 \times 10^9\ \text{yr Hubble}\\ \text{time}\end{pmatrix} \geq 1.5 \begin{pmatrix}\sim 10 \times 10^9\ \text{yr; actual time}\\ \text{back to start of expansion}\end{pmatrix}.$$

It implies a curve for dimensions as a function of time not bending down, as in diagram **B**, but bending up. On some sides the proposal was made to regard the actual curve as rising exponentially. Thus began an era of "theories of continuous creation of matter," all outside the context of Einstein's standard geometrodynamics.

---

*H_0 is predicted to be independent of the choice of galaxy insofar as local motions are unimportant, and insofar as the difference between recession velocity now and recession velocity at the time when the light was emitted is unimportant. The latter condition is well fulfilled by galaxies close enough to admit of the necessary measurement of distance, for they have redshifts only of the order of $z \sim 0.1$ and less (little lapse of time between emission of light and its reception on earth; therefore little change in recession velocity between then and now; see §29.3 and Box 29.4 for a fuller analysis).

Box 27.1 ***(continued)***

This era ended when, for the first time, the distinction between stellar populations of classes I and II was recognized and as a result Cepheid variables were correctly identified, by Baade (1952, 1956) and when Sandage (1958) discovered that Hubble had misidentified as bright stars the HII regions in distant galaxies. Then the scale of galactic distances was set straight. Thus ended the second great cycle of an apparent contradiction to general relativity, then test, and then dramatic vindication.

The mystery of the missing matter marks a third cycle of doubt and test with the final decision yet to come. It follows from equation (2) that, if Einstein's closure boundary condition is correct, then the density of mass-energy must exceed a certain lower limit given by the equation

$$\rho \geq \rho_H = \frac{3}{8\pi} H_o^2 \tag{11}$$

("critical amount of mass-energy required to curve up the geometry of the universe into closure"). A Hubble expansion rate of $H_o = 55$ km/sec Megaparsec implies a lower limit to the density of

$$\rho_H = \frac{3}{8\pi} \frac{1}{(1.7 \times 10^{28}\ \mathrm{cm})^2} \text{ or } \rho_{H,\mathrm{conv}} = 5 \times 10^{-30}\ \mathrm{g/cm^3} \tag{12}$$

as compared to $\rho \sim 2 \times 10^{-31}$ g/cm^3 of "luminous matter" observed in galaxies (§29.6) and more being searched for today in the space between the galaxies.

A fuller treatment of cosmology deals with conditions back in the past corresponding to redshifts of 10,000 or more and dimensions 10,000 times less than they are today, when radiation could not be neglected, and even dominated (§27.10). It also considers even earlier conditions, when anisotropy oscillations of the geometry of the universe as a whole (analogous to the transformation from a cigar to a pancake and back again) may conceivably have dominated (Chapter 30). More broadly, it takes up the evolution of the universe into its present state (Chapter 28) and the present state and future evolution of the universe (Chapter 29). The present chapter examines the basic assumptions that underlie the simple standard cosmology thus traced out, and §27.11 examines what kinds of qualitative changes would result if one or another of these assumptions were to be relaxed.

1
2

2

§27.2. STRESS-ENERGY CONTENT OF THE UNIVERSE—THE FLUID IDEALIZATION

The rest of this chapter, *except* for Box 27.4, is Track 2.
No earlier track-2 material is needed as preparation for it, but it is needed as preparation for Chapter 29 (Present state and future evolution of the universe).

By taking the large-scale viewpoint, one can treat galaxies as "particles" of a "gas" that fills the universe. These particles have internal structure (stars, globular clusters, etc.); but one ignores it. The "particles" cluster on a small scale (clusters of galaxies of size $\lesssim 3 \times 10^7$ light years); but one ignores the clustering. To simplify calculations, one even ignores the particulate nature of the "gas" [though one can take it into account, if one wishes, by adopting a kinetic-theory description; see §22.6 for kinetic theory, and Ehlers, Geren, and Sachs (1968) for its application to cosmology]. One removes the particulate structure of the gas from view by treating it in the perfect-fluid approximation. Thus, one characterizes the gas by a *4-velocity*, $\boldsymbol{u}$ (the 4-velocity of an observer who sees the galaxies in his neighborhood to have *no* mean motion), by a *density of mass-energy*, ρ (the smoothed-out density of mass-energy seen in the frame with 4-velocity $\boldsymbol{u}$; this includes the rest mass plus kinetic energy of the galaxies in a unit volume, divided by the volume), and by a *pressure* p (the kinetic pressure of the galaxies). The stress-energy tensor for this "fluid of galaxies" is the familiar one

Idealization of matter in universe as a perfect fluid ("fluid of galaxies")

$$\boldsymbol{T} = (\rho + p)\boldsymbol{u} \otimes \boldsymbol{u} + \boldsymbol{g}p, \tag{27.1}$$

where $\boldsymbol{g}$ is the metric tensor.

Astronomical observations reveal that the rest-mass density of the galaxies is much greater than their density of kinetic energy. The typical ordinary velocities of the galaxies—and of stars in them—relative to each other are

Large-scale conditions in universe today:

$$\langle v \rangle \sim 200 \text{ km/sec} \sim 10^{-3}. \tag{27.2}$$

Consequently, the ratios of kinetic-energy density and of pressure to rest-mass density are

(1) kinetic energy and pressure of stars and galaxies

$$\begin{aligned} \varepsilon_{\text{kin}}/\rho_{\text{rm}} &= \frac{1}{2}\langle v^2 \rangle \approx 10^{-6}, \\ p/\rho_{\text{rm}} &= \frac{1}{3}\langle v^2 \rangle \approx 10^{-6}. \end{aligned} \tag{27.3}$$

At least, these are the ratios today. Very early in the life of the universe, conditions must have been quite different.

The total density of mass-energy, ρ, is thus very nearly the rest-mass density of the galaxies, ρ_{rm}. Astronomical observations yield for ρ_{rm} today

(2) density of mass in galaxies

$$\rho_{\text{rm}} \gtrsim 2 \times 10^{-31} \text{ g/cm}^3 \tag{27.4}$$

(see §29.6).

(3) cosmic-ray density

Not all the matter in the universe is tied up in galaxies; there is also matter in cosmic rays, with an averaged-out density of mass-energy

$$\rho_{\mathrm{cr}} \lesssim 10^{-33}\ \mathrm{g/cm^3}, \tag{27.5}$$

(4) density of intergalactic gas

and, perhaps, gas in intergalactic space with

$$\rho_{\mathrm{ig}} \lesssim 10^{-28}\ \mathrm{g/cm^3}. \tag{27.6}$$

[Delineating more sharply the value of ρ_{ig} is one of the most important goals of current cosmological research. For a review of this question as of 1971, see "The mean mass density of the universe," pp. 56–120 in Peebles (1971).] These sources of mass density, and the associated pressures, one can lump together with the galaxies into the "cosmological fluid," with stress-energy tensor (27.1).

(5) magnetic fields

Not all the stress-energy in the universe is in the form of matter. There are also magnetic fields, with mean energy density that almost certainly does not exceed the limit

$$\rho_{\mathrm{mag}} \lesssim 10^{-35}\ \mathrm{g/cm^3} \tag{27.7}$$

(6) radiation density

(corresponding to $B_{\mathrm{avg}} \lesssim 10^{-6}\ G$), and radiation (electromagnetic radiation, neutrino radiation, and perhaps gravitational radiation) totaling, one estimates,

$$\rho_{\mathrm{rad}} \approx 10^{-33}\ \mathrm{g/cm^3}. \tag{27.8}$$

The magnetic fields will be ignored in this chapter; they are unimportant for large-scale cosmology, except perhaps very near the "big-bang beginning" of the universe—if they existed then. However, the radiation cannot be ignored, for it plays a crucial role.

The cosmic microwave radiation

Most of the radiation density is in the form of "cosmic microwave radiation," which was discovered by Penzias and Wilson (1965) [see also Dicke, Peebles, Roll, and Wilkinson (1965)], and has been studied extensively since then [for a review, see Partridge (1969)]. The evidence is very strong that this cosmic microwave radiation is a remnant of the big-bang beginning of the universe. This interpretation will be accepted here.

The cosmic microwave radiation has just the form one would expect if the earth were enclosed in a box ("black-body cavity") with temperature 2.7K. The spectrum is a Planck spectrum with this temperature, and the radiation is isotropic [Boughn, Fram, and Partridge (1971)]. Consequently, its pressure and density of mass-energy are given by the formula,

$$\begin{aligned}\rho_{\mathrm{microwave}} = 3p_{\mathrm{microwave}} &= aT^4 \\ &= 4 \times 10^{-34}\ \mathrm{g/cm^3}.\end{aligned} \tag{27.9}$$

Thermodynamic considerations (§27.10) suggest that the universe should also be filled with neutrino radiation and perhaps gravitational radiation that have Planck spectra at approximately the same temperature ($\sim$3K). However, they are not detectable with today's technology.

To high accuracy ($\lesssim 300$ km/sec) the mean rest frame of the cosmic microwave radiation near Earth is the same as the mean rest frame of the galaxies in the neighborhood of Earth [Boughn, Fram and Partridge (1971)]. Consequently, the radiation can be included, along with the matter, in the idealized cosmological fluid.

Summary: From the large-scale viewpoint, the stress-energy of the universe can be idealized as a perfect fluid with 4-velocity $\boldsymbol{u}$, density of mass-energy ρ, pressure p, and stress-energy tensor

Summary of fluid idealization of matter in universe

$$\boldsymbol{T} = (\rho + p)\boldsymbol{u} \otimes \boldsymbol{u} + p\boldsymbol{g}. \tag{27.10}$$

The 4-velocity $\boldsymbol{u}$ at a given event $\mathcal{P}$ in spacetime is the mean 4-velocity of the galaxies near $\mathcal{P}$; it is also the 4-velocity with which one must move in order to measure an isotropic intensity for the cosmic microwave radiation. The density ρ is made up of material density (rest mass plus negligible kinetic energy of galaxies; rest mass plus kinetic energy of cosmic rays; rest mass plus thermal energy of intergalactic gas—all "smeared out" over a unit volume), and also of radiation energy density (electromagnetic radiation, neutrino radiation, gravitational radiation). The pressure p, like the density ρ, is due to both matter and radiation. Today the pressure of the matter is much less than its mass-energy density,

$$p_{\text{matter}} \ll \rho_{\text{matter}} \text{ today,} \tag{27.11a}$$

but this strong inequality cannot have held long ago. Always the pressure of the radiation is $\frac{1}{3}$ its mass-energy density:

$$p_{\text{radiation}} = \frac{1}{3}\rho_{\text{radiation}} \text{ always.} \tag{27.11b}$$

§27.3. GEOMETRIC IMPLICATIONS OF HOMOGENEITY AND ISOTROPY

This chapter will idealize the universe to be *completely* homogeneous and isotropic. This idealization places tight constraints on the geometry of spacetime and on the motion of the cosmological fluid through it. In order to discover these constraints, one must first give precise mathematical meaning to the concepts of homogeneity and isotropy.

Homogeneity means, roughly speaking, that the universe is the same everywhere at a given moment of time. A given moment of what time? Whose time? This is the crucial question that the investigator asks.

In Newtonian theory there is no ambiguity about the concept "a given moment of time." In special relativity there is some ambiguity because of the nonuniversality of simultaneity, but once an inertial reference frame has been specified, the concept becomes precise. In general relativity there are no global inertial frames (unless spacetime is flat); so the concept of "a given moment of time" is completely ambiguous. However, another, more general concept replaces it: the concept of a three-dimensional spacelike hypersurface. This hypersurface may impose itself on one's

Spacelike hypersurface as generalization of "moment of time" attention by reason of natural symmetries in the spacetime. Or it may be selected at the whim or convenience of the investigator. He may find it more convenient to explore spacetime here and there than elsewhere, and to push the hypersurface forward accordingly ("many-fingered time"; the dramatically new conception of time that is part of general relativity). At each event on a spacelike hypersurface, there is a local Lorentz frame whose surface of simultaneity coincides locally with the hypersurface. Of course, this Lorentz frame is the one whose 4-velocity is orthogonal to the hypersurface. These Lorentz frames at various events on the hypersurface do not mesh to form a global inertial frame, but their surfaces of simultaneity do mesh to form the spacelike hypersurface itself.

The intuitive phrase "at a given moment of time" translates, in general relativity, into the precise phrase "on a given spacelike hypersurface." The investigator can go further. He can "slice up" the entire spacetime geometry by means of a "one-parameter family" of such spacelike surfaces. He can give the parameter that distinguishes one such slice from the next the name of "time." Such a one-parameter family of slices through spacetime is not required in the Regge calculus of Chapter 42. However, such a "slicing" is a necessity in most other practical methods for analyzing the dynamics of the geometry of the universe (Chapters 21, 30, and 43). The choice of slicing may dissolve away the difficulties of the dynamic analysis or may merely recognize those difficulties. The successive slices of "moments of time" may shine with simplicity or may only do a tortured legalistic bookkeeping for the dynamics. Which is the case depends on whether the typical spacelike hypersurface is distinguished by natural symmetries or, instead, is drawn arbitrarily.

"Homogeneity of universe" defined in terms of spacelike hypersurfaces

Homogeneity of the universe means, then, that through each event in the universe there passes a spacelike "hypersurface of homogeneity" (physical conditions identical at every event on this hypersurface). At each event on such a hypersurface the density, ρ, and pressure, p, must be the same; and the curvature of spacetime must be the same.

The concept of isotropy must also be made precise. Clearly, the universe cannot look isotropic to all observers. For example, an observer riding on a 10^{20} eV cosmic ray will see the matter of the universe rushing toward him from one direction and receding in the opposite direction. Only an observer who is moving with the cosmological fluid can possibly see things as isotropic. One considers such observers in defining isotropy:

"Isotropy of universe" defined

Isotropy of the universe means that, at any event, an observer who is "moving with the cosmological fluid" cannot distinguish one of his space directions from the others by any local physical measurement.

Isotropy of the universe actually implies homogeneity; of this one can convince oneself by elementary reasoning (exercise 27.1).

Isotropy implies fluid world lines orthogonal to homogeneous hypersurfaces

Isotropy guarantees that the world lines of the cosmological fluid are orthogonal to each hypersurface of homogeneity. This one sees as follows. An observer "moving with the fluid" can discover by physical measurements on which hypersurface through a given event conditions are homogeneous. Moreover, he can measure his own ordinary velocity relative to that hypersurface. If that ordinary velocity is nonzero, it provides the observer with a way to distinguish one space direction in his rest frame from all others—in violation of isotropy. Thus in an isotropic universe, where the concept of "observer moving with the fluid" makes sense, each such observer must discover that he is at rest relative to the hypersurface of homogeneity. His world line is orthogonal to that hypersurface.

Exercise 27.1. ISOTROPY IMPLIES HOMOGENEITY **EXERCISE**

Use elementary thought experiments to show that isotropy of the universe implies homogeneity.

---

§27.4. COMOVING, SYNCHRONOUS COORDINATE SYSTEMS FOR THE UNIVERSE

Construction of a "comoving, synchronous" coordinate system for the universe

The results of the last section enable one to set up special coordinate systems in the spacetime manifold of an isotropic model universe (Figure 27.1). Choose a hypersurface of homogeneity $\mathcal{S}_I$. To all the events on it assign coordinate time, t_I. Lay out, in any manner desired, a grid of space coordinates (x^1, x^2, x^3) on $\mathcal{S}_I$. "Propagate" these coordinates off $\mathcal{S}_I$ and throughout all spacetime by means of the world lines of the cosmological fluid. In particular, assign to every event on a given world line the space coordinates (x^1, x^2, x^3) at which that world line intersects $\mathcal{S}_I$. This assignment has a simple consequence. The fluid is always at rest relative to the space coordinates. In other words, *the space coordinates are "comoving";* they are merely labels for the world lines of the fluid. For the time coordinate t of a given event $\mathcal{P}$, use the lapse of proper time, $\int d\tau$, of $\mathcal{P}$ from $\mathcal{S}_I$, as measured along the fluid world line that passes through $\mathcal{P}$, plus t_I ("standard of time" on the initial hypersurface $\mathcal{S}_I$); thus,

$$t(\mathcal{P}) = t_I + \left(\int_{\mathcal{S}_I}^{\mathcal{P}} d\tau\right)_{\text{along world line of fluid}} \tag{27.12}$$

The surfaces t = constant of such a coordinate system will coincide with the hypersurfaces of homogeneity of the universe. This one sees by focusing attention on observations made by two different observers, A and B, who move with the fluid along different world lines. At coordinate time t_I (on $\mathcal{S}_I$) the universe looks the same to B as to A. Let A and B make observations again after their clocks have ticked away the same time interval $\Delta\tau$. Homogeneity of the initial hypersurface $\mathcal{S}_I$, plus the deterministic nature of Einstein's field equations, guarantees that A and B will again see identical physics. (Identical initial conditions on $\mathcal{S}_I$, plus identical lapses of proper time during which Einstein's equations govern the evolution of the universe near A and B, guarantee identical final conditions.) Therefore, after time lapse $\Delta\tau$, A and B are again on the same hypersurface of homogeneity—albeit a different

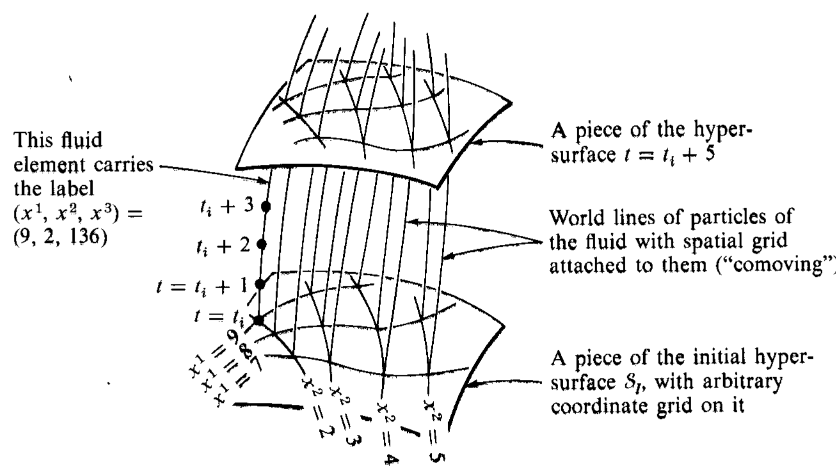


Figure 27.1.
Comoving, synchronous coordinate system for the universe, as constructed in §27.4 of the text. Key features of such a coordinate system are as follows (see §§27.4 and 27.5). (1) The spatial coordinates move with the fluid, and the time coordinate is proper time along the fluid world lines; i.e., the coordinate description of a particular fluid world line is

$$(x^1, x^2, x^3) = \text{constant}, \; x^0 \equiv t = \underbrace{\tau}_{\text{proper time measured along world line}} + \text{constant}.$$

(2) Any surface of constant coordinate time is a hypersurface of homogeneity of the universe. Every such hypersurface is orthogonal to the world lines of all particles of the fluid. (3) The spatial grid on some initial hypersurface $\mathcal{S}_I$ is completely arbitrary. (4) If $\gamma_{ij}\, dx^i\, dx^j$ is the metric on the initial hypersurface in terms of its arbitrary coordinates (with γ_{ij} a function of x^1, x^2, x^3), then the metric of spacetime in terms of the comoving, synchronous coordinate system is

$$ds^2 = -dt^2 + a^2(t)\gamma_{ij}\, dx^i\, dx^j.$$

Thus, the entire dynamics of the geometry of the universe is embodied in a single function of time, $a(t) \equiv$ "expansion factor"; while the shape (but not size) of the hypersurfaces of homogeneity is embodied in the spatial 3-metric $\gamma_{ij}\, dx^i\, dx^j$.

one from $\mathcal{S}_I$, where they began. By virtue of definition (27.12) of coordinate time, the time coordinate at the intersection of B's world line with this new hypersurface is $t = t_I + \Delta\tau$; and similarly for A. Moreover, observers A and B were arbitrary. Consequently the new hypersurface of homogeneity, like $\mathcal{S}_I$, is a hypersurface of constant coordinate time. Q.E.D.

Because the hypersurfaces of homogeneity are given by t = constant, the basis vectors $\partial/\partial x^i$ at any given event $\mathcal{P}$ are tangent to the hypersurface of homogeneity that goes through that event. On the other hand, the time basis vector, $\partial/\partial t$, is tangent to the world line of the fluid through $\mathcal{P}$, since that world line has x^i = constant along it. Consequently, orthogonality of the world line to the hypersurface guarantees orthogonality of $\partial/\partial t$ to $\partial/\partial x^i$:

$$(\partial/\partial t) \cdot (\partial/\partial x^i) = 0 \text{ for } i = 1, 2, 3. \tag{27.13a}$$

The time coordinate has another special property: it measures lapse of proper time along the world lines of the fluid. Because of this, and because $\partial/\partial t$ is tangent to the world lines, one can write

$$\begin{aligned}\partial/\partial t &= (d/d\tau)_{\text{along fluid's world lines}} \\ &= \boldsymbol{u},\end{aligned} \tag{27.13b}$$

where $\boldsymbol{u}$ is the 4-velocity of the "cosmological fluid." The 4-velocity always has unit length,

$$(\partial/\partial t)\cdot(\partial/\partial t) = -1. \tag{27.13c}$$

Conditions (27.13a,c) reveal that, in the comoving coordinate frame [where $g_{\alpha\beta} \equiv (\partial/\partial x^\alpha)\cdot(\partial/\partial x^\beta)$], the line element for spacetime reads

Form of the line element in this coordinate system

$$ds^2 = -dt^2 + g_{ij}\,dx^i\,dx^j. \tag{27.14}$$

Any coordinate system in which the line element has this form is said to be *"synchronous"* (1) because the coordinate time t measures proper time along the lines of constant x^i (i.e., $g_{tt} = -1$), and (2) because the surfaces $t =$ constant are (locally) surfaces of simultaneity for the observers who move with $x^i =$ constant [i.e., $g_{ti} = (\partial/\partial t)\cdot(\partial/\partial x^i) = 0$]; it is also called a "Gaussian normal coordinate system" (*cf.* Figure 21.6).

A hypersurface of homogeneity, $t =$ constant, has a spatial, three-dimensional geometry described by equation (27.14) with $dt = 0$:

$$\begin{aligned}(ds^2)_{\text{on hypersurface of homogeneity}} &= d\sigma^2 \\ &= [g_{ij}]_{t=\text{const}}\,dx^i\,dx^j.\end{aligned} \tag{27.15}$$

To know everything about the 3-geometry on each of these hypersurfaces is to know everything about the geometry of spacetime.

EXERCISE

Exercise 27.2. SYNCHRONOUS COORDINATES IN GENERAL

In an arbitrary spacetime manifold (not necessarily homogeneous or isotropic), pick an initial spacelike hypersurface $\mathcal{S}_I$, place an arbitrary coordinate grid on it, eject geodesic world lines orthogonal to it, and give these world lines the coordinates

$$(x^1, x^2, x^3) = \text{constant}, \qquad x^0 \equiv t = t_I + \tau,$$

where τ is proper time along the world line, beginning with $\tau = 0$ on $\mathcal{S}_I$. Show that in this coordinate system the metric takes on the synchronous (Gaussian normal) form (27.14).

---

§27.5. THE EXPANSION FACTOR

To determine the 3-geometry, $d\sigma^2 = g_{ij}(t, x^k)\, dx^i\, dx^j$, of each of the hypersurfaces of homogeneity, split the problem into two parts: (1) the nature of the 3-geometry on an arbitrary initial hypersurface (dealt with in next section); and (2) the evolution of the 3-geometry as time passes, i.e., as attention moves from the initial hypersurface to a subsequent hypersurface, and another, and another, . . . (dealt with in this section).

Proof that, aside from an over-all "expansion factor," all homogeneous hypersurfaces in the universe have the same 3-geometry

Assume that one knows the initial 3-geometry—i.e., the coefficients in the space part of the metric,

$$\gamma_{ij}(x^k) \equiv g_{ij}(t_I, x^k), \tag{27.16}$$

on the initial hypersurface $\mathcal{S}_I$—in its arbitrary but explicitly chosen coordinate system. What form will the metric coefficients $g_{ik}(t, x^k)$ have on the other hypersurfaces of homogeneity? This question is easily answered by the following argument: Consider two adjacent world lines, $\mathcal{A}$ and $\mathcal{B}$, of the cosmological fluid, with coordinates (x^1, x^2, x^3) and $(x^1 + \Delta x^1, x^2 + \Delta x^2, x^3 + \Delta x^3)$. At time t_I (on surface $\mathcal{S}_I$) they are separated by the proper distance

$$\Delta\sigma(t_I) = (\gamma_{ij}\, \Delta x^i\, \Delta x^j)^{1/2}. \tag{27.17}$$

At some later time t (on surface $\mathcal{S}$), they will be separated by some other proper distance $\Delta\sigma(t)$. Isotropy of spacetime guarantees that the ratio of separations $\Delta\sigma(t)/\Delta\sigma(t_I)$ will be independent of the direction from $\mathcal{A}$ to $\mathcal{B}$ (no shearing motion of the fluid). For any given direction, the additivity of small separations guarantees that $\Delta\sigma(t)/\Delta\sigma(t_I)$ will be independent of $\Delta\sigma(t_I)$. Thus $\Delta\sigma(t)/\Delta\sigma(t_I)$ must be the same for all pairs of world lines near a given world line. Finally, homogeneity guarantees that this scalar ratio will be independent of position on the initial surface $\mathcal{S}_I$—i.e., independent of x^1, x^2, x^3. Define $a(t)$ to be this spatially constant ratio,

$$a(t) \equiv \Delta\sigma(t)/\Delta\sigma(t_I). \tag{27.18}$$

Thus, $a(t)$ is the factor by which the separations of world lines expand between time t_I and time t. In other words, the function $a(t)$ is a universal *"expansion factor,"* or *"scale factor."*

By combining equations (27.17) and (27.18), one obtains for the separation of adjacent world lines at time t

$$\Delta\sigma(t) = a(t)[\gamma_{ij}(x^k)\, \Delta x^i\, \Delta x^j]^{1/2}.$$

This corresponds to the spatial metric at time t,

$$d\sigma^2 = a^2(t)\gamma_{ij}(x^k)\, dx^i\, dx^j, \tag{27.19}$$

and to the spacetime metric,

$$ds^2 = -dt^2 + a^2(t)\gamma_{ij}(x^k)\, dx^i\, dx^j. \tag{27.20}$$

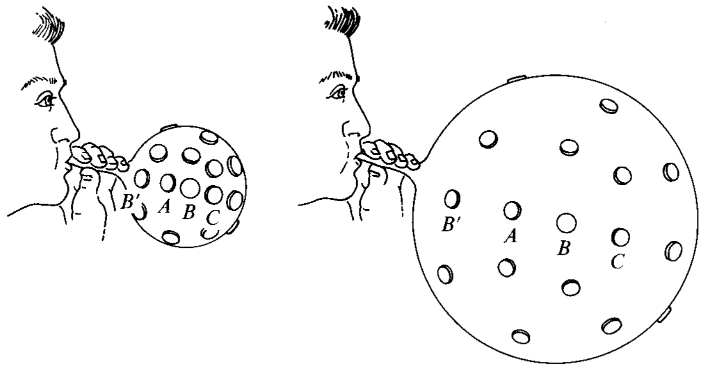


Figure 27.2.
Inflation of a balloon covered with pennies as a model for the expansion of the universe. Each penny A may well consider itself to be the center of the expansion because the distance from A to any neighbor B or C increases the more the more remote that neighbor was to begin with ("the Hubble relation"). The pennies themselves do not expand (constancy of sun-Earth distance, no expansion of a meter stick, no increase of atomic dimensions). The spacing today between galaxy and galaxy ($\sim 10^6$ lyr) is roughly ten times the typical dimension of a galaxy ($\sim 10^5$ lyr).

Notice that the coefficients $\gamma_{ij}(x^k)$ describe the shape not only of the initial hypersurface, but also of all other hypersurfaces of homogeneity. All that changes in the geometry from one hypersurface to the next is the scale of distances. All distances between spatial grid points (fluid world lines) expand by the same factor $a(t)$, leaving the shape of the hypersurface unchanged. This is a consequence of homogeneity and isotropy; and it is precisely true only if the model universe is precisely homogeneous and isotropic.

What expands in the universe, and what does not

Of all the disturbing implications of "the expansion of the universe," none is more upsetting to many a student on first encounter than the nonsense of this idea. The universe expands, the distance between one cluster of galaxies and another cluster expands, the distance between sun and earth expands, the length of a meter stick expands, the atom expands? Then how can it make any sense to speak of any expansion at all? Expansion relative to what? Expansion relative to nonsense! Only later does he realize that the atom does not expand, the meter stick does not expand, the distance between sun and earth does not expand. Only distances between clusters of galaxies and greater distances are subject to the expansion. Only at this gigantic scale of averaging does the notion of homogeneity make sense. Not so at smaller distances. No model more quickly illustrates the actual situation than a rubber balloon with pennies affixed to it, each by a drop of glue. As the balloon is inflated (Figure 27.2) the pennies increase their separation one from another but not a single one of them expands! [For mathematical detail see, e.g., Noerdlinger and Petrosian (1971).]

2

EXERCISE

Exercise 27.3. ARBITRARINESS IN THE EXPANSION FACTOR

How much arbitrariness is there in the definition of the expansion factor $a(t)$? Civilization A started long ago at time t_A. For it, the expansion factor is

$$\frac{\begin{pmatrix}\text{proper distance between}\\ \text{two particles of the “cos-}\\ \text{mological fluid” at time } t\end{pmatrix}}{\begin{pmatrix}\text{proper distance between}\\ \text{same two particles}\\ \text{at time } t_A\end{pmatrix}} = a_A(t).$$

Subsequently men planted civilization B at time t_B on a planet in a nearby galaxy. [At this time, the expansion factor a_A had the value $a_A(t_B)$]. Civilization B defines the expansion factor relative to the time of its own beginning:

$$\frac{\begin{pmatrix}\text{proper distance between}\\ \text{two particles of the “cos-}\\ \text{mological fluid” at time } t\end{pmatrix}}{\begin{pmatrix}\text{proper distance between}\\ \text{the same two particles}\\ \text{at time } t_B\end{pmatrix}} = a_B(t).$$

At two subsequent events, C and D, of which both civilizations are aware, they assign to the universe in their bookkeeping by no means identical expansion factors,

$$a_A(t_C) \neq a_B(t_C),$$
$$a_A(t_D) \neq a_B(t_D).$$

Show that the relative expansion of the model universe in passing from stage C to stage D in its evolution is nevertheless the same in the two systems of bookkeeping:

$$\frac{a_A(t_D)}{a_A(t_C)} = \begin{pmatrix}\text{relative expansion}\\ \text{from } C \text{ to } D\end{pmatrix} = \frac{a_B(t_D)}{a_B(t_C)}.$$

§27.6. POSSIBLE 3-GEOMETRIES FOR A HYPERSURFACE OF HOMOGENEITY

Riemann tensor for homogeneous, isotropic hypersurfaces

Turn now to the 3-geometry $\gamma_{ij}\, dx^i\, dx^j$ for the arbitrary initial hypersurface $\mathcal{S}_I$. This 3-geometry must be homogeneous and isotropic. A close scrutiny of its three-dimensional Riemann curvature must yield no "handles" to distinguish one point on $\mathcal{S}_I$ from any other, or to distinguish one direction at a given point from any other. "No handles" means that $^{(3)}$***Riemann*** must be constructed algebraically from pure numbers and from the only "handle-free" tensors that exist: the 3-metric γ_{ij} and the three-dimensional Levi-Civita tensor ε_{ijk}. (All other tensors pick out preferred directions or locations.) One possible expression for $^{(3)}$***Riemann*** is

$$^{(3)}R_{ijkl} = K(\gamma_{ik}\gamma_{jl} - \gamma_{il}\gamma_{jk});\ K = \text{“curvature parameter”} = \text{constant.} \tag{27.21}$$

Trial and error soon convince one that this is the *only* expression that both has the correct symmetries for a curvature tensor and can be constructed solely from constants, γ_{ij}, and ε_{ijk}. Hence, this must be the 3-curvature of $\mathcal{S}_I$. [One says that any manifold with a curvature tensor of this form is a manifold of *“constant curvature.”*]

Metric for homogeneous, isotropic hypersurfaces: three possibilities—positive, zero, or negative spatial curvature

As one might expect, the metric for $\mathcal{S}_I$ is completely determined, up to coordinate transformations, by the form (27.21) of its curvature tensor. (See exercise 27.4 below). With an appropriate choice of coordinates, the metric reads (see exercise 27.5 below),

$$\begin{aligned}
d\sigma^2 &= \gamma_{ij}\,dx^i\,dx^j = K^{-1}[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\,d\phi^2)] \text{ if } K > 0,\\
d\sigma^2 &= \gamma_{ij}\,dx^i\,dx^j = d\chi^2 + \chi^2(d\theta^2 + \sin^2\theta\,d\phi^2) \text{ if } K = 0,\\
d\sigma^2 &= \gamma_{ij}\,dx^i\,dx^j = (-K)^{-1}[d\chi^2 + \sinh^2\chi(d\theta^2 + \sin^2\theta\,d\phi^2)] \text{ if } K < 0.
\end{aligned} \tag{27.22}$$

Absorb the factor $K^{-1/2}$ or $(-K)^{-1/2}$ into the expansion factor $a(t)$ [see exercise 27.3], and define the function

$$\begin{aligned}
\Sigma &\equiv \sin\chi, && \text{if } k \equiv K/|K| = +1 \text{ (“positive spatial curvature”),}\\
\Sigma &\equiv \chi, && \text{if } k \equiv K = 0 \text{ (“zero spatial curvature”),}\\
\Sigma &\equiv \sinh\chi, && \text{if } k \equiv K/|K| = -1 \text{ (“negative spatial curvature”).}
\end{aligned} \tag{27.23}$$

Thus write the full spacetime geometry in the form

$$\begin{aligned}
ds^2 &= -dt^2 + a^2(t)\gamma_{ij}\,dx^i\,dx^j,\\
\gamma_{ij}\,dx^i\,dx^j &= d\chi^2 + \Sigma^2(d\theta^2 + \sin^2\theta\,d\phi^2),
\end{aligned} \tag{27.24}$$

and the 3-curvatures of the homogeneous hypersurfaces in the form

$$^{(3)}R_{ijkl} = [k/a^2(t)][\gamma_{ik}\gamma_{jl} - \gamma_{il}\gamma_{jk}]. \tag{27.25a}$$

The curvature parameter K, after this renormalization, is evidently

$$K = k/a^2(t). \tag{27.25b}$$

Significance of normalization of the expansion factor

Why is the word “renormalization” appropriate? Previously $a(t)$ was a scale factor describing expansion of linear dimensions relative to the linear dimensions as they stood at some arbitrarily chosen epoch; but the choice of that fiducial epoch was a matter of indifference. Now $a(t)$ has lost that arbitrariness. It has been normalized so that its value here and now gives the curvature of a spacelike hypersurface of homogeneity here and now. Previously the factor $a(t)$ was conceived as dimensionless. Now it has the dimensions of a length. This length is called the “radius of the model universe” when the curvature is positive. Even when the curvature is negative one sometimes speaks of $a(t)$ as a “radius.” Only for zero curvature does the normalization of $a(t)$ still retain its former arbitrariness. Thus, for zero-curvature, consider two choices for $a(t)$, one of them $a(t)$, the other $\bar{a}(t) = 2a(t)$. Then with $\bar{\chi} = \frac{1}{2}\chi$, one can write proper distances in the three directions of interest with perfect indifference in either of two ways:

$$\begin{pmatrix}\text{proper distance}\\ \text{in the direction}\\ \text{of increasing } \chi\end{pmatrix} = a(t)\,d\chi = \bar{a}(t)\,d\bar{\chi},$$

$$\begin{pmatrix}\text{proper distance}\\ \text{in the direction}\\ \text{of increasing } \theta\end{pmatrix} = a(t)\chi\,d\theta = \bar{a}(t)\bar{\chi}\,d\theta,$$

$$\begin{pmatrix}\text{proper distance}\\ \text{in the direction}\\ \text{of increasing } \phi\end{pmatrix} = a(t)\chi\sin\theta\,d\phi = \bar{a}(t)\bar{\chi}\sin\theta\,d\phi,$$

No such freedom of choice is possible when the model universe is curved, because then the χ's in the last two lines are replaced by a function, $\sin\chi$ or $\sinh\chi$, that is not linear in its argument.

Despite the feasibility in principle of determining the absolute value of the "radius" $a(t)$ of a curved universe, in practice today's accuracy falls short of what is required to do so. Therefore it is appropriate in many contexts to continue to regard $a(t)$ as a factor of relative expansion, the absolute value of which one tries to keep from entering into any equation exactly because it is difficult to determine. This motivation will account for the way much of the analysis of expansion is carried out in what follows, with calculations arranged to deal with ratios of a values rather than with absolute a values.

Box 27.2 explores and elucidates the geometry of a hypersurface of homogeneity.

EXERCISES

Exercise 27.4. UNIQUENESS OF METRIC FOR 3-SURFACE OF CONSTANT CURVATURE

Let γ_{ij} and $\gamma_{i'j'}$ be two sets of metric coefficients, in coordinate systems $\{x^i\}$ and $\{x^{i'}\}$, that have Riemann curvature tensors [derived by equations (8.22) and (8.42)] of the constant-curvature type (27.21). Let it be given in addition that the curvature parameters K and K' are equal. Show that γ_{ij} and $\gamma_{i'j'}$ are related by a coordinate transformation. [For a solution, see §8.10 of Robertson and Noonan (1968), or §§10 and 27 of Eisenhart (1926).]

Exercise 27.5. METRIC FOR 3-SURFACE OF CONSTANT CURVATURE

(a) Show that the following metric has expression (27.21) as its curvature tensor

$$\gamma_{ij} = \left(1 + \frac{1}{4}K\,\delta_{kl}x^k x^l\right)^{-2}\delta_{ij}. \tag{27.26}^*$$

*With this choice of spatial coordinates, the spacetime metric reads

$$ds^2 = -dt^2 + \frac{(dx^2 + dy^2 + dz^2)}{[1 + \frac{1}{4}K(x^2 + y^2 + z^2)]^2}.$$

This is often called the *"Robertson-Walker line element,"* because Robertson (1935, 1936) and Walker (1936) gave the first proofs that it describes the most general homogeneous and isotropic spacetime geometry.

(b) By transforming to spherical coordinates (R, θ, ϕ) and then changing to a Schwarzschild radial coordinate ($2\pi r$ = "proper circumference"), transform this metric into the form

$$d\sigma^2 = \frac{dr^2}{1 - Kr^2} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{27.27}$$

(c) Find a further change of radial coordinate that brings the metric into the form (27.22).

Exercise 27.6. PROPERTIES OF THE 3-SURFACES

Verify all statements made in Box 27.2.

Exercise 27.7. ISOTROPY IMPLIES HOMOGENEITY

Use the contracted Bianchi identity ${}^{(3)}G^{ik}{}_{|k} = 0$ (where the stroke indicates a covariant derivative based on the 3-geometry alone) to show (1) that ${}^{(3)}\nabla K = 0$ in equation (27.21), and therefore to show (2) that direction-independence of the curvature [isotropy; curvature of form (27.21)] implies and demands homogeneity (K constant in space).

(*continued on page 726*)

Box 27.2 THE 3-GEOMETRY OF HYPERSURFACES OF HOMOGENEITY

A. Universe with Positive Spatial Curvature '''Spatially Closed Universe'')

Metric of each hypersurface is

$$d\sigma^2 = a^2[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)]. \quad (1)$$

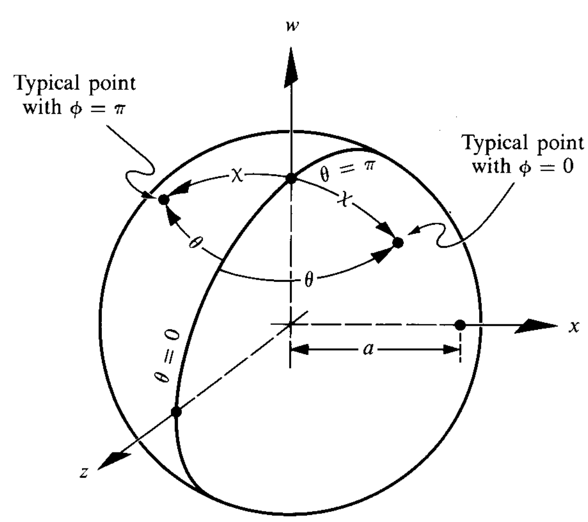
To visualize this 3-geometry, imagine embedding it in a four-dimensional Euclidean space (such embedding possible here; not possible for general three-dimensional manifold; only four freely disposable functions $[w, x, y, z]$ of three variables $[\alpha, \beta, \gamma]$ are at one's disposal to try to reproduce six prescribed functions $[g_{mn}(\alpha, \beta, \gamma)]$ of those same three variables).

The embedding is achieved by

$$w = a\cos\chi, \qquad z = a\sin\chi\cos\theta,$$
$$x = a\sin\chi\sin\theta\cos\phi, \quad (2)$$
$$y = a\sin\chi\sin\theta\sin\phi,$$

since it follows that

$$d\sigma^2 \equiv dw^2 + dx^2 + dy^2 + dz^2$$
$$= a^2[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)]. \quad (3)$$



A 3-surface of positive curvature embedded in four-dimensional Euclidean space. One rotational degree of freedom is suppressed by setting $\phi = 0$ and π ("slice through pole," 3-sphere in 4-space looks like a 2-sphere in 3-space).

Equations (2) for the embedded surface imply that

$$w^2 + x^2 + y^2 + z^2 = a^2; \tag{4}$$

i.e., the surface is a 3-dimensional sphere in 4-dimensional Euclidean space.

To verify homogeneity and isotropy, one need only notice that rotations in the four-dimensional embedding space can move any given point [any given (w, x, y, z) on the 3-sphere] and any given direction at that point into any other point and direction—while leaving unchanged the line element

$$d\sigma^2 = dw^2 + dx^2 + dy^2 + dz^2.$$

The above equations and the picture show that

(1) The 2-surfaces of fixed χ (which look like circles in the picture, because one rotational degree of freedom is suppressed) are actually 2-spheres of surface area $4\pi a^2 \sin^2\chi$; and (θ, ϕ) are standard spherical coordinates on these 2-spheres.

(2) As χ ranges from 0 to π, one moves outward from the "north pole" of the hypersurface, through successive 2-spheres ("shells") of area $4\pi a^2 \sin^2\chi$ (2-spheres look like circles in picture). The area of these shells increases rapidly at first and then more slowly as one approaches the "equator" of the hypersurface, $\chi = \pi/2$. Beyond the equator the area decreases slowly at first, and then more rapidly as one approaches the "south pole", ($\chi = \pi$; area $= 0$).

(3) The entire hypersurface is swept out by

$$0 \le \chi \le \pi,$$
$$0 \le \theta \le \pi,$$
$$0 \le \phi \le 2\pi$$

(ϕ is cyclic; $\phi = 0$ is same as $\phi = 2\pi$);

its 3-volume is

$$\mathcal{V} = \int (a\,d\chi)(a \sin\chi\,d\theta)(a \sin\chi \sin\theta\,d\phi)$$
$$= \int_0^\pi 4\pi a^2 \sin^2\chi (a\,d\chi) = 2\pi^2 a^3. \tag{5}$$

B. Universe with Zero Spatial Curvature ("Spatially Flat Universe")

Metric of each hypersurface is

$$d\sigma^2 = a^2[d\chi^2 + \chi^2(d\theta^2 + \sin^2\theta\,d\phi^2)]. \tag{6}$$

This is a perfectly flat, three-dimensional, Euclidean space described in spherical coordinates. In Cartesian coordinates

$$x = a\chi \sin\theta \cos\phi,$$
$$y = a\chi \sin\theta \sin\phi, \tag{7}$$
$$z = a\chi \cos\theta,$$

the metric is

$$d\sigma^2 = dx^2 + dy^2 + dz^2. \tag{8}$$

The entire hypersurface is swept out by

$$0 \le \chi < \infty,$$
$$0 \le \theta \le \pi, \tag{9}$$
$$0 \le \phi \le 2\pi;$$

and its volume is infinite.

C. Universe with Negative Spatial Curvature ("Spatially open Universe")

Metric of each hypersurface is

$$d\sigma^2 = a^2[d\chi^2 + \sinh^2\chi(d\theta^2 + \sin^2\theta\,d\phi^2)]. \tag{10}$$

This 3-geometry cannot be embedded in a four-dimensional Euclidean space; but it can be embedded in a flat Minkowski space

$$d\sigma^2 = -dw^2 + dx^2 + dy^2 + dz^2. \tag{11}$$

To achieve the embedding, set

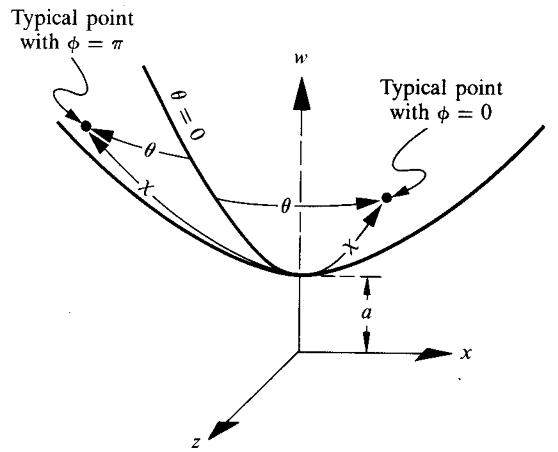
$$w = a\cosh\chi, \qquad z = a\sinh\chi\cos\theta,$$
$$x = a\sinh\chi\sin\theta\cos\phi, \tag{12}$$
$$y = a\sinh\chi\sin\theta\sin\phi;$$

insert this into equation (11), and thereby obtain (10).

Equations (12) for the embedded surface imply that

$$w^2 - x^2 - y^2 - z^2 = a^2; \tag{13}$$

i.e., the surface is a three-dimensional hyperboloid in four-dimensional Minkowski space. (It has the



A 3-surface of negative curvature embedded in four-dimensional Minkowski space. One rotational degree of freedom is suppressed by setting $\phi = 0$ and π ("slice through pole"; 3-hyperboloid in 4-space looks like 2-hyperboloid in 3-space).

same form as a mass hyperboloid in momentum space; see Box 22.5.)

To verify homogeneity and isotropy, one need only notice that "Lorentz transformations" in the embedding space can move any given point on the 3-hyperboloid and any direction through that point into any other point and direction—while leaving unchanged the line element

$$d\sigma^2 = -dw^2 + dx^2 + dy^2 + dz^2.$$

The above equations and the picture show that

(1) The 2-surfaces of fixed χ (which look like circles in the picture because one rotational degree of freedom is suppressed) are actually 2-spheres of surface area $4\pi a^2 \sinh^2\chi$; and (θ, ϕ) are standard spherical coordinates on these 2-spheres.

(2) As χ ranges from 0 to ∞, one moves outward from the (arbitrarily chosen) "pole" of the hypersurface, through successive 2-spheres ("shells") of ever increasing area $4\pi a^2 \sinh^2\chi$. For large χ, surface area increases far more rapidly than it would if the hypersurface were flat

$$\frac{\text{(proper surface area)}}{4\pi\,\text{(proper distance)}^2} = \frac{A}{4\pi\,\ell^2} = \frac{4\pi a^2 \sinh^2\chi}{4\pi a^2 \chi^2} \approx \left(\frac{e^{\ell/a}}{2\,\ell/a}\right)^2 \longrightarrow \infty. \tag{14}$$

The entire hypersurface is swept out by

$$0 \le \chi < \infty, \quad 0 \le \theta \le \pi, \quad 0 \le \phi \le 2\pi \tag{15}$$

(ϕ is cyclic; $\phi = 0$ is same as $\phi = 2\pi$).

The volume of the hypersurface is infinite.

D. Nonuniqueness of Topology

Warning: Although the demand for homogeneity and isotropy determines completely the local geometric properties of a hypersurface of homogeneity up to the single disposable factor K, it leaves the global topology of the hypersurface undetermined. The above choices of topology are the most straightforward. But other choices are possible.

This arbitrariness shows most simply when the hypersurface is flat ($k = 0$). Write the full spacetime metric in Cartesian coordinates as

$$ds^2 = -dt^2 + a^2(t)[dx^2 + dy^2 + dz^2]. \tag{16}$$

Then take a cube of coordinate edge L

$$0 < x < L, \qquad 0 < y < L, \qquad 0 < z < L,$$

and identify opposite faces (process similar to rolling up a sheet of paper into a tube and gluing its edges together; see last three paragraphs of §11.5 for detailed discussion). The resulting geometry is still described by the line element (16), but now all three spatial coordinates are "cyclic," like the ϕ coordinate of a spherical coordinate system:

$$(t, x, y, z) \text{ is the same event as } (t, x + L, y + L, z + L).$$

The homogeneous hypersurfaces are now "3-toruses" of finite volume

$$V = a^3L^3,$$

analogous to the 3-toruses which one meets under the name "periodic boundary conditions" when analyzing electron waves and acoustic waves in solids and electromagnetic waves in space.

Another example: The 3-sphere described in part A above (case of "positive curvature") has the same geometry, but not the same topology, as the manifold of the rotation group, $SO(3)$ [see exercises 9.12, 9.13, 10.16, and 11.12]. For detailed discussion, see for example Weyl (1946), Coxeter (1963), and Auslander and Markus (1959).

§27.7. EQUATIONS OF MOTION FOR THE FLUID

After the above analysis of any one hypersurface of homogeneity, return to the dynamics of the universe. Examine, first, the evolution of the fluid, as governed by the law $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$.

Recall (§22.3 and 23.5) that for a perfect fluid the equations of motion split into two parts. The component along the 4-velocity, $\boldsymbol{u} \cdot (\boldsymbol{\nabla} \cdot \boldsymbol{T}) = 0$, reproduces the first law of thermodynamics

$$(d/d\tau)(\rho V) = -p(dV/d\tau), \tag{27.28a}$$

where V is the volume of any fluid element. The part orthogonal to the 4-velocity, $(\boldsymbol{g} + \boldsymbol{u} \otimes \boldsymbol{u}) \cdot (\boldsymbol{\nabla} \cdot \boldsymbol{T}) = 0$, gives the force equation ("Euler equation")

$$(\rho + p) \times (\text{4-acceleration}) = -(\text{component of } \boldsymbol{\nabla} p \text{ orthogonal to } \boldsymbol{u}). \tag{27.28b}$$

Euler equation is vacuous for a homogeneous universe

For a static star (§23.5) the first law of thermodynamics was vacuous, but the force equation was crucial. For a homogeneous universe, the converse is true; the force equation is vacuous (no accelerations), but the first law of thermodynamics is crucial.

To see that the force equation is vacuous, notice that isotropy guarantees the vanishing of both sides of equation (27.28b). If either side were nonzero at any event $\mathscr{P}$, it would distinguish a direction in the homogeneous hypersurface at $\mathscr{P}$.

In applying the first law of thermodynamics (27.28a) to cosmology, divide the density and pressure into contributions due to matter and contributions due to radiation:

$$\rho = \rho_m + \rho_r; \qquad p = p_m + p_r. \tag{27.29}$$

"Equations of state" for matter and radiation

First discuss the density of mass-energy. Today $\rho_m(\gtrsim 10^{-31}\ \text{g/cm}^3)$ dominates over $\rho_r(\sim 10^{-33}\ \text{g/cm}^3)$. Matter did not always dominate. Therefore, one cannot set $\rho_r = 0$. Now discuss the pressure. During that epoch of the universe when pressure was significant cosmologically, p_r dominated over p_m. Consequently, one can neglect p_m at all times, and one can use the "equation of state" for radiation, $p_r = \frac{1}{3}\rho_r$, to write

$$\rho = \rho_m + \rho_r; \qquad p = \frac{1}{3}\rho_r. \tag{27.30}$$

When (27.30) is inserted into the first law of thermodynamics (27.28a), it yields the result

$$(d/d\tau)(\rho_m V) + (d/d\tau)(\rho_r V) = -\frac{1}{3}\rho_r\, dV/d\tau. \tag{27.31}$$

Energy exchange between matter and radiation is negligible

One cannot integrate this equation until one knows how mass-energy is fed back and forth between matter and radiation—i.e., until one knows another relationship between $\rho_m V$ and $\rho_r V$. All estimates indicate that, except in the first few seconds of the life of the universe, the energy exchanged between radiation and matter was

2

negligible compared to $\rho_m V$ and $\rho_r V$ individually (see §28.1). Under these conditions, equation (27.31) can be split into two parts:

First law of thermodynamics used to express densities of radiation and matter in terms of expansion factor

$$(d/d\tau)(\rho_m V) = 0, \tag{27.32a}$$

and

$$(d/d\tau)(\rho_r V) + \frac{1}{3}\rho_r \, dV/d\tau = 0. \tag{27.32b}$$

The solutions are simple:

$$\rho_m V = \text{constant (conservation of matter)} \tag{27.33a}$$

and

$$\rho_r V^{4/3} = \text{const} = \frac{\rho_r}{\underbrace{V^{-1/3}}_{\substack{\text{energy } hc/\lambda \text{ of} \\ \text{one photon, up} \\ \text{to a factor of} \\ \text{proportionality}}}} V \left(\begin{array}{c}\text{constancy of number} \\ \text{of photons}\end{array}\right) \tag{27.33b}$$

Now what is V? It is the volume of any fluid element. It has the value

$$V = a^3 \Sigma^2 \sin\theta \, \Delta\chi \, \Delta\theta \, \Delta\phi$$

for a fluid element with edges $\Delta\chi$, $\Delta\theta$, $\Delta\phi$. Here χ, θ, ϕ are constant along each world line of the fluid (comoving coordinates). Therefore the element of hyperspherical solid angle $\Sigma^2 \sin\theta \, \Delta\chi \, \Delta\theta \, \Delta\phi$ (or pseudohyperspherical solid angle for the model of an open universe) is constant throughout all time for any fluid element. Therefore the volume of the fluid element grows in direct proportion to the cube of the expansion parameter a; thus,

$$V/a^3 = \text{constant}.$$

Combining this result with the constancy of $\rho_m V$ and $\rho_r V^{4/3}$, one sees that

$$\rho_m a^3 = \text{constant}, \qquad \rho_r a^4 = \text{constant}. \tag{27.34}$$

Let ρ_{mo} be the density of matter today, ρ_{ro} be the density of radiation today, and a_o be the expansion factor for the universe today. Then, at any time in the past,

$$\rho(t) = \rho_{mo}\frac{a_o^3}{a^3(t)} + \rho_{ro}\frac{a_o^4}{a^4(t)} \tag{27.35a}$$

and

$$p(t) = \frac{1}{3}\rho_{ro}\frac{a_o{}^4}{a^4(t)}. \tag{27.35b}$$

These results were based on two key claims, which will be justified in detail later (Chapter 28): the claim that in the epoch when pressure was important p_m was much smaller than p_r; and the claim that exchange of mass-energy between radiation and matter was always negligible (except in the first few seconds after the "creation").

§27.8. THE EINSTEIN FIELD EQUATION

Once the time evolution of the expansion factor, $a(t)$, is known, one can read off the time evolution of the density and pressure directly from equations (27.35). The density and pressure, in turn, determine how the expansion proceeds in time, via Einstein's field equations. Thus the field equations "close the logic loop" and give one a closed mathematical system from which to determine all three quantities, $a(t)$, $p(t)$ and $\rho(t)$.

One can readily calculate the components of the Einstein tensor for the model universe using the orthonormal basis one-forms,

$$\boldsymbol{\omega}^{\hat{t}} \equiv \boldsymbol{d}t, \qquad \boldsymbol{\omega}^{\hat{\chi}} \equiv a(t)\, \boldsymbol{d}\chi, \qquad \boldsymbol{\omega}^{\hat{\theta}} \equiv a(t)\Sigma\, \boldsymbol{d}\theta, \qquad \boldsymbol{\omega}^{\hat{\phi}} \equiv a(t)\Sigma \sin\theta\, \boldsymbol{d}\phi. \tag{27.36}$$

Evaluation of the Einstein field equation for a homogeneous universe:

The result [see equations (5) of Box 14.5] is

$$G_{\hat{t}\hat{t}} = 3\left(\frac{a_{,t}}{a}\right)^2 + \frac{3k}{a^2}, \tag{27.37a}$$

$$G_{\hat{\chi}\hat{\chi}} = G_{\hat{\theta}\hat{\theta}} = G_{\hat{\phi}\hat{\phi}} = -2\frac{a_{,tt}}{a} - \left(\frac{a_{,t}}{a}\right)^2 - \frac{k}{a^2}, \tag{27.37b}$$

$$G_{\hat{\mu}\hat{\nu}} = 0 \text{ if } \mu \neq \nu. \tag{27.37c}$$

(With foresight, one will notice ahead of time that isotropy guarantees the equality $G_{\hat{\chi}\hat{\chi}} = G_{\hat{\theta}\hat{\theta}} = G_{\hat{\phi}\hat{\phi}}$, and similar equalities for the Riemann tensor; and one will calculate only $G_{\hat{\chi}\hat{\chi}}$, the component that is most easily calculated.)

The basis one-forms, $\boldsymbol{\omega}^{\hat{t}}$, $\boldsymbol{\omega}^{\hat{\chi}}$, $\boldsymbol{\omega}^{\hat{\theta}}$, $\boldsymbol{\omega}^{\hat{\phi}}$, are the orthonormal basis carried along by an observer who moves with the "cosmological fluid." Consequently, $T_{\hat{t}\hat{t}}$ is the mass-energy density, ρ, that he measures; $T_{\hat{j}\hat{j}}$ is the pressure, p; $T_{\hat{t}\hat{j}}$ vanishes, because he sees no energy flux (no momentum density); and $T_{\hat{i}\hat{j}}$ vanishes for $i \neq j$ because he sees no shear stresses:

$$T_{\hat{t}\hat{t}} = \rho, \tag{27.38a}$$

$$T_{\hat{\chi}\hat{\chi}} = T_{\hat{\theta}\hat{\theta}} = T_{\hat{\phi}\hat{\phi}} = p, \tag{27.38b}$$

$$T_{\hat{\mu}\hat{\nu}} = 0 \text{ when } \mu \neq \nu. \tag{27.38c}$$

Equate the Einstein ("moment of rotation") tensor of equations (27.37) to the stress-energy tensor of equations (27.38). And if one insists, include the so-called "Λ-term" or "cosmological term" in the field equations [Einstein (1970): "the biggest blunder of my life"). Thus obtain two nonvacuous field equations. The first is an "initial value equation," which relates $a_{,t}$ to a and ρ at any initial moment of time:

$$\left(\frac{a_{,t}}{a}\right)^2 = -\frac{k}{a^2} + \underbrace{\frac{\Lambda}{3}}_{\text{omit}} + \frac{8\pi}{3}\rho. \tag{27.39a}$$

(1) initial value equation

The second is a "dynamic equation," which gives the second time-derivative of the expansion factor, and thereby governs the dynamic evolution away from the initial moment of time,

$$2\frac{a_{,tt}}{a} = -\left(\frac{a_{,t}}{a}\right)^2 - \frac{k}{a^2} + \underbrace{\Lambda}_{\text{omit}} - 8\pi p. \tag{27.39b}$$

(2) dynamic equation

Why the dynamic equation is superfluous

If (27.39b) is to be compared with anything in Newtonian mechanics, it is to be compared with an equation for acceleration (equation of motion), and in the same spirit (27.39a) is to be compared with a first integral of the equation of motion; that is, an equation of energy. In accordance with this comparison, note that one only has to differentiate (27.39a) and combine it with the relation satisfied by the pressure,

$$(\rho a^3)_{,t} = -p(a^3)_{,t}$$

("law of conservation of energy") to get the acceleration equation (27.39b). Without any loss of information, one can therefore ignore the "acceleration equation" or "dynamic equation" (27.39b) henceforth, and work with the analog of an energy equation or what is more properly known as an "initial-value equation" (details of initial-value problem for Track-2 readers in Chapter 21).

Side remarks about initial value equations, dynamic equations, and Bianchi identities in more general contexts

What shows up here in the limited context of Friedmann cosmology is appropriately viewed in the wider context of general geometrodynamics. Conservation of energy plus one field equation have just been seen to reproduce the other field equations. Conversely, by accepting both field equations, one can derive the law of conservation of energy in the form just stated. Thus, the very act of writing the field equation $\boldsymbol{G} = 8\pi \boldsymbol{T}$ (or, if one insists upon the "cosmological term,"

$$\boldsymbol{G} + \underbrace{\Lambda \boldsymbol{g}}_{\text{omit}} = 8\pi \boldsymbol{T})$$

was encouraged by and founded on the automatic vanishing of the divergence $\boldsymbol{\nabla} \cdot \boldsymbol{G}$ (or the vanishing of the divergences of $\boldsymbol{G}$ and $\boldsymbol{g}$), because one knew to begin with that energy and momentum are conserved, $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$. It is not surprising, then, that there should be a redundancy between the conservation law, $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$, and the field equations. Neither is it surprising in the dynamics of the Friedmann universe that one can use what is here the one and only interesting component of the conservation law, plus the one and only interesting initial value component ($G_{\hat{t}\hat{t}}$ component) of the field equations, to obtain the one and only interesting dynamic component ($G_{\hat{x}\hat{x}}$ component) of the field equations.

In a similar way, in more general problems that lack symmetry, one can always eliminate *some* of the dynamic field equations, but when gravitational radiation is present, one cannot eliminate them all. The dynamic field equations that cannot be eliminated, even in principle, govern the propagation of the gravitational waves. No gravitational waves are present in a perfectly homogeneous and isotropic cosmological model; its high degree of symmetry—in particular, its spherical (2-sphere!) symmetry about $\chi = 0$—is incompatible with gravitational waves.

Now turn back from general dynamics to Friedmann cosmology. To determine the time evolution of the expansion factor, a, insert into the initial-value equation (27.39a) the expression for the density of mass-energy given in (27.35a), and arrive at an equation ready for integration,

Differential equation for expansion factor

$$\left(\frac{a_{,t}}{a}\right)^2 = -\frac{k}{a^2} + \underbrace{\frac{\Lambda}{3}}_{\text{omit}} + \underbrace{\frac{(8\pi\rho_{mo}a_o^3/3)}{a^3} + \frac{(8\pi\rho_{ro}a_o^4/3)}{a^4}}_{(8\pi/3)\rho(a)} \tag{27.40}$$

When one has completed the integration of this equation for $a = a(t)$, one turns back to equation (27.35a,b) to get $\rho(t)$ and $p(t)$, and to expression (27.24) to get the geometry,

$$ds^2 = -dt^2 + a^2(t)[d\chi^2 + \Sigma^2(d\theta^2 + \sin^2\theta\, d\phi^2)], \tag{27.41}$$

thus completing the solution of the problem.

§27.9. TIME PARAMETERS AND THE HUBBLE CONSTANT

To the analysis of this dynamic problem, many investigators have contributed over the years, beginning with Friedmann himself in 1922. They discovered, among other results, that there are three natural choices of time variable, the one of greatest utility depending on the application that one has at hand.

Three choices of time parameter for universe:

(1) proper time, t

First is t, the original time variable. This quantity gives directly proper time elapsed since the start of the expansion. This is the time available for the formation of galaxies. It is also the time during which radioactive decay and other physical processes have been taking place.

(2) expansion factor, a

Second is $a(t)$, the expansion factor, which grows with time, which therefore serves to distinguish one phase of the expansion from another, and which consequently can be regarded as a parametric measure of time in its own right. The ratio of $a(t)$ at two times gives the ratio of the dimensions of the universe (cube root of volume) at those two times. It also gives the ratio $(1 + z)$ of wavelengths at those two times (see §29.2). A knowledge of the red shift, z, experienced in time past by radiation received today is equivalent to a knowledge of $a(t)/a_o$, where a_o is the expansion factor today. Specifically, radiation coming in with $z = 999$ is radiation coming in from a time in the history of the universe when it had 10^{-3} of its present dimensions and 10^{-9} of its present volume. During the interval of time while the expansion parameter is increasing from a to $a + da$, the lapse of proper time, according to (27.40), is

$$dt = \frac{da}{[-k + (8\pi/3)a^2\rho(a) + \underbrace{(\Lambda/3)a^2}_{\text{omit}}]^{1/2}}. \qquad (27.42)$$

In terms of a as a new time parameter, it follows from this formula that the metric takes the form [Hughston (1969)]

$$ds^2 = \frac{-(da)^2}{-k + (8\pi/3)a^2\rho(a) + \underbrace{(\Lambda/3)a^2}_{\text{omit}}} + a^2[d\chi^2 + \Sigma^2(d\theta^2 + \sin^2\theta\, d\phi^2)]. \qquad (27.43)$$

Third is $\eta(t)$, the "arc-parameter measure of time." During the interval of time dt, a photon traveling on a hypersphere of radius $a(t)$ covers an arc measured in radians equal to

(3) arc parameter, η

$$d\eta = \frac{dt}{a(t)}. \qquad (27.44)$$

When the model universe is open instead of closed, the same parameter lets itself be defined. Only the words "hypersphere" and "arc" have to be replaced by the corresponding words for a flat hypersurface of homogeneity ($k = 0$) or a hyperboloidal hypersurface ($k = -1$). In all three cases, the "arc parameter" is defined by the integral of this expression from the start of the expansion:

$$\eta = \int_0^t \frac{dt}{a(t)}; \qquad (27.45)$$

thus small values of the "arc parameter time," η, mean early times; and larger values mean later times. In terms of this "arc-parameter measure of time," the metric takes the form

$$ds^2 = a^2(\eta)[-d\eta^2 + d\chi^2 + \Sigma^2(d\theta^2 + \sin^2\theta\, d\phi^2)]. \qquad (27.46)$$

Let a photon start at the "North Pole" of the 3-sphere ($\chi = 0$; any θ and ϕ) at the "arc parameter time" $\eta = \eta_1$. Then, by the "arc parameter time" $\eta = \eta_2$, the photon has traveled to a new point on the hypersphere and encountered a new set of particles of the "cosmological fluid." They lie at the hyperpolar angle

$$\chi = \eta_2 - \eta_1.$$

When one makes a spacetime diagram on a piece of paper to show what is happening when an effect propagates from one point to another in the universe, one finds it most convenient to take (1) the space coordinate to be χ (the life histories of distinct particles of the "cosmological fluid" thus being represented by distinct vertical lines), and (2) the time coordinate to be η (so that photons are described by lines inclined at $\pm 45°$). No time parameter is more natural to use than η when one is tracing out the course of null geodesics. For an example, see the treatment of the cosmological redshift in §29.2. It also turns out that it is simpler analytically (when Λ is taken to be zero) to give $a = a(\eta)$ and $t = t(\eta)$ than to give a directly as a function of time. Thus one gets the connection between the dimension a and the "arc-parameter time" η from the formula

$$\eta = \int d\eta = \int \frac{dt}{a(t)} = \int \frac{da}{[-ka^2 + (8\pi/3)a^4\rho(a) + \underbrace{(\Lambda/3)a^4}_{\text{omit}}]^{1/2}}. \tag{27.47}$$

From a knowledge of the dimension a as a function of this time parameter, one immediately gets proper time itself in terms of this time parameter, from the formula

$$dt = a(\eta)\, d\eta. \tag{27.48}$$

An equation (27.40) for the expansion factor and a choice of parameters for marking out time have now set the stage for a detailed analysis of idealized cosmology, and some of the relevant questions have even been asked: How does the characteristic dimension, a, of the geometry (radius of 3-sphere, in the case of closure) change with time? What is the spacetime geometry? How do geodesics, especially null geodesics, travel in this geometry? However, additional questions are equally important: Is the expansion of the universe decelerating and, if so, how fast? How do density and pressure of matter and radiation vary with time? And finally, for the simplest and most immediate tie between theory and observation, what is the expansion rate?

Hubble constant and Hubble time

In speaking of expansion rate, one refers to the "Hubble constant," the fractional rate of increase of distances,

$$H \equiv \frac{\dot{a}(t)}{a(t)}, \tag{27.49}$$

which is normally evaluated today $H(\text{today}) \equiv H_o$, but is in principle defined as a function of time for every phase of the history of the universe. The reciprocal of H is the "Hubble time," H^{-1}. This quantity represents the time it would have taken for the galaxies to attain their present separations, starting from a condition of infinite compaction, if they had maintained for all time their present velocities ("time for expansion with dimensions linearly extrapolated back to the start"). For the conversion from astrophysical to geometric units and to years, take the currently accepted value, $H_o = 55$ km/sec megaparsec (Box 29.4), as an illustration:

$$\begin{aligned} H_o &= \frac{55 \text{ km/sec}}{(299{,}793 \text{ km/sec})(3.0856 \times 10^{24} \text{ cm or } 3.2615 \times 10^6 \text{ yr of time})} \\ &= 0.59 \times 10^{-28} \text{ per cm of light-travel time} \\ &\qquad \text{or } 5.6 \times 10^{-11} \text{ fractional expansion per yr,} \end{aligned} \tag{27.50}$$

$$H_o^{-1} = 1.7 \times 10^{28} \text{ cm of light-travel time or } 18 \times 10^9 \text{ yr.}$$

§27.10. THE ELEMENTARY FRIEDMANN COSMOLOGY OF A CLOSED UNIVERSE

Features of a closed Friedmann universe with $\Lambda = 0$:

Take the simplest cosmological model, an isotropic homogeneous closed universe with $\Lambda = 0$, and trace out its features in all detail in the two limiting cases where matter dominates and where radiation dominates. The term "Friedmann universe" is used here for both cases, although the matter-dominated model is sometimes referred to as the Friedmann universe and the radiation-dominated one as the Tolman universe. In this analysis, it will be appropriate to let the variable $a(t)$ represent the radius of the universe, as measured in cm, because only by reference to this radius does one have the tool in hand to discuss all the interesting geometric effects that in principle lend themselves to observation. After this discussion, it will be enough, in dealing with other models, to summarize their principal parts and comment on their differences from this simple model, without repeating the full investigation. Any reference to an open universe or any so-called "cosmological constant" or its effects will therefore be deferred to a brief final section, §27.11. There the variable $a(t)$ will sometimes be taken to represent only a parameter of relative expansion, as is appropriate for discussions reaching out only to, say, $z = 0.1$, where global geometric issues are not taken up.

(1) radius as function of time

Rewrite the controlling component (27.40) of Einstein's field equation in the form

$$\left(\frac{da}{dt}\right)^2 - \frac{8\pi\rho_{mo}a_o^3/3}{a} - \frac{8\pi\rho_{ro}a_o^4/3}{a^2} = -1. \tag{27.51}$$

According as one neglects the radiation term or the matter term in this equation, the equation idealizes to

$$\left(\frac{da}{dt}\right)^2 - \frac{a_{\max}}{a} = -1, \qquad \begin{pmatrix}\text{27.52; matter}\\ \text{dominates}\end{pmatrix}$$

or

$$\left(\frac{da}{dt}\right)^2 - \frac{a^{*2}}{a^2} = -1. \qquad \begin{pmatrix}\text{27.52; radiation}\\ \text{dominates}\end{pmatrix}$$

In both cases, the problem lends itself to comparison to the problem of particle motion in Newtonian mechanics with "total energy" -1 and with an "effective potential energy" of the qualitative form shown in diagram **A** of Box 27.1—apart from minor differences in shape according as the potential goes as $-1/a$ or as $-1/a^2$. The principal features of the solution are collected in Box 27.3.

(2) early era, when radiation dominates: types of radiation

It is a striking feature of the radiation-dominated era of the early Friedmann universe that the density of the radiation depends on time according to a simple universal law,

$$\rho_r = 3/32\pi t^2 \tag{27.53}$$

(final line and final column of Box 27.3). This circumstance may someday provide

(*continued on page 736*)

Box 27.3 SOLUTIONS FOR THE ELEMENTARY FRIEDMANN COSMOLOGY OF A CLOSED UNIVERSE IN THE TWO LIMITING CASES IN WHICH (1) MATTER DOMINATES AND RADIATION IS NEGLIGIBLE, AND (2) RADIATION DOMINATES AND MATTER IS NEGLIGIBLE

Idealization for dynamics of 3-sphere	*Matter dominated*	*Radiation dominated*
Model relevant when?	back into past to redshift $z \sim 10{,}000$; through today and through phase of maximum expansion, and recontraction down to dimensions $\sim 10{,}000$-fold smaller than today	very early phase of expansion, for redshifts $z \sim 1{,}000$ and greater; and corresponding phase in late stages of recontraction; not directly relevant today.
Effective "potential" in $\left(\dfrac{da}{dt}\right)^2 + V(a) = -1$	$V(a) = -\dfrac{a_{\max}}{a}$	$V(a) = -\dfrac{a^{*2}}{a^2}$
Value of constant in this "potential" in terms of conditions at some standard epoch	$a_{\max} = \dfrac{8\pi}{3} a_o^3 \rho_{mo}$	$a^{*2} = \dfrac{8\pi}{3} a_o^4 \rho_{ro}$
Solution of dynamic equation expressed parametrically in terms of "arc parameter" η (radians of arc distance on 3-sphere covered by a photon travelling ever since start of expansion)	$a = \dfrac{a_{\max}}{2}(1 - \cos\eta)$ $t = \dfrac{a_{\max}}{2}(\eta - \sin\eta)$	$a = a^* \sin\eta$ $t = a^*(1 - \cos\eta)$
Range of η from start of expansion to end of recontraction	2π (one trip around the universe)	π (gets only as far as antipodal point of universe)
Nature of curve relating radius a to time t	cycloid	semicircle
Hubble time $H^{-1} = \dfrac{a}{(da/dt)} = \dfrac{a^2}{(da/d\eta)}$	$\dfrac{a_{\max}}{2}\dfrac{(1-\cos\eta)^2}{\sin\eta}$	$a^* \dfrac{\sin^2\eta}{\cos\eta}$

Idealization for dynamics of 3-sphere	*Matter dominated*	*Radiation dominated*
Inequality between Hubble or "extrapolated" time and actual time back to start of expansion	$H^{-1} \geq 1.5t$	$H^{-1} \geq 2t$
Density of mass-energy	$\rho_m = \dfrac{3}{\pi a_{\max}^2 (1 - \cos\eta)^3}$	$\rho_r = \dfrac{3}{8\pi a^{*2} \sin^4\eta}$
This density expressed in terms of Hubble expansion rate	$\rho_m = \dfrac{3H^2}{8\pi} \dfrac{2}{1 + \cos\eta}$	$\rho_r = \dfrac{3H^2}{8\pi} \dfrac{1}{\cos^2\eta}$
Inequality satisfied by density	$\rho_m \geq \dfrac{3H^2}{8\pi}$	$\rho_r \geq \dfrac{3H^2}{8\pi}$
Analysis of magnification of distant galaxy by curvature of intervening space	§29.5 and Figure 29.2	§29.5
Limiting form of law of expansion for early times	$\begin{cases} t = \dfrac{a_{\max}}{12}\eta^3 \\ a = \dfrac{a_{\max}}{4}\eta^2 \\ a = \left(\dfrac{9a_{\max}t^2}{4}\right)^{1/3} \end{cases}$	$\begin{cases} t = \dfrac{a^*}{2}\eta^2 \\ a = a^*\eta \\ a = (2a^*t)^{1/2} \end{cases}$
Other features of expansion at early times	$\begin{cases} H^{-1} = \dfrac{a_{\max}}{8}\eta^3 = 1.5t \\ \rho_m = \dfrac{a_{\max}}{(8\pi a^3/3)} \\ \quad = \dfrac{1}{6\pi t^2} = \dfrac{3H^2}{8\pi} \end{cases}$	$\begin{cases} H^{-1} = a^*\eta^2 = 2t \\ \rho_r = 3p_r = \dfrac{a^{*2}}{(8\pi a^4/3)} \\ \quad = \dfrac{3}{32\pi t^2} = \dfrac{3H^2}{8\pi} \end{cases}$

2

a tool to tell how many kinds of radiation contributed to ρ_r in the early universe; or, in other words, to learn about field physics from observational cosmology. Express the density of radiation in the form

$$\rho_r(\mathrm{cm}^{-2}) = \frac{G}{c^4}\,\rho_{r,\mathrm{conv}}(\mathrm{erg/cm^3}) = \frac{Gf\pi^2}{c^4 120}\,\frac{(kT)^4}{\hbar^3 c^3}. \tag{27.54}$$

It would be surprising if electromagnetism made the sole contribution to the radiation density, since the following additional mechanisms are available to sop up thermal energy from a violently radiating source:

electromagnetic radiation (already considered),	$f_{em} = 8$;
gravitational black body radiation,	$f_g = 8$;
neutrino plus antineutrino radiation of the electron-neutrino type [its contribution depends on the chemical potential of the neutrinos, on which see Brill and Wheeler (1957); a zero value is assumed here for that potential],	$f_{e\nu} = 7$;
neutrino plus antineutrino radiation of the muon-neutrino type [with the same assumptions as for ν_e's],	$f_{\mu\nu} = 7$;
pairs of positive and negative electrons produced out of the vacuum when temperatures are of the order of $T = mc^2/k = 0.59 \times 10^{10}$ K and higher, evaluated in the approximation in which these particles are treated as overwhelmingly more numerous than the unpaired electrons that one sees today,	$f_{e^+e^-} = 14$;
other particles such as mesons created out of the vacuum when temperatures are two orders of magnitude higher ($\sim 10^{12}$ K), and baryon-antibaryon pairs created out of the vacuum when temperatures are of the order of $\sim 10^{13}$ K and higher,	$f_{\mu^+\mu^-}, f_\pi, \ldots$;
sum of these f-values,	f. (27.55)

As the expansion proceeds and temperatures drop below 10^{13} K, then 10^{12}, then 10^{10}, the various particle pairs presumably annihilate and disappear [see, however, Alfvén and Klein (1962), Alfvén (1971), Klein (1971), and Omnes (1969)]. One is left with the radiations of zero rest mass, and only these radiations, contributing to the specific heat of the vacuum. At the phases of baryon-antibaryon and electron-positron annihilation, the thermal gravitational radiation present has already effectively decoupled itself from the matter, according to all current estimates. Therefore the energy set free by annihilation of matter and antimatter is expected to pour at first into the other two carriers of energy: neutrinos and electromagnetic radiation. However, the neutrinos also decouple early (after baryon-antibaryon annihilation; before full electron-positron annihilation), because the mean free path for neutrinos rises rapidly with expansion. The energy of the subsequent annihilations goes almost exclusively into electromagnetic radiation. Thus the temperatures of the three radiations at the present time are expected to stand in the order

$$T_{\rm em} > T_\nu > T_g. \tag{27.56}$$

$T_{\rm em}$ has been measured to be 2.7 K; T_ν is calculated to be $(4/11)^{1/3}\, T_{\rm em} = 1.9$ K, and T_g has been calculated to be 1.5 K [Matzner (1968)] in a model where gravitons decouple during an early, quark-dominated era.

Decoupled radiation, once in a Planck spectrum, remains in a Planck spectrum (see Box 29.2). Expansion leaves constant the product $\rho_{r,\rm decoupled}\, a^4$ or the product $T^4_{r,\rm decoupled}\, a^4$. Compare the temperature of this particular radiation now to the temperature of the same radiation at any chosen fiducial time $t_{\rm fid}$ after its era of decoupling. Find

$$T_{r,\rm fid} = \frac{a_{\rm now}}{a_{\rm fid}}\, T_{r,\rm now} = (1+z)T_{r,\rm now}. \tag{27.57}$$

Here z represents the red shift of any "tracer" spectral line, given off at the fiducial time, and observed today, relative to the standard wavelength of the same transition as observed in the laboratory.

If the three radiations could be catalyzed into thermodynamic equilibrium, then all radiations could be treated on the same footing during the radiation-dominated era of cosmology. Their individual f values could be added directly to give $f = 8 + 8 + 7 + 7 = 30$. Temperature and time would then be connected by the formula

$$(T/10^{10}\, K)^2(t/1\ \text{sec}) = 1.19. \tag{27.58a}$$

This formula together with (27.57) implies the relation

$$\left[\left(\frac{T_{r,\rm now}}{10^{10}\,\rm K}\right)(1+z)\right]^2\left(\frac{t_{\rm fid}}{1\ \rm sec}\right) = 1.19. \tag{27.58b}$$

This relation concerns two radiations: (1) the actual electromagnetic radiation with Planck spectrum (a continuum); and (2) the redshift and time of emission of a "tracer radiation" (a line spectrum). A measured departure from this relation could serve as one potential (indirect) indication that, in accordance with standard theory, neutrinos and gravitational radiation today are cooler than electromagnetic radiation.

(3) later era, when matter dominates

Turn now from the radiation-dominated era of cosmology to the matter-dominated era. Numbers sometimes elicit more response from the imagination than formulas. Therefore idealize to a matter-dominated cosmology, and for the moment arbitrarily adopt 20×10^9 yr and 10×10^9 yr as Hubble time and actual time, respectively, back to the start of the expansion. It is certain that future work will show both numbers to require revision, but probably not by more than a factor 2, in the opinion of observational cosmologists. Since any judgment on the best numbers is subject to uncertainty, one can pick the numbers to be simple as well as reasonable. From Box 27.3, one then deduces the present value of the arc parameter time η,

$$\frac{20 \times 10^9 \text{ lyr}}{10 \times 10^9 \text{ lyr}} = \frac{H^{-1}}{t} = \frac{\dfrac{a_{\max}}{2} \dfrac{(1 - \cos\eta)^2}{\sin\eta}}{\dfrac{a_{\max}}{2}(\eta - \sin\eta)} \tag{27.59}$$

or

$$\eta = 1.975 \text{ (or } 113.2^\circ) \tag{27.60}$$

(arc traveled by a photon on the 3-sphere from the start of the expansion to today.) This fixed, all other numbers emerge as shown in Box 27.4.

Box 27.4 A TYPICAL COSMOLOGICAL MODEL COMPATIBLE WITH ASTRONOMICAL OBSERVATIONS AND WITH EINSTEIN'S CONCEPTION OF COSMOLOGY ($\Lambda = 0$; Universe Closed)

Radius at phase of maximum expansion,	18.94×10^9 lyr;
Time from start to maximum,	29.76×10^9 yr;
Time from start to final recontraction,	59.52×10^9 yr;
Time from start to today (adopted value),	10×10^9 yr;
Radius today,	13.19×10^9 lyr;
Hubble time today (adopted value),	20×10^9 yr;
Hubble expansion rate today,	49.0 km/sec Megaparsec;
Deceleration parameter today, q_o [equation (29.1b)]	1.7
Density today $(3/8\pi a_o{}^2) + (3H_o^2/8\pi)$,	$(7.67 + 3.33) \times 10^{-58}\text{ cm}^{-2}$ $= 11.00 \times 10^{-58}\text{ cm}^{-2}$ or $14.8 \times 10^{-30}\text{ g/cm}^3$;
Volume today, $2\pi^2 a_o{}^3$,	$38.3 \times 10^{84}\text{ cm}^3$;
Density at maximum $(3/8\pi a^2) + (3H^2/8\pi)$,	$(3.70 + 0.00) \times 10^{-58}\text{ cm}^{-2}$ $= 5.0 \times 10^{-30}\text{ g/cm}^3$;
Volume at maximum,	$114 \times 10^{84}\text{ cm}^3$;
Rate of increase of radius today,	$13.19 \times 10^9\text{ lyr}/20 \times 10^9\text{ yr}$ $= 0.66$ lyr/yr;
Rate of increase of volume today,	$1.82 \times 10^{68}\text{ cm}^3/\text{sec}$;
Amount of matter,	5.68×10^{56} g;
Equivalent number of solar masses,	2.86×10^{23};
Equivalent number of baryons,	3.39×10^{80}.
Fraction visible today	0.74

It must be emphasized that these numbers do not deserve the title of "canonical," however convenient that adjective may be for describing them; they can at most be called illustrative.

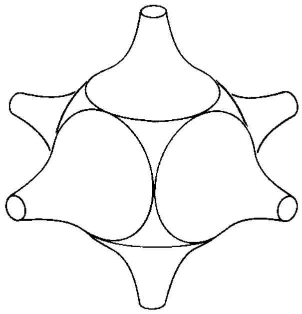


Figure 27.3.
Many Schwarzschild zones are fitted together to make a closed universe. This universe is dynamic because a test particle at the interface between two zones rises up against the gravitational attraction of each and falls back under the gravitational attraction of each. Therefore the two centers themselves have to move apart and move back together again. The same being true for all other pairs of centers, it follows that the lattice universe itself expands and recontracts, even though each Schwarzschild geometry individually is viewed as static. This diagram is taken from Lindquist and Wheeler (1957).

(4) "Where is the new space created during expansion?"—a meaningless question

If every five seconds a volume of space is added to the universe, a volume equivalent to a cube 10^5 lyr ($= 0.95 \times 10^{23}$ cm) on an edge, about equal to the volume occupied by the Milky Way, where does that volume make its entry? Rather than look for an answer, one had better reexamine the question. Space is not like water. The outpouring of fresh water beneath the ocean at the Jesuit Spring off Mount Desert Island can be detected and measured by surrounding the site with flowmeters. There is no such thing as a flowmeter to tell "how fast space is streaming past." The very idea that "space flows" is mistaken. There is no way to define a flow of space, not least because there is no way to measure a flow of space. Water, yes; space, no. Life is very different for the flowmeter, according as it is stationary or moving with respect to the water. For a particle in empty space, however, physics is indistinguishable regardless of whether the particle is at rest or moves at high velocity relative to some chosen inertial frame. To try to pinpoint where those cubic kilometers of space get born is a mistaken idea, because it is a meaningless idea.

One can get a fresh perspective on what is going on in expansion and recontraction by turning from a homogeneous isotropic closed universe to a Schwarzschild lattice closed universe. [Lindquist and Wheeler (1957)]. In the former case, the mass is idealized as distributed uniformly. In the latter, the mass is concentrated into 120 identical Schwarzschild black holes. Each is located at the center of its own cell, of dodecahedral shape, bounded by 12 faces, each approximately a pentagon; and space is empty. The dynamics is easy to analyze in the approximation in which each lattice cell is idealized as spherical, a type of treatment long familiar in solid-state physics as the "Wigner-Seitz approximation" (references in Lindquist and Wheeler). In this approximation, the geometry inside each lattice cell is treated as having exactly the Schwarzschild character (Figure 27.3); a test particle placed midway between black hole A and black hole B rises against the attraction of each, and ultimately falls back toward each, according to the law developed in Chapter 25 [equation (25.28) with a shift of π in the starting point for defining η],

$$\begin{aligned} r &= \frac{R}{2}(1 - \cos\eta), \\ \tau &= \frac{R}{2}\left(\frac{R}{2M}\right)^{1/2}(\eta - \sin\eta). \end{aligned} \tag{27.61}$$

Accordingly, the two masses in question must fall toward each other; and so it is with all the masses. One comes out in this way with the conclusion that the lattice universe follows the same law of expansion and recontraction as the Friedmann universe to an accuracy of better than 4 per cent [Lindquist and Wheeler; Wheeler (1964a), pp. 370–381]. Now ask again the same meaningless question about where the cubic kilometers of space pour into the universe while it is expanding, and where they pour out while it is recontracting. Receive a fuller picture why the question is meaningless. Surrounding each center of mass, the geometry is and remains the Schwarzschild geometry (until eventually the black holes come so close together that they coalesce). The situation inside each cell is therefore static. Moreover, the interface between cell and cell is defined in imagination by a sprinkling of test particles so light that they have no influence on the geometry or its dynamics. The matchup between the geometry in one cell and the next is smooth ("tangency between the two geometries"). There is nothing abnormal whatsoever in the spacetime on and near the interface. One has as little right to say those cubic kilometers are "created" here as anywhere else. To speak of the "creation" of space is a bad way of speaking, and the original question is a bad question. The right way of speaking is to speak of a dynamic geometry. So much for one question!

In charting the dynamics of the geometry of a Friedmann universe, one often finds that it simplifies things to take as space coordinate the hyperpolar angle χ, measured from some chosen world line (moving with the "cosmological fluid") as standard of reference; and to take as time coordinate the arc-parameter measure of time, η, as illustrated in Figure 27.4.

(5) causal isolation of various regions of universe from each other

Inspection of the (χ, η)-diagram makes it clear that photons emitted from matter at one point cannot reach, in a limited time, any matter except that which is located in a limited fraction of the 3-sphere. In a short time t, according to Box 27.3, a photon can cover an arc distance on the 3-sphere equal only to $\eta = (2t/a^*)^{1/2}$. Moreover, what is true of photons is true of other fields, forces, pressures, energies and influences: they cannot reach beyond this limit. Evidently the 3-sphere at time t is divided into a number of "zones,"

$$N = \begin{pmatrix}\text{number of}\\ \text{“zones”}\end{pmatrix} = \frac{\begin{pmatrix}\text{hyperspherical solid}\\ \text{angle of entire 3-sphere}\end{pmatrix}}{\begin{pmatrix}\text{hyperspherical solid}\\ \text{angle of one zone}\end{pmatrix}} = \frac{2\pi^2}{4\pi\chi^3/3} = \frac{3\pi}{2^{5/2}}\left(\frac{a^*}{t}\right)^{3/2}, \tag{27.62}$$

effectively decoupled one from the other. As time goes on, there are fewer separate zones, and ultimately every particle has been subjected to influences from every other particle in the model universe.

EXERCISES

Exercise 27.8. MATTER-DOMINATED AND RADIATION-DOMINATED REGIMES OF FRIEDMANN COSMOLOGY

Derive the results listed in the last two columns of Box 27.3, except for the focusing properties of the curved space.

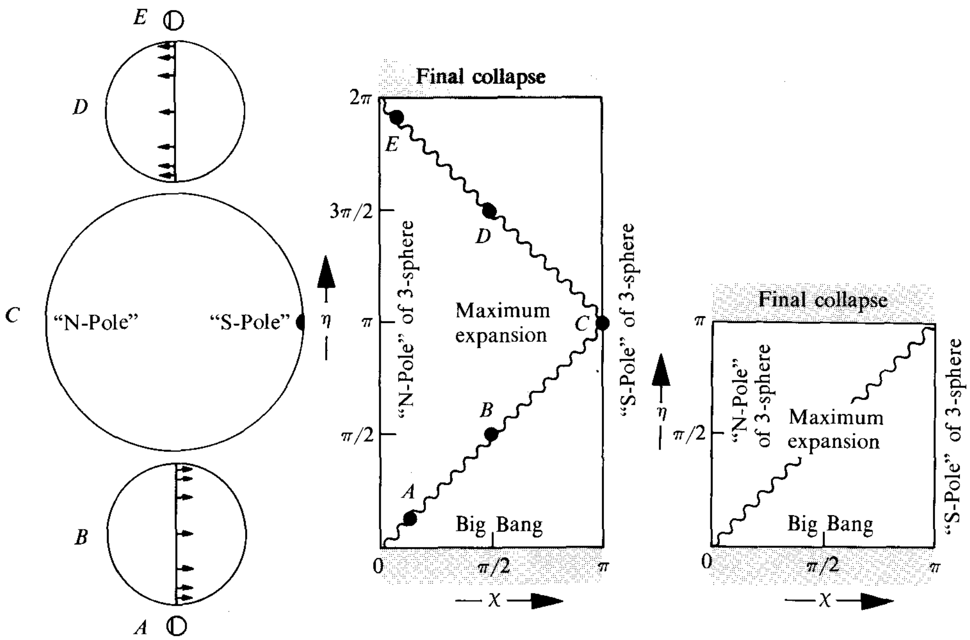


Figure 27.4.
Use of "arc parameter" η as a time coordinate and hyperpolar angle χ as a space coordinate to describe travel of a photon ($\pm 45°$ line) in a Friedmann universe that is matter-dominated (center) or radiation-dominated (right). The burst of photons is emitted from the "*N*-pole" of the 3-sphere at a time very little after the big bang, and the locus of the cloud of photons at subsequent stages of the expansion and recontraction is indicated by sections of the 3-sphere in the diagrams at the left. The matter-dominated Friedmann universe appears to be a reasonable model for the physical universe, except when its dimensions have fallen to the order of one ten-thousandth of those at maximum expansion or less ("radiation regime").

Exercise 27.9. TRANSITION FROM RADIATION-DOMINATED REGIME TO MATTER-DOMINATED REGIME

Including both the radiation and the matter terms in equation (27.51), restate the equation in terms of the arc parameter η (with $d\eta = dt/a$) as independent variable, and integrate to find

$$a = (a_{\max}/2) - [(a_{\max}/2)^2 + a^{*2}]^{1/2} \cos(\eta + \delta), \tag{27.63}$$

$$t = (a_{\max}/2)\eta - [(a_{\max}/2)^2 + a^{*2}]^{1/2}[\sin(\eta + \delta) - \sin\delta], \tag{27.64}$$

where

$$\delta = \arctan[a^*/(a_{\max}/2)]. \tag{27.65}$$

(a) Verify that under appropriate conditions these expressions reduce at early times to a "circle" relation between radius and time and to a "cycloid" relation later.

(b) Assign to a^{*2} the value $a_0 a_{\max}/10{,}000$ (why?) and construct curves for the dimensionless measures of density,

$$\log_{10}\left[(8\pi a_{\max}{}^2/3)\begin{Bmatrix}\rho_m\\ \rho_r\\ \rho_m+\rho_r\end{Bmatrix}\right],$$

as a function of the dimensionless measure of time,

$$\log_{10}(t/a_{\max}).$$

What conclusions emerge from inspecting the logarithmic slope of these curves?

Exercise 27.10. THE EXPANDING AND RECONTRACTING SPHERICAL WAVE FRONT

An explosion takes place at the "N-pole" of the matter-dominated Friedmann model universe at the value of the "arc parameter time" $\eta = \pi/3$, when the radius of the universe has reached half its peak value. The photons from the explosion race out on a spherical wave front. Through what fraction of the "cosmological fluid" has this wave front penetrated at that instant when the wave front has its largest proper surface area?

§27.11. HOMOGENEOUS ISOTROPIC MODEL UNIVERSES THAT VIOLATE EINSTEIN'S CONCEPTION OF COSMOLOGY

Open Friedmann universe with $\Lambda = 0$:

(1) expansion factor as function of time

(2) early stage—same as for closed universe

It violates Einstein's conception of cosmology (Box 27.1)—though not the equations of his theory—to replace the closed 3-sphere of radius a by the open hyperboloidal geometry of equation (27.22) with the same scale length a. Even so, the results of Box 27.3 continue to apply in the two limiting regimes of matter-dominated and radiation-dominated dynamics when the following changes are made. (1) Change the constant -1 on the righthand side of the analog of a "Newtonian energy equation" to $+1$, thus going over from a bound system (maximum expansion) to an open system (forever expanding). (2) Replace $(1 - \cos\eta)$ by $(\cosh\eta - 1)$, $\sin\eta$ by $\sinh\eta$, $\cos\eta$ by $\cosh\eta$, and $(\eta - \sin\eta)$ by $(\sinh\eta - \eta)$. (3) The range of the "arc parameter" η now extends from 0 to ∞, and the curve relating "radius" a to time t changes from cycloid or circle to an ever-rising curve. (4) The listed inequalities on the Hubble time (as related to the actual time of expansion) and on the density (as related to $3H_0^2/8\pi$) no longer hold. (5) The formulas given in Box 27.3 for conditions at early times continue to hold, for a simple reason: at early times the curvature of spacetime "in the direction of increasing time" [the extrinsic curvature $(6/a^2)(da/dt)^2$ as it appears in Box 27.1, equation (2)] is overwhelmingly more important than the curvature within any hypersurface of homogeneity, $\pm 6/a^2$ (the intrinsic curvature); therefore it makes no detectable difference at early times whether the sign is plus or minus, whether the space is closed or open, or whether the geometry of space is spherical or hyperboloidal.

Why doesn't it make a difference? Not why mathematically, but why physically, doesn't it make a difference in early days whether the space is open or closed?

Because photons, signals, pressures, forces, and energies cannot get far enough to "smell out" the difference between closure and openness. The "zones of influence" of (27.62) are too small for any one by itself to sense or to respond significantly to any difference between a negative space curvature $-6/a^2$ and a positive space curvature $+6/a^2$. Therefore the simple power-law time-dependence of the density of the mass-energy of radiation given in Box 27.3 for a closed universe holds equally well in the earliest days of a radiation-dominated, open, isotropic model universe; thus,

$$\rho_r = 3/32\pi t^2. \tag{27.66}$$

Only at a later stage of the expansion, when the "extrinsic curvature" term [equation (2), Box 27.1], $(6/a^2)(da/dt)^2$ (initially varying as $1.5t^{-2}$, according to Box 27.3) has fallen to a value of the same order of magnitude as the "intrinsic curvature" term $\pm 6/a^2$ (initially varying as $\pm 3a^{*-1}t^{-1}$), does the sign of the intrinsic curvature begin to matter. Only then do the differences in rate of expansion begin to show up that distinguish the open model universe from the closed one.

The open model goes on expanding forever. Therefore the density of mass-energy, whether matter-dominated and proportional to $a_{\max}/a^3$, or radiation-dominated and proportional to a^{*2}/a^4, or some combination of the two, (1) ultimately falls to a level that is negligible in comparison with the intrinsic curvature, $-6/a^2$, and (2) thereafter can be neglected. Under these circumstances, the only term left to balance the intrinsic curvature is the extrinsic curvature. The important component of the field equation (after removal from all terms of a common factor 3) now reads

(3) late stage—expansion forever

$$\frac{1}{a^2}\left(\frac{da}{dt}\right)^2 - \frac{1}{a^2} = 0. \tag{27.67}$$

For a closed universe, the two terms (one sixth the extrinsic curvature and one sixth the intrinsic curvature) have the same sign, and any equation like (27.67) leads to an impossibility. Here, however, rather than impossibility, one has the remarkably simple solution

$$a = t, \tag{27.68}$$

and the corresponding metric

$$ds^2 = -dt^2 + t^2[d\chi^2 + \sinh^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)]. \tag{27.69}$$

Write

$$\begin{aligned} r &= t \sinh \chi, \\ t_{\text{new}} &= t \cosh \chi, \end{aligned} \tag{27.70}$$

and find that (27.69), solution as it is of Einstein's empty-space field equation, is identical with the Lorentz-Minkowski metric of flat spacetime,

$$ds^2 = -dt_{\text{new}}^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{27.71}$$

(see Box 27.2C). This geometry had acquired the flavor of an expanding universe because the cosmological fluid, too thinly spread to influence the dynamics of the geometry, and serving only to provide marker points, was flying out in all directions [for a fuller discussion of this "expanding Minkowski universe," see, for example, Chapter 16 of Robertson and Noonan (1968)]. The typical spacelike hypersurface of homogeneity looks to have a curved 3-space geometry, and does have a curved geometry (instrinsic curvature), because the slice (27.70) through flat spacetime is itself curved (extrinsic curvature).

Homogeneous cosmologies with $\Lambda \neq 0$:

Turn now to a second violation of Einstein's conception of cosmology: a cosmological term in the field equation (27.39),

(1) equation for evolution of expansion factor

$$\left(\frac{da/dt}{a}\right)^2 + \frac{k}{a^2} - \frac{\Lambda}{3} = \frac{8\pi}{3}\rho(a) = \frac{8\pi\rho_{mo}a_o^3/3}{a^3} + \frac{8\pi\rho_{ro}a_o^4/3}{a^4}. \tag{27.72}$$

In analyzing the implications of this broadened equation, turn attention from the "radius" $a(t)$ itself, which was the focus of interest in the previous section, §27.10, on Friedmann cosmology. Recognize that present measurements have not yet provided a good, direct handle on the absolute dimension $a(t)$ of the universe. However, they do give good figures for the redshift z and therefore for the ratio between a at the time of emission and $a = a_o$ now,

$$a_o/a = 1 + z \tag{27.73}$$

For any comparison with observations designed to fix limits on k (Einstein value: $k = +1$) and on Λ (expected to be zero), it is therefore appropriate to rewrite the foregoing equations so that they refer as much as possible only to ratios. Thus one rephrases (27.72) as the "generalized Friedmann equation,"

$$\left[\frac{d}{dt}\left(\frac{a(t)}{a_o}\right)\right]^2 + V(a/a_o) = -\frac{k}{{a_o}^2} \equiv -K_o. \tag{27.74}$$

Here the quantity

$$V(a/a_o) \equiv -\frac{8\pi}{3}\left[\rho_{mo}\left(\frac{a_o}{a}\right) + \rho_{ro}\left(\frac{a_o}{a}\right)^2\right] - \frac{1}{3}\Lambda\left(\frac{a}{a_o}\right)^2 \tag{27.75}$$

acts as an "effective potential" for the dynamics of the expansion. The constant term K_o represents one sixth of the intrinsic curvature of the model universe today. Its negative, $-K_o$, plays the role of an "effective energy" in the generalized Friedmann equation (Box 27.5). All the qualitative features of the cosmology can be read off from the curve for the effective potential as a function of (a/a_o) and from the value of K_o.

(2) qualitative features of evolution

For a quantitative analysis, the log-log diagram of Figure 27.5 is often more useful than the straight linear plot of V against (a/a_o) of Box 27.5.

All the limiting features shown in the varied types of cosmology have been encountered before in the analysis of the elementary Friedmann cosmology (big bang out of a configuration of infinite compaction; reaching a maximum expansion at a turning point, or continued expansion to a Minkowski universe; recollapse to infinite density) or lend themselves to simple visualization (static but unstable Einstein universe; "hesitation" model; "turnaround" model), except for the even more rapid expansion that occurs when Λ is positive and the dimension a has surpassed a certain critical value. In this expansion, a eventually increases as $\exp[(\Lambda/3)^{1/2}t]$ irrespective of the openness or closure of the universe ($k = 0, \pm 1$). This expansion dominates every other feature of the cosmology. Therefore, in discussing it, it is appropriate to suppress every other feature of the cosmology, take the density of matter to be negligible, and take $k = 0$ (hypersurfaces of homogeneity endowed with flat 3-space geometry). In this limit, one has the following empty-space solution of Einstein's field equation with cosmological constant:

(3) de Sitter universe

$$ds^2 = -dt^2 + a_o^2 e^{2(\Lambda/3)^{1/2}t}[d\chi^2 + \chi^2(d\theta^2 + \sin^2\theta\, d\phi^2)]. \tag{27.76}$$

This "de Sitter universe" [de Sitter (1917a,b)] may be regarded as a four-dimensional surface,

$$-(z^0)^2 + (z^1)^2 + (z^2)^2 + (z^3)^2 + (z^4)^2 = 3/\Lambda, \tag{27.77}$$

in a five-dimensional space endowed with the metric

$$(ds)^2 = -(dz^0)^2 + (dz^1)^2 + (dz^2)^2 + (dz^3)^2 + (dz^4)^2. \tag{27.78}$$

The correctness of this description may be checked directly by making the substitutions

$$\begin{aligned}
z^0 &= (3/\Lambda)^{1/2}\sinh[(\Lambda/3)^{1/2}t] + \frac{1}{2}(\Lambda/3)^{1/2}e^{(\Lambda/3)^{1/2}t}a_o^2\chi^2,\\
z^4 &= (3/\Lambda)^{1/2}\cosh[(\Lambda/3)^{1/2}t] - \frac{1}{2}(\Lambda/3)^{1/2}e^{(\Lambda/3)^{1/2}t}a_o^2\chi^2,\\
z^1 &= a_o e^{(\Lambda/3)^{1/2}t}\chi\sin\theta\cos\phi,\\
z^2 &= a_o e^{(\Lambda/3)^{1/2}t}\chi\sin\theta\sin\phi,\\
z^3 &= a_o e^{(\Lambda/3)^{1/2}t}\chi\cos\theta.
\end{aligned} \tag{27.79}$$

Because of its beautiful group-theoretical properties and invariance with respect to $5 \times 4/2 = 10$ independent rotations, the de Sitter geometry has been the subject of scores of mathematical investigations. The physical implications of a cosmology following the de Sitter model are described for example by Robertson and Noonan (1968, especially their §16.2). The de Sitter model is the only model obeying Einstein's equations (with $\Lambda \neq 0$) which (1) continually expands and (2) looks the same to any observer who moves with the cosmological fluid, regardless of his position or his time. Any model of the universe satisfying condition (2) is said to obey the so-called "perfect cosmological principle." This phrase arose in the past in studying models that lie outside the framework of general relativity, models in which matter is envisaged as continuously being created, and to which the name of "steady-state universe" has been given. Any such model has been abandoned by most investigators today, not least because it gives no satisfactory account of the 2.7 K background radiation.

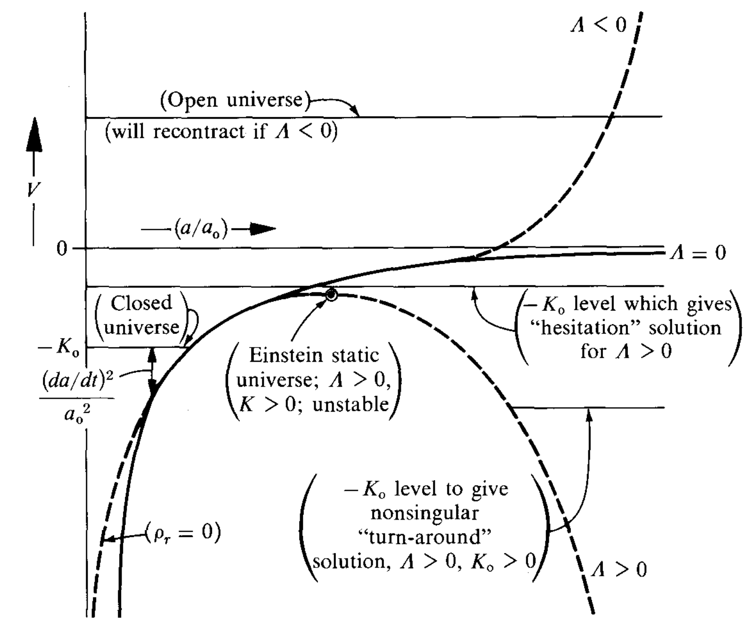
Other non-Einsteinian cosmologies:

(1) steady-state model

(*continued on page 748*)

Box 27.5 EFFECT OF VALUE OF COSMOLOGICAL CONSTANT AND OF INTRINSIC CURVATURE OF MODEL UNIVERSE "TODAY" ON THE PREDICTED COURSE OF COSMOLOGY

The "effective potential" V in the generalized Friedmann equation (27.74) is represented schematically here as a function of the expansion ratio a/a_0. The diagram illustrates the influence on the cosmology of (1) the cosmological constant Λ (determines the behavior of the effective potential at large values of a/a_0; see dashed curves) and (2) the value adopted for K_0 = (one sixth of the intrinsic curvature of 3-space at the present epoch). The value of the quantity $-K_0$ determines the "effective energy level" and is shown in the diagram as a horizontal line. The difference between this horizontal line and the effective potential determines $(a_0^{-1}\, da/dt)^2$. Regions where this difference is negative are inaccessible. From the diagram one can read off the histories of 3-space on the facing page.



The diagram is schematic, not quantitative. Representative values might be $\Lambda_{\text{conv}} = 0$ or $\pm 3 \times 10^{-28}\,\text{g/cm}^3$; $\rho_{mo,\text{conv}} = 10^{-30}\,\text{g/cm}^3$ or $\rho_{mo,\text{conv}} = 10^{-28}\,\text{g/cm}^3$; and $(a_0^{-1}\, da/dt)^2 = H_0^2 = (1/20 \times 10^9\,\text{yr})^2$ or $3.8 \times 10^{-29}\,\text{g/cm}^3$. At small values of a/a_0 the cosmological term $-(\Lambda/3)(a/a_0)^2$ is negligible. Not negligible at small values of (a/a_0) is the difference between a model universe curved only by the density of matter (the dashed curve in the diagram) and one curved also by a density of radiation (the full curve). The different dependence of "radius" and density on time at early times in these two cases of a matter-dominated cosmology and a radiation-dominated cosmology is spelled out in the last part of Box 27.3, giving in the one case $\rho = 1/6\pi t^2$ and in the other $\rho = 3/32\pi t^2$.

Intrinsic curvature of space today	Λ	*Cosmology*
Hyperbolic; K_o negative	negative	Universe starts in a condition of infinite density. It expands to a maximum extent (or minimum density) governed by the value of Λ. It then recontracts at an ever increasing rate to a condition of infinite density.
Hyperbolic; K_o negative	zero	Universe starts in a condition of infinite density. It expands. Ultimately the rate of expansion reaches a steady rate, $da/dt = 1$. The 4-geometry is Minkowski flat spacetime. Only the curvature of the spacelike slices taken through this flat 4-geometry gives the 3-geometry its hyperbolic character [see equation (27.70)].
Closed; K_o positive	zero	Standard Friedmann cosmology: expansion from infinite compaction to a finite radius and recontraction and collapse.
Closed; K_o positive	negative	Qualitatively same as foregoing. Quantitatively a slightly smaller radius at the phase of maximum expansion and a slightly shorter time from start to end.
Closed; K_o positive	Λ more positive than a certain critical value: $\Lambda > \Lambda_{\text{crit}}$	"Summit" of "effective potential" is reduced to a value slightly less than $-K_o$. The closed universe once again starts its expansion from a condition of infinite compaction. This expansion once again slows down as the expansion proceeds and then looks almost as if it is going to stop ("moment of hesitation"). However, the representative point slowly passes over the summit of the potential. Thereafter the expansion gathers more and more speed. It eventually follows the exponential law $$a = \text{constant} \times \exp\,[(\Lambda/3)^{1/2}t].$$
Closed; K_o positive	Λ positive and exactly equal to the critical value, $\Lambda = \Lambda_{\text{crit}}$, that puts the "summit of the potential" into coincidence with $-K_o$	Situation similar to that of a pencil with its tip dug into the table and provided with just enough energy to rise asymptotically in infinite time to the vertical position. Universe starts from a compact configuration and expanding approaches a certain radius ("Einstein radius", a_E) according to a law of the form $$a = a_E - \text{constant} \times \exp\,(-\alpha t).$$ Or (Einstein's original proposal, when he thought that the universe is static, and added the "cosmological term," against his will, to the field equation to permit a static universe) the representative point sits forever at the "summit of the effective potential" (Einstein universe). Aside from contradicting present-day evidence on expansion, this configuration has the same instability as does a pencil trying to stand on its tip. The least disturbance will cause it to "fall" either way, toward collapse or toward accelerating expansion, in the expansion case ultimately approaching the law $$a = \text{constant} \times \exp\,[(\Lambda/3)^{1/2}t].$$
Closed; K_o positive	Λ less positive than the critical value: $0 < \Lambda < \Lambda_{\text{crit}}(K_o)$	Motion on the large a side of the "potential barrier." Far back in the past the model universe has enormous dimensions, but is also contracting with enormous rapidity, in approximate accord with the formula $$a = \text{constant} \times \exp\,[-(\Lambda/3)^{1/2}t].$$ The radius a reaches a minimum value and thereafter the universe reexpands ("turn-around solution"), ultimately approaching the asymptotic law $$a = \text{constant} \times \exp\,[(\Lambda/3)^{1/2}t].$$

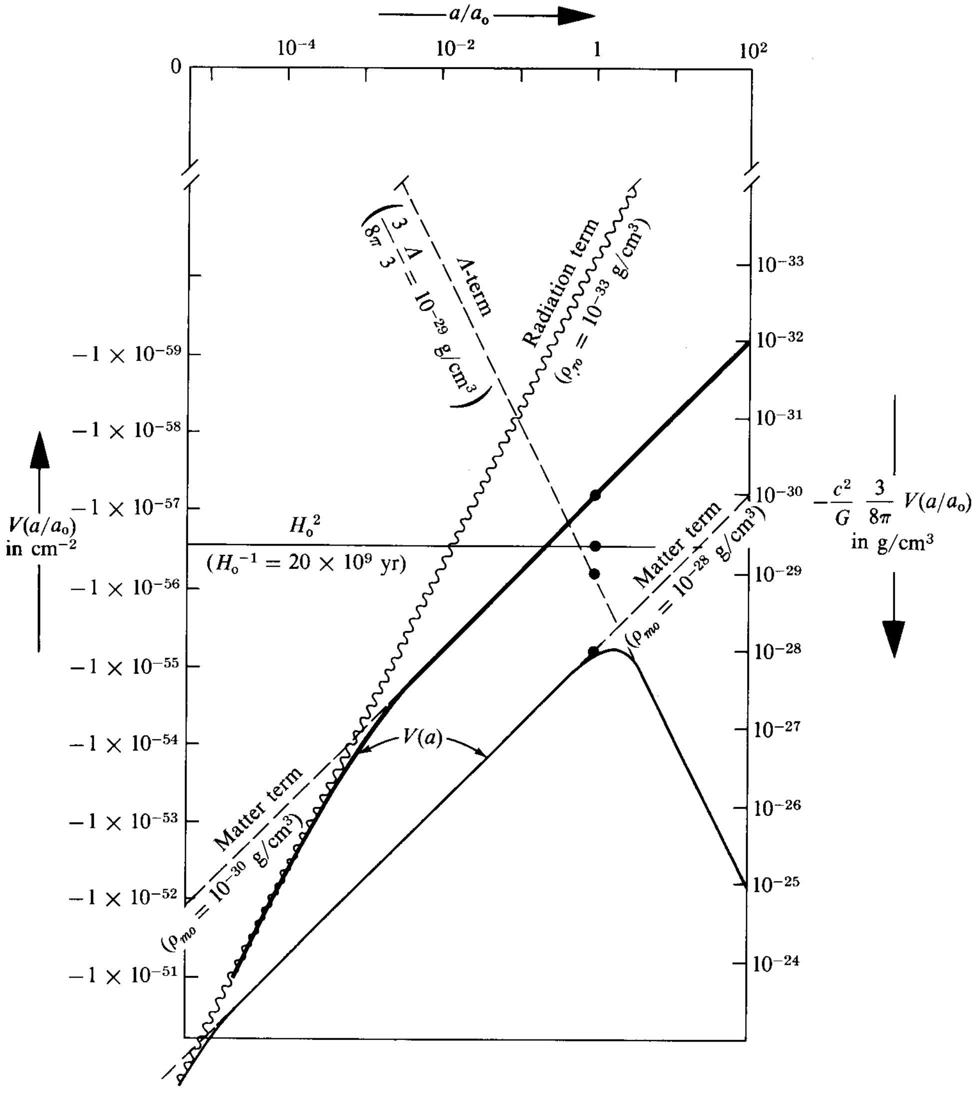
Figure 27.5. ***(facing page)***
Log-log plot of the effective potential $V(a)$ of equation (27.75) and Box 27.5 as it enters the generalized Friedmann equation

$$\left[\frac{d}{dt}\left(\frac{a(t)}{a_o}\right)\right]^2 + V\left(\frac{a}{a_o}\right) = -\frac{k}{a_o^2} = -K_o.$$

Horizontally is given the expansion ratio referred to $(a/a_o)_{\text{today}} = 1$ as standard of reference. Vertically is given the value of $V(a/a_o)$ in the geometric units of cm^{-2}. The supplementary scale at the right translates to $-(c^2/G)(3/8\pi)V(a/a_o)$ as an equivalent density, expressed in g/cm^3. The contribution of radiation density to the effective potential is indicated by the wavy line in the diagram. It is normalized to a value of the radiation density today of $\rho_{ro} = 10^{-33}\ \text{g/cm}^3$ and has a logarithmic slope of two. The contribution of matter density to the "effective potential" has a logarithmic slope of unity. Two choices are made for it, corresponding to a density of matter today of $\rho_{mo} = 10^{-30}\ \text{g/cm}^3$ and $\rho_{mo} = 10^{-28}\ \text{g/cm}^3$ (dashed lines in the diagram). The total effective potential in the two cases is also indicated in the diagram: a heavy line for the case $\rho_{mo} = 10^{-30}\ \text{g/cm}^3$ (no cosmological term included) and a light line for the case $\rho_{mo} = 10^{-28}\ \text{g/cm}^3$. In this second case, a cosmological term is included, with the cosmological constant given by $(3/8\pi)(\Lambda/3) = 10^{-29}\ \text{g/cm}^3$. The line describing the contribution of this term has a negative slope of magnitude two (dashed line). The horizontal or "level line" is drawn for a value of the Hubble expansion rate today, H_o, equal to $1/(20 \times 10^9$ years). The vertical separation on the log plot between the potential curve and the level line gives the ratio $-V/H_o^2$. This ratio as evaluated at any time t has the value $\dot{a}^2(t)/\dot{a}_o^2 + K_oH_o^{-2}$, where $\dot{a} \equiv da/dt$. As evaluated "today" $(a/a_o = 1)$ this ratio has the value $1 + K_oH_o^{-2}$. Knowing the Hubble expansion rate H_o^2 today, and knowing (or trying a certain set of parameters for) the potential curve, one can therefore deduce from the spread between the two the value of $1 + K_oH_o^{-2}$, hence the value of $K_oH_o^{-2}$, hence the present value, K_o, of the curvature factor. As an example, consider the case of the low-density universe (heavy line) and read off "today's" value, $1 + K_oH_o^{-2} = 0.223$. From this follows $K_o = -0.777\ H_o^2$ (open or hyperbolic universe), hence $k = -1$ and $a_o = (k/K_o)^{1/2} = (1/0.777)^{1/2}\ 20 \times 10^9\ \text{yr} = 22.7 \times 10^9$ yr. For the high-density model universe, with $\rho_{mo} = 10^{-28}\ \text{g/cm}^3$, one similarly finds $1 + K_oH_o^{-2} = 24.5$, hence $K_o = +23.5H_o^2$, hence $k = +1$ (closed universe) and $a_o = (k/K_o)^{1/2} = (1/23.5)^{1/2}\ 20 \times 10^9\ \text{yr} = 4.12 \times 10^9$ yr. Expansion stops, if and when it stops, at that stage when the ratio $-V/H_o^2$ between the potential and the level line, or $\dot{a}^2(t)/\dot{a}_o^2 + K_oH_o^{-2}$, falls from its "present value" of $1 + K_oH_o^{-2}$ to $0 + K_oH_o^{-2}$; that is, from 0.223 to -0.777 in the one case, and from 24.5 to 23.5 in the other case. This log-log plot should be replaced by the linear plot of Box 27.5 when $\Lambda < 0$.

(2) hierarchic model

However great a departure it is from Einstein's concept of cosmology to give any heed to a cosmological constant or an open universe, it is a still greater departure to contemplate a "hierarchic model" of the universe, in which clusters of galaxies, and clusters of clusters of galaxies, in this part of the universe are envisaged to grade off in density with distance, with space at great distances becoming asymptotically flat [Alfvén and Klein (1962), Alfvén (1971), Klein (1971), Moritz (1969), de Vaucouleurs (1971), Steigman (1971)]. The viewpoint adopted here is expressed by Oskar Klein in these words, "Einstein's cosmology was adapted to the discovery by Hubble that the observed part is expanding; the so-called cosmological postulate has been used as a kind of an axiomatic background which, when analyzed, makes it appear that this expansion is shared by a very big, but still bounded system. This implies that our expanding metagalaxy is probably just one of a type of stellar objects in different phases of evolution, some expanding and some contracting."



The contrast between the hierarchic cosmology and Einstein's cosmology [Einstein (1931) advocates a closed Friedmann cosmology] appears nowhere more strongly than here, that the one regards asymptotically flat spacetime as a requirement; the other, as an absurdity. "Only the genius of Riemann, solitary and uncomprehended," Einstein (1934) puts it, "had already won its way by the middle of the last century to a new conception of space, in which space was deprived of its rigidity, and in which its power to take part in physical events was recognized as possible." That statement epitomizes cosmology today.

But today's view of cosmology, as dominated by Einstein's boundary condition of closure ($k = +1$) and his belief in $\Lambda = 0$, need not be accepted on faith forever. Einstein's predictions are clear and definite. They expose themselves to destruction. Observational cosmology will ultimately confirm or destroy them, as decisively as it has already destroyed the 1920 belief in a static universe and the 1948 steady-state models (see Box 27.7 on the history of cosmology).

EXERCISES

Exercise 27.11. ON SEEING THE BACK OF ONE'S HEAD

Can a being at rest relative to the "cosmological fluid" ever see the back of his head by means of photons that travel all the way around a closed model universe that obeys the Friedmann cosmology and has a non-zero cosmological constant (see the entries in Box 27.3 for the case of a zero cosmological constant)?

Exercise 27.12. DO THE CONSERVATION LAWS FORBID THE PRODUCTION OF PARTICLE-ANTIPARTICLE PAIRS OUT OF EMPTY SPACE BY TIDAL GRAVITATION FORCES?

Find out what is wrong with the following argument: "The classical equations

$$G_{\alpha\beta} = 8\pi T_{\alpha\beta}$$

are not compatible with the production of pairs, since they lead to the identity $T_\alpha{}^\beta{}_{;\beta} \equiv 0$. Let the initial state be vacuum, and let $T_{\alpha\beta}$ and its derivative be equal to zero on the hypersurface $t = \text{const}$ or $t = -\infty$. It then follows from $T_\alpha{}^\beta{}_{;\beta} = 0$, that the vacuum is always conserved." [*Answer:* See Zel'dovich (1970, 1971, 1972). Also see §30.8.]

Exercise 27.13. TURN-AROUND UNIVERSE MODEL NEGLECTING MATTER DENSITY

If turn-around (minimum radius) occurs far to the right (large a) of the maximum of the potential $V(a)$ in equation (27.75), the matter terms will be negligible. Let $\rho_{mo} = \rho_{ro} = 0$. Then (what signs of k, Λ are needed for turn-around?), solve to show that $\Lambda = 3(a_{\min})^{-2}$, $H = (a_{\min})^{-1} \tanh(t/a_{\min})$ near turnaround ($t = 0$) and that the deceleration parameter $q \equiv -(1/H^2a)(d^2a/dt^2)$ has the value

$$q = -a^2(a^2 - a_{\min}^2)^{-1} < -1.$$

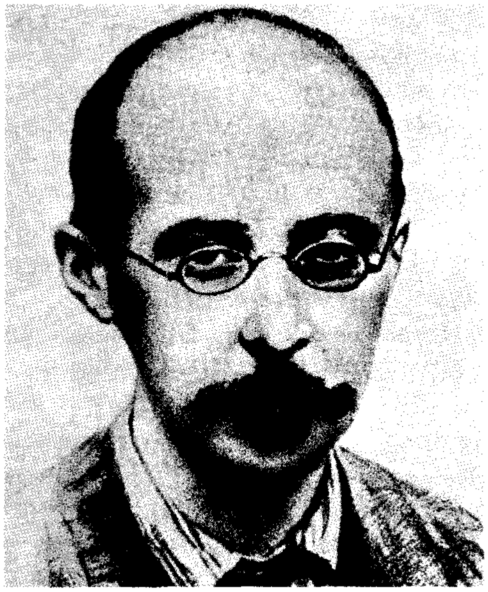
Exercise 27.14. "HESITATION" UNIVERSE

Neglect radiation in equation (27.75) but assume K_o and Λ to be chosen so that the universe spent a very long time with $a(t)$ near a_h (a_h measures location of highest point of the barrier, or the size of the universe at which the universe is most sluggish). Choose $a_h = a_o/3$ to produce an abnormally great number of quasar redshifts near $z = 2$ [as Burbidge and Burbidge (1969a,b) believe to be the case, though their opinion is not shared by all observers]. Show (a) that the density of matter now would account for only 10 per cent of the value of $H_o^2 = (\dot{a}/a)^2_{\text{now}}$ in equation (27.75) ["missing matter", i.e., K_o and Λ terms, account for 90 per cent], (b) that $a_h \simeq 20^{1/2}H_o^{-1}$, and (c) that the deceleration parameter defined in the previous exercise, as evaluated "today", has the value $q_o = -13/10$.

Exercise 27.15. UNIVERSE OPAQUE TO BLACK-BODY RADIATION AT A NONSINGULAR PAST TURN-AROUND REQUIRES IMPOSSIBLE PARAMETERS

From a plot like that in Box 27.5, construct a model of the universe that contains 2.7 K black-body radiation at the present, but, with $k = +1$ and $\Lambda > 0$, had a past turn-around (minimum radius) at which the blackbody temperature reached 3,000 K where hydrogen would be ionized. Try to use values of H_o^{-1} and ρ_{mo} that are as little as possible smaller than presently accepted values.

Box 27.6 ALEXANDER ALEXANDROVITCH FRIEDMANN
St. Petersburg, June 17, 1888—Leningrad, September 15, 1925



Graduated from St. Petersburg University, 1909; doctorate, 1922; 1910, mathematical assistant in the Institute of Bridges and Roads; 1912, lecturer on differential calculus in the Institute of Mines; 1913, physicist in the Aerological Institute of Pavlov; dirigible ascent in preparation for observing eclipse of the sun of August 1914; volunteer in air corps on war front near Osovets, 1914; head of military air navigation service, 1916–1917; professor of mechanics at Perm University, 1918; St. Petersburg University, 1920; lectures in hydrodynamics, tensor analysis; author of books, *Experiments in the Hydromechanics of Compressible Liquids* and *The World as Space and Time,* and the path-breaking paper, *On the Curvature of Space,* 1922; a director of researches in the department of theoretical meteorology of the Main Geophysical Laboratory, Leningrad, and, from February 1925 until his death of typhoid fever seven months later, director of that Laboratory; with L. V. Keller "introduced the concept of coupling moments, i.e., mathematical expectation values for the products of pulsations of hydrodynamic elements at different points and at different instants . . . to elucidate the physical structure of turbulence" [condensed from Polubarinova-Kochina (1963), which also contains a bibliography of items by and about Friedmann].

Box 27.7 SOME STEPS IN COSMOLOGY ON THE WAY TO WIDER PERSPECTIVES AND FIRMER FOUNDATIONS [For general reference on the history of cosmology, see among others Munitz (1957), Nasr (1964), North (1965), Peebles (1971), Rindler (1969), and Sciama (1971); and especially see Peebles and Sciama for bibliographical references to modern developments listed below in abbreviated form.]

A. Before the Twentieth Century

Concepts of very early Indian cosmology [summarized by Zimmer (1946)]: "One thousand mahāyugas—4,320,000,000 years of human reckoning—constitute a single day of Brahmā, a single kalpa. . . . I have known the dreadful dissolution of the universe. I have seen all perish, again and again, at every cycle. At that terrible time, every single atom dissolves into the primal, pure waters of eternity, whence all originally arose."

Plato, ca. 428 to ca. 348 B.C. [from the *Timaeus,* written late in his life, as translated by Cornford (1937)]: "The world [universe] has been fashioned on the model of that which is comprehensible by rational discourse and understanding and is always in the same state. . . . this world came to be . . . a living creature with soul and reason. . . . its maker did not make two worlds nor yet an indefinite number; but this Heaven has come to be and is and shall be hereafter one and unique. . . . he fashioned it complete and free from age and sickness. . . . he turned its shape rounded and spherical . . . It had no need of eyes, for nothing visible was left outside; nor of hearing, for there was nothing outside to be heard. . . . in order that Time might be brought into being, Sun and Moon and five other stars—'wanderers,' as they are called—were made to define and preserve the numbers of Time. . . . the generation of this universe was a mixed result of the combination of Necessity and Reason . . . we must also bring in the Errant Cause. . . . that which is to receive in itself all kinds [all forms] must be free from all characters [all form] . . . . For this reason, then, the mother and Receptacle of what has come to be visible and otherwise sensible must not be called earth or air or fire or water . . . but a nature invisible and characterless, all-receiving, partaking in some very puzzling way of the intelligible, and very hard to apprehend."

Aristotle, 384–322 B.C. [from *On the Heavens,* as translated by Guthrie (1939)]: "Throughout all past time, according to the records handed down from generation to generation, we find no trace of change either in the whole of the outermost heaven or in any one of its proper parts. . . . the shape of the heaven must be spherical. . . . From these considerations [motion invariably in a straight line toward the center; regularity of rising and setting of stars; natural motion of earth toward the center of the universe] it is clear that the earth does not move, neither does it lie anywhere but at the center. . . . the earth . . . must have grown in the form of a sphere. This [shape of segments cut out of moon at time of eclipse of moon; and ability to see in Egypt stars not visible in more northerly lands] proves both that the earth is spherical and that its periphery is not large . . . Mathematicians who try to calculate the circumference put it at 400,000 stades [1 stade = 600 Greek feet = 606 English feet; thus 24.24×10^7 ft/(6080.2 ft/nautical mile) = 39,900 nautical miles—the oldest recorded calculation of the earth's circumference, and reportedly known to Columbus—85 per cent more than the true circumference, $60 \times 360 = 21{,}600$ nautical miles]."

Aristotle [from the *Metaphysics,* as translated by Warrington (1956)]: "Euxodus [of Cnidos, 408–355 B.C.] supposed that the motion of the sun and moon involves, in each case, three spheres. . . . He further assumed that the motion of the planets involves, in each case, four spheres. . . . Calippus [of Cyzicus, flourished 330 B.C.] . . . considered that, in the light of observation, two more spheres should be added to the sun, two to the moon, and one more to each of the other planets."

Eratosthenes, 276–194 B.C. [a calculation attributed to him by Claudius Ptolemy, who observed at Alexandria from 127 to 141 or 151 A.D., in his *Almagest,* I, §12; see the translation by Taliaferro (1952)]:

$$\text{(Maximum distance of moon from earth)} = (64\tfrac{1}{6})\ \text{(radius of earth)};$$

$$\text{(Minimum distance of sun from earth)} = (1{,}160)\ \text{(radius of earth)}.$$

Abū 'Alī al-Husain ibn 'Abdallāh ibn Sīnā, otherwise known as **Avicenna**, 980–1037; physician, philosopher, codifier of Aristotle, and one of the most influential of those who preserved Greek learning and thereby made possible its transmission to mediaeval Europe [quoted in Nasr (1964), p. 225]: "Time is the measure of motion."

From the *Rasā'il,* a 51-treatise encyclopedia, sometimes known as the *Koran after the Koran,* of the **Ikhwān al-Safā'** or Brothers of Sincerity, whose main center was at Basra, Iraq, roughly A.D. 950–1000 [as quoted by Nasr (1964), p. 64; see p. 78 for a list of distances to the planets (in terms of Earth radii) taken from the *Rasā'il,* as well as sizes of planets and the motions of rotation of the various Ptolemaic carrier-spheres]: Space is "a form abstracted from matter and existing only in the consciousness."

Abū Raihān al-Bīrūnī, 973–1030, a scholar, but concerned also with experiment, observation, and measurement, who calculated the circumference of the Earth from measurements he made in India as 80,780,039 cubits (about 4 per cent larger than the value accepted today), and gave a table of distances to the planets [as quoted in Nasr (1964), pp. 120 and 130]: "Both [kinds of eclipses] do not happen together except at the time of the total collapse of the universe."

Étiene Tempier, Bishop of Paris, in 1277, to settle a controversy then dividing much of the French theological community, ruled that one could not deny the power of God to create as many universes as He pleases.

Roger Bacon, 1214–1294, in his *Opus Majus* (1268), gave the diameter of the sphere that carries the stars, on the authority of Alfargani, as 130,715,000 Roman miles

Box 27.7 (*continued*)

[mile equal to 1,000 settings down of the right foot]; the volume of the sun, 170 times that of the Earth; first-magnitude star, 107 times; sixth-magnitude, 18 times Earth.

Nicolas Cusanus, 1401–1464 [from *Of Learned Ignorance* (1440), as translated by Heron (1954)]: "Necessarily all parts of the heavens are in movement. . . . It is evident from the foregoing that the Earth is in movement . . . the world [universe], its movement and form . . . will appear as a wheel in a wheel, a sphere in a sphere without a center or circumference anywhere. . . . It is now evident that this Earth really moves, though to us it seems stationary. In fact, it is only by reference to something fixed that we detect the movement of anything. How would a person know that a ship was in movement, if . . . the banks were invisible to him and he was ignorant of the fact that water flows?"

Nicolaus Copernicus, February 19, 1473, to May 24, 1543 [from *De Revolutionibus Orbium Coelestrum* (1543), as translated by Dobson and Brodetsky (1947)]: "I was induced to think of a method of computing the motions of the spheres by nothing less than the knowledge that the mathematicians are inconsistent in these investigations. . . . they cannot explain or observe the constant length of the seasonal year. . . . some use only concentric circles, while others eccentrics and epicycles. . . . Nor have they been able thereby to discern or deduce the principal thing—namely the shape of the universe and the unchangeable symmetry of its parts. . . .

"I found first in Cicero that Nicetas had realized that the Earth moved. Afterwards I found in Plutarch [~A.D. 46–120] . . . 'The rest hold the Earth to be stationary, but Philolaus the Pythagorean [born ~480 B.C.] says that she moves around the (central) fire on an oblique circle like the Sun and Moon. Heraclides of Pontus [flourished in 4th century B.C.] and Ecphantus the Pythagorean also make the Earth to move, not indeed through space but by rotating round her own center as a wheel on an axle from West to East.' Taking advantage of this I too began to think of the mobility of the Earth. . . .

"Should we not be more surprised if the vast Universe revolved in twenty-four hours, than that little Earth should do so? . . . Idle therefore is the fear of Ptolemy that Earth and all thereon would be disintegrated by a natural rotation. . . . That the Earth is not the center of all revolutions is proved by the apparently irregular motions of the planets and the variations in their distances from the Earth. . . . We therefore assert that the center of the Earth, carrying the Moon's path, passes in a great orbit among the other planets in an annual revolution round the Sun; that near the Sun is the center of the Universe; and that whereas the Sun is at rest, any apparent motion of the Sun can be better explained by motion of the Earth. . . . Particularly Mars, when he shines all night, appears to rival Jupiter in magnitude, being distinguishable only by his ruddy color; otherwise he is scarce equal to a star of the second magnitude, and can be recognized only when his movements are carefully followed. All these phenomena proceed from the same cause, namely Earth's motion. . . . That there are no such phenomena for the fixed stars proves their immeasurable distance, compared to which even the size of the Earth's orbit is negligible and the parallactic effect unnoticeable."

Thomas Digges, 1546–1595 [in *a Perfit Description of the Caelestiall Orbes according to the most aunciente doctrine of the Pythagoreans, latelye reuiued by Copernicus and by Geometricall Demonstrations approued* (1576), the principal vehicle by which Copernicus reached England, as quoted in Johnson (1937)]: "Of whiche lightes Celestiall it is to bee thoughte that we onely behoulde sutch as are in the inferioure partes of the same Orbe, and as they are hygher, so seeme they of lesse and lesser quantity, euen tyll our sighte beinge not able farder to reach or conceyue, the greatest part rest by reason of their wonderfull distance inuisible vnto vs."

Giordano Bruno, born ca. 1548, burned at the stake in the Campo dei Fiori in Rome, February 17, 1600 [from *On the Infinite Universe and Worlds,* written on a visit to England in 1583–1585, as translated by Singer (1950)]: "Thus let this surface be what it will, I must always put the question, what is beyond? If the reply is NOTHING, then I call that the VOID or emptiness. And such a Void or Emptiness hath no measure and no outer limit, though it hath an inner; and this is harder to imagine than is an infinite or immense universe. . . . There are then innumerable suns, and an infinite number of earths revolve around those suns, just as the seven we can observe revolve around this sun which is close to us."

Johann Kepler established the laws of elliptic orbits and of equal areas (1609), and established the connection between planetary periods and semimajor axes (1619).

Galileo Galilei observed the satellites of Jupiter and realized they provided support for Copernican theory, and interpreted the Milky Way as a collection of stars (1610). In 1638 he wrote:

"*Salvati.* Now what shall we do, Simplicio, with the fixed stars? Do we want to sprinkle them through the immense abyss of the universe, at various distances from any predetermined point, or place them on a spherical surface extending around a center of their own so that each of them will be at the same distance from that center?

"*Simplicio.* I had rather take a middle course, and assign to them an orb described around a definite center and included between two spherical surfaces . . ."

Isaac Newton (1687): "Gravitation toward the sun is made up out of the gravitations toward the several particles of which the body of the sun is composed, and in receding from the sun decreases accurately as the inverse square of the distances as far as the orbit of Saturn, as evidently appears from the quiescence of the aphelion of the planets."

Isaac Newton [in a letter of Dec. 10, 1692, to Richard Bentley, quoted in Munitz (1957)]: "If the matter of our sun and planets and all the matter of the universe were evenly scattered throughout all the heavens, and every particle had an innate gravity toward all the rest, and the whole space throughout which this matter was

Box 27.7 *(continued)*

scattered was but finite, the matter on the outside of this space would, by its gravity, tend toward all the matter on the inside and, by consequence, fall down into the middle of the whole space and there compose one great spherical mass. But if the matter was evenly disposed throughout an infinite space, it could never convene into one mass; but some of it would convene into one mass and some into another, so as to make an infinite number of great masses scattered at great distances from one to another throughout all that infinite space. And thus might the sun and fixed stars be formed."

Christiaan Huygens, 1629–1695 [in his posthumously published *Cosmotheoros* (1698)]: "Seeing then that the stars . . . are so many suns, if we do but suppose one of them [Sirius, the Dog-star] equal to ours, it will follow [details, including telescope directed at sun; thin plate; hole in it; comparison with Sirius] . . . that his distance to the distance of the sun from us is as 27,664 to 1. . . . Indeed it seems to me certain that the universe is infinitely extended."

Edmund Halley (1720): "If the number of the Fixt Stars were more than finite, the whole superficies of their apparent Sphere [i.e., the sky] would be luminous" [by today's reasoning the same temperature as the surface of the average star; this is known today as Olber's paradox, or the paradox of P. L. de Chéseaux (1744) and Heinrich Wilhelm Matthias Olbers (1826)].

Thomas Wright of Durham (1750): "To . . . solve the Phaenomena of the Via Lactea . . . granted . . . that the *Milky Way* is formed of an infinite number of small Stars . . . imagine a vast infinite gulph, or medium, every way extended like a plane, and inclosed between two surfaces, nearly even on both sides. . . . Now in this space let us imagine all the Stars scattered promiscuously, but at such an adjusted distance from one another, as to fill up the whole medium with a kind of regular irregularity of objects. [Considering its appearance] "to an eye situated . . . anywhere about the middle plane" . . . all the irregularity we observe in it at the Earth, I judge to be entirely owing to our Sun's position . . . and the diversity of motion . . . amongst the stars themselves, which may here and there . . . occasion a cloudy knot of stars."

Immanuel Kant, 1724–1804 (1755): "It was reserved for an Englishman, Mr. Wright of Durham, to make a happy step . . . we will try to discover the cause that has made the positions of the fixed stars come to be in relation to a common plane. . . . granted . . . that the whole host of [the fixed stars] are striving to approach each other through their mutual attraction . . . ruin is prevented by the action of the centrifugal forces . . . the same cause [centrifugal force] . . . has also so directed their orbits that they are all related to one plane. . . . [The needed motion is calculated to be] one degree [or less] in four thousand years; . . . careful observers . . . will be required for it. . . . Mr. Bradley has observed almost imperceptible displacements of the stars" [known from later work to be caused by aberration (effect of observer velocity) rather than real parallax (effect of position of observer)].

Asks for the first time how a very remote galaxy would appear: "circular if its plane is presented directly to the eye, and elliptical if it is seen from the side or obliquely. The feebleness of its light, its figure, and the apparent size of its diameter will clearly distinguish such a phenomenon when it is presented, from all the stars that are seen single. . . . this phenomenon . . . has been distinctly perceived by different observers [who] . . . have been astonished at its strangeness. . . . Analogy thus does not leave us to doubt that these systems [planets, stars, galaxies] have been formed and produced . . . out of the smallest particles of the elementary matter that filled empty space."

Goes on to consider seriously "the successive expansion of the creation [of planets, stars, galaxies] through the infinite regions of space that have the matter for it. . . . attraction is just that universal relation which unites all the parts of nature in one space. It reaches, therefore, to . . . all the distance of nature's infinitude."

Johann Heinrich Lambert, 1728–1777 (1761): "The fixed stars obeying central forces move in orbits. The Milky Way comprehends several systems of fixed stars. . . . Each system has its center, and several systems taken together have a common center. Assemblages of their assemblages likewise have theirs. In fine, there is a universal center for the whole world round which all things revolve." [First spelling out of a "hierarchical model" for the universe, later taken up by C. V. I. Charlier and by H. Alfvén and O. Klein (1962); see also O. Klein (1966 and 1971)].

Auguste Comte (1835) concluded that it is meaningless to speak of the chemical composition of distant stars because man will never be able to explore them; "the field of positive philosophy lies wholly within the limits of our solar system, the study of the universe being inaccessible in any positive sense."

The first successful determination of the parallax [1 second of parallax: 1 pc $=$ 3.08×10^{18} cm $= 3.26$ lyr] of any star was made in 1838 (for α Centauri by Henderson, for α Lyrae by Struve, and for 61 Cygni by Bessel).

B. The Twentieth Century

Derivation by James Jeans in 1902 of the critical wavelength that separates short-wavelength acoustical modes of vibration of a hot primordial gas and longer wavelength modes of commencement of gravitational condensation of this gas. Application of these considerations by P. J. E. Peebles and R. H. Dicke in 1968 to explain why globular star clusters have masses of the order of $10^5 M_\odot$.

Investigations of cosmic rays from first observation by V. F. Hess and W. Kolhörster in 1911–1913 to date; determination that the energy density in interstellar space (in this galaxy) is about 1 eV/cm^3 or 10^{-12} erg/cm^3, comparable to the density of energy of starlight, to the kinetic energy of clouds of ionized interstellar gas, averaged over the galaxy, and to the energy density of the interstellar

Box 27.7 *(continued)*

magnetic field ($\sim 10^{-5}$ gauss). In connection with this equality, see especially E. Fermi (1949).

Discovery by Henrietta Leavitt in 1912 that there is a well-defined relation between the period of a Cepheid variable and its absolute luminosity.

First determination of the radial velocity of a galaxy by V. M. Slipher in 1912: Andromeda approaching at 200 km/sec. Thirteen galaxies investigated by him by 1915; all but two receding at roughly 300 km/sec.

Albert Einstein (1915d): Interpreted gravitation as a manifestation of geometry; gave final formulation of the law that governs the dynamic development of the geometry of space with the passage of time.

Albert Einstein (1917): Idealized the universe as a 3-sphere filled with matter at effectively uniform density; the radius of this 3-sphere could not be envisaged as static without altering his standard 1915 geometrodynamic law; for this reason Einstein introduced a so-called "cosmological term," which he later dropped as "the biggest blunder" in his life [Gamow (1970)].

Formulation by W. de Sitter in 1917 of a cosmological model in which (1) the universe is everywhere isotropic (and therefore homogeneous) and (2) the universe does not change with time, so that the mean density of mass-energy and the mean curvature of space are constant, but in which perforce (3) a cosmological term ("repulsion") of the Einstein type has to be added to balance the attraction of the matter. Observation by de Sitter that he could obtain another static model by removing all the matter from the original model, but that the Λ-term would cause test particles to accelerate away from one another.

From 1917 to 1920, debate about whether spiral nebulae are mere nebulous objects (Harlow Shapley) or are "island universes" or galaxies similar to but external to the Milky Way (H. D. Curtis).

Discovery by Harlow Shapley in 1918, by mapping distribution of about 100 globular clusters of this Galaxy (10^4 to 10^6 stars each) in space that center is in direction of Sagittarius (present value of distance from sun ~ 10 kpc).

Independent derivation of evolving homogeneous and isotropic cosmological models [also leading to the relation $v = H \cdot$ (distance)] by A. Friedmann in 1922 and G. Lemaître in 1927, with Lemaître tieing in his theoretical analysis with the then-ongoing Mt. Wilson work, to become the "father of the big-bang cosmology". (Universe, however, taken to expand smoothly away from Einstein's static $\Lambda > 0$ solution in Lemaître's original paper).

Remark by H. Weyl in 1923 that test particles in de Sitter model will separate at a rate given by a formula of the form $v = H \cdot$ (distance).

In 1924, resolution of debate about nature of spiral nebulae by Edwin P. Hubble with Mount Wilson 100-inch telescope; discovery of Cepheid variables in Andromeda and other spiral nebulae, and consequent determination of distances to these nebulae.

Determination by Jan Oort in 1927 of characteristic pattern of radial velocities of stars near sun,

$$\delta v_r = Ar \cos 2(\theta - \delta),$$

showing that: (1) axis of rotation of stars in Milky Way is perpendicular to disc; (2) sun makes a complete revolution in $\sim 10^8$ yr; and (3) the effective mass pulling on the sun required to produce a revolution with this period is of the order $\sim 10^{44}$ g or $\sim 10^{11} M_\odot$.

Age of a *uranium ore* as established from lead-uranium ratio: greatest value found up to 1927, 1.3×10^9 yr (A. Holmes and R. W. Lawson). Age of the lead in the "average" *surface rocks* of the earth as calculated from time required to produce this lead from the uranium in the same surface rocks, 2×10^9 yr to 6×10^9 yr. Age of *elemental uranium* as estimated by Rutherford from time required for U^{235} and U^{238} to decay from assumed roughly equal ratio in early days to known very unequal ratio today, $\sim 3 \times 10^9$ yr.

Establishment by Hubble in 1929 that out to 6×10^6 lyr the velocity of recession of a galaxy is proportional to its distance.

Note by A. S. Eddington in 1930 that Einstein $\Lambda > 0$ static universe is unstable against any small increase or decrease in the radius of curvature.

Recommendation from Einstein in 1931 hereafter to drop the so-called cosmological term.

Proposal by Einstein and de Sitter in 1932 that one tentatively adopt the simplest assumption that $\Lambda = 0$, that pressure is negligible, and that the reciprocal of the square of the radius of curvature of the universe is neither positive nor negative (spherical or hyperbolic universe) but zero ("cosmologically flat"), thus leading to the relation $\rho = 3H^2/8\pi$ (in geometric units).

Evidence uncovered by Grote Reber in 1934 for the existence of a discrete radio source in Cygnus; evidence for this source, Cygnus A, firmed up by J. S. Hey, S. J. Parsons, and J. W. Phillips in 1946; six other discrete radio sources, including Taurus A and Centaurus A, discovered by J. G. Bolton in 1948.

Discovery by E. A. Milne and W. H. McCrea in 1934 of close correspondence between Newtonian dynamics of a large gas cloud and Einstein theory of a dynamic universe, with the scale factor of the expansion satisfying the same equation in both theories, so long as pressure is negligible.

Demonstration by H. P. Robertson and by A. G. Walker, independently, in 1935 that the Lemaître type of line element provides the most general Riemannian geometry compatible with homogeneity and isotropy.

Classification of nebulae as spiral, barred spiral, elliptical, and irregular by Hubble in 1936.

First detailed theory of thermonuclear energy generation in the sun, H. A. Bethe, 1939.

Box 27.7 ***(continued)***

Reasoning by George Gamow in 1946 that matter in the early universe was dense enough and hot enough to undergo rapid thermonuclear reaction, and that energy densities were radiation-dominated.

Proposal of so-called "steady-state cosmology" by H. Bondi, T. Gold, and F. Hoyle in 1948, lying outside the framework of Einstein's standard general relativity, with "continuous creation of matter" taking place throughout the universe, and the mean age of the matter present being equal to one third of the Hubble time.

Prediction by R. A. Alpher, H. A. Bethe, and G. Gamow in 1948 that the blackbody radiation that originally filled the universe should today have a Planck spectrum corresponding to a temperature of 25 K. Independent conception of same idea by R. H. Dicke in 1964 and start of an experimental search for this primordial cosmic-fireball radiation. Discovery of unwanted and unexpected 7 cm background radiation in 1965 by A. A. Penzias and R. W. Wilson with a temperature of about 3.5 K; immediate identification of this radiation by Dicke, P. J. E. Peebles, P. G. Roll, and D. T. Wilkinson as the expected relict radiation.

Radio sources Taurus A, Virgo A, and Centaurus A tentatively and, as it later proved, correctly identified with the Crab Nebula and with the galaxies NGC 4486 and NGC 5128 by J. G. Bolton, G. J. Stanley, and O. B. Slee in 1949.

Analysis by Lemaître in 1950 of big-bang expansion approaching very closely the Einstein static universe ($\Lambda > 0$) and then, at first slowly, subsequently more and more rapidly, going into exponential expansion.

Discovery by Walter Baade in 1952 that there are two types of Cepheid variables with different period-luminosity relations; consequent increase in Hubble distance scale by factor of about 2.6, and a corresponding increase in the original value (roughly 2×10^9 yr) of the Hubble time, H_0^{-1}.

Identification of radio source Cygnus A by W. Baade and R. Minkowski in 1954 with the brightest member of a faint cluster of galaxies, contrary to the then widely held view that the majority of radio sources lie within the Milky Way. Determination of redshift in the optical spectrum of $\delta\lambda/\lambda = z = 0.057$ by Minkowski, implying for Cygnus A a distance of 170 Mpc and a radio luminosity of 10^{45} erg/sec, 10^7 times the radio power and ten times the optical power of a normal galaxy.

Resolution of radio source Cygnus A in 1956 into two components symmetrically located on either side of the optical galaxy, the first indication that most radio sources are double. Still unsolved is the mystery of the explosion or other mechanism that caused this and other double sources.

Calculation by G. R. Burbidge in 1956 of the kinetic energy in the electrons giving off synchrotron radiation in a radio galaxy and the energy of the magnetic field that holds these electrons in orbit; minimization of the sum of these two energies;

determination that this minimum is of the order of 10^{60} ergs (energy of annihilation of half a million suns) for Hercules A, for example.

Solar system determined to have an age of 4.55×10^9 yr or more from relative abundances of $Pb^{204,206,207}$ and $U^{235,238}$ in meteorites and oceanic sediments by C. Patterson in 1956; and by others in 1965 and 1969 from evidence on the processes $Rb^{87} \longrightarrow Sr^{87}$ and $K^{40} \longrightarrow A^{40}$ in meteorites.

Discovery by Allen Sandage in 1958 that what Hubble had identified in distant galaxies as bright stars were H II regions, clumps of hot stars surrounded by a plasma ionized by stars, and consequent upping of Hubble's distance scale by a further factor of about 2.2.

Estimation by Jan Oort in 1958, from luminosity of other galaxies, that matter in galaxies contributes to the density of mass-energy in the universe roughly 3×10^{-31} g/cm^3 [see Peebles (1971) for updated analysis], this being one or two orders of magnitude less than that called for by Einstein's concept that the universe is curved up into closure, and thereby giving rise to "the mystery of the missing matter," the focus of much present-day research.

Discovery of celestial (nonsolar) X-rays in 1962 by Giacconi, Gursky, Paolini, and Rossi. Majority of sources in plane of the Milky Way, presumably local to this galaxy, as is the Crab nebula. Extragalactic sources include the radio galaxy Virgo A and the quasar 3C273.

Revised "3C-catalog" of radio sources published in 1962 by A. S. Bennett, containing 328 sources, nearly complete in coverage between declinations $-5°$ and $+90°$ for sources brighter than 9 flux units (9×10^{-26} watt/m^{2}Hz) at 178 MHz.

Identification of the first quasistellar object (QSO) by Maarten Schmidt at Mt. Palomar in 1963: radio-position determination of 3C273 to better than 1 second of arc by C. Hazard, M. B. Mackey, and A. J. Shimmins in 1962, followed by Schmidt's taking an optical spectrum of the star-like source and, despite all presumptions that it was a star in this galaxy, trying to fit it, and succeeding, with a redshift of the magnitude (unprecedented for a "star") of $\delta\lambda/\lambda = z = 0.158$. Distance implied by Hubble relation, 1.5×10^9 lyr; optical brightness, 100 times brightest known galaxy. Largest redshift of any QSO known in 1972, $z = 2.88$ (4C05.34; C. R. Lynds). Such a source detectable even if it had a redshift of 3; but no QSO's known in 1972 with such redshifts. See Box 28.1.

Reasoning by Dennis Sciama in 1964 [see also Sciama (1971)] that intergalactic hydrogen can best escape observation if at a temperature between 3×10^5 K and 10^6 K. With as many as 10^{-5} protons and 10^{-5} electrons per cm^3 and a temperature lower than 3×10^5 K, the number density of neutral atoms would be great enough and the resulting absorption of Lyman α from a distant galaxy ($z = 2$) would be strong enough to show up, contrary to observation.

In 1964 J. E. Gunn and B. A. Peterson, E. J. Wampler, and others determined that, at a temperature greater than 10^6 K, the intensity of 0.25 keV or 50 Å x-rays

Box 27.7 (*continued*)

from intergalactic space would be too high to be compatible with the observations.

Emphasis by Wheeler (1964a) that the dynamic object in Einstein's general relativity is 3-geometry, not 4-geometry, and that this dynamics, both classical and quantum, unrolls in the arena of superspace.

Discovery by Sandage in 1965 of quasistellar galaxies (radio-quiet QSO's).

Discovery by E. M. Burbidge, G. R. Burbidge, C. R. Lynds, and A. N. Stockton in 1965 of a QSO, 3C191, with numerous absorption lines, implying the coexistence of several redshifts in one spectrum.

Fraction (by mass) of matter converted to helium in early few minutes of universe nearly independent of the relative numbers of photons and baryons, over a 10^6 range in values of this number ratio, so long as the universe at 10^{10} K is still radiation-dominated. Value of this plateau helium abundance (following earlier work of others) first accurately calculated as 27 per cent by P. J. E. Peebles in 1966 and by R. V. Wagoner, W. A. Fowler, and F. Hoyle in 1967.

Proposal by C. W. Misner in 1968 to consider as an important part of early cosmology the anisotropy vibrations of the geometry of space previously brought to attention by E. Kasner and by I. M. Khalatnikov and E. Lifshitz. [Misner's hope to account naturally in this way for the otherwise so puzzling homogeneity of the universe was later dashed.]

Proof on the basis of standard general relativity by S. W. Hawking, G. F. R. Ellis, and R. Penrose in 1968 and 1969 [see also related work of earlier investigators cited in Chapter 44] that a model universe presently expanding and filled with matter and radiation obeying a physically acceptable equation of state must have been singular in the past, however wanting in symmetry it is today.

Discovery of pulsars in 1968 by Hewish, Bell, Pilkington, Scott, and Collins, and their interpretation as spinning neutron stars (see Chapter 24).

"No poet, nor artist of any art, has his complete meaning alone. His significance, his appreciation, is the appreciation of his relation to the dead poets and artists. You cannot value him alone; you must set him, for contrast and comparison, among the dead . . . when a new work of art is created . . . something . . . happens simultaneously to all the works of art which preceded it. The existing monuments form an ideal order among themselves, which is modified by the introduction of the new (the really new) work of art among them."

T. S. ELIOT (1920).

CHAPTER **28**

EVOLUTION OF THE UNIVERSE INTO ITS PRESENT STATE

Cosmology . . . restrains the aberrations of the mere undisciplined imagination.

ALFRED NORTH WHITEHEAD (1929, p. 21)

§28.1. THE "STANDARD MODEL" OF THE UNIVERSE

Evolution of universe according to "standard hot big-bang model":

Since the discovery of the cosmic microwave radiation in 1965, extensive theoretical research has produced a fairly detailed picture of how the universe probably evolved into its present state. This picture, called the "standard hot big-bang model" of the universe, is sketched in the present chapter, and its main features appear in Figure 28.1. Gravitation dominates the over-all expansion; but otherwise most details of the evolution are governed much less by gravitation than by the laws of thermodynamics, hydrodynamics, atomic physics, nuclear physics, and high-energy physics. This fact, and the existence of three excellent recent books on the subject [Sciama (1971); Peebles (1972); Zel'dovich and Novikov (1974)], encourage brevity here.

The past evolution of the universe is qualitatively independent of the nature of the homogeneous hypersurfaces ($k = -1, 0$, or $+1$) and qualitatively independent of the cosmological constant, since the contributions of k and Λ to the evolution are not important in early stages of the history (small a/a_o) [see equation (27.40) and Figure 27.5]. One crucial assumption underlies the standard hot big-bang model: that the universe "began" in a state of rapid expansion from a very nearly homogeneous, isotropic condition of infinite (or near infinite) density and temperature.

(1) initial state

During the first second after the beginning, according to this analysis, the temperature of the universe was so high that there was complete thermodynamic equilib-

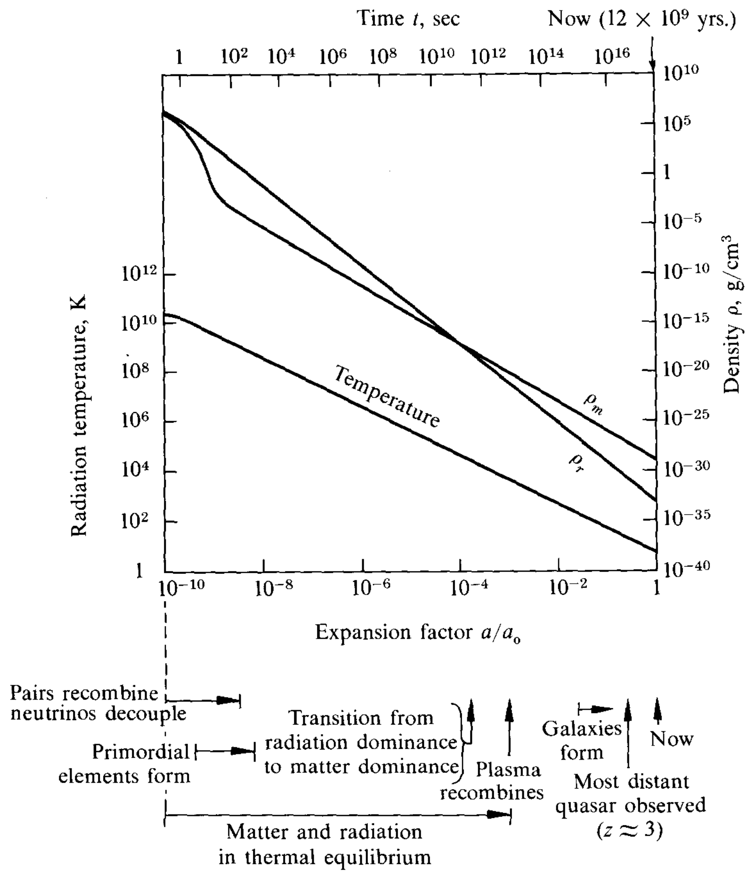


Figure 28.1.
Evolution of the universe into its present state, according to the standard hot big-bang model. The curves are drawn assuming

$$\rho_{mo} = 5 \times 10^{-30}\ \mathrm{g/cm^3}, \qquad \rho_{ro} = 10^{-33}\ \mathrm{g/cm^3}, \qquad k = 0;$$

but for other values of ρ_{mo}, ρ_{ro}, and k within the limits of observation, the curves are virtually the same (see exercise 28.1). See text and Box 28.1 for detailed discussion of the processes described at the bottom of the figure. [This figure is adapted from Dicke, Peebles, Roll, and Wilkinson (1965).]

(2) thermal equilibrium, decay of particles, recombination of pairs ($0 < t \lesssim 10$ sec.)

(3) decoupling and free propagation of gravitons and neutrinos ($t \lesssim 1$ sec.)

rium between photons, neutrinos, electrons, positrons, neutrons, protons, various hyperons and mesons, and perhaps even gravitons (gravitational waves) [see, e.g., Kundt (1971) and references cited therein]. However, by the time the universe was a few seconds old, its temperature had dropped to about 10^{10} K and its density was down to $\sim 10^5$ g/cm^3; so all nucleon-antinucleon pairs had recombined, all hyperons and mesons had decayed, and all neutrinos and gravitons had decoupled from matter. The universe then consisted of freely propagating neutrinos, and perhaps gravitons, with black-body spectra at temperatures $T \sim 10^{10}$ K, plus electron-positron pairs in the process of recombining, plus electrons, neutrons, protons, and photons all in thermal equilibrium at $T \sim 10^{10}$ K.

Since that early state, the gravitons (if present) and neutrinos have continued to propagate freely, maintaining black-body spectra; but their temperatures have been redshifted by the expansion of the universe in accordance with the law

$$T \propto 1/a \tag{28.1}$$

(Box 29.2). Consequently, today their temperatures should be roughly 3 K, and they should still fill the universe. Unfortunately, today's technology is far from being able to detect such a "sea" of neutrinos or gravitons. However, if and when they can be detected, they will provide *direct* observational information about the first one second of the life of the universe!

(4) primordial element formation (2 sec. $\lesssim t \lesssim$ 1,000 sec.)

As the universe continued to expand after the first few seconds, it entered a period lasting from $t \sim 2$ seconds to $t \sim 1{,}000$ seconds ($T \sim 10^{10}$ to $\sim 10^9$ K, $\rho \sim 10^{+5}$ to 10^{-1} g/cm^3), during which primordial element formation occurred. Before this period, there were so many high-energy protons around that they could blast apart any atomic nucleus (e.g., deuterium or tritium or He^3 or He^4) the moment it formed; after this period, the protons were too cold (had kinetic energies too low) to penetrate each others' coulomb barriers, and all the freely penetrating neutrons from the earlier, hotter stage had decayed into electrons plus protons. Only during the short, crucial period from $t \sim 2$ seconds to $t \sim 1{,}000$ seconds were conditions right for making elements. Calculations by Gamow (1948), by Alpher and Hermann (1948a,b; 1950), by Fermi and Turkevitch (1950), by Peebles (1966), and by Wagoner, Fowler, and Hoyle (1967) reveal that about 25 per cent of the baryons in the universe should have been converted into He^4 (alpha particles) during this period, and about 75 per cent should have been left as H^1 (protons). Traces of deuterium, He^3, and Li should have also been created, but essentially no heavy elements. All the heavy elements observed today must have been made later, in stars [see, e.g., Fowler (1967) or Clayton (1968)]. Current astronomical studies of the abundances of the elements give some support for these predictions; but the observational data are not yet very conclusive [see, e.g., Danziger (1968) and pp. 268–275 of Peebles (1971)].

(5) thermal interaction of matter and radiation (1,000 sec. $\lesssim t \lesssim 10^5$ years)

After primordial element formation, the matter and radiation continued to interact thermally through frequent ionization and recombination of atoms, keeping each other at the same temperature. Were the temperatures of radiation and matter not locked together, the radiation would cool more slowly than the matter (for adiabatic expansion, $T_r \propto 1/a$, but $T_m \propto 1/a^2$). Thus thermal equilibrium was maintained only by a constant transfer of energy from radiation to matter. But the heat capacity of the radiation was far greater than that of the matter. Therefore the energy transfer had a negligible effect on ρ_r, p_r, and T_r. It held up the temperature of the matter ($T_m = T_r$) without significantly lowering the temperature of the radiation. On the other hand, the total mass-energy of matter was and is dominated by rest mass. Therefore the energy transfer had negligible influence on ρ_m. [This circumstance justifies the approximation of ignoring energy transfer when passing from equation (27.31) to (27.32).]

(6) plasma recombination and transition to matter dominance ($t \sim 10^5$ yrs.)

When the falling temperature reached a few thousand degrees ($a/a_o \sim 10^{-3}$, $\rho \sim 10^{-20}$ g/cm^3, $t \sim 10^5$ years), two things of interest happened: the universe ceased to be radiation-dominated and became matter-dominated [$\rho_m = \rho_{mo}(a_o/a)^3$ came to exceed $\rho_r = \rho_{ro}(a_o/a)^4$]; and the photons ceased to be energetic enough to keep hydrogen atoms ionized, so the electrons and protons quickly recombined. That these two events were roughly coincident is a result of the specific, nearly conserved value that the entropy per baryon has in our universe:

$$s \equiv \text{entropy per baryon} \sim \frac{\text{(number of photons in universe)}}{\text{(number of baryons in universe)}} \sim 10^8.$$

Why the universe began with this value of s, rather than some other value (e.g. unity), nobody has been able to explain.

(7) subsequent propagation of photons ($t \gtrsim 10^5$ yrs.)

Recombination of the plasma at $t \sim 10^5$ years was crucial, because it brought an end to the interaction and thermal equilibrium between radiation and matter ("decoupling"). Thereafter, with very few free electrons off which to scatter, and with Rayleigh scattering off atoms and molecules unimportant, the photons propagated almost freely through space. Unless energy-releasing processes reionized the intergalactic medium sometime between $a/a_0 \sim 10^{-3}$ and $a/a_0 \sim 0.1$, the photons have been propagating freely ever since the plasma recombined. Even if reionization occurred, the photons have been propagating freely at least since $a/a_0 \sim 0.1$.

The expansion of the universe has redshifted the temperature of the freely propagating photons in accordance with the equation $T \propto 1/a$ (see Box 29.2). As a consequence, today they have a black-body spectrum with a temperature of 2.7 K. They are identified with the cosmic microwave radiation that was discovered in 1965, and they give one direct information about the nature of the universe at the time they last interacted with matter ($a/a_0 \sim 10^{-3}$, $t \sim 10^5$ years if reionization did not occur; $a/a_0 \sim 0.1$, $t \sim 5 \times 10^8$ years if reionization did occur.)

(8) condensation of stars, galaxies and clusters (10^8 yrs. $\lesssim t \lesssim 10^9$ yrs.)

Return to the history of matter. Before plasma recombination, the photon pressure ("elasticity of the cosmological fluid") prevented the uniform matter (25 per cent He^4, 75 per cent H) from condensing into stars, galaxies, or clusters of galaxies. However, after recombination, the photon pressure was gone, and condensation could begin. Small perturbations in the matter density, perhaps dating back to the beginning of expansion, then began to grow larger and larger. Somewhere between $a/a_0 \sim 1/30$ and $a/a_0 \sim 1/10$ (10^8 years $\lesssim t \lesssim 10^9$ years) these perturbations began developing into stars, galaxies, and clusters of galaxies. Slightly later, at $a/a_0 \sim 1/4$, quasars probably "turned on," emitting light which astronomers now receive at Earth (see Box 28.1).

EXERCISE

Exercise 28.1. UNCERTAINTY IN EVOLUTION

Current observations, plus the assumption of complete homogeneity and isotropy at the beginning of expansion, plus the assumption that the excess of leptons over antileptons is less than or of the order of the excess of baryons over antibaryons, place the following limits on the cosmological parameters today:

Matter density today $= \rho_{mo}$, between 10^{-28} and 2×10^{-31} g/cm^3;
$k = 0$ or $+1$ or -1;
temperature of electromagnetic radiation today $= 2.7 \pm 0.1$ K.
Total radiation density [observed photons, plus neutrinos and gravitons that presumably originated in big bang in thermal equilibrium with photons] $= \rho_{ro}$, between 0.7×10^{-33} and 1.2×10^{-33} g/cm^3.

(*continued on page 769*)

Box 28.1 EVOLUTION OF THE QUASAR POPULATION

That the large-scale, average properties of the universe are changing markedly with time one can infer from quasar data. In brief, there appear to have been about 50 times more quasars in the universe at a redshift $z \approx 2$ than at $z \approx 0.5$; and there may well have been fewer, or none, at redshifts $z > 3$. (On the use of redshift to characterize time since the big bang, see Box 29.3.) In greater detail, Schmidt (1972) gives the following analysis of the data:*

1. Schmidt assumes from the outset that quasar redshifts are cosmological in origin [redshift = (Hubble constant) $\times$ (distance); §29.2]. The evidence for this is
   a. Observational: Some quasars are located in clusters of galaxies [as evidenced both by position on sky and by quasar having same redshift as galaxies in cluster; see Gunn (1971)]. Since the evidence for the cosmological distance-redshift relation for galaxies is overwhelming (Boxes 29.4 and 29.5), the redshifts of these particular quasars *must* be cosmological.
   b. Theoretical: Observed quasar redshifts of $z \sim 1$ to 3 cannot be gravitational in origin; objects with gravitational redshifts larger than $z \approx 0.5$ are unstable against collapse (see Chapters 24 and 26 and Box 25.9). Nor are the quasar redshifts likely to be Doppler; how could so massive an object be accelerated to $v \approx 1$ without complete disruption? The only remaining possibility is a cosmological redshift. For this reason, opponents of the cosmological hypothesis usually feel pressed to invoke in the quasars a breakdown of the laws of physics as one understands them today. [See, e.g., Arp (1971) and references cited therein. These references also describe evidence against the cosmological assumption, evidence that a few prominent investigators find compelling, but that most do not as of 1972.]
2. Schmidt then asks how many quasars, N, there were in the universe at a time corresponding to the redshift z, and with absolute luminosity per unit frequency, $L_\nu(2{,}500\ \text{Å})$ at the wavelength 2500 Å as measured in the quasar's local Lorentz frame.
3. The data on quasars available in 1972 are not at all sufficient to determine $N[z, L_\nu(2{,}500\ \text{Å})]$ uniquely. But they *are* sufficient to show unequivocally that:
   a. There *must* have been evolution; $N(z, L_\nu)$ cannot be independent of z.
   b. The evolution cannot have resided primarily in the luminosities: the total number of quasars,

$$N_{\text{tot}}(z) \equiv \sum_{L_\nu(2{,}500\text{Å})} N(z, L_\nu)$$

   must have changed markedly with time (with z).

*Our version of Schmidt's (1972) argument is oversimplified. The reader interested in greater precision should consult his original paper.

Box 28.1 *(continued)*

c. If the evolution was primarily in the total number, $N_{tot}(z)$, i.e., if the changes in the relative luminosity distribution at 2,500 Å

$$f(z, L_\nu) \equiv [1/N_{tot}(z)]N(z, L_\nu)$$

were negligible, and if the universe today is characterized by $\sigma_0 = q_0 = 1$ (see Chapter 29 for notation), then the data show

$$N_{tot}(z = 2) \approx 50 N_{tot}(z = 0.5).$$

This steep increase in number as one goes backward in time—and all other basic features of the observed quasar redshift and magnitude distributions for $z \lesssim 2$—can be fit in a universe with $\sigma_0 = q_0 = 1$ by either of the evolution laws

$$N_{tot}(z) \propto (1 + z)^6,$$
$$N_{tot}[z(t)] \propto 10^{5(t_0 - t)/t_0}.$$

Here t_0 is the current age of the universe and t was the age at redshift z.

d. These evolution laws, when extrapolated beyond a redshift $z \approx 2$ and when combined with the observed relative luminosity function $f(z, L_\nu)$ for quasars near apparent magnitude 18, predict that an observer on Earth should *see* the following fractions of nineteenth and twentieth-magnitude quasars to have redshifts greater than 2.5:

evolution law	fraction with $z > 2.5$ $m = 19$	$m = 20$
$(1 + z)^6$	29%	49%
$10^{5(t_0 - t)/t_0}$	12%	14%

In 1972 about 30 quasars fainter than $m = 18.5$ are known, and of these only 1 (3%) has $z > 2.5$. This shows, in Schmidt's words, "that the density law $(1 + z)^6$ cannot persist beyond a redshift of around 2.5." Schmidt regards the $10^{5(t_0 - t)/t_0}$ law (which becomes nearly constant at $z > 2.5$) to be also in apparent conflict with the observations, but he says that "further spectroscopic work on faint quasars is needed to confirm this suspicion."

One reason for caution is the difficult problem of removing "observational selection effects" from the data. Schmidt, Sandage, and others have independently searched for selection effects that might produce an artificial apparent decrease in the number of quasars at $z > 2.5$. None have been found. In the words of Sandage (1972d) "The apparent cutoff in quasar redshifts near $z = 2.8$ [has been] examined for selection effects that could produce it artificially. If the cutoff is real, it may be the time of the birth of the first quasars, although the suggested redshift is unexpectedly small. At $z = 3$ in a $q_0 = 1$ universe, the look-back time is 89 per cent of the Friedmann age. Assessment of the observational selection effects shows that none are positively established that could produce the cutoff artificially."

(The uncertainties taken into account in ρ_{r0} are uncertainty about whether quadrupole moments at early times were sufficient to create gravitons at the full level corresponding to thermal equilibrium, and uncertainty about the number and statistical weights of particle species in equilibrium at the time gravitons decoupled.) Use the equations in §27.10 to calculate the uncertainties in the evolutionary history (Figure 28.1) caused by these uncertainties in the present state of the universe.

§28.2. STANDARD MODEL MODIFIED FOR PRIMORDIAL CHAOS

The standard hot big-bang model is remarkably powerful and accords well with observations (primordial helium abundances; existence, temperature, and isotropy of cosmic microwave radiation; homogeneity and isotropy of universe in the large; close accord between age of universe as measured by expansion and ages of oldest stars; . . .). However, in 1972 it encounters apparent difficulty with one item: the origin of galaxies. In a universe that is initially homogeneous and isotropic it is not clear that random fluctuations will give rise (after plasma recombination) to perturbations in the density of matter of sufficient amplitude to condense into galaxies. The perturbations that eventually form galaxies might have to reside in the initial, exploding state of the universe. [See Zel'dovich and Novikov (1974) for detailed review and discussion; see also references cited in §30.1.]

Is it reasonable to assume a small amount of initial inhomogeneity? Is it not much more reasonable to assume either perfect homogeneity (one extreme) or perfect chaos (the other extreme)?

What if the universe began chaotic?

Thus, if perfect initial homogeneity turns out to be incompatible with the origin of galaxies, it is attractive to try "perfect initial chaos"—i.e., completely random initial conditions, with a full spectrum of fluctuations in density, entropy, and local expansion rate [Misner (1968, 1969b)]. It is conceivable, but far from proved, that during its subsequent evolution such a model universe will quickly smooth itself out by natural processes (Chapter 30) such as "Mixmaster oscillations," neutrino-induced viscosity [see, e.g., Matzner and Misner (1971)], and gravitational curvature-induced creation of particle pairs [Zel'dovich (1972)]. Will one be left, after a few seconds or less, with a nearly homogeneous and isotropic, Friedmann universe, containing just enough remaining perturbations to condense eventually into galaxies? Theoretical calculations have not yet been carried far enough to give a clear answer. Of course, after the initial chaos subsides, if it subsides, such a model universe will evolve in accord with the standard big-bang model of the last section.

§28.3. WHAT "PRECEDED" THE INITIAL SINGULARITY?

No problem of cosmology digs more deeply into the foundations of physics than the question of what "preceded" the "initial state" of infinite (or near infinite) density, pressure, and temperature. And, unfortunately, no problem is farther from solution in 1973.

The initial singularity and quantum gravitational effects

General relativity predicts, inexorably, that even if the "initial state" was chaotic rather than smooth, it must have involved a spacetime "singularity" of some sort [see Hawking and Ellis (1968); also §34.6 of this book]. And general relativity is incapable of projecting backward through the singularity to say what "preceded" it. Perhaps only by coming to grip with quantum gravitational effects (marriage of quantum theory with classical geometrodynamics) will one ever reach a clear understanding of the initial state and of what, if anything, "preceded" it [see Misner (1969c), Wheeler (1971c)]. For further discussion of these deep issues, see §§34.6, 43.4, the final section of Box 30.1, and Chapter 44.

§28.4. OTHER COSMOLOGICAL THEORIES

Cosmologies that violate general relativity

This book confines attention to the cosmology of general relativity. If one were to abandon general relativity, one would have a much wider set of possibilities, including (1) the steady-state theory [Hoyle (1948); Bondi and Gold (1948)], which has not succeeded in accounting for the cosmic microwave radiation or in explaining observed evolutionary effects in radio sources and quasars [Box 28.1]; (2) the Klein-Alfvén "hierarchic cosmology" of matter in an asymptotically flat spacetime [Alfvén and Klein (1962), Alfvén (1971), Klein (1971), Moritz (1969), de Vaucouleurs (1971)], which disagrees with cosmic-ray and gamma-ray observations [Steigman (1971)]; and the Brans-Dicke cosmologies [Dicke (1968), Greenstein (1968a,b), Morganstern (1973)], which are qualitatively the same and quantitatively almost the same as the standard hot big-bang model. However, no motivation or justification is evident for abandoning general relativity. The experimental basis of general relativity has been strengthened substantially in the past decade (Chapters 38–40); and the standard big-bang model of the universe predicted by general relativity accords remarkably well with observations—far better than any other model ever proposed!

CHAPTER 29

PRESENT STATE AND FUTURE EVOLUTION OF THE UNIVERSE

§29.1. PARAMETERS THAT DETERMINE THE FATE OF THE UNIVERSE

This chapter is entirely Track 2. Chapter 27 (idealized cosmological models) is needed as preparation for it, but this chapter is not needed as preparation for any later chapter.

Will the universe continue to expand forever; or will it slow to a halt, reverse into contraction, and implode back to a state of infinite (or near infinite) density, pressure, temperature, and curvature? The answer is not yet known for certain. To discover the answer is one of the central problems of cosmology today.

Expansion forever vs. recontraction of universe

The only known way to discover the answer is to measure, observationally, the present state of the universe; and then to calculate the future evolution using Einstein's field equations. The field equations have already been solved in §§27.10 and 27.11. From those solutions one reads off the following correlation between the present state of the universe and its future.

If $\Lambda = 0$ [in accord with Einstein's firmly held principle of simplicity]:

Expansion forever $\Longleftrightarrow$ negative or zero spatial curvature for hypersurfaces of homogeneity, i.e., $k/a_o^2 \leq 0$ ("open" or "flat");

Recontraction $\Longleftrightarrow$ positive spatial curvature for homogeneous hypersurfaces, i.e., $k/a_o^2 > 0$ ("closed").

If $\Lambda \neq 0$:

$$\text{Expansion forever} \Longleftrightarrow \Lambda \geq \Lambda_{\text{crit}} \equiv \begin{cases} 0 & \text{if } k \leq 0, \\ (4\pi\rho_{mo}a_o{}^3)^{-2} & \text{if } k > 0; \end{cases}$$

$$\text{Recontraction} \Longleftrightarrow \Lambda < \Lambda_{\text{crit}}.$$

Parameters required to predict future of universe:

(1) "relativity parameters" Λ, k/a_o^2, ρ_{mo}

Evidently three parameters are required to predict the future: the cosmological constant, Λ; the curvature parameter today for the hypersurface of homogeneity, k/a_o^2; and the density of matter today, ρ_{mo}. (To extrapolate into the past, as was done in the last chapter, one needs, besides these quantities, the radiation density today, ρ_{ro}. But ρ_{ro} is so small now and is getting smaller so fast ($\rho_r \propto a^{-4}$; $\rho_m \propto a^{-3}$) that it can have no influence on the decision between the possibilities just listed.

The task of predicting the future, then, reduces to the task of measuring the "relativity parameters" Λ, k/a_o^2, and ρ_{mo}.

(2) "observational parameters" H_o, q_o, σ_o

In tackling this task, observational cosmologists prefer to replace the three "relativity parameters," which have immediate significance for relativity theory, by parameters that are more directly observable. One parameter close to the observations is the Hubble expansion rate *today,* i.e., the "*Hubble constant,*"

$$H_o \equiv (a_{,t}/a)_o. \tag{29.1a}$$

Another is the dimensionless "*deceleration parameter*" today, q_o, defined by

$$q_o \equiv -\frac{a_{,tt}}{a}\frac{1}{H_o^2} = -\left(\frac{aa_{,tt}}{a_{,t}^2}\right)_o. \tag{29.1b}$$

And a third is the dimensionless "*density parameter,*" today,

$$\sigma_o \equiv \frac{4\pi\rho_{mo}}{3H_o^2}. \tag{29.1c}$$

(3) relationship between relativity parameters and observational parameters

The relationships between these three "observational parameters" and the three "relativity parameters" Λ, k/a_o^2, and ρ_{mo} (together making six "cosmological parameters") can be calculated by combining definitions (29.1) with the Einstein field equations (27.39), which, evaluated today, say

$$\begin{aligned} H_o^2 &= -\frac{k}{a_o^2} + \frac{\Lambda}{3} + \frac{8\pi}{3}\rho_{mo}, \\ -2q_oH_o^2 &= -H_o^2 - \frac{k}{a_o^2} + \Lambda. \end{aligned} \tag{29.2}$$

By combining these equations, one finds the relationships shown in Box 29.1, where the implications of several values of σ_o and q_o are also shown.

EXERCISE

Exercise 29.1. IMPLICATIONS OF PARAMETER VALUES

Derive the results quoted in Box 29.1.

---

§29.2. COSMOLOGICAL REDSHIFT

Observed features of cosmological redshift

One of the key pieces of observational data used in measurements of H_o, q_o, and σ_o is the cosmological redshift: spectral lines emitted by galaxies far from Earth and received at Earth are found to be shifted in wavelength toward the red. For example, the [O II]λ3727 line, when both emitted *and* observed in an Earth-bound laboratory, has a wavelength of 3727 Å. However, when it is emitted by a star in the galaxy

Box 29.1 OBSERVATIONAL PARAMETERS COMPARED TO RELATIVITY PARAMETERS

A. Relativity Parameters

1. Matter density today,

$$\rho_{mo}$$

2. Curvature of hypersurface of homogeneity today,

$$k/a_o^2$$

3. Cosmological constant,

$$\Lambda$$

4. Radiation density today, ρ_{ro} (unimportant for the present dynamics of the universe, and therefore ignored in this chapter)

B. Observational Parameters

1. Hubble constant (Hubble expansion rate today),

$$H_o \equiv (a_{,t}/a)_o$$

2. Deceleration parameter,

$$q_o \equiv -\frac{a_{,tt}}{a}\frac{1}{H_o^2}$$

3. Density parameter,

$$\sigma_o \equiv \frac{4\pi\rho_{mo}}{3H_o^2}$$

C. Observational Parameters in Terms of Relativity Parameters

$$H_o^2 = (8\pi/3)\rho_{mo} - k/a_o^2 + \Lambda/3, \tag{1}$$

$$q_o = \frac{(4\pi/3)\rho_{mo} - \Lambda/3}{(8\pi/3)\rho_{mo} - k/a_o^2 + \Lambda/3}, \tag{2}$$

$$\sigma_o = \frac{(4\pi/3)\rho_{mo}}{(8\pi/3)\rho_{mo} - k/a_o^2 + \Lambda/3}. \tag{3}$$

D. Relativity Parameters in Terms of Observational Parameters

$$\rho_{mo} = (3/4\pi)H_o^2\sigma_o, \tag{4}$$

$$k/a_o^2 = H_o^2(3\sigma_o - q_o - 1), \tag{5}$$

$$\Lambda = 3H_o^2(\sigma_o - q_o). \tag{6}$$

E. Implications of Specific Parameter Values

1. $\Lambda = 0$ (in accord with Einstein's point of view) if and only if $\sigma_o = q_o$.
2. Sign of Λ is same as sign of $\sigma_o - q_o$.

2

Box 29.1 *(continued)*

3. If $\Lambda = 0$

(a) $q_o > \dfrac{1}{2} \Longleftrightarrow \rho_{mo} > \rho_{\text{crit}} \equiv \dfrac{3}{8\pi} H_o^2 \Longleftrightarrow k > 0$ $\begin{pmatrix}\text{positive curvature;}\\ \text{“closed” universe}\end{pmatrix}$

$\Longleftrightarrow$ universe will eventually recontract;

(b) $q_o = \dfrac{1}{2} \Longleftrightarrow \rho_{mo} = \rho_{\text{crit}} \equiv \dfrac{3}{8\pi} H_o^2 \Longleftrightarrow k = 0$ $\begin{pmatrix}\text{zero curvature;}\\ \text{“flat” universe}\end{pmatrix}$

$\Longrightarrow$ universe will expand forever;

(c) $q_o < \dfrac{1}{2} \Longleftrightarrow \rho_{mo} < \rho_{\text{crit}} \equiv \dfrac{3}{8\pi} H_o^2 \Longleftrightarrow k < 0$ $\begin{pmatrix}\text{negative curvature;}\\ \text{“open” universe}\end{pmatrix}$

$\Longrightarrow$ universe will expand forever.

4. If $\Lambda \neq 0$

(a) $\sigma_o > \dfrac{1}{3}(q_o + 1) \Longleftrightarrow k > 0$ $\begin{pmatrix}\text{positive curvature;}\\ \text{“closed” universe}\end{pmatrix}$,

and in this case,

$$\sigma_o - q_o \geq \frac{1}{\sigma_o^2}\left(\sigma_o - \frac{q_o + 1}{3}\right)^3 \Longleftrightarrow \text{universe will expand forever,}$$

$$\sigma_o - q_o < \frac{1}{\sigma_o^2}\left(\sigma_o - \frac{q_o + 1}{3}\right)^3 \Longleftrightarrow \text{universe will eventually recontract;}$$

(b) $\sigma_o = \dfrac{1}{3}(q_o + 1) \Longleftrightarrow k = 0$ $\begin{pmatrix}\text{zero curvature;}\\ \text{“flat” universe}\end{pmatrix}$,

and in this case,

$\sigma_o \geq q_o \Longleftrightarrow$ universe will expand forever,

$\sigma_o < q_o \Longleftrightarrow$ universe will eventually recontract;

(c) $\sigma_o < \dfrac{1}{3}(q_o + 1) \Longleftrightarrow k < 0$ $\begin{pmatrix}\text{negative curvature;}\\ \text{“open” universe}\end{pmatrix}$,

and in this case,

$\sigma_o \geq q_o \Longleftrightarrow$ universe will expand forever,

$\sigma_o < q_o \Longleftrightarrow$ universe will eventually recontract.

3C 295 (presumably with the same wavelength, $\lambda_{em} = 3727$ Å) and received at Earth, it is measured here to have the wavelength $\lambda_{rec} = 5447$ Å. The fractional change in wavelength is

$$z \equiv (\lambda_{rec} - \lambda_{em})/\lambda_{em} = 0.4614 \text{ for 3C 295.} \tag{29.3}$$

The cosmological redshift is observed to affect all spectral lines alike, and not only lines in the visible spectrum. Thus, the 21-cm line of hydrogen, with 400,000 times the wavelength of the central region of the visible, undergoes a redshift that agrees (within the errors of the measurements) with the redshifts of lines in the visible for recession velocities of the order of $v \sim 0.005$, according to observation of thirty objects by Dieter, Epstein, Lilley, and Roberts (1962) and further observations by Roberts (1965).

Why redshift cannot be due to "tired light"

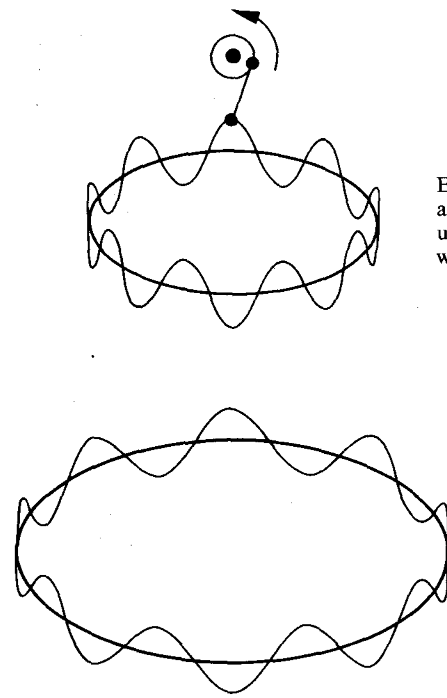
No one has ever put forward a satisfactory explanation for the cosmological redshift other than the expansion of the universe (see below). The idea has been proposed at various times by various authors that some new process is at work ("tired light") in which photons interact with atoms or electrons on their way from source to receptor, and thereby lose bits and pieces of their energy. Ya. B. Zel'dovich (1963) gives a penetrating analysis of the difficulties with any such ideas:

(1) "If the energy loss is caused by an interaction with the intergalactic matter, it is accompanied by a transfer of momentum; that is, there is a change of the direction of motion of the photon. There would then be a smearing out of images; a distant star would be seen as a disc, not a point, and that is not what is observed." (2) "Let us suppose that the photon decays, $\gamma \longrightarrow \gamma' + k$, giving up a small part of its energy to some particle, k. It follows from the conservation laws that k must move in the direction of the photon (this, by the way, avoids a smearing out), and must have zero rest mass. Because of the statistical nature of the process, however, some photons would lose more energy than others, and there would be a spectral broadening of the lines, which is also not observed."

(3) If there does exist any such decay process, then simple arguments of special relativity that Zel'dovich attributes to M. P. Bronshtein, and spells out in detail, demand the relationship

$$\begin{pmatrix}\text{probability per}\\ \text{second of}\\ \text{“photon decay”}\end{pmatrix} = \frac{\begin{pmatrix}\text{a universal constant with}\\ \text{the dimensions sec}^{-2}\end{pmatrix}}{(\text{frequency of photon in sec}^{-1})}.$$

"Thus," Zel'dovich concludes, "if the decay of photons is possible at all, those in radio waves must decay especially rapidly! This would mean that the Maxwell equation for a static electric field would have to be changed . . . . There is no experimental indication of such effects: the radio-frequency radiation from distant sources is transmitted to us not a bit more poorly than visible light, and the red shift measured in different parts of the spectrum is exactly the same . . . . Thus, suggestions that there is an explanation of the red shift other than Friedmann's fail completely."



Emission:
atom excites n-node standing wave;
universe small, $a(t_e) = a_{\rm em}$;
wavelengths small, $\lambda(t_e) = \lambda_{\rm em}$.

Reception:
universe larger, $a(t_r) = a_{\rm rec}$;
wavelengths larger, $\lambda(t_r) = \lambda_{\rm rec}$;
number of nodes in standing wave unchanged;

$$n = \text{constant} = \frac{2\pi a_{\rm rec}}{2\pi\lambda_{\rm rec}} = \frac{a_{\rm em}}{\lambda_{\rm em}}.$$

Figure 29.1.
Redshift as an effect of standing waves. The ratio of wavelengths, $\lambda_{\rm rec}/\lambda_{\rm em}$, is identical with the ratio of dimensions, $a_{\rm rec}/a_{\rm em}$ in any closed spherically symmetrical (Friedmann) model universe. The atom excites an n-node standing wave in the universe. The number n stays constant during the expansion. Therefore wavelengths increase in the same proportion as the dimensions of the universe. One sees immediately in this way that the redshift is independent of all such details as (1) why the expansion came about (spherical symmetry, but arbitrary equation of state); (2) the rate—uniform or nonuniform—at which it came about; and (3) the distance between source and receptor at emission, at reception, or at any time in-between. The reasoning in the diagram appears to depend on the closure of the universe (standing waves; $k = +1$ rather than 0 or -1). That closure is not required for this simple result is seen from the further analysis given in the text.

Not the least among the considerations that lead one to accept the general recession of the galaxies as the explanation for the redshift is the circumstance that this general recession was predicted [Friedmann (1922)] before the redshift was observed [Hubble (1929)].

Derivation of redshift formula:
$$\lambda \propto \begin{pmatrix}\text{expansion}\\ \text{factor}\end{pmatrix}$$

The cosmological redshift is easily understood (Figure 29.1) in terms of the standard big-bang model for the universe. A detailed analysis focuses attention on three processes: emission of the light, propagation of the light through curved spacetime from emitter to receiver, and reception of the light. Emission and reception occur in the proper reference frames (orthonormal tetrads) of the emitter and receiver; they are special-relativistic phenomena. Propagation, by contrast, is a general-relativistic process; it is governed by the law of geodesic motion in curved spacetime.

In calculating all three processes—emission, propagation, and absorption—one needs a coordinate system. Use the coordinates (t, χ, θ, ϕ) or $(\eta, \chi, \theta, \phi)$ introduced in Chapter 27; and orient the space coordinates in such a way that the paths of the light rays through the coordinate system are simple. This is best done by putting the origin of the coordinate system ($\chi = 0$) at the Earth. Then the emitting galaxy will lie at some "radius" χ_e and some angular position (θ_e, ϕ_e). The cosmological line element

$$\begin{aligned} ds^2 &= -dt^2 + a^2(t)[d\chi^2 + \Sigma^2(d\theta^2 + \sin^2\theta\, d\phi^2)] \\ &= a^2(\eta)[-d\eta^2 + d\chi^2 + \Sigma^2(d\theta^2 + \sin^2\theta\, d\phi^2)], \end{aligned} \tag{29.4a}$$

$$\Sigma = \begin{cases} \sin\chi & \text{if } k = +1, \\ \chi & \text{if } k = 0, \\ \sinh\chi & \text{if } k = -1, \end{cases} \tag{29.4b}$$

is spherically symmetric about $\chi = 0$ (i.e., about the Earth) whether $k = -1, 0$, or $+1$. Consequently, the geodesics (photon world lines) that pass through both Earth and the emitting galaxy must all be radial

$$\theta = \theta_e, \qquad \phi = \phi_e, \qquad \chi = \chi(t). \tag{29.5}$$

(One who wishes to forego any appeal to symmetry can examine the geodesic equation in the (t, χ, θ, ϕ) coordinate system, and discover that if $d\theta/d\lambda = d\phi/d\lambda = 0$, then $d^2\theta/d\lambda^2 = d^2\phi/d\lambda^2 = 0$. Consequently a geodesic that is initially radial will always remain radial.)

Consider, now, emission. A galaxy at rest (moving with the "cosmological fluid") at $(\chi_e, \theta_e, \phi_e)$ emits two successive crests, A and B, of a wave train toward Earth at coordinate times t_{Ae} and t_{Be}. It has been arranged that proper time as measured on the galaxy is the same as coordinate time ($t = \tau +$ const. was part of the construction process for the coordinate system in §27.4). Consequently the period of the radiation as seen by the emitter is $P_{em} = t_{eB} - t_{eA}$; and the wavelength is the same as the period when geometrized units are used:

$$\lambda_{\text{em}} = t_{eB} - t_{eA}. \tag{29.6}$$

Next examine propagation. Wave crests A and B propagate along null geodesics. This fact enables one to read the world lines of the wave crests, $\chi_A(t)$ and $\chi_B(t)$, directly from the line element (29.4): $ds^2 = 0$ guarantees that $a(t)\, d\chi = -dt$ ($-$, not $+$, because the light propagates toward the Earth at $\chi = 0$). Consequently, the world lines are

$$\chi_e - \chi_A(t \text{ or } \eta) = \eta - \eta_{eA} = \int_{t_{eA}}^{t} a^{-1}\, dt, \tag{29.7}$$

$$\chi_e - \chi_B(t \text{ or } \eta) = \eta - \eta_{eB} = \int_{t_{eB}}^{t} a^{-1}\, dt.$$

Finally, examine reception. The receiver on Earth moves with the "cosmological fluid," just as does the distant emitter. (Ignore the Earth's "peculiar motion" relative to the fluid—motion around the sun, motion around center of our Galaxy, etc.; it can be taken into account by an ordinary Doppler correction.) Thus, for receiver as for emitter, proper time is the same as coordinate time, and

$$\lambda_{\rm rec} = t_{rB} - t_{rA}, \tag{29.8}$$

where t_{rB} and t_{rA} are the times of reception of the successive wave crests.

Now combine equations (29.6), (29.7), and (29.8) to obtain the redshift. The receiver is at $\chi = 0$. Therefore equations (29.7) say

$$\begin{aligned} 0 &= \chi_e - \int_{t_{eA}}^{t_{rA}} a^{-1}\,dt, \\ 0 &= \chi_e - \int_{t_{eB}}^{t_{rB}} a^{-1}\,dt. \end{aligned} \tag{29.9}$$

Subtract these equations from each other to obtain

$$\begin{aligned} 0 &= \int_{t_{eB}}^{t_{rB}} a^{-1}\,dt - \int_{t_{eA}}^{t_{rA}} a^{-1}\,dt \\ &= \int_{t_{rA}}^{t_{rB}} a^{-1}\,dt - \int_{t_{eA}}^{t_{eB}} a^{-1}\,dt \approx \frac{t_{rB} - t_{rA}}{a(t_r)} - \frac{t_{eB} - t_{eA}}{a(t_e)}; \end{aligned}$$

and combine with (29.6) and (29.8) to discover

$$\frac{\lambda_{\rm rec}}{a(t_r)} = \frac{\lambda_{\rm em}}{a(t_e)}; \tag{29.10}$$

i.e.,

$$z \equiv \Delta\lambda/\lambda = a(t_r)/a(t_e) - 1. \tag{29.11}$$

These redshift equations confirm the simple result of Figure 29.1: As the light ray propagates, its wavelength (as measured by observers moving with the "fluid") increases in direct proportion to the linear expansion of the universe. *The ratio of the wavelength to the expansion factor, λ/a, remains constant.* For important applications of this result, see Boxes 29.2 and 29.3.

EXERCISES

Exercise 29.2. ALTERNATIVE DERIVATION OF REDSHIFT

Notice that the only part of the line element that is relevant for the light ray is

$$ds^2 = -dt^2 + a^2(t)\,d\chi^2,$$

since $d\theta = d\phi = 0$ along its world line (spherical symmetry!). Regard the light ray as made

2

Box 29.2 COSMOLOGICAL REDSHIFT OF THE PRIMORDIAL RADIATION

As an important application of the redshift formula

$$\lambda/a = \text{constant} \tag{1}$$

[equation (29.10)], consider the radiation emerging from the hot big bang. Because it is initially in thermal equilibrium with matter, this primordial radiation initially has a Planck black-body spectrum. Subsequent interactions with matter cannot change the spectrum, because the matter remains in thermal equilibrium with the radiation so long as interactions are occurring. The cosmological redshift can and does change the spectrum, however. It was shown in exercise 22.17, using kinetic theory, that radiation with a Planck spectrum as viewed by one observer has a Planck spectrum as viewed by all observers; but the observed temperature is redshifted in precisely the same manner as the frequency of an individual photon is redshifted. Consequently, as seen by observers at rest in the "fluid," the temperature of the primordial radiation is redshifted

$$T \propto 1/a. \tag{2}$$

This is true after plasma recombination, when the radiation and matter are decoupled, as well as before recombination, when they are interacting. And it is true not only for the primordial photons but also for thermalized neutrinos and gravitons emerging from the hot big bang.

There is another way to derive the redshift equation (2). Combine the equation

$$\rho_r \propto T^4 \tag{3}$$

for the energy density of black-body radiation in terms of temperature, with the equation

$$\rho_r \propto (\text{volume})^{-4/3} \propto (a^3)^{-4/3} \propto a^{-4} \tag{4}$$

for the decrease of energy density with adiabatic expansion.

Box 29.3 USE OF REDSHIFT TO CHARACTERIZE DISTANCES AND TIME

Distance: When discussing objects within the Earth's cluster of galaxies, astronomers typically describe distances in units of lightyears or parsecs. But when dealing with more distant objects (galaxies, quasars, etc.), astronomers find it more convenient to describe distance in terms of what is actually observed: redshift. For example, the statement "the galaxy 3C 295 is at a redshift of 0.4614" means that "3C 295 is at that distance from Earth [given by equation (29.16)] which corresponds to a redshift of $z = 0.4614$."

Time: When discussing events that occurred during the last few 10^9 years, astronomers usually measure time in units of years. Example: "The solar system condensed out of interstellar gas 4.6×10^9 years ago" [see Wasserburg and Burnett (1968)]. But when dealing with events much nearer the beginning of the universe, all of which have essentially the same age, of about 12×10^9 years, astronomers find it more convenient to describe time in terms of redshift. Example: "The primordial plasma recombined at a redshift of 1,000" means that "If a photon had been emitted at the time of plasma recombination, and had propagated freely ever since, it would have experienced a total redshift between then and now of $z = 1{,}000$." Equivalently, since $1 + z = (a_o/a)$ [see equation (29.11)], "the plasma recombined when the universe was a factor of $1 + z \approx 1{,}000$ smaller than it is today." [Application: In Figure 28.1, where the past evolution of the universe is summarized, one can freely replace the horizontal scale a/a_o by $1/(1 + z)$, and thereby see that primordial element formation occurred at a redshift of $z \approx 10^9$.) The conversion from redshift units to time units is strongly dependent on the parameters ρ_{mo}, ρ_{ro}, and $k/a_o{}^2$ [see §§27.10 and 27.11; also equation (29.15)].

of photons with 4-momenta $\boldsymbol{p}$. From the geodesic equation (or, for the reader who has studied chapter 25, from arguments about Killing vectors), show that

$$p_\chi \equiv \boldsymbol{p} \cdot (\partial/\partial\chi)$$

is conserved along the photon's world line. Use this fact, the fact that a photon's 4-momentum is null, $\boldsymbol{p} \cdot \boldsymbol{p} = 0$, and the equation $E = -\boldsymbol{p} \cdot \boldsymbol{u}$ for the energy measured by an observer with 4-velocity $\boldsymbol{u}$, to derive the redshift equation (29.11).

Exercise 29.3. REDSHIFT OF PARTICLE DE BROGLIE WAVELENGTHS

A particle of finite rest mass μ moves along a geodesic world line through the expanding cosmological fluid. Let

$$p \equiv (\boldsymbol{p} \cdot \boldsymbol{p})^{1/2} \equiv \frac{\mu v}{(1 - v^2)^{1/2}}$$

be the spatial 4-momentum of the particle as measured by observers at rest in the fluid. (The ordinary velocity they measure in their proper reference frames is v.) The associated "de Broglie wavelength" of the particle is $\lambda \equiv h/p$.

(a) Show that this de Broglie wavelength is redshifted in precisely the same manner as a photon wavelength:

$$\lambda/a = \text{constant}.$$

(b) Employing this result, show that, for the molecules of an ideal gas that fills the universe, their mean kinetic energy decreases in inverse proportion to a^2 when the gas is nonrelativistic and (like photon energies) in inverse proportion to a when the gas is highly relativistic.

§29.3. THE DISTANCE-REDSHIFT RELATION; MEASUREMENT OF THE HUBBLE CONSTANT

Equation (29.11) expresses the redshift in terms of the change in expansion factor between the event of emission and the event of reception. For "nearby" emitters (emitters at distances much less than $1/H_o$, the "Hubble length") it is more convenient to express the redshift in terms of the distance between the emitter and Earth. That distance ("present distance") is defined on the hypersurface of homogeneity that passes through Earth today, since that hypersurface agrees locally with the surface of simultaneity of the receiver today, and it is also, locally, a surface of simultaneity for any observer moving today with the "cosmological fluid."

Derivation of distance-redshift relation

The distance between emitter and observer today [the distance along the spatial geodesic of constant (t, θ, ϕ) connecting $(t_r, 0, \theta_e, \phi_e)$ and $(t_r, \chi_e, \theta_e, \phi_e)$] can be read directly from the line element (29.4):

$$\ell = a(t_r)(\chi_e - \chi_r) = a(t_r)\chi_e. \tag{29.12}$$

Using expression (29.9) for χ_e, one finds

$$\ell = a(t_r)\int_{t_e}^{t_r} a^{-1}\, dt. \tag{29.12'}$$

In the recent past, $a(t)$ was given by

$$\begin{aligned} a(t) &= a(t_r) + (a_{,t})_{t_r}(t - t_r) + \frac{1}{2}(a_{,tt})_{t_r}(t - t_r)^2 + \cdots \\ &= a(t_r)[1 + H_o(t - t_r) - \frac{1}{2}q_o H_o^2(t - t_r)^2 + \cdots], \end{aligned} \tag{29.13}$$

where definitions (29.1) for the Hubble constant H_o and the deceleration parameter q_o have been used. Putting this expression into equation (29.12′) and integrating, one finds for the distance the expression

$$\ell = (t_r - t_e) + \frac{1}{2}H_o(t_r - t_e)^2 + \cdots$$

or, equivalently,

$$t_r - t_e = \ell - \frac{1}{2}H_o \ell^2 + \cdots. \tag{29.14}$$

The redshift [equation (29.11)] can be expressed as a power series in $t_r - t_e$ by using equation (29.13):

$$\begin{aligned} z &= \frac{a(t_r) - a(t_e)}{a(t_e)} = \frac{a(t_r)[H_o(t_r - t_e) + \frac{1}{2}q_o H_o^2(t_r - t_e)^2 + \cdots]}{a(t_r)[1 - H_o(t_r - t_e) + \cdots]} \\ &= H_o(t_r - t_e) + H_o^2\left(1 + \frac{1}{2}q_o\right)(t_r - t_e)^2 + \cdots. \end{aligned} \tag{29.15}$$

Combining this with equation (29.14) for $t_r - t_e$ in terms of ℓ, one finally obtains

Result for distance-redshift relation

$$z = H_o \ell + \frac{1}{2}(1 + q_o)(H_o \ell)^2 + O([H_o \ell]^3). \tag{29.16}$$

This is the "*distance-redshift relation*" for the standard big-bang model of the universe.

Measurement of Hubble constant H_o

By comparing this distance-redshift relation with astronomical observations (see Box 29.4, which is best read after the next section), Allan Sandage (1972a) obtains a Hubble constant of

$$H_o = 55 \pm 7 \text{ km sec}^{-1} \text{ Mpc}^{-1}; \tag{29.17}$$

i.e.,

$$H_o^{-1} = (18 \pm 2) \times 10^9 \text{ years}. \tag{29.18}$$

(Note: 1 Mpc $\equiv$ one Megaparsec is 3.26×10^6 light years, or 3.08×10^{24} cm.) The uncertainty of ± 7 km sec^{-1} Mpc^{-1} quoted here is the "one-sigma" statistical uncertainty associated with the distance-redshift data. Systematic errors, not now understood, might be somewhat larger; but the true value of H_o almost certainly is within a factor 2 of Sandage's value, 55 km sec^{-1} Mpc^{-1}.

Note that, if $\Lambda = 0$, then the "critical density" marking the dividing line between a "closed" universe and an "open" universe—i.e., between eventual recontraction and expansion forever—is

Value of critical density

$$\rho_{\text{crit}} = \frac{3}{8\pi} H_o^2 = 5 \times 10^{-30}\ \text{g/cm}^3. \tag{29.19}$$

(As described in Box 29.1, $\rho > \rho_{\text{crit}} \Longleftrightarrow$ "closed" $\Longleftrightarrow$ recontraction; $\rho < \rho_{\text{crit}} \Longleftrightarrow$ "open" $\Longleftrightarrow$ expansion forever.) Comparison with the actual density will be delayed until §29.6.

The distance measurements are not accurate enough to yield useful information about the deceleration parameter, q_o.

§29.4. THE MAGNITUDE-REDSHIFT RELATION; MEASUREMENT OF THE DECELERATION PARAMETER

Information about q_o is best obtained by comparing the apparent magnitudes of galaxies with their redshifts.

Apparent magnitude defined

In astronomy one defines the apparent (bolometric) magnitude, m, of an object by the formula

$$\begin{aligned} m &= -2.5 \log_{10}(S/2.52 \times 10^{-5}\ \text{erg cm}^{-2}\ \text{sec}^{-1}) \\ &= -2.5 \log_{10} S + \text{constant}, \end{aligned} \tag{29.20}$$

where S is the flux of energy (energy per unit time per unit area) that arrives at Earth from the object. [Of course, one cannot measure the flux over the entire wavelength range $0 < \lambda < \infty$; so one distinguishes various apparent magnitudes (m_U, m_B, $m_V, \ldots$) corresponding to fluxes in various wavelength ranges ("U" $\equiv$ "ultraviolet"; "B" $\equiv$ "blue"; "V" $\equiv$ "visual"). However, these subtleties are too far from gravitation physics to be treated here.]

Derivation of magnitude-redshift relation

Calculate the apparent magnitude for a galaxy of intrinsic luminosity L and redshift z. To simplify the calculation, put the emitter at the origin of the space coordinates ($\chi_e = 0$); and put the Earth at (χ_r, θ_r, ϕ_r). (Note the reversal of locations compared to redshift calculation of §29.2.) On Earth, place a photographic plate of area A perpendicular to the incoming light. Then at time t_r the plate is a tiny segment of a spherical two-dimensional surface ($t = t_r, \chi = \chi_r$; θ and ϕ vary) about the emitting galaxy. The total area of the 2-sphere surrounding the galaxy is

$$\mathcal{A} = 4\pi[a(t_r)\Sigma(\chi_r)]^2. \tag{29.21}$$

Therefore, the ratio of the area of the plate to the area of the 2-sphere is given by

$$\frac{A}{\mathcal{A}} = \frac{A}{4\pi[a(t_r)\Sigma(\chi_r)]^2}. \tag{29.22}$$

The plate catches a fraction $A/\mathcal{A}$ of the energy that pours out through the 2-sphere.

If there were no redshift, the power crossing the entire 2-sphere at time t_r would be precisely the luminosity of the emitter at time t_e. However, the redshift modifies this result in two ways. (1) The energy of each photon that crosses the 2-sphere is smaller, as measured in the local Lorentz frame of the fluid there, than it was as measured by the emitter:

$$E_{\text{rec}}/E_{\text{em}} = \lambda_{\text{em}}/\lambda_{\text{rec}} = 1/(1+z). \tag{29.23}$$

(2) Two photons with the same θ and ϕ, which are separated by a time Δt_r as measured by an observer stationary with respect to the "cosmological fluid" at the 2-sphere, were separated by a shorter time Δt_e as measured by the emitter:

$$\Delta t_r/\Delta t_e = \lambda_r/\lambda_e = 1+z. \tag{29.24}$$

The luminosity, L, as measured at the source, is the sum of the energies $E_{\text{em}J}$ of the individual photons (labeled with the index J) emitted in a time interval Δt_e, divided by Δt_e:

$$L = (1/\Delta t_e)\sum_J E_{\text{em}J}. \tag{29.25}$$

The power that crosses the 2-sphere a time $t_r - t_e$ later, as measured by the fluid at the 2-sphere, is

$$P = (1/\Delta t_r)\sum_J E_{\text{rec}J}, \tag{29.26}$$

where the summation runs over the same set of photons.

Combining equations (29.23) to (29.26), one sees that the power crossing the 2-sphere is

$$P = L/(1+z)^2.$$

Of this, a fraction,

$$P_A = \frac{A}{\mathcal{A}}P = \frac{AL}{4\pi[(1+z)a(t_r)\Sigma(\chi_r)]^2},$$

crosses the photographic plate; so the flux measured at the Earth is

$$S = \frac{P_A}{A} = \frac{L}{4\pi R^2(1+z)^2}, \tag{29.27}$$

where R is the "radius of curvature" of the 2-sphere surrounding the emitter and passing through the receiver at the time of reception,

$$R \equiv a_o\Sigma(\chi_r - \chi_e) = \begin{cases} a_o \sinh(\chi_r - \chi_e) & \text{if } k = -1, \\ a_o[\chi_r - \chi_e] & \text{if } k = 0, \\ a_o \sin(\chi_r - \chi_e) & \text{if } k = +1 \end{cases} \tag{29.28}$$

[recall: χ_e is 0 according to the present conventions, and $a_o = a(t_r)$]. The corresponding apparent magnitude [equation (29.20)] is

$$m = +5 \log_{10}[(1 + z)R] - 2.5 \log_{10}L + \text{constant}. \tag{29.29}$$

In order to relate the apparent magnitude to the redshift of the emitter, one must express the quantity R in terms of z. From equation (29.7) for the photon propagation (with sign reversed because positions of receiver and emitter have been reversed), one knows that

$$\chi_r - \chi_e = \int_{t_e}^{t_r} a^{-1}\, dt = \int_1^{a(t_r)/a(t_e)} \left[\frac{a}{a(t_r)}\right]\left[\frac{dt}{da}\right] d\left[\frac{a(t_r)}{a}\right], \tag{29.30}$$

and from equation (29.11) one knows that

$$z = a(t_r)/a(t_e) - 1.$$

Hence

$$\chi_r - \chi_e = \int_1^{1+z} \left[\frac{a}{a(t_r)}\right]\left[\frac{dt}{da}\right] d\left[\frac{a(t_r)}{a}\right]. \tag{29.31}$$

Equations (4) to (6) of Box 29.1, and (27.40), determine the function dt/da in terms of $a/a(t_r)$ and the constants H_o, q_o, σ_o. By inserting that result into equation (29.31) and integrating, one obtains $\chi_r - \chi_e$ in terms of the redshift z and the cosmological parameters H_o, q_o, σ_o:

$$\chi_r - \chi_e = |1 + q_o - 3\sigma_o|^{1/2} \int_1^{1+z} \frac{du}{[2\sigma_o u^3 + (1 + q_o - 3\sigma_o)u^2 + \sigma_o - q_o]^{1/2}}. \tag{29.32a}$$

The 2-sphere radius of curvature R is obtained by inserting this expression into the equation

$$R = \frac{H_o^{-1}}{|1 + q_o - 3\sigma_o|^{1/2}} \Sigma(\chi_r - \chi_e) \tag{29.32b}$$

[equation (29.28), with a_o evaluated by equation (5) of Box 29.1].

Equations (29.29) and (29.32) determine the apparent magnitude, m, in terms of redshift, z.

For the case of vanishing cosmological constant ($\sigma_o = q_o$; $\Lambda = 0$), the integral (29.32a) can be expressed in terms of elementary functions, yielding

$$\begin{aligned} R &= \frac{H_o^{-1}}{q_o^2(1 + z)}[-q_o + 1 + q_o z + (q_o - 1)(2q_o z + 1)^{1/2}], \\ &\approx H_o^{-1} z\left[1 - \frac{1}{2}(1 + q_o)z + O(z^2)\right], \end{aligned} \tag{29.33}$$

so that

Result for magnitude-redshift relation

$$\begin{aligned} m &= 5 \log_{10} [1 - q_o + q_o z + (q_o - 1)(2q_o z + 1)^{1/2}] - 2.5 \log_{10} L + \text{const.} \\ &\approx 5 \log_{10} z + 1.086(1 - q_o)z + O(z^2) - 2.5 \log_{10} L + \text{const.} \quad \text{for } z \ll 1. \end{aligned} \tag{29.34}$$

(Note: the factor 1.086 is actually 2.5/ln 10.) A power-series solution for nonzero Λ (for $\sigma_o \neq q_o$) reveals a dependence on σ_o only at $O(z^2)$ and higher:

$$R \approx H_o^{-1}z\left[1 - \frac{1}{2}(1 + q_o)z + (\text{corrections of } O(z^2) \text{ depending on } \sigma_o \text{ and } q_o)\right], \tag{29.35a}$$

$$m \approx 5 \log_{10} z + 1.086(1 - q_o)z + O(z^2) - 2.5 \log_{10} L + \text{const.} \tag{29.35b}$$

Sheldon (1971) gives the exact solution for $\Lambda \neq 0$ in terms of the Weierstrass elliptic function. Refsdal *et al.* (1967) tabulate and plot the exact solution.

Measurement of deceleration parameter, q_o

By comparing the theoretical magnitude-redshift relation (29.35b) with observations of the brightest galaxies in 82 clusters, Allan Sandage (1972a,c,d) obtains the following value for the deceleration parameter:

$$q_o = 1.0 \pm 0.5, \qquad \text{if } \sigma_o = q_o (\text{i.e. } \Lambda = 0). \tag{29.36}$$

(Note: 0.5 is the "one-sigma" uncertainty. Sandage estimates with 68 per cent confidence that $0.5 < q_o < 1.5$, and with 95 per cent confidence that $0 < q_o < 2$—providing unknown evolutionary effects are negligible.) The observations leading to this result and the uncertainties due to evolutionary effects are described in Box 29.4. Box 29.5 gives a glimpse of Edwin Hubble, the man who laid the foundations for such cosmological measurements.

(*continued on page 794*)

Box 29.4 MEASUREMENT OF HUBBLE CONSTANT AND DECELERATION PARAMETER

I. *Hubble Constant, H_o*

A. *Objective:* To measure the constant H_o by comparing observational data with the distance-redshift relation

$$z = H_o\ell + \frac{1}{2}(1 + q_o)(H_o\ell)^2 + O([H_o\ell]^3).$$

Here ℓ is distance from Earth to source today; and z is redshift of source as measured at Earth.

B. *Key Difficulty:* This distance-redshift relation does not apply to stars in our Galaxy: the Galaxy is gravitationally bound and therefore is impervious to the universal expansion. Nor does the distance-redshift relation apply to the separations between our Galaxy and nearby galaxies (the "local group"); gravitational attraction between our Galaxy and its neighbors is so great it perturbs their motions substantially away from universal expansion. Only on

Box 29.4 (*continued*)

scales large enough to include many galaxies (scales where each galaxy or cluster of galaxies can be thought of as a "grain of dust," with the grains distributed roughly homogeneously)—only on such large scales should the distance-redshift relation hold with good accuracy. But it is very difficult to obtain reliable measurements of the distances ℓ to galaxies that are so far away!

C. *Procedure by which H_o has been measured* [*Sandage and Tamman, as summarized in Sandage* (*1972a*)]:

1. Cepheid variables are pulsating stars with pulsation periods (as measured by oscillations in light output) that are very closely correlated with their luminosities L—or, equivalently, with their absolute (bolometric) magnitudes, M:

$$M \equiv \begin{pmatrix}\text{apparent magnitude star would have were it at a} \\ \text{distance of 10 parsecs} = 32.6 \text{ light years}\end{pmatrix} \tag{1}$$
$$= -2.5 \log_{10}(L/3.0 \times 10^{35} \text{ erg sec}^{-1})$$

[see equation (29.20).] By measurements within our Galaxy, astronomers have obtained the "period-luminosity relation" for cepheid variables.

2. Cepheid variables are clearly visible in galaxies as far away as $\sim$4 Mpc (4 Megaparsecs $\equiv 4 \times 10^6$ parsecs). In each such galaxy one measures the periods of the cepheids; one then infers their absolute magnitudes M from the period-luminosity relation; one measures their apparent magnitudes m; and one then calculates their distances ℓ from Earth using the relation

$$m - M = 5 \log_{10}(\ell/10 \text{ pc}). \tag{2}$$

By this means one obtains the distances ℓ to all galaxies within $\sim$4 Mpc of our own. Unfortunately, such galaxies are too close to participate cleanly in the universal expansion. (They include only the "local group," the "M81 group," and the "south polar group.") Thus, one must push the distance scale out still farther before attempting to measure H_o.

3. Galaxies of types Sc, Sd, Sm, and Ir within $\sim$4 Mpc contain huge clouds of ionized hydrogen, which shine brightly in "Hα light." These clouds, called "H II regions," exhibit a very tight correlation between diameter D of the H II region and luminosity L of the galaxy (or, equivalently, between D and absolute magnitude of galaxy, M). In fact, for a given galaxy luminosity L, the fractional spread in H II diameters is $\sigma(\Delta D/D) \approx 0.12$. Using (a) the distances ($\lesssim$4 Mpc) to these galaxies as determined via cepheid variables, (b) the apparent magnitudes of the galaxies, and (c) the angular diameters of H II regions in the galaxies, one calculates

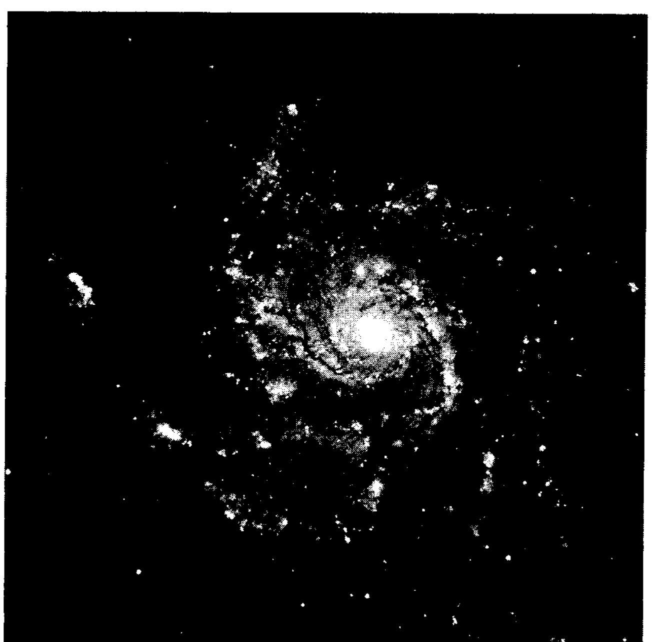
the actual H II diameters D and galaxy luminosities L, and thereby obtains the "diameter-luminosity relation" $D(L)$.

4. H II regions are large enough to be seen clearly in galaxies as far away as $\sim$60 Mpc. By measuring the H II angular diameters $\alpha = D/\ell$ and galaxy apparent (bolometric) magnitudes

$$m = -2.5 \log_{10}\left(\frac{L/4\pi\ell^2}{2.52 \times 10^{-5}\,\text{erg cm}^{-2}\,\text{sec}^{-1}}\right), \tag{3}$$

and by combining with the diameter-luminosity relation, one obtains the distances ℓ to all galaxies of types Sc, Sd, Sm, and Ir which possess H II regions and lie within $\sim$60 Mpc of Earth. Unfortunately, this is *still* not far enough away for local motions to be negligible compared with the universal expansion.

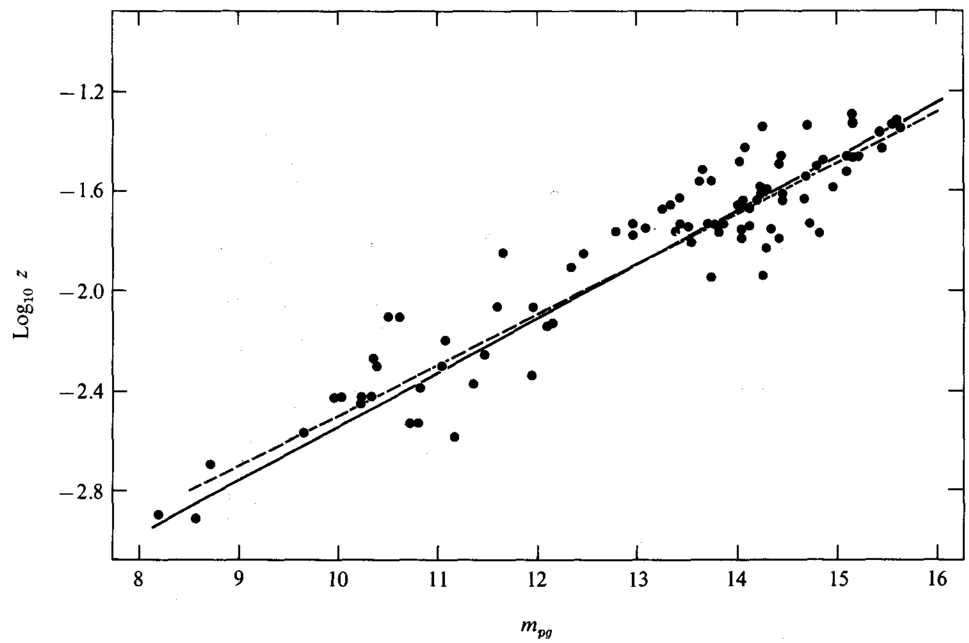
5. Within $\sim$60 Mpc reside enough galaxies of type Sc I for one to discover that their luminosities (absolute magnitudes) are rather constant (difference in L from one Sc I galaxy to another $\lesssim$50 per cent). Using the distances to such Sc I galaxies, as measured via H II regions, and using measurements of their apparent magnitudes, one calculates their universal absolute magnitude (measured photographically) to be $M_{pg} = -21.2$.



The Sc I galaxy M101 at a distance $\ell \sim 3$ Mpc from Earth, as photographed with the 200-inch telescope. (Courtesy of Hale Observatories)

Box 29.4 ***(continued)***

6. One then examines all known Sc I galaxies with distances greater than ~70 Mpc. For each, one measures the apparent magnitude and compares it with the universal absolute magnitude to obtain the distance ℓ from Earth. And for each, one measures the redshift $z = \Delta\lambda/\lambda$ of the spectral lines. From the resulting redshift-distance relation—and taking into account the statistical uncertainties in all steps leading up to it—Sandage and Tamman (work carried out in 1965–1972) obtain the value $H_o = dz/d\ell = 55 \pm 7$ (km/sec) Mpc^{-1} $= 1/[(18 \pm 2) \times 10^9$ years]. [For a review see Sandage (1972a).] The quoted error is purely statistical. Systematic errors are surely larger—but they almost surely do not exceed a factor 2 [i.e., $30 < H_o < 110$ (km/sec) Mpc^{-1}].



Magnitude-redshift relation for Sc I galaxies at distances $\gtrsim 70$ Mpc. Solid line is a least-squares fit to the data; dotted line has the theoretical slope of 5. [From Sandage and Tamman.]

II. *Deceleration Parameter, q_o.*

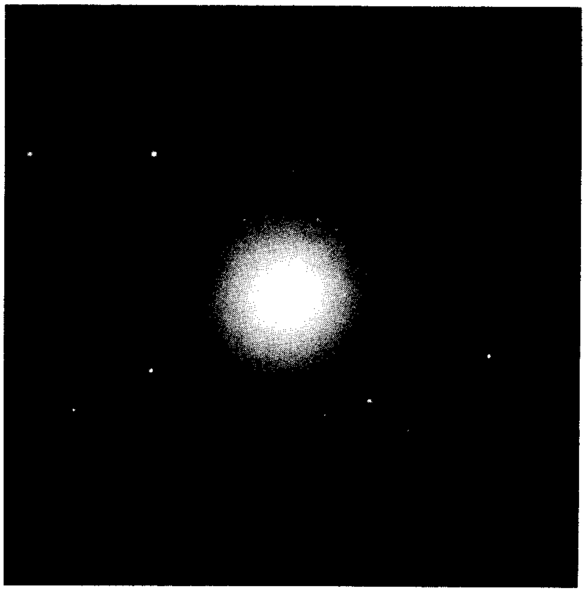
A. *Objective:* To measure the constant q_o by comparing observational data with the magnitude-redshift relation:

$$m = 5 \log_{10} z + 1.086(1 - q_o)z + O(z^2) - 2.5 \log_{10} L + \text{const.} \tag{4}$$

[Note: This relation is valid even if the cosmological constant is nonzero, i.e., even if $\sigma_o \neq q_o$. Dependence on σ_o occurs only at $O(z^2)$ and higher.]

B. *Key Difficulty:* One must use data for objects with the same absolute luminosity L ("standard candles"). But one cannot measure L at distances great enough for the effects of q_o to show up.

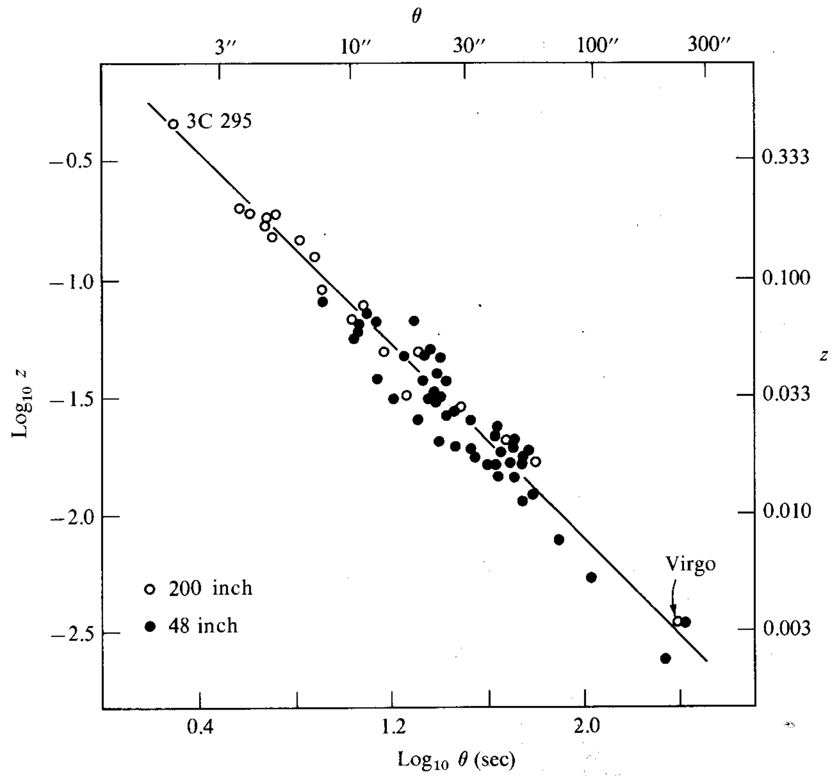
C. *The Search for a Standard Candle:* One obvious choice for the standard candle would be Sc I galaxies, since they were found to all have nearly the same L (see above). But they are not bright enough to be seen at distances great enough for effects of q_o to show up. An alternative choice, quasars, are bright enough to be seen at very large redshifts (z as large as ~ 3). But their absolute luminosities have enormous scatter—or so one infers from the failure of quasars to fall on a straight line, even at small z, in the magnitude-redshift diagram. The best choice is the brightest type of object that has small scatter in L. Sandage (1972a,b,c) chooses the brightest galaxy in "recognized regular clusters of galaxies." Such clusters are composed predominately of E-type galaxies, and the brightest members are remarkably similar from one cluster



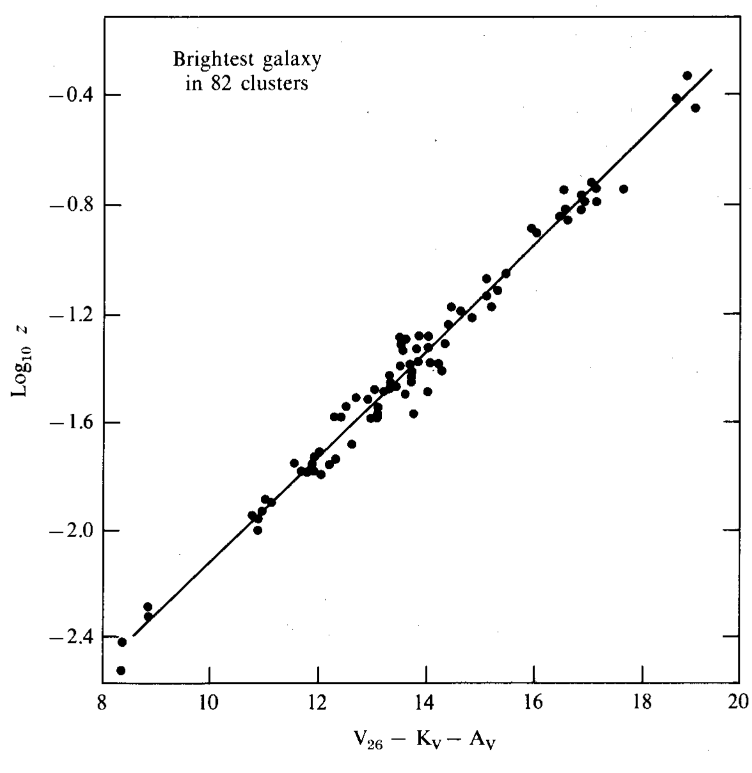
The E-type galaxy M87 at a distance $\ell \sim 11$ Mpc from Earth, as photographed with the 200-inch telescope. (Courtesy of Hale Observatories)

to another (scatter in L is ~ 25 per cent). The similarity shows up in their spectra and in the very precise straight lines they give when one plots angular diameter versus redshift (next page), or apparent magnitude versus redshift (next page), or angular diameter versus apparent magnitude.

Box 29.4 (continued)



Angular diameter versus redshift for brightest galaxy in recognized regular clusters. From Sandage (1972a,b). [These data are not sufficiently precise to yield useful information about q_o and σ_o; but improvements in 1973 may bring the needed precision; see §29.5.]



Magnitude versus redshift for brightest galaxy in recognized regular clusters. $V_{26} - K_V - A_V$ is the apparent magnitude with certain corrections taken into account. The line plotted corresponds to $\sigma_o = q_o = 1$ (straight line of slope 5). From Sandage (1972a).

D. *Procedure by which q_o has been measured* [*Sandage* (*1972a,c*)]:

1. Data on magnitude versus redshift have been gathered for the brightest galaxy in 82 recognized regular clusters (see above).
2. The data, when fitted with a straight line, show a slope of

$$dm/d\log_{10} z = 5.150 \pm 0.268 \text{ (rms)}, \tag{5}$$

by comparison with a theoretical slope of 5.

3. The data, when fitted to the theoretical relation

$$m = 5 \log_{10} z + 1.086(1 - q_o)z + O(z^2) + \text{const}, \tag{6}$$

[with the correct $O(z^2)$ and higher terms included; see equations (29.29), (29.32), and (29.34)] yield

$$\left.\begin{aligned} q_o &= 1 \pm 0.5 \text{ (one-sigma)} \\ &= 1 \pm 1 \quad \text{(two-sigma)} \end{aligned}\right\} \text{ if } \sigma_o = q_o(\Lambda = 0). \tag{7}$$

The data are inadequate to determine σ_o and q_o simultaneously. [The $O(z^2)$ terms, which depend on σ_o, play a significant role in the fit to the data. For a graphical depiction of their theoretical effects see Figure 2 of Refsdal *et. al* (1967).]

E. *Evolutionary uncertainties*

1. Sandage's fit of data to theory assumes that the luminosities of his "standard candles" are constant in time. If, because of evolution of old stars and formation of new ones, his galaxies were to dim by 0.09 magnitudes per 10^9 years, then galaxies 10^9 light-years away, which one sees as they were 10^9 years ago, would be 0.09 magnitudes brighter intrinsically than identical nearby galaxies. Correction for this effect would lower the most probable value of q_o from 1 to 0 [Sandage (1972c)].
2. Knowledge of the evolution of galaxies in 1972 is too rudimentary to confirm or rule out such an effect. [See references cited by Sandage (1972c).]

2

Box 29.5 EDWIN POWELL HUBBLE
November 20, 1889, Marshfield Missouri—
September 28, 1953, Pasadena, California

Edwin Hubble, at age 24, earned a law degree from Oxford University and began practicing law in Louisville, Kentucky. After a year of practice he became fed up and, in his own words, "chucked the law for astronomy, and I knew that even if I were second-rate or third-rate it was astronomy that mattered." He chose the University of Chicago and Yerkes Observatory as the site for his astronomy education, and three years later (1917) completed a Ph.D. thesis on "Photographic Investigations of Faint Nebulae."

When Hubble entered astronomy, it was suspected that some nebulae lie outside the Galaxy, but the evidence was exceedingly weak. During the subsequent two decades, Hubble, more than anyone else, was responsible for opening to man's purview the extragalactic universe. Working with the 60-inch and 100-inch telescopes at Mount Wilson, Hubble developed irrefutable evidence of the extragalactic nature of spiral nebulae, elliptical nebulae, and irregular nebulae (now called galaxies). He devised the classification scheme for galaxies which is still in use today. He systematized the entire subject of extragalactic research: determining distance scales, luminosities, star densities, and the peculiar motion of our Galaxy; and obtaining extensive evidence that the laws of physics outside the Galaxy are the same as near Earth (in Hubble's words: "verifying the principle of the uniformity of nature"). He discovered and quantified the large-scale homogeneity of the universe. And—his greatest triumph of all!—he discovered the expansion of the universe.

The details of Hubble's pioneering work are best sketched in his own words:

"Extremely little is known of the nature of nebulae; and no classification has yet been suggested. . . . The agreement [between the velocity of escape from a spiral nebula and that from our galaxy] is such as to lend some color to the hypothesis that the spirals are stellar systems at distances to be measured often in millions of light years."

(1920; Ph.D. THESIS; PUBLICATION DELAYED 3 YEARS BY WORLD WAR I)

This box is based largely on the biography of Hubble by Mayall (1970).

"The present investigation [using Cepheid variables for the first time as an indicator of distances beyond the Magellanic clouds] identifies NGC 6822 as an isolated system of stars and nebulae of the same type as the Magellanic clouds, although somewhat smaller and much more distant. A consistent structure is thus reared on the foundation of the Cepheid criterion, in which the dimensions, luminosities, and densities, both of the system [NGC 6822] as a whole and its separate members, are of orders of magnitude which are thoroughly familiar. The distance is the only quantity of a new order. The principle of the uniformity of nature thus seems to rule undisturbed in this remote region of space."

(1925)

"Critical tests made with the 100-inch reflector, the highest resolving power available, show no difference between the photographic images of the so-called condensations in Messier 33 and the images of ordinary galactic stars. . . . The period-luminosity relation is conspicuous among the thirty-five Cepheids and indicates a distance about 8.1 times that of the Small Magellanic Cloud. Using Shapley's value for the latter, the distance of the spiral is about 263,000 parsecs."

(1926a)

"[To the present paper (1926b)] is prefaced a general classification of nebulae . . . the various types [of extragalactic nebulae] are homogeneously distributed over the sky. . . . The data are now available for deriving a value for the order of the density of space. This is accomplished by means of the formulae for the numbers of nebulae to a given limiting magnitude and for the distance in terms of the magnitude. [The result is]

$$\rho = 1.5 \times 10^{-31} \text{ grams per cubic centimeter.}$$

This must be considered as a lower limit, for loose material scattered between the systems is entirely ignored. The mean density of space can be used to determine the dimensions of the finite but boundless universe of general relativity . . .

$$R = \frac{c}{\sqrt{4\pi k}} \frac{1}{\sqrt{\rho}} = \cdots = 2.7 \times 10^{10} \text{ parsecs."}$$

(1926b)

"The data . . . indicate a linear correlation between distances and velocities [for extragalactic nebulae]. Two solutions have been made, one using the 24 nebulae individually, the other combining them into 9 groups according the proximity in direction and distance. The results are . . . 24 objects: $K = 465 \pm 50$ km/sec per 10^6 parsecs; 9 groups: $K = 513 \pm 60$ km/sec per 10^6 parsecs. . . . The outstanding feature, however, is the possibility that the velocity-distance relation may represent the de Sitter effect, and hence that numerical data may be introduced into discussions of the general curvature of space."

(1929)*

*Hubble's value of K (the "Hubble constant") was later revised downward by the work of Baade and Sandage; see section titled *The Hubble Time* in Box 27.1.

2

Box 29.5 (*continued*)

"The velocity-distance relation is re-examined with the aid of 40 new velocities. . . . The new data extend out to about eighteen times the distance available in the first formulation of the velocity-distance relation, but the form of the relation remains unchanged except for [Shapley's 10 per cent] revision of the unit of distance."

(1931), WITH M. L. HUMASON

"Many ways of producing such effects [redshifts in extragalactic nebulae] are known, but of them all, only one will produce large redshifts without introducing other effects which should be conspicuous but actually are not found. This one known permissible explanation interprets redshifts as due to actual motion away from the observer."

(1934a)

"We now have a hasty sketch of some of the general features of the observable region as a unit. The next step will be to follow the reconnaissance with a survey—to repeat carefully the explorations with an eye to accuracy and completeness. The program, with its emphasis on methods, will be a tedious series of successive approximations."

(1934b)

Most of the remainder of Hubble's career was dedicated to this "tedious series of successive approximations." Shortly before Hubble's death the 200-inch telescope went into operation at Palomar Mountain; and Hubble's student, Alan Sandage, began using it in a continuation of Hubble's quest into the true nature of the universe. (See Box 29.4).

EXERCISES

Exercise 29.4. m(z) DERIVED USING STATISTICAL PHYSICS

Derive the magnitude-redshift relation using a statistical description of the photon distribution [cf. eq. (22.49) and associated discussion].

Exercise 29.5. DOPPLER SHIFT VERSUS COSMOLOGICAL REDSHIFT

(a) Consider, in flat spacetime, a galaxy moving away from the Earth with velocity v, and emitting light that is received at Earth. Let the distance between Earth and galaxy, as measured in the Earth's Lorentz frame at some specific moment of emission, be r; and let the Doppler shift of the radiation when it is eventually received be $z = \Delta\lambda/\lambda$. Show that the flux of energy S received at the Earth is related to the galaxy's intrinsic luminosity L by

$$S = \frac{L}{4\pi r^2(1+z)^4}. \tag{29.37}$$

[Track-2 readers will find it most convenient to use the statistical formalism of equation (22.49).]

(b) Compare this formula for the flux with formula (29.27), where the redshift is of cosmological origin. Why is the number of factors of $1 + z$ different for the two formulas? [Mathematical answer: equation (6.28a) of Ellis (1971).]

§29.5. SEARCH FOR "LENS EFFECT" OF THE UNIVERSE

Curved space should act as a lens of great focal length. The curving of light rays has little effect on the apparent size of nearby objects. However, distant galaxies—galaxies from a quarter of the way up to halfway around the universe—are expected to have greatly magnified angular diameters [Klauder, Wakano, Wheeler, and Willey (1959)]. To see a normal galaxy at such a distance by means of an optical telescope seems out of the question. However, radio telescopes resolve features in quasistellar sources and other radiogalaxies at redshifts of $z = 2$ or more. Moreover, paired radio telescopes at intercontinental distances (for example, Goldstone, California, and Woomera, Australia) resolve distant sources to better than $0''.001$ or 4.8×10^{-9} radians or 15 lightyears for an object at a distance of 3×10^9 lightyears (Euclidean geometry temporarily being assumed). A radio telescope in space paired with a radio telescope on earth will be able to do even better on angular resolution. Will one be able to find any fiducial distance characteristic of any one class of objects that will serve as a natural standard of length, for very great distances ($z = 2$ to $z = 3$) as well as for galaxies closer at hand? Perhaps not. However, it would seem unwise to discount this possibility, with all the advantages it would bring, in view of the demonstrated ability of skilled observers to find regularities elsewhere where one had no right to expect them in advance.

The hope for a fiducial length in distant objects

Let L denote the actual length of a fiducial element (if any be found) in a galaxy; and let $\delta\theta$ (radians!) denote the apparent length of the object, idealized as perpendicular to the line of sight, as seen by the observer. The ratio of these two quantities defines the "angle effective distance" of the source,

Angle effective distance defined

$$(\text{angle effective distance}) = r_{\text{aed}} = L/\delta\theta. \qquad (29.38\text{a})$$

In flat space and for objects with zero relative velocity, this distance is to be identified with the actual distance, r, to the source or with the actual time of flight, t, of light from source to observer. The situation is changed in an expanding universe.

To calculate the angle effective distance as a function of redshift, place the Earth (receiver) at $\chi_r = 0$; and place the object under study (emitter) at χ_e. Let the fiducial length L lie on the sphere at χ_e (perpendicular to line of sight), and let it run from θ_e to $\theta_e + \delta\theta$ [one end of fiducial element at $(\chi_e, \theta_e, \phi_e)$; other at $(\chi_e, \theta_e + \delta\theta, \phi_e)$]. Then

$$L = a(t_e)\Sigma(\chi_e)\,\delta\theta,$$

and

$$r_{\text{aed}} = a(t_e)\Sigma(\chi_e) = [a(t_r)\Sigma(\chi_e - \chi_r)]a(t_e)/a(t_r);$$

i.e. [see equation (29.28), with χ_r and χ_e reversed],

$$r_{\text{aed}} = R/(1 + z). \qquad (29.38\text{b})$$

Angle effective distance as function of redshift

Here R is given as a function of redshift of source, z, and cosmological parameters H_o, q_o, σ_o, by equations (29.32) in general, or by (29.33) if $\Lambda = 0$. [Equation (29.38b) is modified if the beam preferentially traverses regions of low mass density ("vacuum between the galaxies"); see equation (22.37) and Gunn (1967).]

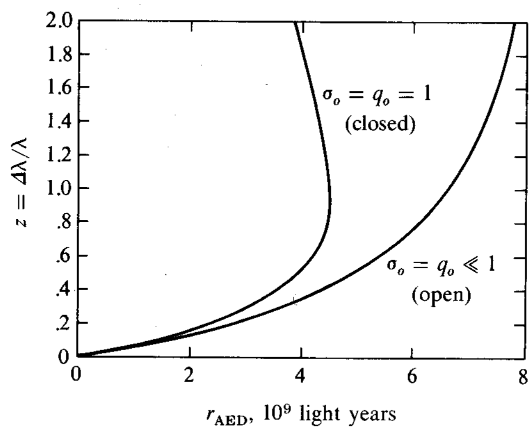


Figure 29.2.
Angle effective distance versus redshift for two typical cosmological models—one open ($0 < \sigma_o = q_o \ll 1$); the other closed ($\sigma_o = q_o = 1$); both with zero cosmological constant; both with $H_o^{-1} = 18 \times 10^9$ lyr.

Angle effective distance as a tool for determining whether universe is closed

Figure 29.2 shows angle effective distance as a function of redshift for a few selected choices of the relevant parameters. It is evident that the angle effective distance has a maximum for a redshift roughly of the order $z \sim 1$, provided that the universe is closed. However, there is a big difference if the universe is open (Figure 29.2). The rapid improvements taking place in radio astronomy make increasingly attractive the possibility it provides for testing whether the universe is closed, as Einstein argued it should be [Einstein (1950), pp. 107–108]. Moreover, even with optical telescopes, in 1973 one may be on the verge of measuring q_o by studies of angle effective distance: preliminary studies [Sandage (1972b)] suggest that the optical size of the brightest E-type galaxies may be a usable fiducial length.

§29.6. DENSITY OF THE UNIVERSE TODAY

Measurements of mean mass density of universe:

(1) luminous matter in galaxies

It is exceedingly difficult to measure the mean density ρ_{mo} of the universe today. A large amount of matter may be in forms that astronomers have not yet managed to observe (intergalactic matter, black holes, etc.). Therefore, the best one can do is to add up all the luminous matter in galaxies and regard the resulting number as a lower limit on ρ_{mo}. Even adding up the luminous matter is a difficult and risky task, so difficult that even today no analysis is more definitive than the classic work of Oort (1958). [See, however, the very detailed review of the problem in Chapter 4 of Peebles (1971)]. Oort's result is

$$\rho_{\text{luminous matter}} \sim (2 \times 10^{-31}\ \text{g/cm}^3)(H_o/55\ \text{km sec}^{-1}\ \text{Mpc}^{-1})^2, \quad (29.39)$$

corresponding to

$$\sigma_o \gtrsim 0.02 \qquad \text{(independent of the value of } H_o\text{)}. \quad (29.40)$$

As an example (albeit an atypical one) of the danger inherent in any such estimate, Oort points to the Virgo cluster of galaxies. If the Virgo cluster is not gravitationally bound, then its ~2,500 galaxies will go flying apart, destroying any semblance of a cluster, in about one billion years. If it is gravitationally bound, then the mean velocity of its galaxies relative to each other, when combined with the virial theorem, yields an estimate of the cluster's total mass. That estimate is 25 times larger than the value one gets by Oort's method of adding up the luminous mass of the cluster.

(2) matter in intergalactic space

Although one has no *definitive* evidence for or against large amounts of matter (enough to close the universe) in intergalactic space, one has tentative indirect limits: (1) If $\Lambda = 0$ (in accord with Einstein), then $\sigma_o = q_o$; so Sandage's value of $q_o \lesssim 1$—stretched to $q_o < 10$ under the most wild of assumptions about galaxy evolution—implies

$$\rho_{ig} < 10^{-28}\ \mathrm{g/cm^3} \qquad (\sigma_o = q_o < 10).$$

(2) Gott and Gunn (1971) point out that, if the density of gas in intergalactic space were $\gtrsim 10^{-30}$ g/cm^3 (i.e., if σ_o were $\gtrsim 0.1$), one would expect gas falling into the Coma cluster of galaxies to form a shock wave, which would emit large amounts of X-rays. From the current X-ray observations, one can place a limit on the amount of such infalling matter—and therefrom a limit

$$\rho_{ig} \lesssim 10^{-30}\ \mathrm{g/cm^3} \qquad (\sigma_o \lesssim 0.1)$$

on the density of gas in intergalactic space. But these limits, like others obtained in other ways [see Chapter 4 of Peebles (1971) for a review] are far from definitive; they depend too much on theoretical calculations to make one feel fully comfortable.

§29.7. SUMMARY OF PRESENT KNOWLEDGE ABOUT COSMOLOGICAL PARAMETERS

Summary of observational parameters of universe

The best data available in 1973 [equations (29.18), (29.36), (29.40)] reveal

$$\begin{aligned} H_o^{-1} &= (18 \pm 2) \times 10^9 \text{ years}, \\ q_o = 1 \pm 0.5 \text{ (one-sigma)} &\qquad \text{if } \sigma_o = q_o (\Lambda = 0), \\ \sigma_o &\gtrsim 0.02, \end{aligned} \tag{29.41}$$

for the observational parameters of the universe. These numbers are inadequate to reveal whether the universe is closed or open, and whether it will continue to expand forever or will eventually slow to a halt and recontract.

Some quantitative triumphs of cosmology

If one is disappointed in this lack of knowledge, one can at least be consoled by the following. (1) There is excellent agreement between theory and observation for the linear (low-z) parts of the distance-redshift, magnitude-redshift, and angular diameter-redshift relations (Box 29.4). (2) There is remarkably good agreement between (a) the age of the universe (18 billion years if $q_o = \sigma_o \ll 1$; 12 billion years if $q_o = \sigma_o = \frac{1}{2}$) as calculated from the measured value of H_o; (b) the ages of the oldest stars ($\sim 10 \times 10^9$ years) as calculated by comparing the theory of stellar evolution with the properties of the observed stars; (c) the time ($\sim$9 billion years) since nucleosynthesis of the uranium, thorium, and plutonium atoms that one finds on Earth, as calculated from the measured relative abundances of various nucleides; and (d) the ages (4.6 billion years) of the oldest meteorites and oldest lunar rock samples, as calculated from measured relative abundances of other nucleides. For further detail see, e.g., Sandage (1968, 1970), Wasserburg *et al.* (1969), Wasserburg and Burnett (1968), and Fowler (1972).

The bright prospects for observational cosmology

(3) Observations of the cosmic microwave radiation and measurements of helium abundance are now capable of giving direct information about physical processes in the universe at redshifts $z \gg 1$ (Chapter 28). (4) One may yet find "fiducial lengths" in radio sources, visible out to $z \gtrsim 1$, with which to measure q_o and σ_o by the angle-effective-distance method (§29.5). (5) The enigmas of the nature of quasars and of their peculiar distribution with redshift (great congregation at $z \sim 2$; absence at $z \gtrsim 3$) may yet be cracked and may yield, in the process, much new information about the origin of structure in the universe (Box 28.1). (6) The next decade may well bring as many great observational surprises, and corresponding new insights, as has the last decade.

EXERCISES

Exercise 29.6. SOURCE COUNTS

Suppose that one could find (which one cannot) a family of light or radio sources that (1) are all identical with intrinsic luminosities L, (2) are distributed uniformly throughout the universe, and (3) are born at the same rate as they die so that the number in a unit comoving coordinate volume is forever fixed.

(a) Show that the number of such sources $N(z)$ with redshifts less than z, as observed from Earth today, would be

$$N(z) = (\text{constant}) \cdot z^3\left[1 - \frac{3}{2}(1 + q_o)z + O(z^2)\right]. \tag{29.42}$$

(b) Show that the number of sources $N(S)$ with fluxes greater than S as observed at Earth today would be

$$N(S) = (\text{constant}) \cdot \underbrace{\left(\frac{LH_o{}^2}{4\pi S}\right)}_{= z^2 + O(z^3)}{}^{3/2}\left[1 - \underbrace{3\left(\frac{LH_o{}^2}{4\pi S}\right)^{1/2}}_{\left[\substack{\text{first-order correction} \\ \text{independent of } q_o \text{ and } \sigma_o}\right]} + O\left(\frac{LH_o{}^2}{4\pi S}\right)\right] \tag{29.43}$$

[*Answer:* See §15.7 of Robertson and Noonan (1968).]

Exercise 29.7. COSMIC-RAY DENSITY (Problem devised by Maarten Schmidt)

Suppose the universe has contained the same number of galaxies indefinitely into the past. Suppose further that the cosmic rays in the universe were created in galaxies and that a negligible fraction of them have been degraded or lost since formation. Derive an expression for the average density of energy in cosmic rays in the universe today in terms of: (1) the number density of galaxies, N_o, today; and (2) the nonconstant rate, dE/dz, at which the average galaxy created cosmic-ray energy during the past history of the universe. [At redshift z in range dz, the average galaxy liberates energy $(dE/dz)\,dz$ into cosmic rays.]

Exercise 29.8. FRACTION OF SKY COVERED BY GALAXIES

Assume that the redshifts of quasars are cosmological. Let the number of galaxies per unit physical volume in the universe today be N_o, and assume that no galaxies have been created or destroyed since a redshift of $\gtrsim 7$. Let D be the average angular diameter of a galaxy. Calculate the probability that the light from a quasar at redshift z, has passed through at least one intervening galaxy during its travel to Earth. [For a detailed discussion of this problem, see Wagoner (1967).]

2

CHAPTER 30

ANISOTROPIC AND INHOMOGENEOUS COSMOLOGIES

§30.1. WHY IS THE UNIVERSE SO HOMOGENEOUS AND ISOTROPIC?

This chapter is entirely Track 2.
The main text requires no special preparation, although Chapters 27–29 would be helpful.
Box 30.1 contains more technical sections: ideal preparation for it would be Chapters 4, 9–14, 21, and 27–29, plus §25.2; minimal preparation would be exercises 9.13, 9.14, and 25.2, Chapter 21 through §21.8, and §§27.8, 27.11, and 29.2.
Chapter 30 is not needed as preparation for any later chapter.

The last three chapters studied the Friedmann cosmological models and the relatively satisfactory picture they give of the universe and its evolution. This chapter describes less simplified cosmological models, and uses them to begin answering the question, "Why are the very simple Friedmann models satisfactory?" This question is intended to probe more deeply than the first, obvious answer—namely, that the models are satisfactory because they do not contradict observations. Accepting the agreement with observations, we want to understand *why the laws of physics should demand (rather than merely permit) a universe that is homogeneous and isotropic to high accuracy on large scales*. Because this question cannot be answered definitively in 1972, many readers will prefer to omit this chapter on the first reading and return to it only after they have surveyed the major results in other areas such as black holes (Chapter 33), gravitational waves (Chapters 35–37), and solar-system experiments (Chapter 40).

Motivation for studying inhomogeneous and anisotropic cosmologies: Why is universe so uniform?

The approach described here to the question "Why is the universe so highly symmetric?" is to ask Einstein's equations to describe what would have happened if the universe had started out highly irregular.

The first step in this approach is to ask what would have happened if the universe had started a little bit irregular. This problem can be tackled by analyzing small perturbations away from the high symmetry of the Friedmann models. Such an analysis is most fruitful in its discussion of the beginnings of galaxy formation, and in its ability to relate small upper limits on the present-day anisotropy of the microwave background radiation to limits on density and temperature irregularities that might have existed ten billion years ago, when the radiation was emitted. These studies are described so well in the book by Zel'dovich and Novikov (1974) [see also Field (1973), Peebles (1969), Peebles and Yu (1970), Jones and Peebles (1972), and references cited therein] that we omit them here.

Another approach is to allow large deviations from the symmetry of the Friedmann universes, but to put the asymmetries into only a few degrees of freedom.

§30.2. THE KASNER MODEL FOR AN ANISOTROPIC UNIVERSE

Kasner metric: an example of an anisotropic model universe

The prototype for cosmological models with great asymmetry in a few degrees of freedom is the Kasner (1921a) metric,

$$ds^2 = -dt^2 + t^{2p_1}\,dx^2 + t^{2p_2}\,dy^2 + t^{2p_3}\,dz^2, \tag{30.1}$$

which was first studied as a cosmological model by Schücking and Heckmann (1958). In this metric the p_i are constants satisfying

$$p_1 + p_2 + p_3 = (p_1)^2 + (p_2)^2 + (p_3)^2 = 1. \tag{30.2}$$

Each $t =$ constant hypersurface of this cosmological model is a flat three-dimensional space. The world lines of constant x, y, z are timelike geodesics along which galaxies or other matter, treated as test particles, can be imagined to move. This model represents an expanding universe, since the volume element

$$\sqrt{-g} = \sqrt{{}^{(3)}g} = t$$

is constantly increasing. But it is an anisotropically expanding universe. The separation between two standard (constant x, y, z) observers is $t^{p_1}\,\Delta x$ if only their x-coordinates differ. Thus, distances parallel to the x-axis expand at one rate, $\ell_1 \propto t^{p_1}$, while those along the y-axis can expand at a different rate, $\ell_2 \propto t^{p_2}$. Most remarkable perhaps is the fact that along one of the axes distances contract rather than expand. This contraction shows up mathematically in the fact that equations (30.2) require one of the p's, say p_1, to be nonpositive:

$$-\frac{1}{3} \leq p_1 \leq 0. \tag{30.3}$$

As a consequence, in a universe of this sort, if black-body radiation were emitted at one time t and never subsequently scattered, later observers would see blue shifts near one pair of antipodes on the sky and red shifts in most other directions. In terms of this example, the fundamental cosmological question is why the Friedmann metrics should be a more accurate approximation to the real universe than is this Kasner metric.

§30.3. ADIABATIC COOLING OF ANISOTROPY

Kasner model with matter becomes isotropic in "old age"

In seeking an answer, ask a question. Ask, in particular, what would become of a universe that starts out near $t = 0$ with a form described by the Kasner metric of equation (30.1). This metric is an exact solution of the vacuum Einstein equation $\boldsymbol{G} = 0$. It approximates a situation where the matter terms in the Einstein equations are negligible by comparison with typical non-zero components of the Riemann tensor. Schücking and Heckmann (1958) give solutions with matter included as a pressureless fluid. In this situation, the curvature of empty spacetime dominates both the geometry and the expansion rate at early times, $t \longrightarrow 0$; but after some characteristic time t_m the matter terms become more important, and the metric reduces asymptotically to the homogeneous, isotropic model with $k = 0$.

This example illustrates the possibility that the universe might achieve a measure of isotropy and homogeneity in old age, even if it were born in a highly irregular state. Whether the symmetry of our universe can be explained along these lines is not yet clear in 1972. The model universe just mentioned is only a hint, especially since the critical parameter t_m can be given any value whatsoever.

Anisotropy energy

The standard Einstein general-relativity physics of this model can be described in other language (Misner, 1968) by ascribing to the anisotropic motions of empty spacetime an "effective energy density" ρ_{aniso}, which enters the G_{00} component of the Einstein equation on an equal footing with the matter-energy density, and thereby helps to account for the expansion of the universe:

$$H^2 = \left(\frac{1}{3}\frac{d}{dt}\ln\sqrt{{}^{(3)}g}\right)^2 = \frac{8\pi}{3}(\rho_{\text{aniso}} + \rho_{\text{matter}}). \tag{30.4}$$

The anisotropy energy density is found to have an equation of state

$$\rho_{\text{aniso}} \propto {}^{(3)}g^{-1} = (\text{volume})^{-2},$$

while

$$\rho_{\text{matter}} \propto {}^{(3)}g^{-\gamma/2} = (\text{volume})^{-\gamma}.$$

For pressureless matter $\gamma = 1$; for a radiation fluid $\gamma = 4/3$; for a nonrelativistic ideal gas $\gamma = 5/3$).

Adiabatic cooling of anisotropy

This arrangement of the Einstein equation allows one to think of the anisotropy motions as being adiabatically cooled by the expansion of the universe, just as the thermal motions of an ideal gas would be. Since the adiabatic index for homogeneous anisotropy is $\gamma = 2$, the anisotropy will be the dominant source of "effective energy" in a highly compressed state, whereas the matter will dominate in an expanded state.

§30.4. VISCOUS DISSIPATION OF ANISTROPY

The model universe sketched above can be further elaborated by introducing dissipative mechanisms that convert anisotropy energy into thermal energy. Suppose that such an anisotropic universe were filled at one time with thermal radiation. If the radiation were collisionless or nearly so, the quanta moving parallel to the contracting x-axis would get blueshifted and would develop an energy distribution corresponding to a high temperature. The quanta moving parallel to the other (expanding) axes would be redshifted to an energy distribution corresponding to a low temperature. Any collisions taking place between these two systems of particles would introduce a "thermal contact" between them, and would transfer energy from the hot system to the cold one, with a corresponding large production of entropy. This provides an irreversible dissipative process, which decreases ρ_{aniso} and increases $\rho_{\text{radiation}}$ relative to the values they would have had under conditions of adiabatic expansion. [For further details, see, e.g., Matzner and Misner (1972).]

It is possible that both the adiabatic cooling of anisotropy and the dissipation of anisotropy by its action on a gas of almost collisionless quanta have played significant roles in the evolution of our universe. In particular, neutrinos above 10^{10} K may have undergone sufficient ν-e scattering to have provided strong dissipation during the first few seconds of the life of the universe.

§30.5. PARTICLE CREATION IN AN ANISOTROPIC UNIVERSE

Adiabatic cooling and viscous dissipation might not be the chief destroyers of anisotropy in an expanding universe. More powerful still might be another highly dissipative process, which might occur at still earlier times, very near the initial "singularity." This is a process of particle creation which was first treated by DeWitt (1953), then explored by Parker (1966 and 1969) for isotropic cosmologies and finally by Zel'dovich (1970) in the present context of anisotropic cosmologies. In this process one again turns to the Kasner metric for the simplest example, but now quantum-mechanical considerations enter the picture. One realizes that not only would real quanta propagating in different directions be subject to red shifts and blue shifts, but that virtual quanta must be considered as well. Vacuum fluctuations (zero-point oscillations) entail a certain minimum number of virtual quanta, which are subject to the redshifting and blueshifting action of the strong gravitational fields. Virtual quanta that are blueshifted sufficiently violently can materialize as real particles, thanks to their energy gain. In this context "sufficiently violently" means *not adiabatically.*

Creation of particles by anisotropy of expansion

In an adiabatic expansion, the number of particles does not change, although the energy of each one does. This adiabatic limit is just the geometric-optics approximation to wave equations, which was discussed in §22.5. There one saw that, if spacetime were not flat on the scale of a wavelength, then the wave equation could not be replaced by a particle description with conserved particle numbers. Thus, the adiabatic limit (geometric-optics approximation) is violated in the conditions of high curvature near the singularity at the beginning of the universe.

By studying wave equations in the Kasner background metric, Zel'dovich and Starobinsky (1971) find quantitatively the consequences of the failure of the adiabatic approximation near the singularity. Classically, the amplitudes of waves at frequencies comparable to the Hubble constant for any given epoch increase faster than a simple blue-shift calculation would imply (amplification through parametric resonance). Quantum-mechanically, the same amplification, applied to zero-point oscillations, leads to the creation of particle-antiparticle pairs. The calculations indicate that this effect is very strong at the characteristic time $t_q = \sqrt{G\hbar/c^5} \simeq 10^{-43}$ sec. (All calculations performed thus far are inadequate when the effect becomes strong, thus for $t \lesssim t_q$).

For the creation of massless particles, it is essential that an anisotropically expanding universe be postulated (except for scalar particles, for which particle creation occurs already in the Friedmann universe, *unless* the particle satisfies the conformal-invariant wave equation). The isotropic Friedmann universes are all conformally flat, so that solutions of the wave equation for a field of zero rest mass can be given in terms of solutions for flat-space wave equations where there is no particle creation. There is some particle creation even in the isotropic Friedmann universe when the particle has finite rest mass and low energy. However, the particle-creation process normally uses anisotropy energy as the energy supply that it converts into radiation energy.

Anisotropy might have created the matter content of our universe, damping itself out in the process

The pioneering work by Parker and Zel'dovich suggests that one should study in detail cosmological models in which the initial conditions are a singularity, and in which quantum effects near the time $t = t_q$ dissipate all anisotropies and simultaneously give rise to the matter content of the model. This program of research, which is in its infancy, seems to require extrapolating laws of physics down to the very natural looking but preposterously small dimension $\sqrt{G\hbar/c^5} \simeq 10^{-43}$ sec, or equivalently $\sqrt{G\hbar/c^3} \sim 10^{-33}$ cm.

§30.6. INHOMOGENEOUS COSMOLOGIES

Inhomogeneous cosmological models:

(1) with spherical symmetry

The model universes considered above were all homogeneous although anisotropic. It is also crucial to study inhomogeneous cosmological models, in which the metric has a nontrivial dependence on the space coordinates. One class of such models is spherically symmetric universes, where the matter density, expansion rate, and all other locally measurable physical quantities have spherical symmetry about some preferred origin. Models of this sort were first considered by Lemaitre (1933a,b), Tolman (1934b), and Datt (1938), and were also treated by Bondi in 1947. These models provide a means for studying density perturbations of large amplitude.

(2) with (rather symmetric) gravitational waves

A recent tool is making it possible to study large-amplitude, spatially varying curvature perturbations of other symmetries; this tool is the Gowdy (1971, 1973) metrics. These metrics, which are exact solutions of the Einstein equations, represent closed universes with various topologies ($S^3, S^1 \times S^2, T^3$) containing gravitational waves. The wave form in these solutions is essentially arbitrary, but all the waves propagate along a single preferred direction and have a common polarization.

(3) near a singularity, with few or no symmetries

A rather different approach to understanding the behavior of inhomogeneous and anisotropic solutions of the Einstein equations has been developed by Khalatnikov, Lifshitz, and their colleagues. Rather than truncate the Einstein theory by limiting attention to specialized situations where exact solutions can be obtained, they have sought to study the widest possible class of solutions, but to describe their behavior only in the immediate neighborhood of the singularity. These studies give a greatly enhanced significance to some of the exact solutions, by showing that phenomena found in them are in fact typical of much broader classes of solutions.

Thus, in the first large class of solutions studied [Lifshitz and Khalatnikov (1963)], it was found that near the singularity solutions containing matter showed no features not already found in the vacuum solutions. Furthermore, space derivatives in the Einstein equations became negligible near the singularity in these solutions, with the consequence that a metric of the Kasner form [equation (30.1)] described the local behavior of spacetime near the singularity, but with a different set of p_i values possible at each point of the singular hypersurface. Subsequently, broadened studies of solutions near a singularity [Belinsky and Khalatnikov (1970)] showed that the mixmaster universe [Misner (1969b); Belinsky, Khalatnikov, and Lifshitz (1970)] is a still better homogeneous prototype for singularity behavior than the Kasner metric.

§30.7. THE MIXMASTER UNIVERSE

Mixmaster universe:

(1) "anisotropy oscillations" explained in terms of Kasner model

The simplest example of a mixmaster universe is described in Box 30.1. It shows how, near the singularity, the Kasner exponents p_i can become functions of time. The result is most simply described in terms of the Khalatnikov-Lifshitz parameter u:

$$\begin{aligned} p_1 &= -u/(1 + u + u^2), \\ p_2 &= (1 + u)/(1 + u + u^2), \\ p_3 &= u(1 + u)/(1 + u + u^2). \end{aligned} \tag{30.5}$$

As one extrapolates backward in time toward the singularity, one finds that the expansion rates in the three principal directions correspond to those of the Kasner metric of equation (30.1), with p_i values corresponding to some fixed u parameter. In these mixmaster models, however, the metric is not independent of the space coordinates (the spacelike hypersurfaces can, for instance, have the same 3-sphere topology as the closed Friedmann universes).

The Kasner-like behavior at fixed u can persist through many decades of volume expansion before effects of the spatial derivatives of the metric come into play. The role then played by the space curvature is brief and decisive. The expansion is converted from a type corresponding to a parameter value $u = u_0$ to a type corresponding to the value $u = -u_0$ (which is equivalent, under a relabeling of the axes, to the value $u = u_0 - 1$). Extrapolating still farther back toward the singularity, one finds a previous period with $u = u_0 - 2$. Throughout an entire sequence $u = u_0$, $u_0 - 1, u_0 - 2, u_0 - 3, \ldots$, with $u_0 \gg 1$, nearly the entire volume expansion is due to expansion in the 3-direction, whereas the 1- and 2-directions change very little, alternating at each step between expansion and contraction. Sufficiently far in the past, however, such a sequence leads to a value of u between 0 and 1. This value

2

can be interpreted as the starting point for another, similar sequence, through the transformation $u \longrightarrow 1/u$, which interchanges the names of axes 2 and 3.

The extrapolation of the universe's evolution back toward the singularity at $t = 0$ therefore shows an extraordinarily complex behavior, in which similar but not precisely identical sequences of behavior are repeated infinitely many times. In terms of a time variable which is approximately $\log(\log t^{-1})$, these behaviors are quasi-periodic. In the generic example to which the Khalatnikov-Lifshitz methods lead, one has a metric whose asymptotic behavior near the singularity is at each point of the singular hypersurface described by a mixmaster-type behavior, but with the principal axes of expansion changing their directions as well as their roles (as characterized by the u parameter) at each step, and with the mixmaster parameters spatially variable. [For more details see Belinsky, Lifshitz, and Khalatnikov (1971), and Ryan (1971, 1972).]

(2) as a prototype for generic behavior near singularities

It is not yet (1972) known whether there are important solutions or classes of solutions relevant to the cosmological problem, with asymptotic singularity behavior *not* described by the Khalatnikov-Lifshitz generic case. The difficulty in reaching a definitive assessment here is that Khalatnikov and Lifshitz use essentially local methods, confined to a single coordinate patch, whereas the desired assessment poses an essentially global question. The global approaches (described in Chapter 34) have not, however, provided any comparable description of the nature of the singularity whose necessity they prove. One attempt to bridge these differences in technique and content is the work by Eardley, Liang, and Sachs (1972).

Are there any other generic types of behavior near singularities?

(continued on page 815)

Box 30.1 THE MIXMASTER COSMOLOGY

The Mixmaster Cosmology is a valuable example. As described in §30.7, it shows a singularity behavior which illustrates most of the features of the most general examples known. In particular, it shows how properties of empty space reminiscent of an elastic solid become evident near the cosmological singularity.

The mathematical path to this example, as given in this box, also illustrates several important techniques in using the variational principles for the Einstein equations to elucidate the solution of these equations. The Mixmaster example can also be used to provide simple examples of superspace ideas and of quantum formulations of the laws of gravity [Misner (1972a)].

A Generalized Kasner Model

Two generalizations must be implemented in order to progress from the Kasner example (30.1) of a cosmological singularity to the Mixmaster example. The first is to allow a more general time-dependence while preserving some of the simplicity of the conditions (30.2) on the exponents p_i. Note that these exponents satisfy, e.g., $p_2 \equiv d \ln g_{22}/d \ln g$. Therefore one is led to parametrize the 3×3 spatial metric as

$$g_{ij} = e^{2\alpha}(e^{2\beta})_{ij} \tag{1}$$

or equivalently, $(\ln g)_{ij} = 2\alpha\, \delta_{ij} + 2\beta_{ij}$, where β_{ij} is a traceless 3×3 symmetric matrix, and the exponential is a matrix power series, so $\det e^{2\beta} = 1$ and

$$\sqrt{g} = e^{3\alpha}. \tag{2}$$

For the purposes of this paragraph only, define

$p_{ij} = d(\ln g)_{ij}/d \ln \det g$. Then from equations (1) and (2), one computes

$$p_{ij} = \frac{1}{3}[\delta_{ij} + (d\beta_{ij}/d\alpha)]; \tag{3}$$

so the one Kasner condition

$$1 = \sum_i p_i \equiv \operatorname{trace} p_{ij} = 1 + \frac{1}{3}\operatorname{trace}(d\beta/d\alpha)$$

is an identity in view of trace $\beta_{ij} = 0$. The second condition on the Kasner exponents is trace $(p^2) = 1$, and becomes $(d\beta_{ij}/d\alpha)^2 = 6$ by equation (3). This is not an identity, but a consequence of the Einstein equations in empty space. For the (Bianchi Type I) metric

$$ds^2 = -dt^2 + e^{2\alpha}(e^{2\beta})_{ij}\, dx^i\, dx^j, \tag{4}$$

and in the case when β_{ij} is diagonal, the Einstein equations are,

$$\left(\frac{d\alpha}{dt}\right)^2 = \frac{8\pi}{3}\left[T^{00} + \frac{1}{16\pi}(d\beta_{ij}/dt)^2\right] \tag{5}$$

and

$$e^{-3\alpha}\frac{d}{dt}\left(e^{3\alpha}\frac{d\beta_{ij}}{dt}\right) = 8\pi\left(T_{ij} - \frac{1}{3}\delta_{ij}T_{kk}\right), \tag{6}$$

together with a redundant equation involving T_{kk} and the equation $T_{0k} = 0$. [The stress components here refer to an orthonormal frame with basis 1-forms $\omega^i = e^{\alpha}(e^{\beta})_{ij}\, dx^j$.] From equation (5) one immediately derives

$$\rho_{\text{aniso(I)}} = (c^2/16\pi G)(d\beta_{ij}/dt)^2 \tag{7}$$

as a formula for the effectiveness of Type I anisotropy in contributing to the Hubble constant $H = d\alpha/dt$ on a basis comparable to matter energy, as in equation (30.4). Similarly, for equation (6) in the case of fluid matter (isotropic pressures), the stress terms vanish, and one obtains $\rho_{\text{aniso(I)}}e^{6\alpha} =$ const., as in the equation following (30.4). The Kasner condition $\Sigma p_i{}^2 = 1$ or $(d\beta_{ij}/d\alpha)^2 = 6$ follows from equation (5) whenever $T^{00} \ll \rho_{\text{aniso}}$.

In the diagonal case, β_{ij} has only two independent components, and it is convenient at times to define them explicitly by the parameterization

$$\begin{aligned} \beta_{11} &= \beta_+ + \sqrt{3}\beta_-, \\ \beta_{22} &= \beta_+ - \sqrt{3}\beta_-, \\ \beta_{33} &= -2\beta_+. \end{aligned} \tag{8}$$

For these the Kasner condition $(d\beta_{ij}/d\alpha)^2 = 6$ becomes

$$(d\beta_+/d\alpha)^2 + (d\beta_-/d\alpha)^2 = 1. \tag{9}$$

The $\beta_\pm$ are related to the Kasner exponents p_i or the u parameter of equations (30.5) by

$$\begin{aligned} d\beta_+/d\alpha &= \frac{1}{2}(1 - 3p_3) \\ &= -1 + (3/2)(1 + u + u^2)^{-1} \\ d\beta_-/d\alpha &= \frac{1}{2}\sqrt{3}(p_1 - p_2) \\ &= -\frac{1}{2}\sqrt{3}(1 + 2u)(1 + u + u^2)^{-1}. \end{aligned} \tag{10}$$

Introducing Space Curvature

The first step in generalizing the Kasner metric has focused attention on the "velocity" $\boldsymbol{\beta}' \equiv (d\beta_+/d\alpha, d\beta_-/d\alpha)$ which is a derivative of anisotropy with respect to expansion. The effects of matter or, as will soon appear, space curvature can change the magnitude $\|\boldsymbol{\beta}'\|$ from the Kasner value of unity. The second step of generalization is to introduce space curvature. This one achieves in a simple example by retaining the metric components of equation (1), but employing them in a non-holonomic basis. Use the basis vectors introduced in exercises 9.13 and 9.14 on the rotation group $SO(3)$, whose dual 1-forms are

$$\begin{aligned} \sigma^1 &= \cos\psi\, d\theta + \sin\psi \sin\theta\, d\phi, \\ \sigma^2 &= \sin\psi\, d\theta - \cos\psi \sin\theta\, d\phi, \\ \sigma^3 &= d\psi + \cos\theta\, d\phi, \end{aligned} \tag{11}$$

to form the metric

$$ds^2 = -N^2\, dt^2 + e^{2\alpha}(e^{2\beta})_{ij}\sigma^i\sigma^j, \tag{12}$$

where N, α, and β_{ij} are functions of t only. When

Box 30.1 ***(continued)***

$\alpha = 0 = \beta_{ij}$, the three-dimensional space metric here reduces to the one studied in exercise 13.15, which is the metric of highest symmetry on the group space $SO(3)$. The simply connected covering space has the 3-sphere topology, and is obtained by extending the range of the Euler angle ψ to give it a 4π period [$SU(2)$ or spin $\frac{1}{2}$ covering of the rotation group]. With $N = 1, \frac{1}{2}a = e^{\alpha}$, and $\beta_{ij} = 0$, one obtains from equation (12) the same metric (in different coordinates) as that treated in exercise 14.4 and in Chapter 27 in discussions of the closed Friedmann cosmological model. A non-zero value for β_{ij} allows the 3-sphere to have a different circumference on great circles in each of 3 mutually orthogonal principal directions, thus destroying its isotropy but not its homogeneity.

Let us consider only the case with β_{ij} diagonal, as in equation (8). Then the T^{00} Einstein equation becomes (with $N = 1$ as a time-coordinate condition)

$$3(\dot{\alpha}^2 - \dot{\beta}_+{}^2 - \dot{\beta}_-{}^2) + \frac{1}{2}({}^3R_{\mathrm{IX}}) = 8\pi T^{00}, \quad (13)$$

where only the term

$${}^3R_{\mathrm{IX}} = \frac{1}{2}e^{-2\alpha}\,\mathrm{trace}\,(2e^{-2\beta} - e^{4\beta}) \quad (14)$$

is different from equation (5). This term [see equation (21.92)] is the scalar curvature of a three-dimensional slice, $t =$ const [which has symmetry properties known as "Bianchi Type IX" for the metric of equations (11) and (12)]. If equation (13) is interpreted in terms of an anisotropy energy density contributing, with T^{00}, to the volume expansion $\dot{\alpha}^2$, then there are not only kinetic energy terms $\dot{\beta}^2$ [as in equations (5) and (7)], but also a potential energy term. This term shows that negative scalar curvature, which can be produced by anisotropy ($\beta \neq 0$), is equivalent to a positive potential (or "internal") energy, and suggests that empty space has properties with analogies to an elastic solid and resists shear strains. The more detailed analysis which follows shows that, near the singularity, the scalar curvature is always negligible when positive.

Negative curvatures, however, arise in this closed universe from large shear (β) deformations near the singularity and become large enough to reverse one Kasner shear motion [u-value, etc.; equation (10)] and change it to another.

These conclusions and further details of the time-evolution of the "Mixmaster" metric (11, 12) require, in principle, the study of all the Einstein equations, not just equation (13) for T^{00}. As described in Chapter 21, however, this T^{00} constraint equation is central, and actually contains implicitly the full content of the Einstein equations when formulated properly.

Variational Principles

One adequate formulation, adopted here, involves treating equation (13) not as an energy equation (involving velocities), but as a Hamiltonian (involving momenta). Take the Einstein variational principle (21.15) in ADM form (21.95) and carry out the space integration, using

$$\int \sigma^1 \wedge \sigma^2 \wedge \sigma^3 = \int \sin\theta\, d\phi \wedge d\theta \wedge d\psi = (4\pi)^2,$$

to obtain the action integral in the form

$$I = (\pi)\int\Big\{\pi^{ij}\,dg_{ij} + Ne^{3\alpha}\Big[{}^3R_{\mathrm{IX}} + e^{-6\alpha}\Big(\frac{1}{2}(\pi^l{}_l)^2 - \pi^{ik}\pi_{ik}\Big)\Big]\,dt\Big\}. \quad (15)$$

When introducing the specific form (1) and (8) for g_{ij}, it is convenient also to parameterize the diagonal matrix $\pi^i{}_k$ as follows:

$$\begin{aligned} p_\alpha &= (2\pi)\pi^k{}_k, \\ p^i{}_k &= (2\pi)\Big(\pi^i{}_k - \frac{1}{3}\delta^i{}_k\pi^l{}_l\Big), \end{aligned} \quad (16)$$

with

$$\begin{aligned} 6p^1{}_1 &= p_+ + p_-\sqrt{3}, \\ 6p^2{}_2 &= p_+ - p_-\sqrt{3}, \\ 6p^3{}_3 &= -2p_+ \end{aligned} \quad (17)$$

[see equation (8)]. The result is

$$I = \int p_+ \, d\beta_+ + p_- \, d\beta_- + p_\alpha \, d\alpha - \frac{Ne^{-3\alpha}}{24\pi}[-p_\alpha^2 + p_+{}^2 + p_-{}^2 - 24\pi^2 e^{6\alpha}({}^3R_{\rm IX})] \, dt.$$

This is cleaned up for further study as follows. Write

$$^3R_{\rm IX} = \frac{3}{2} e^{-2\alpha}(1 - V), \tag{18}$$

where

$$V = V(\beta) = \frac{1}{3}\,\mathrm{trace}\,(1 - 2e^{-2\beta} + e^{4\beta}) \tag{19}$$

so $V(0) = 0$; and adjust the zero of α ($\alpha \longrightarrow \alpha - \alpha_0$) so that $e^{2\alpha} \longrightarrow (6\pi)^{-1}e^{2\alpha}$. Then *the metric is*

$$ds^2 = -N^2 \, dt^2 + (6\pi)^{-1} e^{2\alpha}(e^{2\beta})_{ij}\sigma^i\sigma^j, \tag{20}$$

and the variational integral is

$$I = \int p_+ \, d\beta_+ + p_- \, d\beta_- + p_\alpha \, d\alpha - (3\pi/2)^{1/2} Ne^{-3\alpha}\mathcal{H} \, dt, \tag{21}$$

with

$$2\mathcal{H} \equiv -p_\alpha{}^2 + p_+{}^2 + p_-{}^2 + e^{4\alpha}(V - 1). \tag{22}$$

One demands $\delta I = 0$ for arbitrary independent variations of $p_\pm$, p_α, $\beta_\pm$, α, N to obtain the Einstein equations. From varying N, one obtains the fundamental constraint equation $\mathcal{H} = 0$ [which would reduce to the vacuum version of equation (13) when the momenta are replaced by velocities (via equations obtained by varying the p's) if the coordinate condition $N = 1$ were imposed.]

ADM Hamiltonian

The standard ADM prescription for reducing this variational principle to canonical (Hamiltonian) form is to choose one of the field variables or momenta as a time-coordinate, and solve the constraint for its conjugate Hamiltonian. Here an obvious and satisfactory choice is to set $t = \alpha$, and solve $\mathcal{H} = 0$ for

$$H_{\rm ADM} = -p_\alpha = [p_+{}^2 + p_-{}^2 + e^{4\alpha}(V - 1)]^{1/2}. \tag{23}$$

The $\dot\alpha$ equation [vary p_α in equation (21)] is

$$\dot\alpha = -(3\pi/2)^{1/2} Ne^{-3\alpha} p_\alpha \tag{24}$$

and shows that the choice $\alpha = t$ (so $\dot\alpha = 1$) requires

$$N_{\rm ADM} = (2/3\pi)^{1/2} e^{3\alpha}/H_{\rm ADM}. \tag{25}$$

The reduced, canonical, variational principle which results when equation (23) is used to eliminate p_α reads $\delta I_{\rm red} = 0$ with

$$I_{\rm red} = \int p_+ \, d\beta_+ + p_- \, d\beta_- - H_{\rm ADM} \, d\alpha \tag{26}$$

and must be supplemented by equation (25).

Super-Hamiltonian

A more convenient approach here is one more closely related to the Dirac Hamiltonian methods than those of ADM. Note, however, that one does not remove the arbitrariness in the lapse function by taking it to be some specified function $N(t)$ of the coordinates. Instead the procedure adopted here is to eliminate N from the variational principle (21) by choosing it (coordinate condition!) to be some chosen function of the field variables and momenta, $N = N(\alpha, \beta_\pm, p_\alpha, p_\pm)$. Any such choice, inserted in equation (21), leaves a variational integral in canonical Hamiltonian form. The content of this new variational principle becomes equivalent to the original one only when supplemented by the constraint

$$\mathcal{H} = 0, \tag{27}$$

which can no longer be derived from the variational principle. [The other Euler-Lagrange equations for these two principles differ only by terms proportional to $\mathcal{H}$, and thus are equivalent when

Box 30.1 (*continued*)

$\mathcal{H} = 0$ is imposed on the initial conditions.] The choice

$$N = (2/3\pi)^{1/2} e^{3\alpha} \tag{28}$$

is obvious and convenient. It makes $\mathcal{H}$ become a super-Hamiltonian in the resulting variational principle

$$I = \int p_+ \, d\beta_+ + p_- \, d\beta_- + p_\alpha \, d\alpha - \mathcal{H} \, d\lambda, \tag{29}$$

where $t \equiv \lambda$ has been written to label the specific time-coordinate choice that equation (28) implies.

Mixmaster Dynamics

If matter terms with no additional degrees of freedom are included, the super-Hamiltonian in equation (29) is modified simply. For an example, choose

$$T^{\hat{0}\hat{0}} = -T^0{}_0 = (3/4)^2(\mu e^{-3\alpha} + \Gamma e^{-4\alpha}) \tag{30}$$

for the energy density of matter in a frame with time-axis $\boldsymbol{e}_{\hat{0}} = N^{-1}(\partial/\partial t)$. The two terms represent a nonrelativistic perfect fluid ($\rho \propto V^{-1}$) and a radiation fluid ($\rho \propto V^{-4/3}$), respectively, and lead to

$$\begin{aligned} 2\mathcal{H} = -p_\alpha{}^2 + p_+{}^2 + p_-{}^2 + e^{4\alpha}(V - 1) \\ + \mu e^{3\alpha} + \Gamma e^{2\alpha}. \end{aligned} \tag{31}$$

This Hamiltonian, with its simple quadratic momentum dependence, differs in only two ways from the Hamiltonians of elementary mechanics, namely, (1) in the sign of the $p_\alpha{}^2$ term and (2) in the detailed shape of the "potential" term as function of α and $\beta_\pm$, the study of which reduces to a study of the function $V(\beta)$. Hamilton's equations, from varying α, $\beta_\pm$, p_α, and $p_\pm$ in equation (29), yield

$$\frac{d^2\beta_\pm}{d\lambda^2} = -\frac{\partial\mathcal{H}}{\partial\beta_\pm} = -\frac{1}{2} e^{4\alpha} \frac{\partial V}{\partial\beta_\pm} \tag{32}$$

and

$$\frac{d^2\alpha}{d\lambda^2} = +\frac{\partial\mathcal{H}}{\partial\alpha} = 2e^{4\alpha}(V - 1) + \frac{3}{2}\mu e^{3\alpha} + \Gamma e^{2\alpha}. \tag{33}$$

Thus the sign of the $p_\alpha{}^2$ term causes α to accelerate toward (rather than away from) higher values of the "potential" terms $e^{4\alpha}(V - 1) + \mu e^{3\alpha} + \Gamma e^{2\alpha}$. When $|V| \ll 1$ (small anisotropy), equation (33) is identical to its form in the isotropic Friedmann model, and allows a deceleration only when α is large enough that the positive curvature term $(-e^{4\alpha})$ dominates over matter ($\mu e^{3\alpha}$) and radiation ($\Gamma e^{2\alpha}$). Near the singularity ($\alpha \longrightarrow -\infty$), the positive curvature term is always negligible compared to radiation and matter.

For studies of the singularity behavior, it is sufficient to study the simplified super-Hamiltonian

$$2\mathcal{H} \sim -p_\alpha{}^2 + p_+{}^2 + p_-{}^2 + e^{4\alpha}V(\beta), \tag{34}$$

since the other terms obviously vanish for $\alpha \longrightarrow -\infty$. This form retains only the V term in ${}^3R_{\text{IX}} = \frac{3}{2}e^{-2\alpha}(1 - V)$, which dominates when the curvature of this closed universe becomes negative, $V \gg 1$. If the term in $V(\beta)$ were also negligible, then $\mathcal{H} = -p_\alpha{}^2 + p_+{}^2 + p_-{}^2$ would make each p_α, $p_\pm$ constant, giving the Kasner behavior with

$$d\beta_\pm/d\alpha = p_\pm/p_\alpha = \text{const}$$

and $|d\boldsymbol{\beta}/d\alpha|^2 = 1$ as expected (since matter and curvature have been neglected). To proceed further, a study of $V(\beta)$ is required, based on equations (19) and (8), and their immediate consequence:

$$\begin{aligned} V(\beta) = \frac{1}{3} e^{-8\beta_+} - \frac{4}{3} e^{-2\beta_+} \cosh 2\sqrt{3}\beta_- \\ + 1 + \frac{2}{3} e^{4\beta_+}(\cosh 4\sqrt{3}\beta_- - 1). \end{aligned} \tag{35}$$

One finds that $V(\beta)$ is a *positive definite* "potential well" which has the same symmetries as an equilateral triangle in the $\beta_+\beta_-$ plane. Near the origin, $\beta_\pm = 0$, the equipotentials are circles, since

$$V(\beta) = 8(\beta_+{}^2 + \beta_-{}^2) + 0(\beta^3). \tag{36}$$

For large β values, one finds

$$V(\beta) \sim \frac{1}{3} e^{-8\beta_+}, \qquad \beta_+ \longrightarrow -\infty, \tag{37}$$

and

$$V(\beta) \sim 1 + 16\beta_-{}^2 e^{4\beta_+}, \qquad \begin{array}{l} \beta_+ \longrightarrow +\infty, \\ |\beta_-| \ll 1. \end{array} \tag{38}$$

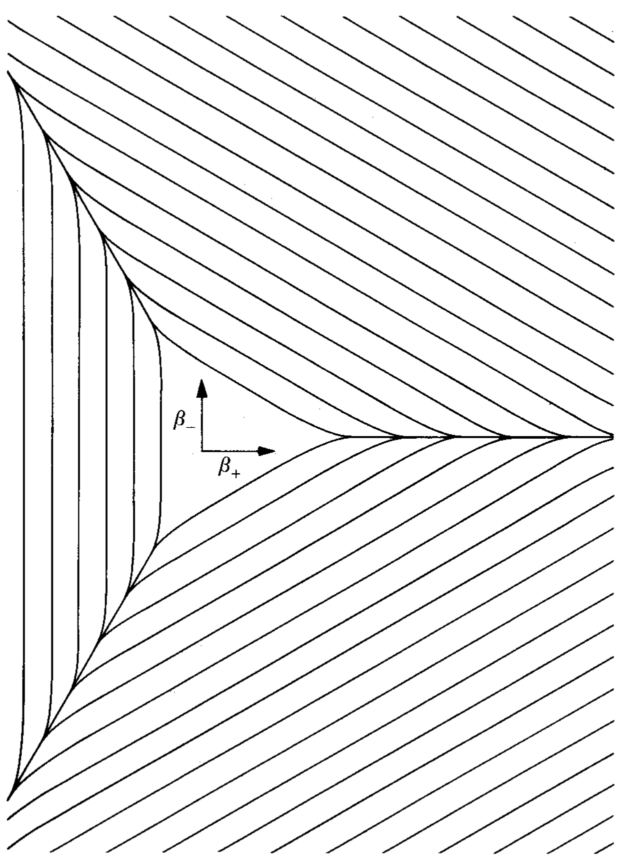
These two asymptotic forms, together with the triangular symmetry, give a complete asymptotic description of $V(\beta)$, as sketched in the figure, where on successive levels separated by $\Delta\beta = 1$, the potential V increases by a *factor* of $e^8 = 3 \times 10^3$.

"Bounce" Interrupts Kasner-like Steps Toward the Singularity

The dominant feature of the $V(\beta)$ potential is evidently its steep (exponential) triangular walls, with equation (37) representing the typical one for study. Under the influence of this potential wall, the evolution of this model universe is governed by the super-Hamiltonian

$$2\mathcal{H} \sim -p_\alpha{}^2 + p_+{}^2 + p_-{}^2 + \frac{1}{3} e^{4(\alpha - 2\beta_+)}. \tag{39}$$

If $\alpha \longrightarrow -\infty$ with $d\beta_+/d\alpha > 1/2$ [recall $d\beta_+/d\alpha$ = const., $|d\boldsymbol{\beta}/d\alpha| = 1$, when the last term in (39) is small], then the potential term grows and will eventually become large enough to influence the motion. A simple "Lorentz" transformation, suggested by the superspace metric (coefficients of the



Some equipotentials, $V(\beta)$ = constant, are shown for the function defined in equation (35). Equipotentials near the origin of the β-plane are closed curves for $V < 1$ and are omitted here.

Box 30.1 *(continued)*

quadratic in the momenta) simplifies the computation further. Set

$$\bar{\beta}_+ = \left(\beta_+ - \frac{1}{2}\alpha\right)/\sqrt{3/4},$$
$$\bar{\alpha} = \left(\alpha - \frac{1}{2}\beta_+\right)/\sqrt{3/4},$$

and find

$$2\mathcal{H} = -\bar{p}_\alpha{}^2 + \bar{p}_+{}^2 + p_-{}^2 + \frac{1}{3}\exp\left(-4\sqrt{3}\bar{\beta}_+\right). \tag{40}$$

For this super-Hamiltonian both $\bar{p}_\alpha$ and p_- are constants of motion, whereas the $\bar{\beta}_+$-Hamiltonian, $\bar{p}_+{}^2 + \frac{1}{3}e^{-4\sqrt{3}\bar{\beta}_+}$, represents a simple bounce against a one-dimensional potential wall with the initial and final values of $\bar{p}_+$ different only in sign. The behavior of the anisotropy parameters $\beta_\pm$ near the singularity thus consists of a simple Kasner step (where $d\beta_\pm/d\alpha =$ const., with the $d\beta_+/d\alpha \geq \frac{1}{2}$, or conditions equivalent by symmetry, satisfied relative to one of the three walls), followed by a bounce against that wall, beginning a new Kasner step with other Kasner parameters. [The most detailed description of this behavior and its relation to more general cosmological models can be found in Belinsky, Khalatnikov, and Lifshitz (1970)—see also the briefer report, Khalatnikov and Lifshitz (1970)—using quite different methods. For detailed developments by Hamiltonian methods, which supercede the partial Lagrangian methods of Misner (1969b), see Misner (1970, 1972a), and Ryan (1972a,b).]

Steady-State, Quasiperiodic Infinity of "Bounces" Approaching the Singularity

Some comprehensive features of the singularity behavior, involving many Kasner-like steps, can be exhibited by another transformation of the parameter space (superspace) of the metric field. The transformation introduces a "radial" t-coordinate out from the origin of $\alpha\beta_\pm$ space, while respecting the metric properties of this superspace implied by the form of the super-Hamiltonian. Thus one defines (for any constant α_0)

$$\begin{aligned} \alpha_0 - \alpha &= e^t \cosh\zeta, \\ \beta_+ &= e^t \sinh\zeta\cos\phi, \\ \beta_- &= e^t \sinh\zeta\sin\phi, \end{aligned} \tag{41}$$

and finds

$$2\mathcal{H} = e^{-2t}[(-p_t{}^2 + p_\zeta{}^2 + p_\phi{}^2 \sinh^{-2}\zeta) + e^{2t}e^{4\alpha}V]. \tag{42}$$

The advantage of this transformation is that in the limit $t \longrightarrow \infty$ ($\alpha \longrightarrow -\infty$, singularity) the potential terms become, in first approximation, independent of t. Thus equation (37) gives, for one potential wall,

$$e^{2t}e^{4\alpha}V \sim \frac{1}{3}e^{2t}\exp\left[4\alpha_0 - 8e^t\left(\sinh\zeta\cos\phi + \frac{1}{2}\cosh\zeta\right)\right]. \tag{43}$$

For $t \longrightarrow \infty$ this expression evidently tends to either zero or infinity, depending on the sign of the expression in parentheses. Therefore define the asymptotic potential walls by

$$\tanh\zeta + \frac{1}{2}\sec\phi = 0 \tag{44}$$

in the sector $|\phi - \pi| < \pi/3$, and equivalent formulae in which ϕ is replaced by $\phi \pm (2\pi/3)$ for the other sides of the triangle. Consequently, an asymptotic approximation to the super-Hamiltonian is

$$2\mathcal{H} = e^{-2t}[-p_t{}^2 + p_\zeta{}^2 + p_\phi{}^2\sinh^{-2}\zeta + V'(\zeta,\phi)], \tag{45}$$

where $V'(\zeta,\phi)$ vanishes inside the asymptotic walls (44) and equals $+\infty$ outside. Because the remaining t-dependence is a common factor in (45), a simple change of independent variable $e^{-2t}\,d\lambda = d\lambda'$ in equation (29)—equivalent to the choice

$$N = (2/3\pi)^{1/2}e^{-2t}\exp\left[3(\alpha_0 - e^t\cosh\zeta)\right] \tag{46}$$

in place of equation (28)—gives a new super-Hamiltonian $\mathcal{H}' = e^{2t}\mathcal{H}$ with the variational integral

$$I = \int p_t\,dt + p_\zeta\,d\zeta + p_\phi\,d\phi - \mathcal{H}'\,d\lambda'. \quad (47)$$

In the asymptotic approximation where

$$2\mathcal{H}' = -p_t^{\,2} + p_\zeta^{\,2} + p_\phi^{\,2}\sinh^{-2}\zeta + V'(\zeta, \phi), \quad (48)$$

one immediately sees that p_t is a constant of motion, and that the "bouncing" of the $\zeta\phi$ values within the asymptotic potential walls is a stationary, quasi-periodic process in this time-coordinate λ' (or t, since $dt/d\lambda' = -p_t = \text{const}$). [More detailed studies based on this asymptotic super-Hamiltonian show that the motion is even ergodic, with $\zeta\phi$ approaching arbitrarily close to any given value infinitely many times as $t \longrightarrow \infty$; see Chitre (1972a).]

Summary

One has found the singularity behavior in this Mixmaster example to be extraordinarily active. In the simple Kasner singularity, two axes collapse, but the third is stretched in a simple tidal deformation accompanied by volume compression. But in the Mixmaster example, every such collapse attempt is defeated by the high negative curvature it implies. Or rather it is diverted to another attempt as compression continues inexorably, but the tidal deformations attempt first one configuration, then another, in an infinitely recurring probing of all possible configurations.

Speculations on Time and the Singularity

The cosmological singularity (in all examples where its character is not known to be unstable) involves infinite curvature and infinite density. One's abhorence of such a theoretical prediction is particularly heightened by the correlative prediction that these infinities occurred at a finite proper time in the past, and would—if they recur—occur again at some finite proper time in the future. The singularity prediction would be more tolerable if the infinite densities could be removed to the infinitely distant past. The universe could then, as now, find its natural state to be one of expansion, so every *finite* density will have been experienced at some suitably remote past time, but *infinite* density becomes a formal abstraction never realized in the course of evolution.

To push infinite curvature out of the finite past might be achieved in two ways. It is not known which, if either, works. One way is to change the physical laws which require the singularity, changing them perhaps only in obvious and desirable ways, such as stating the laws of gravity in a proper quantum language. Computations of quantum geometry are not yet definitive, however, and some (perhaps inadequate) approximations [Misner (1972a)] do not remove the singularity problem.

Another way to discard the singularity is to accept the mathematics of the classical Einstein equations, but reinterpret it in terms of an infinite past time. There are, of course, simple and utterly inadequate ways to do this by arbitrary coordinate transformations such as $t = \ln\tau$ which change a $\tau = 0$ singularity into one at $t = -\infty$. But an arbitrary coordinate is without significance. The problem is that the singularity occurs at a finite *proper* time in the past, and proper time is the most physically significant, most physically real time we know. It corresponds to the ticking of physical clocks and measures the natural rhythms of actual events. To reinterpret finite past time as infinite, one must attack proper time on precisely these grounds, and claim it is inadequately physical. On a local basis, where special relativity is valid, no challenge to the physical significance of proper time can succeed. It is on a more global scale that the physical primacy of proper time needs to be reviewed.

"The cosmological singularity occurred ten thousand million years ago." In this statement, take time to mean the proper time along the world line of the solar system, ephemeris time. Then the statement would have a most direct physical sig-

Box 30.1 ***(continued)***

nificance if it meant that the Earth had completed 10^{10} orbits about the sun since the beginning of the universe. But proper time is not that closely tied to actual physical phenomena. The statement merely implies that those 5×10^9 orbits which the earth may have actually accomplished give a standard of time which is to be extrapolated in prescribed ways, thus giving theoretical meaning to the other 5×10^9 years which are asserted to have preceeded the formation of the solar system.

A hardier standard clock changes the details of the argument, but not its qualitative conclusion. To interpret 10^{10} years in terms of SI (Systeme Internationale) seconds assigns a past history containing some 3×10^{27} oscillations of a hyperfine transition in neutral Cesium. But again the critical early ticks of the clock (needed to locate the singularity in time by actual physical events) are missing. The time needed for stellar nucleosynthesis to produce the first Cesium disqualifies this clock on historical grounds, and the still earlier high temperatures nearer the singularity would have ionized all Cesium even if this element had predated stars.

Thus proper time near the singularity is not a direct counting of simple and actual physical phenomena, but an elaborate mathematical extrapolation. Each actual clock has its "ticks" discounted by a suitable factor—3×10^7 seconds per orbit from the Earth-sun system, 1.1×10^{-10} seconds per oscillation for the Cesium transition, etc. Since no single clock (because of its finite size and strength) is conceivable all the way back to the singularity, a statement about the proper time since the singularity involves the concept of an infinite sequence of successively smaller and sturdier clocks with their ticks then discounted and added. "Finite proper time," then, need not imply that any finite sequence of events was possible. It may describe a necessarily infinite number of events ("ticks") in any physically conceivable history, converted by mathematics into a finite sum by the action of a non-local convergence factor, the "discount" applied to convert "ticks" into "proper time."

Here one has the conceptual inverse of Zeno's paradox. One rejects Zeno's suggestion that a single swing of a pendulum is infinitely complicated—being composed of a half period, plus a quarter period, plus 2^{-n} *ad infinitum*—because the terms in his infinite series are mathematical abstractions, not physically achieved discrete acts in a drama that must be played out. By a comparable standard, one should ignore as a mathematical abstraction the finite sum of the proper-time series for the age of the universe, if it can be proved that there must be an infinite number of discrete acts played out during its past history. In both cases, finiteness would be judged by counting the number of discrete ticks on realizable clocks, not by assessing the weight of unrealizable mathematical abstractions.

Whether the universe is infinitely old by this standard remains to be determined. The quantum influences, in particular, remain to be calculated. The decisive question is whether each present-epoch event is subject to the influence of infinitely many previous discrete events. In that case statistical assumptions (large numbers, random phases, etc.) could enter in stronger ways into theories of cosmology. The Mixmaster cosmological model does have an infinite past history in this sense, since each "bounce" from one Kasner-like motion to another is a recognizable cosmological event, of which infinitely many must be realized between any finite epoch and the singularity.

§30.8. HORIZONS AND THE ISOTROPY OF THE MICROWAVE BACKGROUND

The fundamental cosmological question—"Must a universe that is born chaotic necessarily become as homogeneous and isotropic as our universe is, and do so before life evolves?"—entails one further issue. This issue is *horizons*. As was discussed in §27.10, at any given epoch in the expansion of a Friedmann universe (e.g., the present epoch), there may be significant portions of the universe from which no light signal or other causally propagating influence will have yet reached Earth in the time available since the initial singularity. "If we should live so long," the question would arise, "will the new portions of the universe which first come into view during the next ten billion years look statistically identical to the neighboring portions which are already being seen?"

Horizons in a Friedmann universe

Fortunately, this question need not be posed only for the future. It can be asked as of some past time, and the answer then is yes. Microwave background radiation arrives at the earth from all directions in the sky with very nearly the same temperature. [The data of Boughn, Fram, and Partridge (1971) and of Conklin (1969) show $\Delta T/T \lesssim 0.004$.] The plasma that emitted the microwave radiation coming to earth from one direction in the sky had not been able, before the epoch of emission, to communicate causally with the plasma emitting the radiation that arrives from other directions. If one adopts a Friedmann model of the universe, then different sectors of the microwave sky are disjoint from each other in this sense if they are separated from each other by more than 30°, even if the microwaves were emitted as recently as $z = 7$. (The critical angle is much smaller if the microwaves were last scattered at $z = 1{,}000$.) From this, one concludes that the foundations for the homogeneity and isotropy of the universe were laid long before the universe became approximately Friedmann, for if statistical homogeneity and isotropy of the universe had not already been achieved at the longest wavelengths earlier, these horizon limitations would have prevented any further synchronization of conditions over large scales while the universe was in a nearly Friedmann state, and small amplitude (10%) deviations from isotropy should be observed now.

Observed isotropy of microwave radiation proves foundations for homogeneity were laid before universe became Friedmann-like

The mixmaster universe received its name from the hope that it could contribute to the solution of this problem. The very large u values that occur sporadically an infinite number of times near the singularity in a mixmaster universe give a geometry close to that of the Kasner model with $p_1 = 1$, $p_2 = p_3 = 0$. This model can be written in the form

What made the universe homogeneous and isotropic?

(1) Mixmaster oscillations?—probably not

$$ds^2 = e^{2\eta}(-d\eta^2 + dx^2) + dy^2 + dz^2, \tag{30.6}$$

where $\eta = \ln t$. If this metric is converted into a closed-universe model by interpreting x, y, z as angle coordinates each with period 4π, then one sees that a light ray can circumnavigate the universe in the x-direction in a time interval $\Delta\eta = 4\pi$, which corresponds to a volume expansion by a factor $\sqrt{-g_1}/\sqrt{-g_2} = e^{4\pi}$. Unfortunately, a quantitative analysis of the degree and frequency with which the mixmaster universe achieves this specific Kasner form suggests that the horizon breaking

(2) particle creation near singularity?

is inadequate to explain the present state of the universe [Doroshkevich, Lukash, and Novikov (1971); Chitre (1972)]. It may turn out that particle creation near the singularity can solve this horizon question, as well as provide for the dissipation of anisotropy. Hope is provided by the fact that particle creation, when described in purely classical terms, has some acausal appearances, even though it is a strictly causal process at the quantum level [Zel'dovich (1972)].

PART VII

GRAVITATIONAL COLLAPSE AND BLACK HOLES

Wherein the reader is transported to the land of black holes, and encounters colonies of static limits, ergospheres, and horizons—behind whose veils are hidden gaping, ferocious singularities.

CHAPTER 31

SCHWARZSCHILD GEOMETRY

§31.1. INEVITABILITY OF COLLAPSE FOR MASSIVE STARS

There is no equilibrium state at the endpoint of thermonuclear evolution for a star containing more than about twice the number of baryons in the sun ($A > A_{max} \sim 2A_{\odot}$). This is one of the most surprising—and disturbing—consequences of the discussion in Chapter 24. Stated differently: A star with $A > A_{max} \sim 2A_{\odot}$ must eject all but A_{max} of its baryons—e.g., by nova or supernova explosions—before settling down into its final resting state; otherwise there will be no final resting state for it to settle down into.

What is the fate of a star that fails to eject its excess baryons before nearing the endpoint of thermonuclear evolution? For example, after a very massive supernova explosion, what will become of the collapsed degenerate-neutron core when it contains more than A_{max} baryons? Such a supercritical mass cannot explode, since it is gravitationally bound and it has no more thermonuclear energy to release. Nor can it reach a static equilibrium state, since there exists no such state for so large a mass. There remains only one alternative; the supercritical mass must collapse through its "gravitational radius," $r = 2M$, leaving behind a gravitating "black hole" in space.

The phenomenon of collapse through the gravitational radius, as described by classical general relativity, will be the subject of the next chapter. However, before tackling it, one must understand more fully than heretofore the Schwarzschild spacetime geometry, which surrounds black holes and collapsing stars as well as static stars.

This chapter, on Schwarzschild geometry, is key preparation for understanding gravitational collapse (next chapter) and black holes (following chapter)

This chapter will concern itself with two topics that, at first sight, appear to be disconnected. One is the fall of a test particle in a preexisting Schwarzschild geometry, which is regarded as static, but can also be visualized as all that remains of a star that underwent collapse some time ago. The second topic is the physical character of this geometry, regarded in and by itself. For the exploration of this geometry, the test particle serves as the best of all explorers. But the test particle may also be regarded in another light. It can be viewed as a rag-tag johnny-come-lately piece of the matter of the falling star. Regarded in this way, it provides the simplest of all illustrations of an asymmetry in the distribution of mass of a collapsing star. That this asymmetry irons itself out will therefore give one some preliminary insight into how more complicated asymmetries also iron themselves out. In brief, the motion of the test particle and the dynamics of the Schwarzschild geometry (for this geometry will prove to be dynamic), two apparently different problems, have the happy ability to throw light on each other.

§31.2. THE NONSINGULARITY OF THE GRAVITATIONAL RADIUS

The Schwarzschild spacetime geometry

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 + \frac{dr^2}{1 - 2M/r} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{31.1}$$

The Schwarzschild line element becomes singular at $r = 2M$ ("gravitational radius")

appears to behave badly near $r = 2M$; there g_{tt} becomes zero, and g_{rr} becomes infinite. However, one cannot be sure without careful study whether this pathology in the line element is due to a pathology in the spacetime geometry itself, or merely to a pathology of the (t, r, θ, ϕ) coordinate system near $r = 2M$. (As an example of a coordinate-induced pathology, consider the neighborhood of $\theta = 0$ on one of the invariant spheres, $t =$ const and $r =$ const. There $g_{\phi\phi}$ becomes zero because the coordinate system behaves badly; however, the intrinsic, coordinate-independent geometry of the sphere is well-behaved there. For another example, see Figure 1.4.

The worrisome region of the Schwarzschild geometry, $r = 2M$, is called the "gravitational radius," or the "Schwarzschild radius," or the "Schwarzschild surface," or the "Schwarzschild horizon," or the "Schwarzschild sphere." It is also called the "Schwarzschild singularity" in some of the older literature; but that is a misnomer, since, as will be shown, the spacetime geometry is not singular there.

To determine whether the spacetime geometry is singular at the gravitational radius, send an explorer in from far away to chart it. For simplicity, let him fall freely and radially into the gravitational radius, carrying his orthonormal tetrad with him as he falls. His trajectory through spacetime ["parabolic orbit"; radial geodesic of metric (31.1)] is

$$\begin{aligned} \frac{\tau}{2M} &= -\frac{2}{3}\left(\frac{r}{2M}\right)^{3/2} + \text{constant}, \\ \frac{t}{2M} &= -\frac{2}{3}\left(\frac{r}{2M}\right)^{3/2} - 2\left(\frac{r}{2M}\right)^{1/2} + \ln\left|\frac{(r/2M)^{1/2} + 1}{(r/2M)^{1/2} - 1}\right| + \text{constant}. \end{aligned} \tag{31.2}$$

[See §25.5 and especially equation (25.38) for derivation and discussion.] One obtains the r coordinate of the explorer in terms of the proper time measured on a clock he carries, $r(\tau)$, by inverting the first equation; one finds his r coordinate in terms of coordinate time, $r(t)$, by inverting the second equation.

Of all the features of the traveler's trajectory, one stands out most clearly and disturbingly: to reach the gravitational radius, $r = 2M$, requires a finite lapse of proper time, but an infinite lapse of coordinate time:

An infalling observer reaches $r = 2M$ in finite proper time but infinite coordinate time

$$\begin{aligned} r/2M &= 1 - (\tau + \text{constant})/2M \qquad \text{when near } r = 2M; \\ r/2M &= 1 + \text{constant} \times \exp(-t/2M) \qquad \text{in limit as } t \longrightarrow \infty. \end{aligned} \tag{31.3}$$

(see Fig. 25.5.) Of course, proper time is the relevant quantity for the explorer's heart-beat and health. No coordinate system has the power to prevent him from reaching $r = 2M$. Only the coordinate-independent geometry of spacetime could possibly do that; and equation (31.3) shows it does not!

Let the explorer approach and reach $r = 2M$, then. What spacetime geometry does he measure there? Is it singular or nonsingular? Restated in terms of measurements, do infinite tidal gravitational forces tear the traveler apart and crush him as he approaches $r = 2M$, or does he feel only finite tidal forces which in principle his body can withstand?

The tidal forces felt by the explorer as he passes a given radius r are measured by the components of the Riemann curvature tensor with respect to his orthonormal frame there (equation of geodesic deviation). To calculate those curvature components at r, proceed in two steps. (1) Calculate the components, not in the traveler's frame, but rather in the "static" orthonormal frame

$$\boldsymbol{\omega}^{\hat{t}} = \left(1 - \frac{2M}{r}\right)^{1/2} \boldsymbol{d}t,\ \boldsymbol{\omega}^{\hat{r}} = \frac{\boldsymbol{d}r}{(1 - 2M/r)^{1/2}},\ \boldsymbol{\omega}^{\hat{\theta}} = r\,\boldsymbol{d}\theta,\ \boldsymbol{\omega}^{\hat{\phi}} = r\sin\theta\,\boldsymbol{d}\phi \tag{31.4a}$$

located at the event through which he is passing; the result [obtainable from equations (14.50) and (14.51) by setting $e^{2\Phi} = e^{-2\Lambda} = 1 - 2M/r$] is

$$\begin{aligned} R_{\hat{t}\hat{r}\hat{t}\hat{r}} &= \frac{-2M}{r^3}, & R_{\hat{t}\hat{\theta}\hat{t}\hat{\theta}} &= R_{\hat{t}\hat{\phi}\hat{t}\hat{\phi}} = \frac{M}{r^3}, \\ R_{\hat{\theta}\hat{\phi}\hat{\theta}\hat{\phi}} &= \frac{2M}{r^3}, & R_{\hat{r}\hat{\theta}\hat{r}\hat{\theta}} &= R_{\hat{r}\hat{\phi}\hat{r}\hat{\phi}} = \frac{-M}{r^3}; \end{aligned} \tag{31.4b}$$

all other $R_{\hat{\alpha}\hat{\beta}\hat{\gamma}\hat{\delta}}$ vanish except those obtainable from the above by symmetries of ***Riemann***.

(2) Calculate the components in the explorer's frame by applying to the "static-frame" components (31.4b) the appropriate transformation—for $r > 2M$, a Lorentz boost in the $\boldsymbol{e}_{\hat{r}}$ direction with ordinary velocity $v^{\hat{r}}$; for $r < 2M$, not a "boost," but a transformation given by the standard boost formula (Box 2.4) with $v^{\hat{r}} > 1$. Here

$$v^{\hat{r}} = \frac{(g_{rr})^{1/2}\,dr}{(-g_{tt})^{1/2}\,dt} = \frac{dr/dt}{1 - 2M/r} = -\left(\frac{2M}{r}\right)^{1/2}. \tag{31.5}$$

The amazing result (a consequence of special algebraic properties of the Schwarzschild geometry, and somewhat analogous to what happens—or, rather, does not happen—to the components of the electromagnetic field, $\boldsymbol{E}$ and $\boldsymbol{B}$, when they are both parallel to a boost) is this: all the components of ***Riemann*** are left completely unaffected by the boost. If $\boldsymbol{e}_{\hat{\rho}}$ is the traveler's radial basis vector, and $\boldsymbol{e}_{\hat{\tau}} = \boldsymbol{u}$ is his time basis vector, then

$$\begin{aligned} R_{\hat{\tau}\hat{\rho}\hat{\tau}\hat{\rho}} &= -2M/r^3, & R_{\hat{\tau}\hat{\theta}\hat{\tau}\hat{\theta}} &= R_{\hat{\tau}\hat{\phi}\hat{\tau}\hat{\phi}} = M/r^3, \\ R_{\hat{\theta}\hat{\phi}\hat{\theta}\hat{\phi}} &= 2M/r^3, & R_{\hat{\rho}\hat{\theta}\hat{\rho}\hat{\theta}} &= R_{\hat{\rho}\hat{\phi}\hat{\rho}\hat{\phi}} = -M/r^3. \end{aligned} \tag{31.6}$$

(See exercise 31.1.)

The infalling observer does not feel infinite tidal forces at $r = 2M$

The payoff of this calculation: according to equations (31.6), none of the components of ***Riemann*** in the explorer's orthonormal frame become infinite at the gravitational radius. The tidal forces the traveler feels as he approaches $r = 2M$ are finite; they do not tear him apart—at least not when the mass M is sufficiently great, because at $r = 2M$ the typical non-zero component $R_{\hat{\alpha}\hat{\beta}\hat{\gamma}\hat{\delta}}$ of the curvature tensor is of the order $1/M^2$. The gravitational radius is a perfectly well-behaved, nonsingular region of spacetime, and nothing there can prevent the explorer from falling on inward.

Thus, the spacetime geometry is well behaved at $r = 2M$, but the coordinate system is pathological

By contrast, deep inside the gravitational radius, at $r = 0$, the traveler must encounter infinite tidal forces, independently of the route he uses to reach there. One says that "$r = 0$ is a physical singularity of spacetime." To see this, one need only calculate from equation (31.4b) or (31.6) the "curvature invariant":

$$I \equiv R_{\hat{\alpha}\hat{\beta}\hat{\gamma}\hat{\delta}} R^{\hat{\alpha}\hat{\beta}\hat{\gamma}\hat{\delta}} = 48M^2/r^6. \tag{31.7}$$

Box 31.1 THE "SCHWARZSCHILD SINGULARITY": HISTORICAL REMARKS

Although Eddington (1924) was the first to construct a coordinate system that is nonsingular at $r = 2M$, he seems not to have recognized the significance of his result. Lemaître (1933c, especially p. 82) appears to have been the first to recognize that the so-called "Schwarzschild singularity" at $r = 2M$ is not a singularity. He wrote, "La singularité du champ de Schwarzschild est donc une singularité fictive, analogue à celle qui se présentait a l'horizon du centre dans la forme originale de l'univers de de Sitter". He also provided a coordinate system to go through $r = 2M$. However, his coordinate system, like Eddington's, covered only half of the Schwarzschild geometry: regions I and II of Figure 31.3. Synge (1950) was the first to discover the incompleteness in the Eddington and Lemaître coordinate systems, and to provide coordinates that cover the entire geometry (regions I, II, III, IV of Figure 31.3). Fronsdal (1959), unaware of Synge's work, rediscovered the global structure of the Schwarzschild geometry by means of embedding diagrams and calculations. The coordinate system that provides maximum insight into the Schwarzschild geometry is the one generally known as the Kruskal-Szekeres coordinate system. It was constructed independently by Kruskal (1960) and by Szekeres (1960).

In every local Lorentz frame this will be a sum of products of curvature components, and it will have the same value $48M^2/r^6$. Thus, in every local Lorentz frame, including the traveler's, ***Riemann*** will have one or more infinite components as $r \longrightarrow 0$; i.e., tidal forces will become infinite.

At $r = 0$ the curvature is infinite

EXERCISE

Exercise 31.1. TIDAL FORCES ON INFALLING EXPLORER

(a) Carry out the details of the derivation of the Riemann tensor components (31.6).

(b) Calculate, roughly, the critical mass M_{crit} such that, if $M > M_{\text{crit}}$ the explorer's body (a human body made of normal flesh and bones) can withstand the tidal forces at $r = 2M$, but if $M < M_{\text{crit}}$ his body is mutilated by them. [Answer: $M_{\text{crit}} \sim 1000M_{\odot}$. Evidently, if $M \sim M_{\odot}$ the physicist should transform himself into an ant before taking the plunge! For details see §32.6.]

---

§31.3. BEHAVIOR OF SCHWARZSCHILD COORDINATES AT $r = 2M$

Since the spacetime geometry is well behaved at the gravitational radius, the singular behavior there of the Schwarzschild metric components, $g_{tt} = -(1 - 2M/r)$ and $g_{rr} = (1 - 2M/r)^{-1}$, must be due to a pathology there of the Schwarzschild coordinates t, r, θ, ϕ. Somehow one must find a way to get rid of that pathology—i.e., one must construct a new coordinate system from which the pathology is absent. Before doing this, it is helpful to understand better the precise nature of the pathology.

Nature of the coordinate pathology at $r = 2M$:

The most obvious pathology at $r = 2M$ is the reversal there of the roles of t and r as timelike and spacelike coordinates. In the region $r > 2M$, the t direction, $\partial/\partial t$, is timelike ($g_{tt} < 0$) and the r direction, $\partial/\partial r$, is spacelike ($g_{rr} > 0$); but in the region $r < 2M$, $\partial/\partial t$ is spacelike ($g_{tt} > 0$) and $\partial/\partial r$ is timelike ($g_{rr} < 0$).

(1) t and r reverse roles as timelike and spacelike coordinates

What does it mean for r to "change in character from a spacelike coordinate to a timelike one"? The explorer in his jet-powered spaceship prior to arrival at $r = 2M$ always has the option to turn on his jets and change his motion from decreasing r (infall) to increasing r (escape). Quite the contrary is the situation when he has once allowed himself to fall inside $r = 2M$. Then the further decrease of r represents the passage of time. No command that the traveler can give to his jet engine will turn back time. That unseen power of the world which drags everyone forward willy-nilly from age twenty to forty and from forty to eighty also drags the rocket in from time coordinate $r = 2M$ to the later value of the time coordinate $r = 0$. No human act of will, no engine, no rocket, no force (see exercise 31.3) can make time stand still. As surely as cells die, as surely as the traveler's watch ticks away "the unforgiving minutes," with equal certainty, and with never one halt along the way, r drops from $2M$ to 0.

At $r = 2M$, where r and t exchange roles as space and time coordinates, g_{tt} vanishes while g_{rr} is infinite. The vanishing of g_{tt} suggests that *the surface $r = 2M$, which*

(2) the region $r = 2M$, $-\infty < t < +\infty$ is two-dimensional rather than three

appears to be three-dimensional in the Schwarzschild coordinate system ($-\infty < t < +\infty$, $0 < \theta < \pi$, $0 < \phi < 2\pi$) *has zero volume and thus is actually only two-dimensional, or else is null;* thus,

$$\begin{aligned} &\int_{r=2M} |g_{tt} g_{\theta\theta} g_{\phi\phi}|^{1/2}\, dt\, d\theta\, d\phi = 0; \\ &\int_{(r=2M,\, t=\text{const})} |g_{\theta\theta} g_{\phi\phi}|^{1/2}\, d\theta\, d\phi = 4\pi(2M)^2. \end{aligned} \tag{31.8}$$

The divergence of g_{rr} at $r = 2M$ does *not* mean that $r = 2M$ is infinitely far from all other regions of spacetime. On the contrary, the proper distance from $r = 2M$ to a point with arbitrary r is

$$\int_{2M}^{r} |g_{rr}|^{1/2}\, dr = \begin{cases} [r(r-2M)]^{1/2} + 2M \ln |(r/2M - 1)^{1/2} + (r/2M)^{1/2}| \\ \qquad\qquad \text{when } r > 2M, \\ -2M \cot^{-1}[r^{1/2}/(2M - r)^{1/2}] - [r(2M - r)]^{1/2} \\ \qquad\qquad \text{when } r < 2M, \end{cases} \tag{31.9}$$

which is finite for all $0 < r < \infty$.

Just how the region $r < 2M$ is physically connected to the region $r > 2M$ can be discovered by examining the radial geodesics of the Schwarzschild metric. Focus attention, for concreteness, on the trajectory of a test particle that gets ejected from the singularity at $r = 0$, flies radially outward through $r = 2M$, reaches a maximum radius r_{max} ("top of orbit") at proper time $\tau = 0$ and coordinate time $t = 0$, and then falls back down through $r = 2M$ to $r = 0$. The solution of the geodesic equation for such an orbit was derived in §25.5 and described in Figure 25.3. It has the "cycloid form" (with the parameter η running from $-\pi$ to $+\pi$),

$$r = \frac{1}{2} r_{\text{max}}(1 + \cos \eta), \tag{31.10a}$$

$$\tau = (r_{\text{max}}{}^3/8M)^{1/2}(\eta + \sin \eta), \tag{31.10b}$$

$$\begin{aligned} t = 2M \ln &\left| \frac{(r_{\text{max}}/2M - 1)^{1/2} + \tan(\eta/2)}{(r_{\text{max}}/2M - 1)^{1/2} - \tan(\eta/2)} \right| \\ &+ 2M\left(\frac{r_{\text{max}}}{2M} - 1\right)^{1/2} \left[\eta + \left(\frac{r_{\text{max}}}{4M}\right)(\eta + \sin \eta)\right]. \end{aligned} \tag{31.10c}$$

Figure 31.1 plots this orbit in the r, t-coordinate plane (curve F-F'-F''), along with several other types of radial geodesics.

(3) radial geodesics reveal that the regions $r = 2M$, $t = \pm\infty$ are "finite" parts of spacetime

Every radial geodesic except a "set of geodesics of measure zero" crosses the gravitational radius at $t = +\infty$ (or at $t = -\infty$, or both), according to Figure 31.1 and the calculations behind that figure (exercises for the student! See Chapter 25). One therefore suspects that all the physics at $r = 2M$ is consigned to $t = \pm\infty$ by reason of some unhappiness in the choice of the Schwarzschild coordinates. A better coordinate system, one begins to believe, will take these two "points at infinity" and

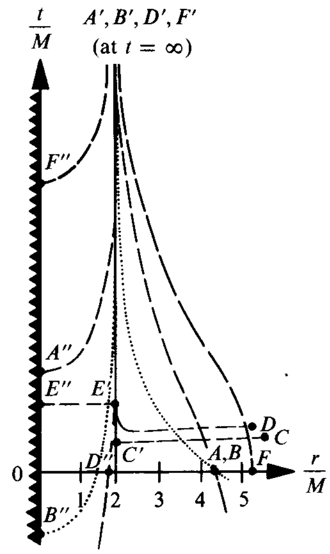


Figure 31.1.
Typical radial geodesics of the Schwarzschild geometry, as charted in Schwarzschild coordinates (schematic). $FF'F''$ [see equations (31.10)] is the timelike geodesic of a test particle that starts at rest at $r = 5.2M$ and falls straight in, arriving in a finite proper time at the singularity $r = 0$ (zig-zag marking). The unhappiness of the Schwarzschild coordinate system shows in two ways: (1) in the fact that t goes to ∞ partway through the motion; and (2) in the fact that t thereafter decreases as τ (not shown) continues to increase. The course of the same trajectory prior to $t = 0$ may be constructed by reflecting the diagram in the horizontal axis ("time inversion"). The time-reversed image of F'' marks the ejection of the test particle from the singularity. $AA'A''$ is a timelike geodesic which comes in from $r = +\infty$. $BB'B''$ is the null geodesic travelled by a photon that falls straight in (no summit; never at rest!). $DD'D''$ is a spacelike radial geodesic. So is CC', but $E'E''$ is timelike. Neither of the latter two ever succeed in crossing $r = 2M$. (Unanswered questions about these geodesics will answer themselves in Figure 31.4, where the same world lines are charted in a "Kruskal-Szekeres diagram").

Described mathematically via equation (31.10), the geodesic $F''_{\text{inverse}} F'_{\text{inverse}} FF'F''$ starts with ejection at

$$r = 0 \text{ at } t = -2\pi M\left(\frac{r_{\max}}{2M} - 1\right)^{1/2}\left(\frac{r_{\max}}{4M} + 1\right), \quad \tau = -\frac{\pi}{2}\left(\frac{r_{\max}{}^3}{2M}\right)^{1/2};$$

it flies outward with increasing proper time τ, but decreasing coordinate time, t, until it reaches the gravitational radius

$$r = 2M \text{ at } t = -\infty, \quad \tau = -\left(\frac{r_{\max}{}^3}{8M}\right)^{1/2}\cos^{-1}\left(\frac{4M}{r_{\max}} - 1\right) - r_{\max}\left(1 - \frac{2M}{r_{\max}}\right)^{1/2};$$

it then continues to fly on outward, but with coordinate time now increasing from $t = -\infty$, until it reaches its maximum radius

$$r = r_{\max} \text{ at } t = 0, \quad \tau = 0 \text{ (event } F \text{ in diagram);}$$

it then falls inward, with t continuing to increase, until it crosses the gravitational radius again

$$r = 2M \text{ at } t = +\infty, \quad \tau = +\left(\frac{r_{\max}{}^3}{8M}\right)^{1/2}\cos^{-1}\left(\frac{4M}{r_{\max}} - 1\right) + r_{\max}\left(1 - \frac{2M}{r_{\max}}\right)^{1/2}$$

(event F' in diagram);

and it finally falls on in with decreasing t (but, of course, still increasing τ) to

$$r = 0 \text{ at } t = +2\pi M\left(\frac{r_{\max}}{2M} - 1\right)^{1/2}\left(\frac{r_{\max}}{4M} + 1\right), \quad \tau = +\frac{\pi}{2}\left(\frac{r_{\max}{}^3}{2M}\right)^{1/2}$$

(event F'' in diagram).

spread them out into a line in a new $(r_{\text{new}}, t_{\text{new}})$-plane; and will squeeze the "line" ($r = 2M$, t from $-\infty$ to $+\infty$) into a single point in the $(r_{\text{new}}, t_{\text{new}})$-plane. One is the more prepared to accept this tentative conclusion and act on it because one has already seen (equation 31.8) that the region covering the (θ, ϕ) 2-sphere at $r = 2M$, and extending from $t = -\infty$ to $t = +\infty$, has zero proper volume. What timelier indication could one want that the "line" $r = 2M$, $-\infty < t < \infty$, is actually a point?

§31.4. SEVERAL WELL-BEHAVED COORDINATE SYSTEMS

Novikov coordinates:
(1) how constructed

The well-behaved coordinate system that is easiest to visualize is one in which the radially moving test particles of equations (31.10) remain always at rest ("comoving coordinates"). Such coordinates were first used by Novikov (1963). Novikov attaches a specific value of his radial coordinate, R^*, to each test particle as it emerges from the singularity of infinite tidal forces at $r = 0$, and insists that the particle carry that value of R^* throughout its "cycloidal life"—up through $r = 2M$ to $r = r_{\text{max}}$, then back down through $r = 2M$ to $r = 0$. For definiteness, Novikov expresses the R^* value for each particle in terms of the peak point on its trajectory by

$$R^* = (r_{\text{max}}/2M - 1)^{1/2}. \tag{31.11}$$

As a time coordinate, Novikov uses proper time τ of the test particles, normalized so $\tau = 0$ at the peak of the orbit. Every particle in the swarm is ejected in such a manner that it arrives at the summit of its trajectory ($r = r_{\text{max}}$, $\tau = 0$) at one and the same value of the Schwarzschild coordinate time; namely, at $t = 0$.

Simple though they may be conceptually, the Novikov coordinates are related to the original Schwarzschild coordinates by a very complicated transformation: (1) combine equations (31.10b) and (31.11) to obtain $\eta(\tau, R^*)$; (2) combine $\eta(\tau, R^*)$ with (31.10a) and (31.11) to obtain $r(\tau, R^*)$; (3) combine $\eta(\tau, R^*)$ with (31.10c) and (31.11) to obtain $t(\tau, R^*)$. The resulting coordinate transformation, when applied to the Schwarzschild metric (31.1), yields the line element

(2) line element

$$ds^2 = -d\tau^2 + \left(\frac{R^{*2}+1}{R^{*2}}\right)\left(\frac{\partial r}{\partial R^*}\right)^2 dR^{*2} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{31.12a}$$

(*"Schwarzschild geometry in Novikov coordinates"*.) Here r is no longer a radial coordinate; it is now a metric function $r(\tau, R^*)$ given implicitly by

$$\frac{\tau}{2M} = \pm (R^{*2}+1)\left[\frac{r}{2M} - \frac{(r/2M)^2}{R^{*2}+1}\right]^{1/2} + (R^{*2}+1)^{3/2}\cos^{-1}\left[\left(\frac{r/2M}{R^{*2}+1}\right)^{1/2}\right]. \tag{31.12b}$$

Figure 31.2 shows the locations of several key regions of Schwarzschild spacetime in this coordinate system. The existence of two distinct regions with $r = 0$ (singularities) and two distinct regions with $r \longrightarrow \infty$ (asymptotically flat regions; recall that $4\pi r^2$ = surface area!) will be discussed in §31.5.

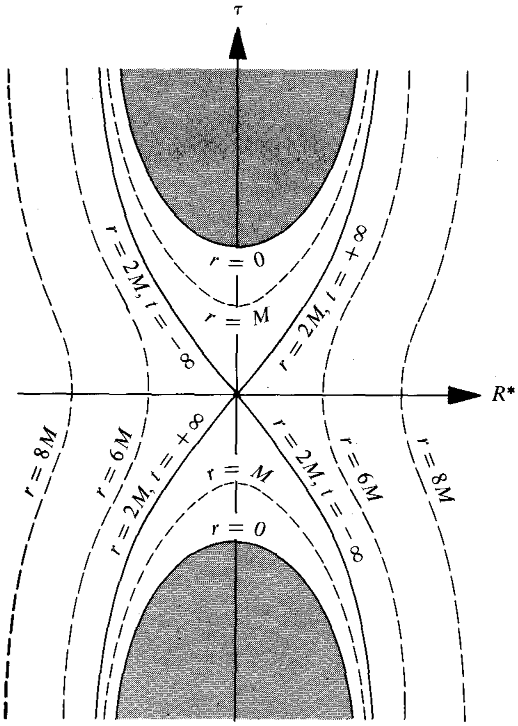


Figure 31.2.
The Novikov coordinate system for Schwarzschild spacetime (schematic). The dashed curves are curves of constant r (recall: $4\pi r^2$ = surface area about center of symmetry). The region shaded gray is not part of spacetime; it corresponds to $r < 0$, a region that cannot be reached because of the singularity of spacetime at $r = 0$. Notice that the "line" ($r = 2M$, $-\infty < t < +\infty$) of the Schwarzschild coordinate diagram (Figure 31.1) has been compressed into a point here, in accordance with the discussion at the end of §31.3.

Although Novikov's coordinate system is very simple conceptually, the mathematical expressions for the metric components in it are rather unwieldy. Simpler, more usable expressions have been obtained in a different coordinate system ("Kruskal-Szekeres coordinates") by Kruskal (1960), and independently by Szekeres (1960).

Kruskal-Szekeres coordinates

Kruskal and Szekeres use a dimensionless radial coordinate u and a dimensionless time coordinate v related to the Schwarzschild r and t by

$$\left.\begin{aligned} u &= (r/2M - 1)^{1/2} e^{r/4M} \cosh(t/4M) \\ v &= (r/2M - 1)^{1/2} e^{r/4M} \sinh(t/4M) \end{aligned}\right\} \text{when } r > 2M, \tag{31.13a}$$

$$\left.\begin{aligned} u &= (1 - r/2M)^{1/2} e^{r/4M} \sinh(t/4M) \\ v &= (1 - r/2M)^{1/2} e^{r/4M} \cosh(t/4M) \end{aligned}\right\} \text{when } r < 2M. \tag{31.13b}$$

(Motivation for introducing such coordinates is given in Box 31.2.) By making this change of coordinates in the Schwarzschild metric (31.1), one obtains the following line element:

$$ds^2 = (32M^3/r)e^{-r/2M}(-dv^2 + du^2) + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{31.14a}$$

("*Schwarzschild geometry in Kruskal-Szekeres coordinates*"). Here r is to be regarded as a function of u and v defined implicitly by

$$(r/2M - 1)e^{r/2M} = u^2 - v^2 \tag{31.14b}$$

[cf. equations (31.13)].

(*continued on page 833*)

Box 31.2 MOTIVATION FOR KRUSKAL-SZEKERES COORDINATES*

A. EDDINGTON-FINKELSTEIN COORDINATES

The motivation for the Kruskal-Szekeres system begins by introducing a different coordinate system, first devised by Eddington (1924) and rediscovered by Finkelstein (1958). Eddington and Finkelstein use as the foundation of their coordinate system, not freely falling particles as did Novikov, but freely falling photons. More particularly, they introduce coordinates $\tilde{U}$ and $\tilde{V}$, which are labels for outgoing and ingoing, radial, null geodesics. The geodesics are given by

$$ds^2 = 0 = -(1 - 2M/r)\,dt^2 + (1 - 2M/r)^{-1}\,dr^2.$$

Equivalently, outgoing geodesics are given by $\tilde{U} = \text{const}$, where

$$\tilde{U} \equiv t - r^*; \tag{1a}$$

and ingoing geodesics are given by $\tilde{V} = \text{const}$, where

$$\tilde{V} \equiv t + r^*. \tag{1b}$$

Here r^* is the "tortoise coordinate" of §25.5 and Figure 25.4:

$$r^* \equiv r + 2M \ln |r/2M - 1|. \tag{1c}$$

Ingoing Eddington-Finkelstein Coordinates—Adopt r and $\tilde{V}$ as coordinates in place of r and t

The Schwarzschild metric becomes.

$$ds^2 = -(1 - 2M/r)\,d\tilde{V}^2 + 2\,d\tilde{V}\,dr + r^2\,d\Omega^2. \tag{2}$$

The radial light cone, $ds^2 = 0$, has one leg

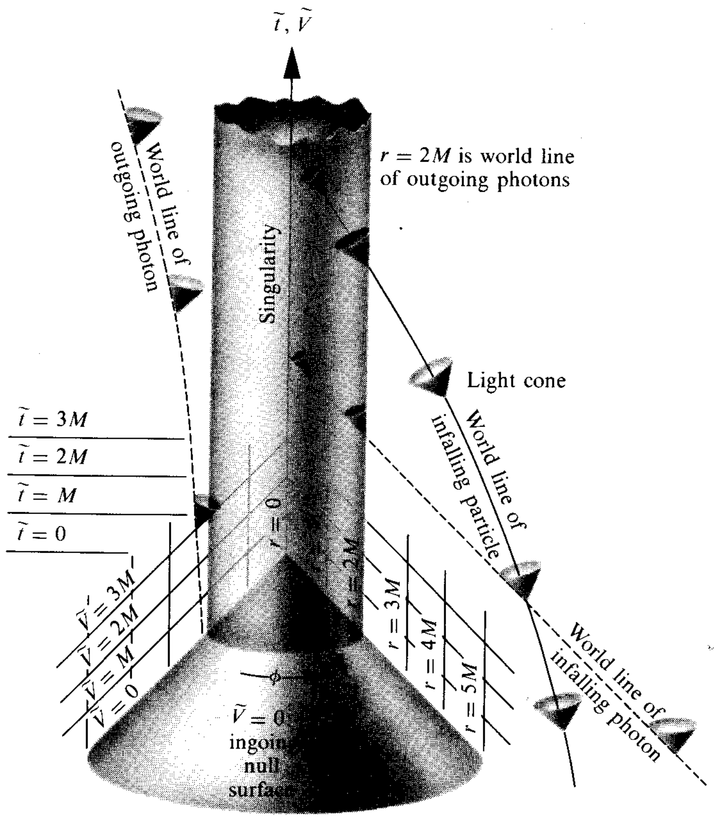
$$d\tilde{V}/dr = 0, \tag{3a}$$

and the other leg

$$\frac{d\tilde{V}}{dr} = \frac{2}{1 - 2M/r}. \tag{3b}$$

From this, and this alone, one can infer all features of the drawing.

*This box is based on Misner (1969a).



Ingoing Eddington-Finkelstein coordinates (one rotational degree of freedom is suppressed; i.e., θ is set equal to $\pi/2$). Surfaces of constant $\tilde{V}$, being ingoing null surfaces, are plotted on a 45-degree slant, just as they would be in flat spacetime. Equivalently, surfaces of constant

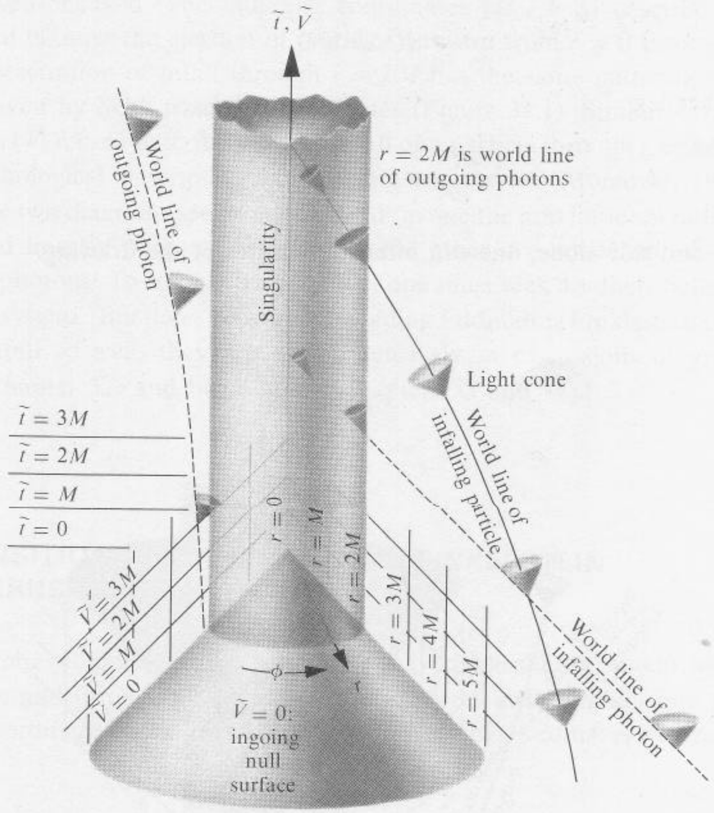
$$\tilde{t} \equiv \tilde{V} - r = t + 2M \ln|r/2M - 1|$$

are plotted as horizontal surfaces.

Outgoing Eddington-Finkelstein Coordinates—Adopt r and $\tilde{U}$ as coordinates in place of r and t

The Schwarzschild metric becomes

$$ds^2 = -(1 - 2M/r)\, d\tilde{U}^2 - 2\, d\tilde{U}\, dr + r^2\, d\Omega^2. \tag{4}$$



Ingoing Eddington-Finkelstein coordinates (one rotational degree of freedom is suppressed; i.e., θ is set equal to $\pi/2$). Surfaces of constant $\tilde{V}$, being ingoing null surfaces, are plotted on a 45-degree slant, just as they would be in flat spacetime. Equivalently, surfaces of constant

$$\tilde{t} \equiv \tilde{V} - r = t + 2M \ln |r/2M - 1|$$

are plotted as horizontal surfaces.

Outgoing Eddington-Finkelstein Coordinates—Adopt *r* and $\tilde{U}$ as coordinates in place of *r* and *t*

The Schwarzschild metric becomes

$$ds^2 = -(1 - 2M/r)\, d\tilde{U}^2 - 2\, d\tilde{U}\, dr + r^2\, d\Omega^2. \tag{4}$$

Box 31.2 (*continued*)

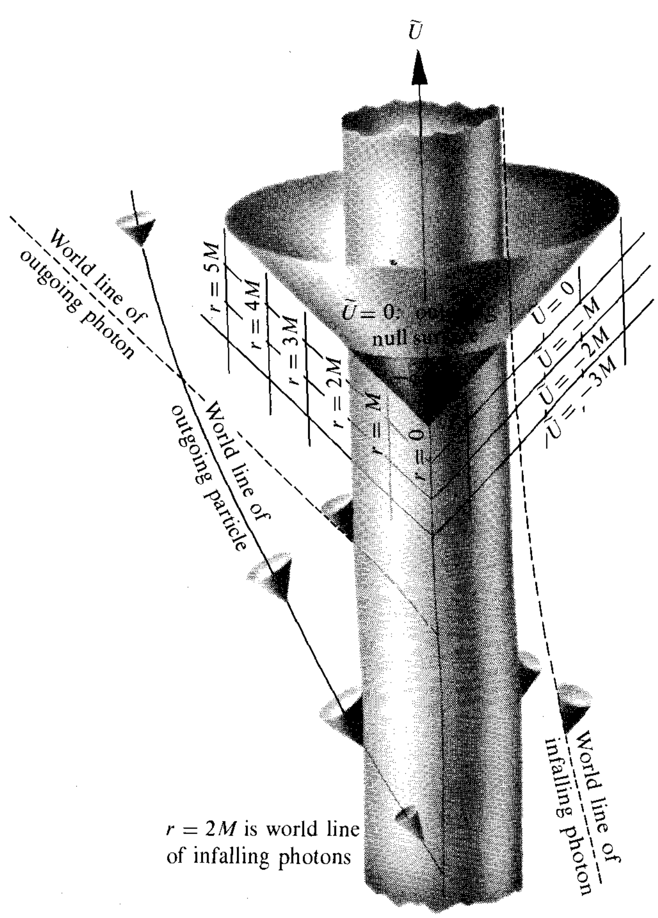
The radial light cone, $ds^2 = 0$, has one leg

$$d\tilde{U}/dr = 0, \tag{5a}$$

and the other leg

$$\frac{d\tilde{U}}{dr} = -\frac{2}{1 - 2M/r}. \tag{5b}$$

From this, and this alone, one can infer all features of the drawing.



Outgoing Eddington-Finkelstein coordinates (one rotational degree of freedom is suppressed). (Surfaces of constant $\tilde{U}$, being outgoing null surfaces, are plotted on a 45-degree slant, just as they would be in flat spacetime.)

Box 31.2 (*continued*)

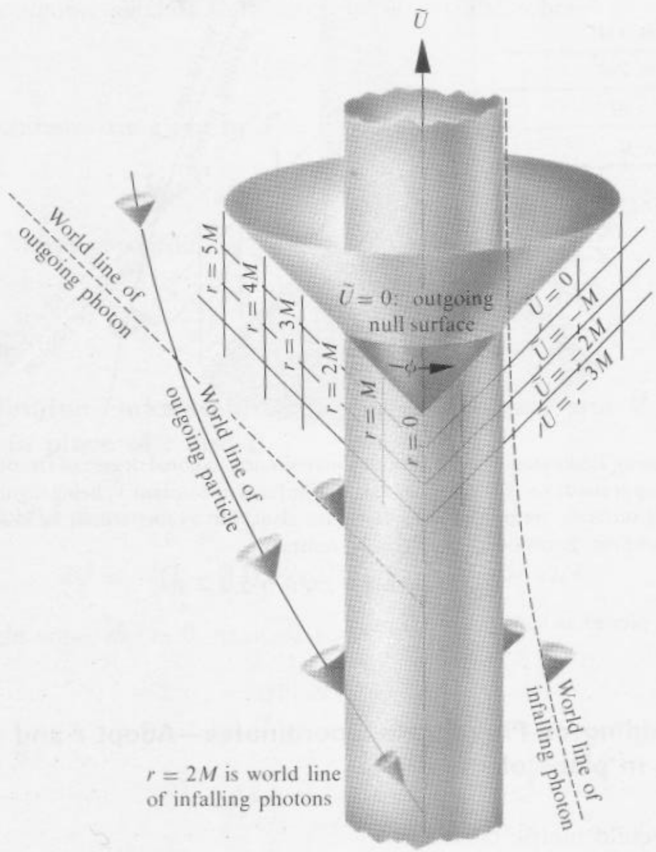
The radial light cone, $ds^2 = 0$, has one leg

$$d\tilde{U}/dr = 0, \tag{5a}$$

and the other leg

$$\frac{d\tilde{U}}{dr} = -\frac{2}{1 - 2M/r}. \tag{5b}$$

From this, and this alone, one can infer all features of the drawing.



Outgoing Eddington-Finkelstein coordinates (one rotational degree of freedom is suppressed). (Surfaces of constant $\tilde{U}$, being outgoing null surfaces, are plotted on a 45-degree slant, just as they would be in flat spacetime.)

Notice that both Eddington-Finkelstein coordinate systems are better behaved at the gravitational radius than is the Schwarzschild coordinate system; but they are not *fully* well-behaved. The outgoing coordinates $(\tilde{U}, r, \theta, \phi)$ describe in a non-pathological manner the ejection of particles outward from $r = 0$ through $r = 2M$; but their description of infall through $r = 2M$ has the same pathology as the description given by Schwarzschild coordinates (Figure 31.1). Similarly, the ingoing coordinates $(\tilde{V}, r, \theta, \phi)$ describe well the infall of a particle through $r = 2M$, but they give a pathological description of outgoing trajectories. Moreover, the contrast between the two diagrams seems paradoxical: in one the gravitational radius is made up of world lines of outgoing photons; in the other it is made up of world lines of ingoing photons! To resolve the paradox, one must seek another, better-behaved coordinate system. [But *note:* because the ingoing Eddington-Finklestein coordinates describe infall so well, they are used extensively in discussions of gravitational collapse (Chapter 32) and black holes (Chapters 33 and 34).]

B. TRANSITION FROM EDDINGTON-FINKELSTEIN TO KRUSKAL-SZEKERES

Perhaps one would obtain a fully well-behaved coordinate system by dropping r from view and using $\tilde{U}$, $\tilde{V}$, as the two coordinates in the radial-time plane. The resulting coordinate system is related to Schwarzschild coordinates by [see equations (1)]

$$\tilde{V} - \tilde{U} = 2r^*, \tag{6a}$$

$$\tilde{V} + \tilde{U} = 2t; \tag{6b}$$

and the line element in terms of the new coordinates reads

$$ds^2 = -(1 - 2M/r)\, d\tilde{U}\, d\tilde{V} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{7}$$

Contrary to one's hopes, this coordinate system is pathological at $r = 2M$.

Second thoughts about the construction reveal the trouble: the surfaces $\tilde{U} =$ constant (outgoing null surfaces) used in constructing it are geometrically well-defined, as are the surfaces $\tilde{V} =$ constant (ingoing null surfaces); but the way of labeling them is not. Any relabeling, $\tilde{u} = F(\tilde{U})$ and $\tilde{v} = G(\tilde{V})$, will leave the surfaces unchanged physically. What one needs is a relabeling that will get rid of the singular factor $1 - 2M/r$ in the line element (7). A successful relabeling is suggested by the equation

$$\exp[(\tilde{V} - \tilde{U})/4M] = \exp(r^*/2M) = (r/2M - 1)\exp(r/2M), \tag{8}$$

which follows from equations (6a) and (1c). Experimenting with this relation quickly reveals that the relabeling

$$\tilde{u} \equiv -e^{-\tilde{U}/4M} = -(r/2M - 1)^{1/2} e^{r/4M} e^{-t/4M}, \tag{9a}$$

$$\tilde{v} \equiv e^{+\tilde{V}/4M} = (r/2M - 1)^{1/2} e^{r/4M} e^{t/4M}, \tag{9b}$$

will remove the offending $1 - 2M/r$ from the metric coefficients. In terms of these new coordinates, the line element reads

$$ds^2 = -(32M^3/r)e^{-r/2M}\, d\tilde{v}\, d\tilde{u} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{10a}$$

Here r is still defined by $4\pi r^2 =$ surface area, but it must be regarded as a function of $\tilde{v}$ and $\tilde{u}$:

$$(r/2M - 1)e^{r/2M} = -\tilde{u}\tilde{v}. \tag{10b}$$

One can readily verify that this equation determines r uniquely (recall: $r > 0$!) in terms of the product $\tilde{u}\tilde{v}$ [details in Misner (1969a)].

The coordinates, $\tilde{u}$, $\tilde{v}$, which label the ingoing and outgoing null surfaces, are null coordinates; i.e.,

$$\partial/\partial\tilde{u} \cdot \partial/\partial\tilde{u} = g_{\tilde{u}\tilde{u}} = 0, \qquad \partial/\partial\tilde{v} \cdot \partial/\partial\tilde{v} = g_{\tilde{v}\tilde{v}} = 0$$

[see equation (10a)]. If one is not accustomed to working with null coordinates, it is helpful to replace $\tilde{u}$ and $\tilde{v}$ by spacelike and timelike coordinates, u and v (Kruskal-Szekeres coordinates!) defined by

$$u \equiv \frac{1}{2}(\tilde{v} - \tilde{u}) = (r/2M - 1)^{1/2} e^{r/4M} \cosh(t/4M), \tag{11a}$$

$$v \equiv \frac{1}{2}(\tilde{v} + \tilde{u}) = (r/2M - 1)^{1/2} e^{r/4M} \sinh(t/4M), \tag{11b}$$

so that

$$dv^2 - du^2 = d\tilde{v}\, d\tilde{u}. \tag{12}$$

In terms of these coordinates, the line element has the Kruskal form (31.14), which is fully well-behaved at the gravitational radius.

Although the Kruskal-Szekeres line element is well behaved at $r = 2M$, the transformation (11) from Schwarzschild to Kruskal-Szekeres is not; it becomes meaningless (u and v "imaginary") when one moves from $r > 2M$ to $r < 2M$. Of course, this is a manifestation of the pathologies of Schwarzschild coordinates. By trial and error, one readily finds a new transformation, to replace (11) at $r < 2M$, leading from Schwarzschild to Kruskal-Szekeres coordinates:

$$u = (1 - r/2M)^{1/2} e^{r/4M} \sinh(t/4M), \tag{11c}$$

$$v = (1 - r/2M)^{1/2} e^{r/4M} \cosh(t/4M). \tag{11d}$$

§31.5. RELATIONSHIP BETWEEN KRUSKAL-SZEKERES COORDINATES AND SCHWARZSCHILD COORDINATES

Kruskal-Szekeres coordinates reveal that Schwarzschild spacetime has two "$r = 0$ singularities" and two "$r \longrightarrow \infty$ exterior regions"

In the Kruskal-Szekeres coordinate system, the singularity $r = 0$ is located at $v^2 - u^2 = 1$. Thus there are actually *two* singularities, not one; both

$$v = +(1 + u^2)^{1/2} \text{ and } v = -(1 + u^2)^{1/2} \text{ correspond to } r = 0! \tag{31.15}$$

This is not the only surprise that lies hidden in the Kruskal-Szekeres line element (31.14). Notice also that $r \gg 2M$ (the region of spacetime far outside the gravitational radius) is given by $u^2 \gg v^2$. Thus there are actually *two* exterior regions*; both

$$u \gg +|v| \text{ and } u \ll -|v| \text{ correspond to } r \gg 2M! \tag{31.16}$$

How can this be? When the geometry is charted in Schwarzschild coordinates, it contains one singularity and one exterior region; but when expressed in Kruskal-Szekeres coordinates, it shows two of each. The answer must be that the Schwarzschild coordinates cover only part of the spacetime manifold; they must be only a local coordinate patch on the full manifold. Somehow, by means of the coordinate transformation that leads to Kruskal-Szekeres coordinates, one has analytically extended the limited Schwarzschild solution for the metric to cover all (or more nearly all) of the manifold.

To understand this covering more clearly, transform back from Kruskal-Szekeres coordinates to Schwarzschild coordinates (see Figure 31.3). The transformation equations, as written down in (31.13) were valid only for the quadrants $u > |v|$ [equation (31.13a)] and $v > |u|$ [equation (31.13b)] of Kruskal coordinates. Denote these quadrants by the numerals I and II; and denote the other quadrants by III and IV (see Figure 31.3). In the other quadrants, one can also transform the Kruskal-Szekeres line element (31.14) into the Schwarzschild line element (31.1); but slightly different transformation equations are needed. One easily verifies that the following sets of transformations work:

Transformation between Schwarzschild coordinates and Kruskal-Szekeres coordinates

$$\text{(I)} \begin{cases} u = (r/2M - 1)^{1/2} e^{r/4M} \cosh(t/4M) \\ v = (r/2M - 1)^{1/2} e^{r/4M} \sinh(t/4M) \end{cases}, \tag{31.17a}$$

$$\text{(II)} \begin{cases} u = (1 - r/2M)^{1/2} e^{r/4M} \sinh(t/4M) \\ v = (1 - r/2M)^{1/2} e^{r/4M} \cosh(t/4M) \end{cases}, \tag{31.17b}$$

$$\text{(III)} \begin{cases} u = -(r/2M - 1)^{1/2} e^{r/4M} \cosh(t/4M) \\ v = -(r/2M - 1)^{1/2} e^{r/4M} \sinh(t/4M) \end{cases}, \tag{31.17c}$$

$$\text{(IV)} \begin{cases} u = -(1 - r/2M)^{1/2} e^{r/4M} \sinh(t/4M) \\ v = -(1 - r/2M)^{1/2} e^{r/4M} \cosh(t/4M) \end{cases}. \tag{31.17d}$$

*The global structure of the Schwarzschild geometry, including the existence of two singularities and two exterior regions, was first discovered by Synge (1950). See Box 31.1.

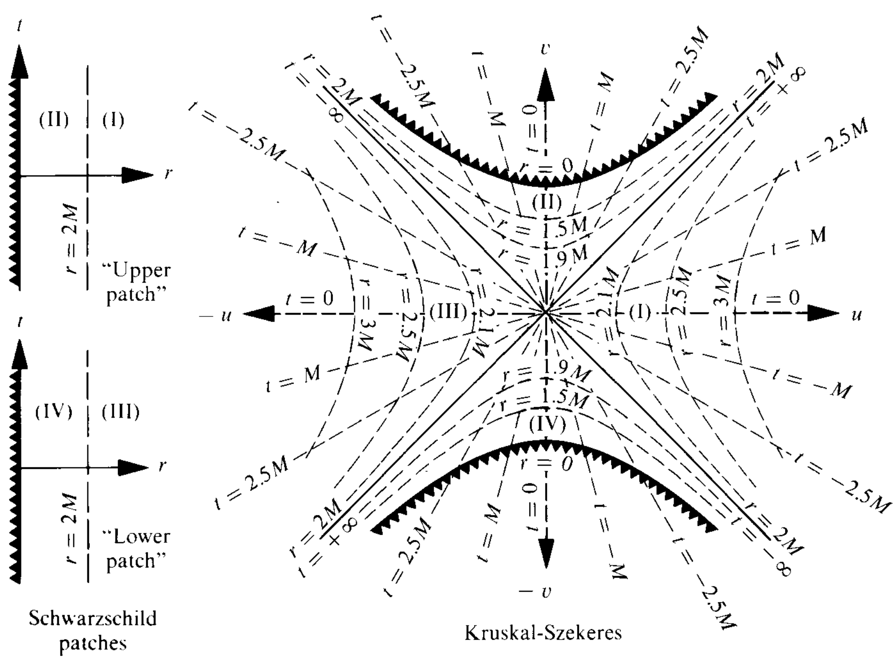


Figure 31.3.
The transformation of the Schwarzschild vacuum geometry between Schwarzschild and Kruskal-Szekeres coordinates. Two Schwarzschild coordinate patches I, II, and III, IV (illustrated in the upper and lower portions of Figure 31.5,a) are required to cover the complete Schwarzschild geometry, whereas a single Kruskal-Szekeres coordinate system suffices. The Schwarzschild geometry consists of four regions I, II, III, IV. Regions I and III represent two distinct, but identical, asymptotically flat universes in which $r > 2M$; while regions II and IV are two identical, but time-reversed, regions in which physical singularities ($r = 0$) evolve. The transformation laws that relate the Schwarzschild and Kruskal-Szekeres coordinate systems to each other are given by equations (31.17) and (31.18). In the Kruskal-Szekeres u,v-plane, curves of constant r are hyperbolae with asymptotes $u = \pm v$, while curves of constant t are straight lines through the origin.

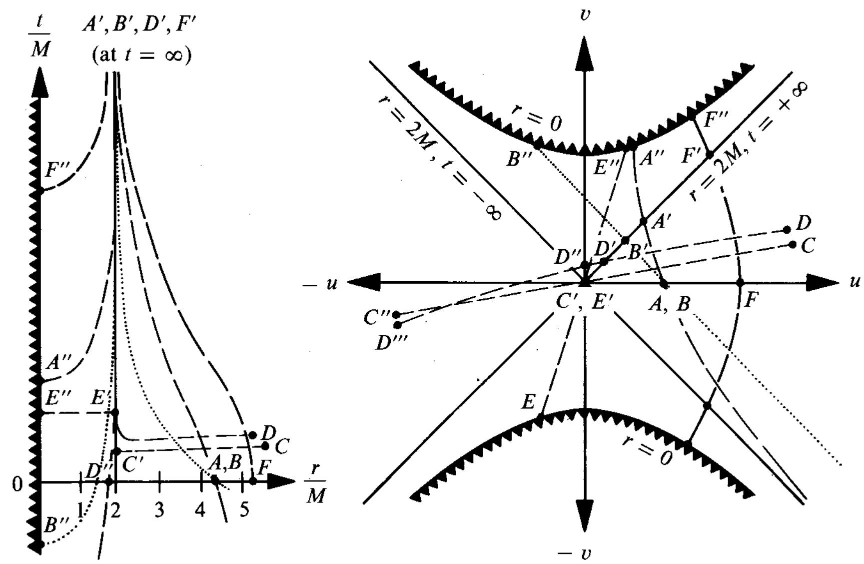
The inverse transformations are

$$(r/2M - 1)e^{r/2M} = u^2 - v^2 \text{ in I, II, III, IV;} \tag{31.18a}$$

$$t = \begin{cases} 4M \tanh^{-1}(v/u) \text{ in I and III,} \\ 4M \tanh^{-1}(u/v) \text{ in II and IV.} \end{cases} \tag{31.18b}$$

Two Schwarzschild coordinate patches are required to cover all of spacetime

These coordinate transformations are exhibited graphically in Figure 31.3. Notice that two Schwarzschild coordinate patches, I, II, and III, IV, are required to cover the entire Schwarzschild geometry; but a single Kruskal coordinate system suffices. Schwarzschild patch I, II, is divided into two regions—region I, which is outside the gravitational radius ($r > 2M$), and region II, which is inside the gravitational radius ($r < 2M$). Similarly, Schwarzschild patch III, IV, consists of an exterior region (III) and an interior region (IV).



(a)
Schwarzschild

(b)
Kruskal-Szekeres

Figure 31.4.
(a) Typical radial timelike (A, E, F), lightlike (B), and spacelike (C, D) geodesics of the Schwarzschild geometry, as seen in the Schwarzschild coordinate system (schematic only). This is a reproduction of Figure 31.1.

(b) The same geodesics, as seen in the Kruskal-Szekeres coordinate system, and as extended either to infinite length or to the singularity of infinite curvature at $r = 0$ (schematic only).

Equations (31.18) reveal that the regions of constant r (constant surface area) are hyperbolae with asymptotes $u = \pm v$ in the Kruskal-Szekeres diagram, and that regions of constant t are straight lines through the origin.

Properties of the Kruskal-Szekeres coordinate system

Several radial geodesics of the complete Schwarzschild geometry are depicted in the Kruskal-Szekeres coordinate system in Figure 31.4. Notice how much more reasonable the geodesic curves look in Kruskal-Szekeres coordinates than in Schwarzschild coordinates. Notice also that *radial, lightlike geodesics (paths of radial light rays) are 45-degree lines in the Kruskal-Szekeres coordinate system.* This can be seen from the Kruskal-Szekeres line element (31.14), for which $du = \pm dv$ guarantees $ds = 0$. Because of this 45-degree property, the radial light cone in a Kruskal-Szekeres diagram has the same form as in the space-time diagram of special relativity. Any radial curve that points "generally upward" (i.e., makes an angle of less than 45 degrees with the vertical, v, axis) is timelike; and curves that point "generally outward" are spacelike. This property enables a Kruskal-Szekeres diagram to exhibit easily the causality relation between one event in spacetime and another (see exercises 31.2 to 31.4).

EXERCISES

Exercise 31.2. NONRADIAL LIGHT CONES

Show that the world line of a photon traveling nonradially makes an angle less than 45 degrees with the vertical v-axis of a Kruskal-Szekeres coordinate diagram. From this, infer that particles with finite rest mass, traveling nonradially or radially, must always move "generally upward" (angle less than 45 degrees with vertical v-axis).

Exercise 31.3. THE CRACK OF DOOM

Use a Kruskal diagram to show the following.

(a) If a man allows himself to fall through the gravitational radius $r = 2M$, there is no way whatsoever for him to avoid hitting (and being killed in) the singularity at $r = 0$.

(b) Once a man has fallen inward through $r = 2M$, there is no way whatsoever that he can send messages out to his friends at $r > 2M$, but he can still receive messages from them (e.g., by radio waves, or laser beam, or infalling "CARE packages").

Exercise 31.4. HOW LONG TO LIVE?

Show that once a man falling inward reaches the gravitational radius, no matter what he does subsequently (no matter in what directions, how long, and how hard he blasts his rocket engines), he will be pulled into the singularity and killed in a proper time of

$$\tau < \tau_{\max} = \pi M = 1.54 \times 10^{-5}(M/M_{\odot}) \text{ seconds.} \tag{31.19}$$

[*Hint:* The trajectory of longest proper time lapse must be a geodesic. Use the mathematical tools of Chapter 25 to show that the geodesic of longest proper time lapse between $r = 2M$ and $r = 0$ is the radial geodesic (31.10a), with $r_{\max} = 2M$, for which the time lapse is πM.]

Exercise 31.5. EDDINGTON-FINKELSTEIN AND KRUSKAL-SZEKERES COMPARED

Use coordinate diagrams to compare the ingoing and outgoing Eddington-Finkelstein coordinates of Box 31.2 with the Kruskal-Szekeres coordinates. Pattern the comparison after that between Schwarzschild and Kruskal-Szekeres in Figures 31.3 and 31.4.

Exercise 31.6. ANOTHER COORDINATE SYSTEM

Construct a coordinate diagram for the $\tilde{U}$, $\tilde{V}$, θ, ϕ coordinate system of Box 31.2 [equations (6) and (7)]. Show such features as (1) the relationship to Schwarzschild and to Kruskal-Szekeres coordinates; (2) the location of $r = 2M$; and (3) radial geodesics.

§31.6. DYNAMICS OF THE SCHWARZSCHILD GEOMETRY

What does the Schwarzschild geometry look like? This question is most readily answered by means of embedding diagrams analogous to those for an equilibrium star (§23.8; Figure 23.1; and end of Box 23.2) and for Friedmann universes of positive and negative spatial curvature [equations (27.23) and (27.24) and Box 27.2].

Examine, first, the geometry of the spacelike hypersurface $v = 0$, which extends from $u = +\infty$ $(r = \infty)$ into $u = 0$ $(r = 2M)$ and then out to $u = -\infty$ $(r = \infty)$. In Schwarzschild coordinates this surface is a slice of constant time, $t = 0$ [see equation (31.18b)]; it is precisely the surface for which an embedding diagram was calculated in equation (23.34b). The embedded surface, with one degree of rotational freedom suppressed, is described by the paraboloid of revolution

$$\bar{r} = 2M + \bar{z}^2/8M \tag{31.20}$$

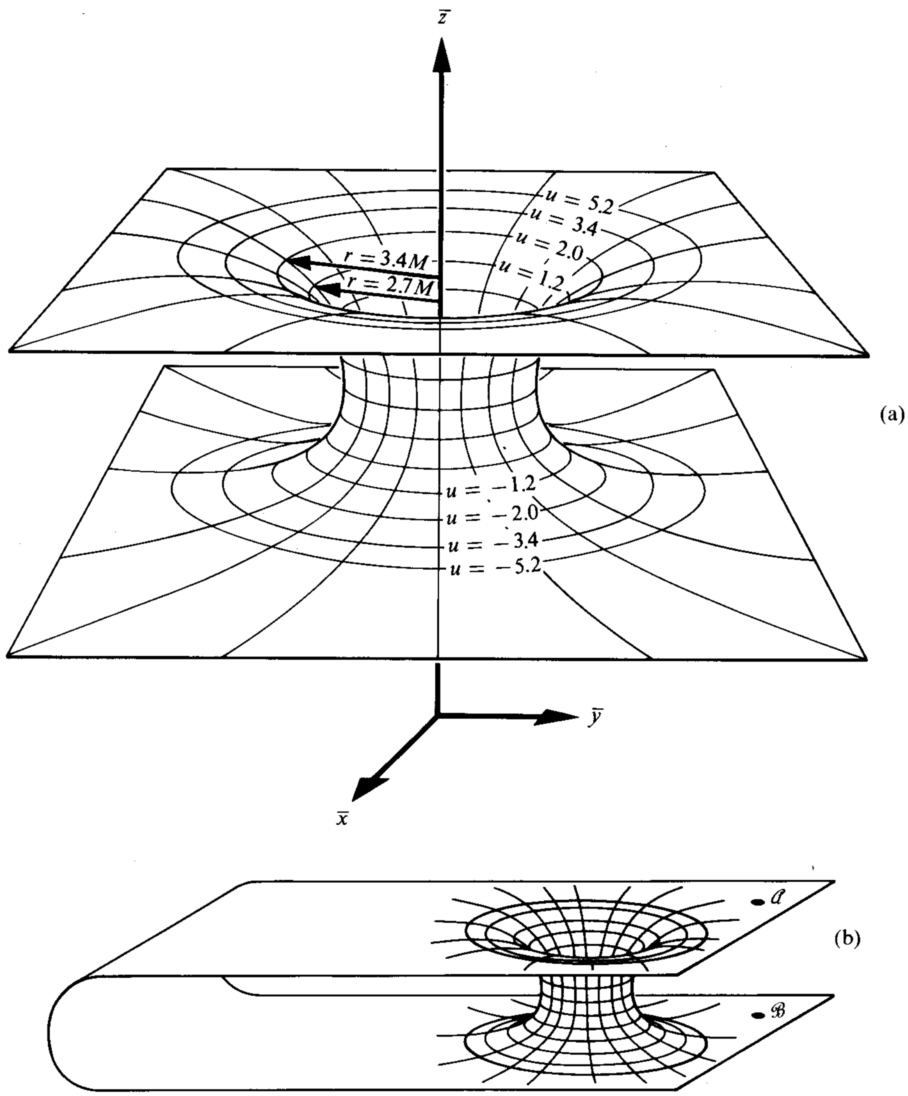


Figure 31.5.

(a) The Schwarzschild space geometry at the "moment of time" $t = v = 0$, with one degree of rotational freedom suppressed ($\theta = \pi/2$). To restore that rotational freedom and obtain the full Schwarzschild 3-geometry, one mentally replaces the circles of constant $\bar{r} = (\bar{x}^2 + \bar{y}^2)^{1/2}$ with spherical surfaces of area $4\pi\bar{r}^2$. Note that the resultant 3-geometry becomes flat (Euclidean) far from the throat of the bridge in both directions (both "universes").

(b) An embedding of the Schwarzschild space geometry at "time" $t = v = 0$, which is geometrically identical to the embedding (a), but which is topologically different. Einstein's field equations fix the local geometry of spacetime, but they do not fix its topology; see the discussion at end of Box 27.2. Here the Schwarzschild "wormhole" connects two distant regions of a single, asymptotically flat universe. For a discussion of issues of causality associated with this choice of topology, see Fuller and Wheeler (1962).

in the flat Euclidean space with metric

$$d\sigma^2 = d\bar{r}^2 + d\bar{z}^2 + \bar{r}^2\, d\bar{\phi}^2. \tag{31.21}$$

(See Figure 31.5.)

Notice from the embedding diagram of Figure 31.5,a, that the Schwarzschild

The 3-surface $v = t = 0$ is a "wormhole" connecting two asymptotically flat universes, or two different regions of one universe

geometry on the spacelike hypersurface t = const consists of a bridge or "wormhole" connecting two distinct, but identical, asymptotically flat universes. This bridge is sometimes called the "Einstein-Rosen bridge" and sometimes the "Schwarzschild throat" or the "Schwarzschild wormhole." If one so wishes, one can change the topology of the Schwarzschild geometry by connecting the two asymptotically flat universes together in a region distant from the Schwarzschild throat [Fuller and Wheeler (1962); Fig. 31.5b]. The single, unique universe then becomes multiply connected, with the Schwarzschild throat providing one spacelike path from point $\mathcal{A}$ to point $\mathcal{B}$, and the nearly flat universe providing another. For concreteness, focus attention on the interpretation of the Schwarzschild geometry, not in terms of Wheeler's multiply connected single universe, but rather in terms of the Einstein-Rosen double universe of Figure 31.5,a.

Schwarzschild geometry is dynamic in regions $r < 2M$

One is usually accustomed to think of the Schwarzschild geometry as static. However, the static "time translations," $t \longrightarrow t + \Delta t$, which leave the Schwarzschild geometry unchanged, are time translations in the strict sense of the words only in regions I and III of the Schwarzschild geometry. In regions II and IV, $t \longrightarrow t + \Delta t$ is a spacelike motion, not a timelike motion (see Fig. 31.3). Consequently, a spacelike hypersurface, such as the surface t = const of Figure 31.5,a, which extends from region I through $u = v = 0$ into region III, is *not* static. As this spacelike hypersurface is pushed forward in time (in the $+v$ direction of the Kruskal diagram), it enters region II, and its geometry begins to change.

In order to examine the time-development of the Schwarzschild geometry, one needs a sequence of embedding diagrams, each corresponding to the geometry of a spacelike hypersurface to the future of the preceding one. But how are the hypersurfaces to be chosen? In Newtonian theory or special relativity, one chooses hypersurfaces of constant time. But in dynamic regions of curved spacetime, no naturally preferred time coordinate exists. This situation forces one to make a totally *arbitrary* choice of hypersurfaces to use in visualizing the time-development of geometry, and to keep in mind how very arbitrary that choice was.

Time evolution of the wormhole: creation; expansion; recontraction; and pinch-off

Figure 31.6 uses two very different choices of hypersurfaces to depict the time-development of the Schwarzschild geometry. (Still other choices are shown in Figure 21.4.) Notice that the precise geometry of the evolving bridge depends on the arbitrary choice of spacelike hypersurfaces, but that the qualitative nature of the evolution is independent of the choice of hypersurfaces. Qualitatively speaking, the two asymptotically flat universes begin disconnected, with each one containing a singularity of infinite curvature ($r = 0$). As the two universes evolve in time, their singularities join each other and form a nonsingular bridge. The bridge enlarges, until it reaches a maximum radius at the throat of $r = 2M$ (maximum circumference of $4\pi M$; maximum surface area of $16\pi M^2$). It then contracts and pinches off, leaving the two universes disconnected and containing singularities ($r = 0$) once again. The formation, expansion, and collapse of the bridge occur so rapidly, that no particle or light ray can pass across the bridge from the faraway region of the one universe to the faraway region of the other without getting caught and crushed in the throat as it pinches off. (To verify this, examine the Kruskal-Szekeres diagram of Figure 31.3, where radial light rays move along 45-degree lines.)

Communication through the wormhole is impossible: it pinches off too fast

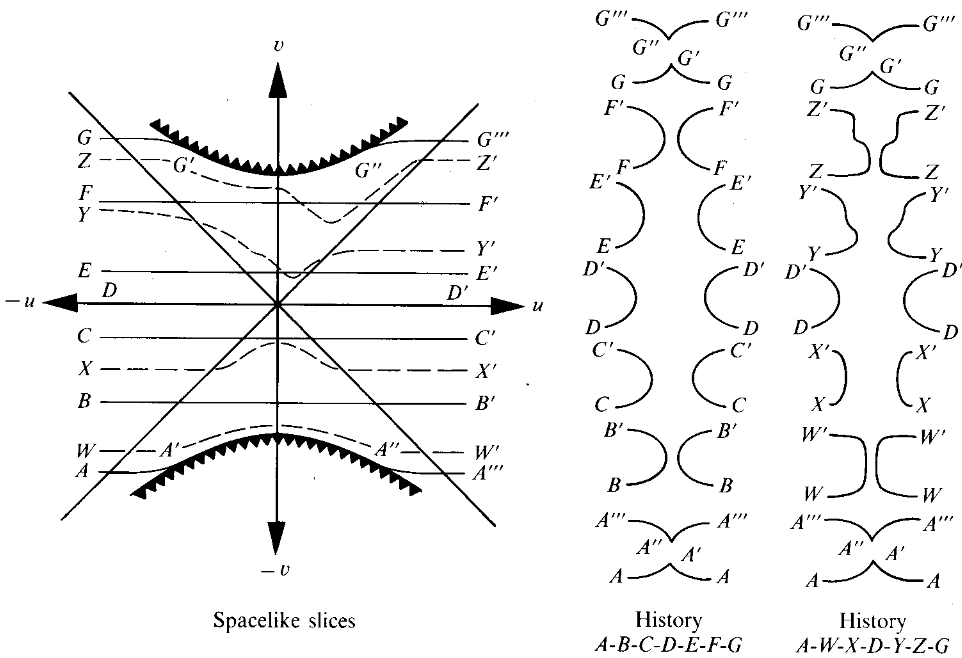


Figure 31.6.
Dynamical evolution of the Einstein-Rosen bridge of the vacuum Schwarzschild geometry (schematic). Shown here are two sequences of embedding diagrams corresponding to two different ways of viewing the evolution of the bridge—History A-B-C-D-E-F-G, and History A-W-X-D-Y-Z-G. The embedding diagrams are skeletonized in that each diagram must be rotated about the appropriate vertical axis in order to become two-dimensional surfaces analogous to Figure 31.5,a. [Notice that the hypersurfaces of which embedding diagrams are given intersect the singularity only tangentially. Hypersurfaces that intersect the singularity at a finite angle in the u,v-plane are not shown because they cannot be embedded in a Euclidean space. Instead, a Minkowski space (indefinite metric) must be used, at least near $r = 0$. For an example of an embedding in Minkowski space, see the discussion of a universe with constant negative spatial curvature in equations (27.23) and (27.24) and Box 27.2C.] Figure 21.4 exhibits embedding diagrams for other spacelike slices in the Schwarzschild geometry.

Creation and destruction in the singularities

From the Kruskal-Szekeres diagram and the 45-degree nature of its radial light rays, one sees that any particle that ever finds itself in region IV of spacetime must have been "created" in the earlier singularity; and any particle that ever falls into region II is doomed to be crushed in the later singularity. Only particles that stay forever in one of the asymptotically flat universes I or III, outside the gravitational radius ($r > 2M$), are forever safe from the singularities.

Nonviable proposals for modifying the topology of Schwarzschild spacetime

Some investigators, disturbed by the singularities at $r = 0$ or by the "double-universe" nature of the Schwarzschild geometry, have proposed modifications of its topology. One proposal is that the earlier and later singularities be identified with each other, so that a particle which falls into the singularity of region II, instead of being destroyed, will suddenly reemerge, being ejected, from the singularity of region IV. One cannot overstate the objections to this viewpoint: the region $r = 0$ is a physical singularity of infinite tidal gravitation forces and infinite Riemann curvature. Any particle that falls into that singularity must be destroyed by those forces. Any attempt to extrapolate its fate through the singularity using Einstein's field equations must fail; the equations lose their predictive power in the face of infinite curvature. Consequently, to postulate that the particle reemerges from the earlier singularity is to make up an *ad hoc* mathematical rule, one unrelated to physics. It is conceivable, but few believe it true, that any object of finite mass will modify the geometry of the singularity as it approaches $r = 0$ to such an extent that it can pass through and reemerge. However, whether such a speculation is correct must be answered not by *ad hoc* rules, but by concrete, difficult computations within the framework of general relativity theory (see Chapter 34).

A second proposal for modifying the topology of the Schwarzschild geometry is this: one should avoid the existence of two different asymptotically flat universes by identifying each point (v, u, θ, ϕ) with its opposite point $(-v, -u, \theta, \phi)$ in the Kruskal-Szekeres coordinate system. Two objections to this proposal are: (1) it produces a sort of "conical" singularity (absence of local Lorentz frames) at $(v, u) = (0, 0)$, i.e., at the neck of the bridge at its moment of maximum expansion; and (2) it leads to causality violations in which a man can meet himself going backward in time.

One good way for the reader to become conversant with the basic features of the Schwarzschild geometry is to reread §§31.1–31.4 carefully, reinterpreting everything said there in terms of the Kruskal-Szekeres diagram.

EXERCISES

Exercise 31.7. SCHWARZSCHILD METRIC IN ISOTROPIC COORDINATES

(a) Show that, rewritten in the isotropic coordinates of Exercise 23.1, the Schwarzschild metric reads

$$ds^2 = -\left(\frac{1 - M/2\bar{r}}{1 + M/2\bar{r}}\right)^2 dt^2 + \left(1 + \frac{M}{2\bar{r}}\right)^4 [d\bar{r}^2 + \bar{r}^2(d\theta^2 + \sin^2\theta\, d\phi^2)]; \qquad (31.22)$$

and derive the transformation

$$r = \bar{r}(1 + M/2\bar{r})^2 \qquad (31.23)$$

between the two radial coordinates.

(b) Which regions of spacetime (I, II, III, IV; see Figure 31.3) are covered by the isotropic coordinate patch, and which are not?

(c) Calculate and construct an embedding diagram for the spacelike hypersurface $t = 0$, $0 < \bar{r} < \infty$.

(d) Find a coordinate transformation that interchanges the region near $\bar{r} = 0$ with the region near $\bar{r} = \infty$, while leaving the metric coefficients in their original form.

Exercise 31.8. REISSNER-NORDSTRØM GEOMETRY

(a) Solve the Einstein field equations for a spherically symmetric, static gravitational field

$$ds^2 = -e^{2\Phi(r)}\, dt^2 + e^{2\Lambda(r)}\, dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2),$$

with no matter present, but with a radial electric field $\boldsymbol{B} = 0$, $\boldsymbol{E} = f(r)\boldsymbol{e}_{\hat{r}}$ in the static orthonormal frame

$$\boldsymbol{\omega}^{\hat{t}} = e^{\Phi}\,\boldsymbol{d}t, \qquad \boldsymbol{\omega}^{\hat{r}} = e^{\Lambda}\,\boldsymbol{d}r, \qquad \boldsymbol{\omega}^{\hat{\theta}} = r\,\boldsymbol{d}\theta, \qquad \boldsymbol{\omega}^{\hat{\phi}} = r\sin\theta\,\boldsymbol{d}\phi.$$

Use as a source in the Einstein field equations the stress-energy of the electric field. [*Answer:*

$$\boldsymbol{E} = (Q/r^2)\boldsymbol{e}_{\hat{r}}, \tag{31.24a}$$

$$ds^2 = -\left(1 - \frac{2M}{r} + \frac{Q^2}{r^2}\right)dt^2 + \left(1 - \frac{2M}{r} + \frac{Q^2}{r^2}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{31.24b}$$

This is called the "Reissner (1916)-Nordstrøm (1918) metric".]

(b) Show that the constant Q is the total charge as measured by a distant observer ($r \gg 2M$ and $r \gg Q$), who uses a Gaussian flux integral, or who studies the coulomb-force-dominated orbits of test charges with charge-to-mass ratio $e/\mu \gg M/Q$. What is the charge-to-mass ratio, in dimensionless units, for an electron? Show that the constant M is the total mass as measured by a distant observer using the Keplerian orbits of electrically neutral particles.

(c) Show that for $Q > M$, the Reissner-Nordstrøm coordinate system is well-behaved from $r = \infty$ down to $r = 0$, where there is a physical singularity and infinite tidal forces.

(d) Explore the nature of the spacetime geometry for $Q < M$, using all the techniques of this chapter (coordinate transformations, Kruskal-like coordinates, studies of particle orbits, embedding diagrams, etc.).

[*Solution:* see Graves and Brill (1960); also Fig. 34.4 of this book.]

(e) Similarly explore the spacetime geometry for $Q = M$. [*Solution:* see Carter (1966b).]

(f) For the case of a large ratio of charge to mass [$Q > M$ as in part (c)], show that the region near $r = 0$ is unphysical. More precisely, show that any spherically symmetric distribution of charged stressed matter that gives rise to the fields (31.24) outside its boundary must modify these fields for $r < r_0 = Q^2/2M$. [Hint: Study the quantity $m(r)$ defined in equations (23.18) and (32.22h), noting its values deduced from equation (31.24), on the one hand, and from the appropriate Einstein equation within the matter distribution, on the other hand. See Figure 26 of Misner (1969a) for a similar argument.]

1

CHAPTER 32

GRAVITATIONAL COLLAPSE

Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run twice as fast as that.

The Red Queen, in *Through the Looking Glass,*
LEWIS CARROLL (1871)

§32.1. RELEVANCE OF SCHWARZSCHILD GEOMETRY

The story that unfolded in the preceeding chapter was fantastic! One began with the innocuous looking Schwarzschild line element

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 + \frac{dr^2}{1 - 2M/r} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2), \tag{32.1}$$

which was derived originally as the external field of a static star. One asked what happens if the star is absent; i.e., one probed the nature of the Schwarzschild geometry when no star is present to generate it. One might have expected the geometry to be that of a point mass sitting at $r = 0$. But it was not. It turned out to represent a "wormhole" connecting two asymptotically flat universes. Moreover, the wormhole was dynamic. It was created by the "joining together" of two "$r = 0$" singularities, one in each universe; it expanded to a maximum circumference of $4\pi M$; it then recontracted and pinched off, leaving the two universes disconnected once again, each with its own "$r = 0$" singularity.

The roles and relevance of the Schwarzschild geometry

As a solution to Einstein's field equations, this expanding and recontracting wormhole must be taken seriously. It is an exact solution; and it is one of the simplest of all exact solutions. But there is no reason whatsoever to believe that such wormholes exist in the real universe! They can exist only if the expanding universe, $\sim 10 \times 10^9$ years ago, was "born" with the necessary initial conditions—with "$r = 0$"

Schwarzschild singularities ready and waiting to blossom forth into wormholes. There is no reason at all to believe in such pathological initial conditions!

Why, then, was so much time and effort spent in Chapter 31 on understanding the Schwarzschild geometry? (1) Because it illustrates clearly the highly non-Euclidean character of spacetime geometry when gravity becomes strong; (2) because it illustrates many of the techniques one can use to analyze strong gravitational fields; and most importantly (3) because, when appropriately truncated, it is the spacetime geometry of a black hole and of a collapsing star—as well as of a wormhole.

This chapter explores the role of the Schwarzschild geometry in gravitational collapse; the next chapter explores its role in black-hole physics.

§32.2. BIRKHOFF'S THEOREM

The uniqueness of the Schwarzschild geometry: Birkhoff's theorem

That the Schwarzschild geometry is relevant to gravitational collapse follows from *Birkhoff's (1923) theorem: Let the geometry of a given region of spacetime (1) be spherically symmetric, and (2) be a solution to the Einstein field equations in vacuum. Then that geometry is necessarily a piece of the Schwarzschild geometry.* The external field of any electrically neutral, spherical star satisfies the conditions of Birkhoff's theorem, whether the star is static, vibrating, or collapsing. Therefore the external field must be a piece of the Schwarzschild geometry.

The physics underlying Birkhoff's theorem

Birkhoff's theorem is easily understood on physical grounds. Consider an equilibrium configuration that is unstable against gravitational collapse and that, like all equilibrium configurations (see §23.6), has the Schwarzschild geometry as its external gravitational field. Perturb this equilibrium configuration in a spherically symmetric way, so that it begins to collapse radially. The perturbation and subsequent collapse cannot affect the external gravitational field so long as exact spherical symmetry is maintained. Just as Maxwell's laws prohibit monopole electromagnetic waves, so Einstein's laws prohibit monopole gravitational waves. There is no possible way for any gravitational influence of the radial collapse to propagate outward.

Proof of Birkhoff's theorem

Not only is Birkhoff's theorem easy to understand, but it is also fairly easy to prove. Consider a spherical region of spacetime. Spherical symmetry alone is sufficient to guarantee that conditions (i), (ii), and (iii) of Box 23.3 are satisfied, and thus to guarantee that one can introduce Schwarzschild coordinates

$$\begin{gathered} ds^2 = -e^{2\Phi}\, dt^2 + e^{2\Lambda}\, dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2), \\ \Phi = \Phi(t, r), \text{ and } \Lambda = \Lambda(t, r). \end{gathered} \tag{32.2}$$

[See Box 23.3 for proof; and notice that: (1) for generality one must allow $g_{tt} = -e^{2\Phi}$ and $g_{rr} = e^{2\Lambda}$ to be positive or negative (no constraint on sign!); (2) at events where the gradient of the "circumference function" r is zero or null, Schwarzschild coordinates cannot be introduced. The special case $(\nabla r)^2 = 0$ is treated in exercise 32.1.]

Impose Einstein's vacuum field equation on the metric (32.2), using the orthonormal components of the Einstein tensor as derived in exercise 14.16:

$$G_{\hat{t}\hat{t}} = r^{-2}(1 - e^{-2\Lambda}) + 2(\Lambda_{,r}/r)e^{-2\Lambda} = 0, \tag{32.3a}$$

$$G_{\hat{t}\hat{r}} = G_{\hat{r}\hat{t}} = 2(\Lambda_{,t}/r)e^{-(\Lambda+\Phi)} = 0, \tag{32.3b}$$

$$G_{\hat{r}\hat{r}} = 2(\Phi_{,r}/r)e^{-2\Lambda} + r^{-2}(e^{-2\Lambda} - 1) = 0, \tag{32.3c}$$

$$\begin{aligned} G_{\hat{\theta}\hat{\theta}} = G_{\hat{\phi}\hat{\phi}} = &+(\Phi_{,rr} + \Phi_{,r}{}^2 - \Phi_{,r}\Lambda_{,r} + \Phi_{,r}/r - \Lambda_{,r}/r)e^{-2\Lambda} \\ &- (\Lambda_{,tt} + \Lambda_{,t}{}^2 - \Lambda_{,t}\Phi_{,t})e^{-2\Phi} = 0. \end{aligned} \tag{32.3d}$$

Equation (32.3b) guarantees that Λ is a function of r only, and equation (32.3a) then guarantees that Λ has the same form as for the Schwarzschild metric:

$$\Lambda = -\frac{1}{2}\ln|1 - 2M/r|. \tag{32.4a}$$

Equations (32.3c,d) then become two equivalent equations for $\Phi(t, r)$—equivalent by virtue of the Bianchi identity, $\boldsymbol{\nabla}\cdot\boldsymbol{G} = 0$—whose solution is

$$\Phi = \frac{1}{2}\ln|1 - 2M/r| + f(t). \tag{32.4b}$$

Here f is an arbitrary function. Put expressions (32.4) into the line element (32.2); thereby obtain

$$ds^2 = -e^{2f(t)}\left(1 - \frac{2M}{r}\right)dt^2 + \frac{dr^2}{1 - 2M/r} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2).$$

Then redefine the time coordinate

$$t_{\text{new}} = \int e^{f(t)}\, dt,$$

and thereby bring the line element into the Schwarzschild form

$$ds^2 = -\left(1 - \frac{2M}{r}\right)dt^2 + \frac{dr^2}{1 - 2M/r} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2).$$

Conclusion: When the spacetime surrounding any object has spherical symmetry and is free of charge, mass, and all fields other than gravity, then one can introduce coordinates in which the metric is that of Schwarzschild. Conclusion restated in coordinate-free language: the geometry of any spherically symmetric vacuum region of spacetime is a piece of the Schwarzschild geometry (Birkhoff's theorem). Q.E.D.

EXERCISE

Exercise 32.1. UNIQUENESS OF REISSNER-NORDSTRØM GEOMETRY [Track 2]

Prove the following generalization of Birkhoff's theorem. Let the geometry of a given region of spacetime (1) be spherically symmetric, and (2) be a solution to the Einstein field equations with an electromagnetic field as source. Then that geometry is necessarily a piece of the Reissner-Nordstrøm geometry [equation (31.24b)] with electric and magnetic fields, as measured in the standard static orthonormal frames

$$\boldsymbol{E} = (Q_e/r^2)\boldsymbol{e}_{\hat{r}}, \qquad \boldsymbol{B} = (Q_m/r^2)\boldsymbol{e}_{\hat{r}}, \qquad Q = (Q_e{}^2 + Q_m{}^2)^{1/2}.$$

[*Hints:* (1) First consider regions of spacetime in which $(\boldsymbol{\nabla} r)^2 \neq 0$, using the same methods as the text uses for Birkhoff's theorem. The result is the Reissner-Nordstrøm solution. (2) Any region of dimensionality less than four, in which $(\boldsymbol{\nabla} r)^2 = 0$ (e.g., the Schwarzschild radius), can be treated as the join between four-dimensional regions with $(\boldsymbol{\nabla} r)^2 \neq 0$. Moreover, the geometry of such a region is determined uniquely by the geometry of the adjoining four-dimensional regions ("junction conditions"; §21.13). Since the adjoining regions are necessarily Reissner-Nordstrøm (step 1), then so are such "sandwiched" regions. (3) Next consider four-dimensional regions in which $\boldsymbol{\nabla} r = \boldsymbol{d}r$ is null and nonzero. Show that in such regions there exist coordinate systems with

$$ds^2 = -2\Psi\, dr\, dt + r^2(d\theta^2 + \sin^2\theta\, d\phi^2),$$

where $\Psi = \Psi(r, t)$. Show further that the Ricci tensor for this line element has an orthonormalized component

$$R_{\hat{\theta}\hat{\theta}} = 1/r^2,$$

whereas the stress-energy tensor for a spherically symmetric electromagnetic field has

$$8\pi T_{\hat{\theta}\hat{\theta}} = 8\pi\left(T_{\hat{\theta}\hat{\theta}} - \frac{1}{2}g_{\hat{\theta}\hat{\theta}}T\right) = Q^2\, r^{4}. \qquad Q = \text{const.}$$

These quantities, $R_{\hat{\theta}\hat{\theta}}$ and $8\pi T_{\hat{\theta}\hat{\theta}}$, must be equal (Einstein's field equation) but cannot be because of their different r-dependence. Thus, an electromagnetic field cannot generate regions with $\boldsymbol{d}r \neq 0$, $\boldsymbol{d}r \cdot \boldsymbol{d}r = 0$. (4) Finally, consider four-dimensional regions in which $\boldsymbol{d}r = 0$. Denote the constant value of r by a, and show that any event can be chosen as the origin of a locally well-behaved coordinate system with

$$ds^2 = a^2(-d\tilde{\tau}^2 + e^{2\lambda}\, dz^2 + d\theta^2 + \sin^2\theta\, d\phi^2),$$
$$\lambda = \lambda(\tilde{\tau}, z), \qquad \lambda(\tilde{\tau} = 0, z) = 0, \qquad \dot{\lambda}(\tilde{\tau} = 0, z) = 0.$$

[Novikov-type coordinate system; see §31.4.] Show that, in the associated orthonormal frame, spherical symmetry demands

$$\boldsymbol{E} = (Q_e/a^2)\boldsymbol{e}_{\hat{z}}, \qquad \boldsymbol{B} = (Q_m/a^2)\boldsymbol{e}_{\hat{z}}, \qquad Q \equiv (Q_e{}^2 + Q_m{}^2)^{1/2},$$

and that the Einstein field equations then require $Q = a$ and $e^{\lambda} = \cos\tilde{\tau}$, so that

$$ds^2 = Q^2(-d\tilde{\tau}^2 + \cos^2\tilde{\tau}\, dz^2 + d\theta^2 + \sin^2\theta\, d\phi^2).$$

(5) This solution of the field equations [sometimes called the "Bertotti (1959)-Robinson (1959a) Electromagnetic Universe," and explored in this coordinate system by Lindquist (1960)] is actually the throat of the Reissner-Nordstrøm solution for the special case $Q = M$. Verify this claim by performing the following coordinate transformation on the Reissner-Nordstrøm throat region [equation (31.24b) with $Q = M$ and $|r - Q| \ll Q$]:

$$r - Q = Qe^{-z}\cos\tilde{\tau}, \qquad t = Qe^{z}\tan\tilde{\tau}.$$

(6) Thus, each possible case leads either to no solution at all, or to a segment of the Reissner-Nordstrøm geometry. Q.E.D.] Note: The missing case, $(\nabla r)^2 = 0$, in the text's proof of Birkhoff's theorem, is resolved by noting that, for $Q = 0$, steps (3) and (4) above lead to no solutions at all. We thank G. F. R. Ellis for pointing out the omission of the case $(\nabla r)^2 = 0$ in the preliminary version of this book.

---

§32.3. EXTERIOR GEOMETRY OF A COLLAPSING STAR

Consider a star that is momentarily static, but will subsequently begin to collapse. Its space geometry at the initial moment of Schwarzschild coordinate time, $t = 0$, has two parts: in the exterior, vacuum region ($r > R > 2M$), it is the Schwarzschild geometry (Birkhoff's theorem!); but in the star's interior, it is some other, totally different geometry. Whatever the interior geometry may be, it has an embedding diagram at time $t = 0$ which is qualitatively like that of Figure 23.1. (For discussion and proof of this, see §23.8.) Notice that the star's space geometry is obtained by discarding the lower universe of the full Schwarzschild geometry (Figure 31.5,a), and replacing it with a smooth "bowl" on which the matter of the star is contained.

Gravitational collapse analyzed by examining the star's exterior, Schwarzschild geometry

To follow the subsequent collapse of this star in the Schwarzschild coordinate system, or in the Kruskal-Szekeres coordinate system, or in an ingoing Eddington-Finkelstein coordinate system, one can similarly discard that part of the coordinate diagram which lies inside the star's surface, and keep only the exterior Schwarzschild region. (See Figure 32.1.) In place of the discarded interior Schwarzschild region, one must introduce some other coordinate system, line element, and diagram that correctly describe the interior of the collapsing star.

From truncated coordinate diagrams (such as Figures 32.1,a,b,c), one can readily discover and understand the various peculiar features of collapse through the gravitational radius.

The gravitational radius as a point of no return, and the "crushing" at $r = 0$

(1) No matter how stiff may be the matter of which a (spherical) star is made, once its surface has collapsed within the gravitational radius, the star will continue to collapse until its surface gets crushed in the singularity at $r = 0$. This one discovers by recalling that the star's surface cannot move faster than the speed of light, so its world line must always make an angle of less than 45 degrees with the v-axis of the Kruskal-Szekeres diagram.

(2) No signal (e.g., photon) emitted from the star's surface after it collapses inside the gravitational radius can ever escape to an external observer. Rather, all signals emitted from inside the gravitational radius get caught and destroyed by the collapse of the surrounding geometry into the singularity at $r = 0$ as space "pinches off" around the star.

(3) Consequently, an external observer can never see the star after it passes the gravitational radius; and he can never see the singularity that terminates its collapse—unless he chooses to fall through the gravitational radius himself and pay the price of death for the knowledge gained.

The redshift of signals emitted just before passage through the gravitational radius

Does this mean that the collapsing star instantaneously and completely disappears from external view as it reaches the gravitational radius? No, not according to the analysis depicted in Figure 31.1,c: Place an astrophysicist on the surface of a collapsing star, and have him send a series of uniformly spaced signals to a distant astronomer, at rest at $r \gg 2M$, to inform him of the progress of the collapse. These signals propagate along null lines in the spacetime diagram of Figure 31.1c. The signals originate on the world line of the stellar surface, and they are received by the distant astronomer when they intersect his world line, $r = \text{constant} \gg M$. As the star collapses closer and closer to its gravitational radius, $R = 2M$, the signals, which are sent at equally spaced intervals according to the astrophysicist's clock, are received by the astronomer at more and more widely spaced intervals. The astronomer does not receive a signal emitted just before the gravitational radius is reached until after an infinite amount of time has elapsed; and he never receives signals emitted after the gravitational radius has been passed. Those signals, like the astrophysicist who sends them, after brief runs get caught and destroyed by the collapsing geometry in the singularity, at $r = 0$. It is not only the star that collapses. The geometry around the star collapses.

Hence, to the distant astronomer, the collapsing star appears to slow down as it approaches its gravitational radius: light from the star becomes more and more red-shifted. Clocks on the star appear to run more and more slowly. It takes an infinite time for the star to reach its gravitational radius; and, as seen by the distant astronomer, the star never gets beyond there.

Optical appearance of the collapsing star

The optical appearance of a collapsing star was first analyzed mathematically, giving main attention to radially propagating photons, by J. R. Oppenheimer and H. Snyder (1939). More recently a number of workers have reexamined the problem [see, e.g., Podurets (1964), Ames and Thorne (1968) and Jaffe (1969)]. The most important quantitative results of these studies are as follows. In the late stages of collapse, when the distant astronomer sees the star to be very near its gravitational radius, he observes its total luminosity to decay exponentially in time

$$L \propto \exp\left(-\frac{2}{3\sqrt{3}}\frac{t}{2M}\right). \tag{32.5}$$

Simultaneously, photons that travel to him along radial trajectories arrive with exponentially increasing redshifts

$$z = \Delta\lambda/\lambda \propto e^{t/4M}. \tag{32.6}$$

However, the light from the star is dominated in these late stages, not by photons flying along radial trajectories from near the gravitational radius, but by photons that were deposited by the star in unstable circular orbits as its surface passed through $r = 3M$ (see §25.6 and Box 25.7). As time passes, these photons gradually leak out the diffuse spherical shell of trapped photons at $r = 3M$ and fly off to the distant observer, who measures them to have redshift $z \approx 2$. Consequently, in the late stages of collapse the star's spectral lines are broadened enormously, but they are brightest at redshift $z \approx 2$.

(*continued on page 850*)

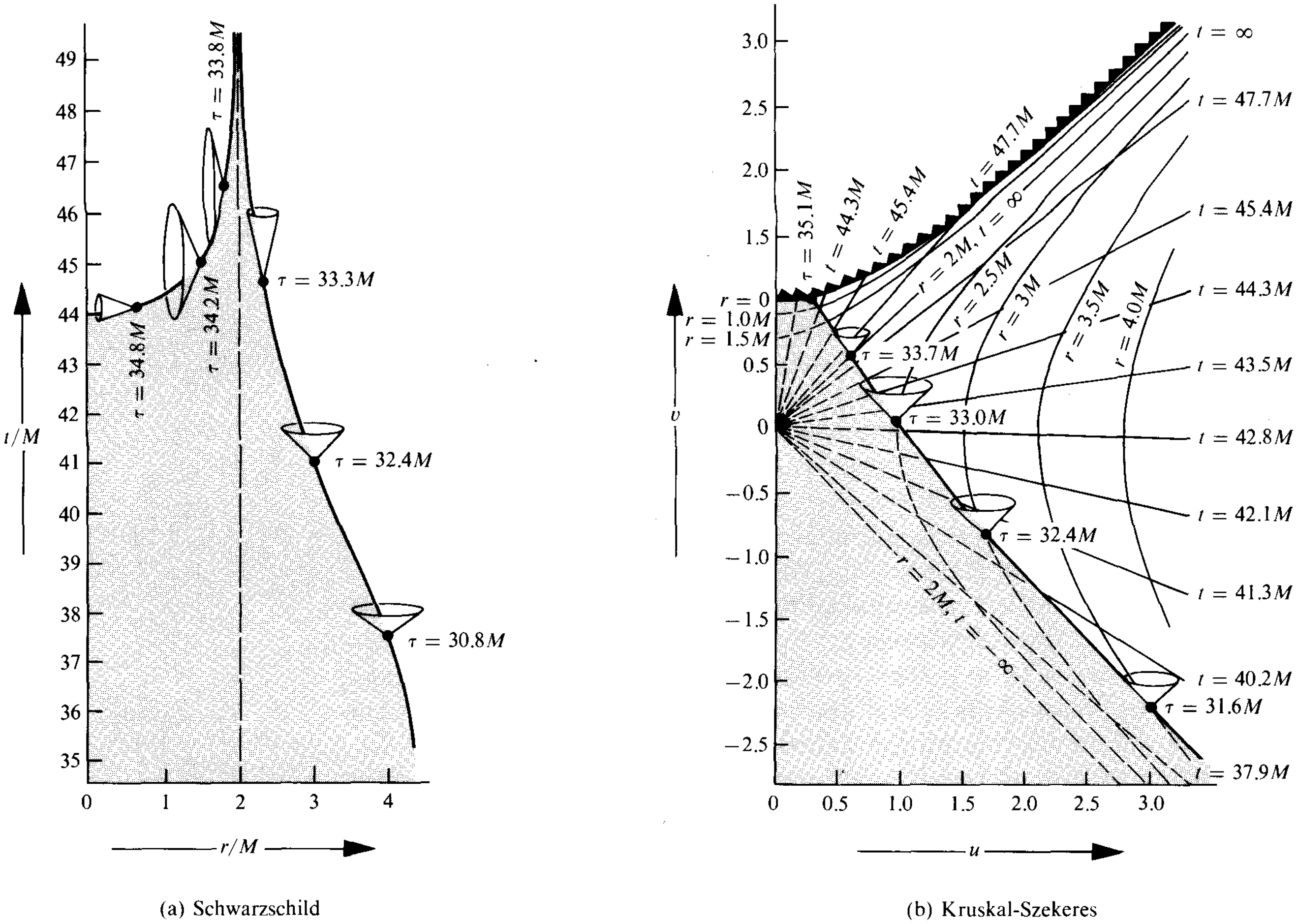


Figure 32.1.
The free-fall collapse of a star of initial radius $R_i = 10\ M$ as depicted alternatively in (a) Schwarzschild coordinates, (b) Kruskal-Szekeres coordinates, and (c) ingoing Eddington-Finkelstein coordinates (see Box 31.2). The region of spacetime inside the collapsing star is grey, that outside it is white. Only the geometry of the exterior region is that of Schwarzschild. The curve separating the grey and white regions is the geodesic world line of the surface of the collapsing star (equations [31.10] or [32.10] with $r_{\max} = R_i = 10\ M$). This world line is parameterized by proper time, τ, as measured by an observer who sits on the surface of the star; the radial light cones, as calculated from $ds^2 = 0$, are attached to it.

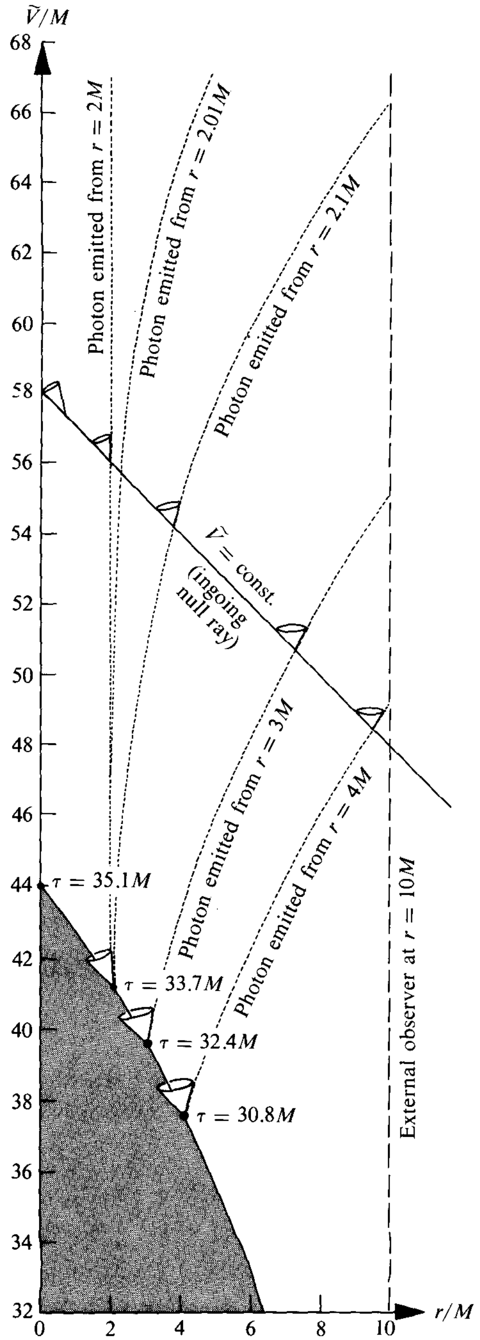
Notice that, although the shapes of the light cones are not all the same relative to Schwarzschild coordinates or relative to Eddington-Finkelstein coordinates, they are all the same relative to Kruskal-Szekeres coordinates. This is because light rays travel along 45-degree lines in the u,v-plane ($dv = \pm du$), but they travel along curved paths in the r,t-plane and $r,\tilde{V}$-plane.

The Kruskal-Szekeres spacetime diagram shown here is related to the Schwarzschild diagram by equations (31.13) plus a translation of Schwarzschild time: $t \longrightarrow t + 42.8\ M$. The Eddington-Finklestein diagram is related to the Schwarzschild diagram by

$$\tilde{V} = t + r^* = t + r + 2\ M \ln |r/2\ M - 1|$$

(see Box 31.2).

It is evident from these diagrams that the free-fall collapse is characterized by a constantly diminishing radius, which drops from $R = 10\ M$ to $R = 0$ in a finite and short comoving proper time interval, $\Delta\tau = 35.1\ M$. The point $R = 0$ and the entire region $r = 0$ outside the star make up a physical "singularity" at which infinite tidal gravitational forces—according to classical, unquantized general relativity—can and do crush matter to infinite density (see end of §31.2; also §32.6).



(c) Eddington-Finkelstein

The Eddington-Finkelstein diagram depicts a series of photons emitted radially from the surface of the collapsing star, and received by an observer at $r = R_{\text{initial}} = 10\ M$. The observer eventually receives all photons emitted radially from outside the gravitational radius; all photons emitted after the star passes through its gravitational radius eventually get pulled into the singularity at $r = 0$; and any photon emitted radially at the gravitational radius stays at the gravitational radius forever.

Non-free-fall collapse is similar to the collapse depicted here. When pressure gradients are present, only the detailed shape of the world line of the star's surface changes.

Notice how short is the characteristic e-folding time for the decay of luminosity and for the radial redshift:

$$\tau_{\text{char}} = 2M \approx 1 \times 10^{-5}(M/M_{\odot})\ \text{sec}$$
$$= \begin{pmatrix}\text{light-travel time across a flat-space} \\ \text{distance equal to the gravitational radius}\end{pmatrix}. \tag{32.7}$$

Here $M_{\odot}$ denotes one solar mass.

EXERCISE

Exercise 32.2. REDSHIFTS DURING COLLAPSE

(a) Let a radio transmitter on the surface of a collapsing spherical star emit monochromatic waves of wavelength λ_e; and let a distant observer, at the same θ, ϕ, as the transmitter, receive the waves. Show that at late times the wavelength received varies as

$$\lambda_{\text{rec}}/\lambda_{\text{em}} \propto e^{t/4M} \tag{32.8a}$$

[equation (32.6)], where t is proper time as measured by the distant observer.

(b) [Track 2] Use kinetic theory for the outgoing photons (conservation of density in phase space: Liouville's theorem; §22.6) to show that the energy flux of the radiation received (ergs/cm^2 sec) varies as

$$F \propto e^{-t/M}. \tag{32.8b}$$

The rest of this chapter is Track 2. No previous Track-2 material is needed as preparation for it, but it is needed as preparation for (1) the Track-2 part of Chapter 33 (black holes), and (2) Chapter 34 (singularities and global methods).

(c) Suppose that nuclear reactions at the center of the collapsing star generate neutrinos of energy E_e, and that these neutrinos flow freely outward (negligible absorption in star). Show that the energy of the neutrinos received by a distant observer decreases at late times as

$$E_{\text{rec}}/E_e \propto e^{-t/4M}. \tag{32.9a}$$

(d) Show that the flux of neutrino energy dies out at late times as

$$F \propto e^{-t/2M}. \tag{32.9b}$$

(e) Explain in elementary terms why the decay laws (32.8a) and (32.9a) for energy are the same, but the decay laws (32.8b) and (32.9b) for energy flux are different.

(f) Let a collapsing star emit photons from its surface at the black-body rate

$$\frac{dN}{d\tau} = \left(1.5 \times 10^{11} \frac{\text{photons}}{\text{cm}^2\ \text{sec}\ \text{K}^3}\right) \times \begin{pmatrix}\text{surface area} \\ \text{of star}\end{pmatrix} \times \begin{pmatrix}\text{temperature} \\ \text{of surface}\end{pmatrix}^3.$$

Let a distant observer count the photons as they pass through his sphere of radius $r \gg M$. Let him begin his count (time $t = 0$) when he sees (via photons traveling radially outward) the center of the star's surface pass through the radius $r = 3M$. Show that, in order of magnitude, the time he and his associates must wait, until the last photon that will ever get out has reached them, is

$$t = (M/M_{\odot})[8 \times 10^{-4} + 5 \times 10^{-5} \log_{10}(T_{11}M/M_{\odot})]\ \text{seconds}, \tag{32.9c}$$

where T_{11} is the star's surface temperature in units of 10^{11} K.

---

§32.4. COLLAPSE OF A STAR WITH UNIFORM DENSITY AND ZERO PRESSURE

When one turns attention to the interior of a collapsing star and to the precise world line that its surface follows in the Schwarzschild geometry, one encounters rather complicated mathematics. The simplest case to treat is that of a "star" with uniform density and zero pressure; and, indeed, until recently that was the only case which had been treated in detail. The original—and very complete—analysis of the collapse of such a uniform-density "ball of dust" was given in the classic paper of Oppenheimer and Snyder (1939). More recently, other workers have discussed it from slightly different points of view and using different coordinate systems. The approach taken here was devised by Beckedorff and Misner (1962).

The collapse, from rest, of a uniform-density ball of "dust":

Because no pressure gradients are present to deflect their motion, the particles on the surface of any ball of dust must move along radial geodesics in the exterior Schwarzschild geometry. For a ball that begins at rest with finite radius, $R = R_i$, at time $t = 0$, the subsequent geodesic motion of its surface is given by equations (31.10):

(1) world line of ball's surface in exterior Schwarzschild coordinates

$$R = (R_i/2)(1 + \cos\eta), \tag{32.10a}$$

$$t = 2M \ln\left|\frac{(R_i/2M - 1)^{1/2} + \tan(\eta/2)}{(R_i/2M - 1)^{1/2} - \tan(\eta/2)}\right| + 2M(R_i/2M - 1)^{1/2}[\eta + (R_i/4M)(\eta + \sin\eta)]. \tag{32.10b}$$

Here R is the Schwarzschild radial coordinate (i.e., $4\pi R^2$ is the star's surface area) at Schwarzschild time t. This world line is plotted in Figure 32.1 for $R_i = 10M$, in terms of Schwarzschild coordinates, Kruskal-Szekeres coordinates, and Eddington-Finkelstein coordinates. The proper time read by a clock on the surface of the collapsing star is given by equation (31.10b):

$$\tau = (R_i^3/8M)^{1/2}(\eta + \sin\eta). \tag{32.10c}$$

Note that the collapse begins when the parameter η is zero ($R = R_i$, $t = \tau = 0$); and it terminates at the singularity ($R = 0, \eta = \pi$) after a lapse of proper time, as measured on any test particle falling with the dust, equal to

$$\Delta\tau = \pi(R_i^3/8M)^{1/2}.$$

It is interesting, though coincidental, that this is precisely the time-lapse required for free-fall collapse to infinite density in Newtonian theory [see equation (25.27′), Figure 25.3, and associated discussion].

What is the behavior of the interior of the ball of dust as it collapses? A variety of different interiors for pressureless dust can be conceived (exercise 32.8). But here attention focuses on the simplest of them: an interior that is homogeneous and isotropic everywhere, except at the surface—i.e., an interior locally identical to a dust-filled Friedmann cosmological model (Box 27.1). Is the Friedmann interior to be "open" ($k = -1$), "flat" ($k = 0$), or "closed" ($k = +1$)? Only the closed case

(2) interior of ball is identical to a portion of a closed Friedmann universe

is appropriate, since one has already demanded [equation (32.10)] that the star be at rest initially (initial rate of change of density equals zero; "moment of maximum expansion").

Using comoving hyperspherical coordinates, χ, θ, ϕ, for the star's interior, and putting the origin of coordinates at the star's center, one can write the line element in the interior in the familiar Friedmann form

$$ds^2 = -d\tau^2 + a^2(\tau)[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)]. \tag{32.11}$$

Here $a(\tau)$ is given by the familiar cycloidal relation,

$$\begin{aligned} a &= \frac{1}{2}a_m(1 + \cos\eta), \\ \tau &= \frac{1}{2}a_m(\eta + \sin\eta); \end{aligned} \tag{32.12}$$

and the density is given by

$$\rho = (3a_m/8\pi)a^{-3} = (3/8\pi a_m^{\ 2})\left[\frac{1}{2}(1 + \cos\eta)\right]^{-3} \tag{32.13}$$

[see equations (1), (9), (4), and (5) of Box 27.1, with η replaced by $\eta + \pi$].

(3) the join between Friedmann interior and Schwarzschild exterior

There is one possible difficulty with this interior solution. In the cosmological case, the solution was homogeneous and isotropic everywhere. Here homogeneity and isotropy are broken at the star's surface—which lies at some radius

$$\chi = \chi_0 \tag{32.14}$$

for all τ, as measured in terms of the hyperspherical polar angle χ, a comoving coordinate (first picture in Box 27.2). At that surface (i.e., three-dimensional world tube enclosing the star's fluid) the interior Friedmann geometry must match smoothly onto the exterior Schwarzschild geometry. If the match cannot be achieved, then the Friedmann line element (32.11) cannot represent the interior of a collapsing star. An example of a case in which the matching could not be achieved is an interior of uniform and nonzero pressure, as well as uniform density. In that case there would be an infinite pressure gradient at the star's surface, which would blow off the outer layers of the star, and would send a rarefaction wave propagating inward toward its center. The uniform distribution of density and pressure would quickly be destroyed.

For the case of zero pressure, the match is possible. As a partial verification of the match, one can examine the separate and independent predictions made by the interior and exterior solutions for the star's circumference, $C = 2\pi R$, as a function of proper time τ at the star's surface. The external Schwarzschild solution predicts the cycloidal relation,

$$\begin{aligned} C = 2\pi R &= 2\pi(R_i/2)(1 + \cos\eta), \\ \tau &= (R_i^{\ 3}/8M)^{1/2}(\eta + \sin\eta) \end{aligned} \tag{32.15}$$

[equations (32.10)]. The interior Friedmann solution predicts a similar cycloidal relation:

$$C = 2\pi R = 2\pi a \sin\chi_0 = 2\pi\left(\frac{1}{2}a_m \sin\chi_0\right)(1 + \cos\eta),$$

$$\tau = \frac{1}{2}a_m(\eta + \sin\eta). \tag{32.16}$$

The two predictions agree perfectly for all time if and only if

$$R_i = a_m \sin\chi_0, \tag{32.17a}$$

$$M = \frac{1}{2}a_m \sin^3\chi_0. \tag{32.17b}$$

A more complete verification of the match is given in exercise 32.4.

For further insight into this idealized model of gravitational collapse, see Box 32.1.

EXERCISES

Exercise 32.3. EMBEDDING DIAGRAMS AND PHOTON PROPAGATION FOR COLLAPSING STAR

Verify in detail the features of homogeneous collapse described in Box 32.1.

Exercise 32.4. MATCH OF FRIEDMANN INTERIOR TO SCHWARZSCHILD EXTERIOR

The Einstein field equations are satisfied on a star's surface if and only if the intrinsic and extrinsic geometries of the surface's three-dimensional world tube are the same, whether measured on its interior or on its exterior (see §21.13 for proof and discussion). Verify that for the collapsing star discussed above, the intrinsic and extrinsic geometries match at the join between the Friedmann interior and the Schwarzschild exterior. [Hints: (a) Use η, θ, ϕ, as coordinates on the world tube of the star's surface, and show that the intrinsic geometry has the same line element

$$ds^2 = a^2(\eta)[-d\eta^2 + \sin^2\chi_0(d\theta^2 + \sin^2\theta\, d\phi^2)], \tag{32.18a}$$

whether measured in the Schwarzschild exterior or in the Friedmann interior. (b) Show that the extrinsic curvature of the world tube has the same components

$$\begin{aligned} K_{\eta\eta} &= K_{\eta\theta} = K_{\eta\phi} = K_{\theta\phi} = 0, \\ K_{\theta\theta} &= K_{\phi\phi}/\sin^2\theta = -a(\eta)\sin\chi_0\cos\chi_0, \end{aligned} \tag{32.18b}$$

whether measured in the Schwarzschild exterior or in the Friedmann interior.]

Exercise 32.5. STARS THAT COLLAPSE FROM INFINITY

(a) Patch together a truncated Schwarzschild geometry and the geometry of a truncated "flat" ($k = 0$), dust-filled Friedmann universe to obtain a model of a star that collapses from rest at an infinite initial radius. [*Hint:* The world line of the star's surface in the Schwarzschild geometry is given by equations (31.2).]

(b) Similarly patch together a truncated Schwarzschild geometry and the geometry of a truncated "open" ($k = -1$), dust-filled Friedmann universe to obtain a star which collapses from infinity with finite initial inward velocity.

(continued on page 857)

Box 32.1 AN IDEALIZED COLLAPSING STAR WITH FRIEDMANN INTERIOR AND SCHWARZSCHILD EXTERIOR

(See §32.4 and exercises 32.3 and 32.4 for justification of the results described here.)

Initial State

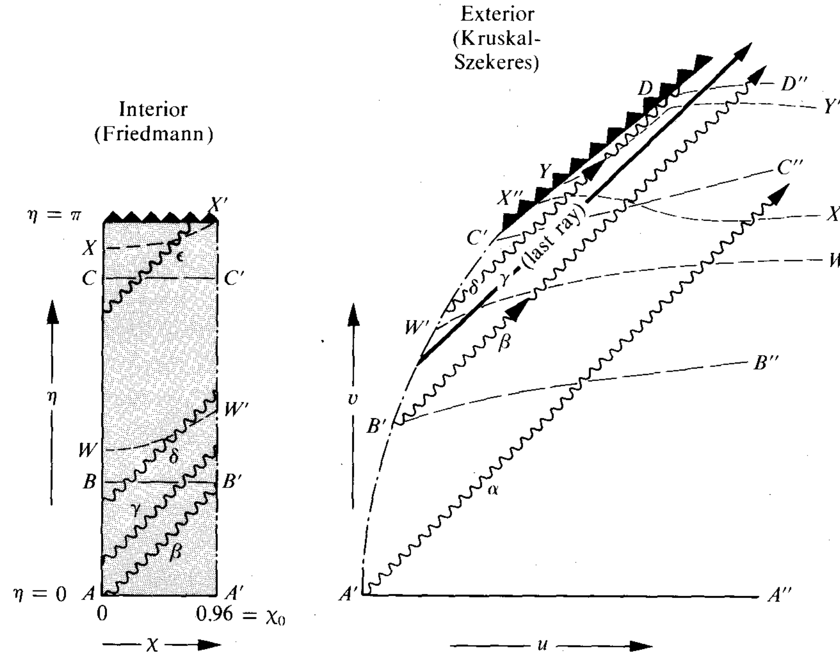
(1) Take a Friedmann universe of radius $a = a_m$ at its moment of maximum expansion, $\eta = 0$; and slice off and discard the region $\chi_0 < \chi \leq \pi$, where χ_0 is some angle less than $\pi/2$. (2) Take a Schwarzschild geometry of mass $M = (a_m/2)\sin^3\chi_0$ at the moment $t = 0$; and slice off and discard the region $r < R_i = a_m \sin\chi_0$. (3) Glue the remaining pieces of Friedmann and Schwarzschild geometry together smoothly along their cut surfaces. The resultant object will be a momentarily static star of uniform density $\rho_i = 3/(8\pi a_m{}^2)$, of mass $M = (a_m/2)\sin^3\chi_0$, and of radius $R_i = a_m \sin\chi_0$.

Subsequent Evolution

Release this star from its intial state, and let it collapse in accord with Einstein's field equations. The interior, truncated Friedmann universe and the exterior, truncated Schwarzschild geometry will evolve just as though they had never been cut up and patched together; and this evolution will preserve the smoothness of the match between interior and exterior!

Details of the Collapse

Probe the details of the collapse using sequences of embedding diagrams (histories *ABCD* and *AWXY*), and using photons that propagate radially outward (photons α, β, γ, δ, ϵ). The example shown here has $\chi_0 = 0.96$ and $R_i/M = 2/\sin^2\chi_0 = 3$.



History of Collapse as Probed by Hypersurfaces ABCD:

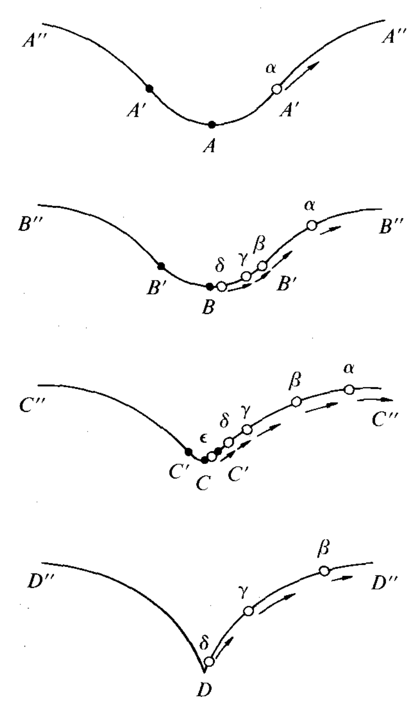
(1) Initial configuration, $A - A' - A''$, is that constructed by cutting and sewing at times $\eta = t = 0$.

(2) Each subsequent configuration has as its interior a slice of constant Friedmann time η.

(3) The interior remains always a spherical cup of half-angle χ_0; but it contracts from radius $R = R_i = a_m \sin \chi_0$ to $R = 0$ as time increases.

(4) The matter in the star is all crushed simultaneously to infinite density when R reaches zero, and the external Schwarzschild "funnel" develops a cusp-like singularity at that point.

(5) As time increases further, this cusp pulls the region $r < 2M$ of the funnel into $r = 0$ so fast that the outward-traveling photon δ is gobbled up and crushed.



These embedding diagrams must be rotated about the vertical axes in order to become 2-dimensional surfaces analogous to Figure 23.1.

Box 32.1 (*continued*)

History as Probed by Hypersurfaces AWXY

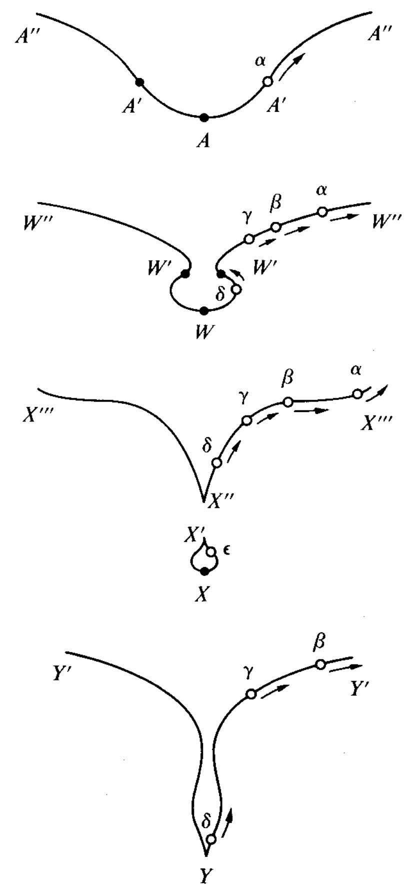
(1) Initial configuration, $A - A' - A''$, is again that constructed by cutting and sewing at $\eta = t = 0$.

(2) Subsequent hypersurfaces are very different from $\eta =$ const.

(3) As time passes, a neck develops in the geometry just outside the surface of the star.

(4) This neck becomes tighter and tighter and then pinches off, leaving the star completely isolated from the rest of the universe, and leaving a deadly cusp-like singularity in the exterior geometry where the star used to be.

(5) The isolated star, in its own little closed universe, continues to contract until it is crushed to infinite density, while the external geometry begins to develop another neck and the cusp quickly gobbles up photon δ.



The extreme difference between histories *ABCD* and *AWXY* dramatizes the "many-fingered time" of general relativity. The hypersurface on which one explores the geometry can be pushed ahead faster in time in one region, at the option of the party of explorers. Thus whether one region of the star collapses first, or another, or the entire star collapses simultaneously, is a function both of the spacetime geometry and of the choice of slicing. The party of explorers has this choice of slicing at their own control, and thus they themselves to this extent govern what kind of spacelike slices they will see as their exploration moves forward in time. The spacetime geometry that they slice, however, is in no way theirs to control or to change. To the extent that their masses are negligible and they serve merely as test objects, they have no influence whatsoever on the spacetime. It was fixed completely by the specification of the initial conditions for the collapse. In brief, spacetime is four-dimensional and slices are only three-dimensional (and in the pictures here look only two-dimensional or one-dimensional). Any one set of slices captures only a one-sided view of the whole story. To see the entire picture one must either examine the dynamics of the geometry as it reveals itself in varied choices of the slicing or become accustomed to visualizing the spacetime geometry as a whole.

2

§32.5. SPHERICALLY SYMMETRIC COLLAPSE WITH INTERNAL PRESSURE FORCES

Spherical collapse with pressure is qualitatively the same as without pressure

So far as the external gravitational field is concerned, the only difference between a freely collapsing star and a collapsing, spherically symmetric star with internal pressure is this: that the surfaces of the two stars move along different world lines in the exterior Schwarzschild geometry. Because the exterior geometry is the same in both cases, *the qualitative aspects of free-fall collapse as described in the last section can be carried over directly to the case of nonnegligible internal pressure.*

An important and fascinating question to ask is this: can large internal pressures in any way prevent a collapsing star from being crushed to infinite density by infinite tidal gravitational forces? From the Kruskal-Szekeres diagram of Figure 32.1,b, it is evident that, once a star has passed inside its gravitational radius ($R < 2M$), no internal pressures, regardless of how large they may be, can prevent the star's surface from being crushed in a singularity. The surface must move along a time-like world line, and all such world lines inside $r = 2M$ hit $r = 0$. Although there is no such theorem now available, one can reasonably conjecture that, if the surface of a spherical configuration is crushed in the $r = 0$ singularity, the entire interior must also be crushed.

The details of the interior dynamics of a spherically symmetric collapsing star with pressure are not so well-understood as the exterior Schwarzschild dynamics. However, major advances in one's understanding of the interior dynamics are now being made by means of numerical computations and analytic analyses [see Misner (1969a) for a review]. In these computations and analyses, no new features (at least, no unexpected ones) have been encountered beyond those that occurred in the simple uniform-density, free-fall collapse of the last section.

EXERCISES

Exercise 32.6. GENERAL SPHERICAL COLLAPSE: METRIC IN COMOVING COORDINATES

Consider an inhomogeneous star with pressure, undergoing spherical collapse. Spherical symmetry alone is enough to guarantee the existence of a Schwarzschild coordinate system (t, r, θ, ϕ) throughout the interior and exterior of the star [see equation (32.2) and preceding discussion]. Label each spherical shell of the star by a parameter a, which tells how many baryons are contained interior to that shell. Then $r(a, t)$ is the world line of the shell with label a. The expression for these world lines can be inverted to obtain $a(t, r)$, the number of baryons interior to radius r at time t. Show that there exists a new time coordinate $\tilde{t}(t, r)$, such that the line element (32.2), rewritten in the coordinates $(\tilde{t}, a, \theta, \phi)$, has the form

$$ds^2 = -e^{2\tilde{\Phi}}\, d\tilde{t}^2 + \left[\frac{(\partial r/\partial a)_{\tilde{t}}\, da}{\Gamma}\right]^2 + r^2\, d\Omega^2, \tag{32.19a}$$

$$\tilde{\Phi} = \tilde{\Phi}(\tilde{t}, a), \qquad r = r(\tilde{t}, a), \qquad \Gamma = \Gamma(\tilde{t}, a). \tag{32.19b}$$

These are "comoving, synchronous coordinates" for the stellar interior.

Exercise 32.7. ADIABATIC SPHERICAL COLLAPSE: EQUATIONS OF EVOLUTION [Misner and Sharp (1964)]

Describe the interior of a collapsing star by the comoving, synchronous metric (32.19), by the number density of baryons n, by the total density of mass-energy ρ, and by the pressure p. The 4-velocity of the star's fluid is

$$\boldsymbol{u} = e^{-\tilde{\Phi}}\,\partial/\partial\tilde{t}, \tag{32.20}$$

since the fluid is at rest in the coordinate system. Let a dot denote a proper time derivative as seen by the fluid—e.g.,

$$\dot{n} \equiv \boldsymbol{u}[n] = e^{-\tilde{\Phi}}(\partial n/\partial\tilde{t})_a;$$

and let a prime denote a partial derivative with respect to baryon number,—e.g.

$$n' \equiv (\partial n/\partial a)_{\tilde{t}}.$$

Denote by U the rate of change of $(1/2\pi) \times$ (circumference of shell), as measured by a man riding in a given shell:

$$U \equiv \dot{r}; \tag{32.21a}$$

and denote by $m(\tilde{t}, a)$ the "total mass-energy interior to shell a at time $\tilde{t}$:

$$m(\tilde{t}, a) \equiv \int_0^a 4\pi r^2 \rho(\tilde{t}, a) r'\, da. \tag{32.21b}$$

(See Box 23.1 for discussion of this method of localizing mass-energy.) Assume that the collapse is adiabatic (no energy flow between adjacent shells; stress-energy tensor entirely that of a perfect fluid).

(a) Show that the equations of collapse [baryon conservation, (22.3); local energy conservation, (22.11a); Euler equation, (22.13); and Einstein field equations (ex. 14.16)] can be reduced to the following eight equations for the eight functions $\tilde{\Phi}$, Γ, r, n, ρ, p, U, m:

$$\dot{r} = U \qquad \text{(dynamic equation for } r\text{);} \tag{32.22a}$$

$$\frac{(nr^2)^{\cdot}}{nr^2} = -\frac{U'}{r'} \qquad \text{(dynamic equation for } n\text{);} \tag{32.22b}$$

$$\frac{\dot{\rho}}{\rho + p} = \frac{\dot{n}}{n}, \qquad \text{except at a shock front, where adiabaticity breaks down (dynamic equation for } \rho\text{);} \tag{32.22c}$$

$$\dot{U} = -\frac{\Gamma^2}{\rho + p}\frac{p'}{r'} - \frac{m + 4\pi r^3 p}{r^2} \qquad \text{(dynamic equation for } U\text{);} \tag{32.22d}$$

$$p = p(n, \rho) \qquad \text{(equation of state);} \tag{32.22e}$$

$$\tilde{\Phi}' = -p'/(\rho + p),\ \tilde{\Phi} = 0 \text{ at star's surface} \qquad \text{(source equation for } \tilde{\Phi}\text{);} \tag{32.22f}$$

$$m' = 4\pi r^2 \rho r', \qquad m = 0 \text{ at } a = 0, \qquad \text{(source equation for } m\text{);} \tag{32.22g}$$

$$\Gamma = \operatorname{sign}(r')(1 + U^2 - 2m/r)^{1/2} \qquad \text{(algebraic equation for } \Gamma\text{).} \tag{32.22h}$$

(b) The preceding equations are in a form useful for numerical calculations. [For particular numerical solutions and for the handling of shocks, see May and White (1966).] For analytic work it is often useful to replace (32.22b) by

$$n = \Gamma/(4\pi r^2 r'), \tag{32.22b'}$$

and (32.22d) by

$$\dot{m} = -4\pi r^2 pU. \tag{32.22d$'$}$$

Derive these equations.

(c) Explain why equations (32.22g) and (32.22d′) justify the remarks made in Box 23.1 about localizability of energy.

Exercise 32.8. ANALYTIC SOLUTIONS FOR PRESSURE-FREE COLLAPSE [Tolman (1934b); Datt (1938)]

Show that the general solution to equations (32.22) in the case of zero pressure can be generated as follows.

(a) Specify the mass inside shell a, $m(a)$; by equation (32.22d′), with $p = 0$, it will not change with time $\tilde{t}$.

(b) Assume that all the dust particles have rest masses μ that depend upon radius, $\mu(a)$; so

$$\rho = \mu n. \tag{32.23a}$$

(c) Calculate Γ from the equation

$$\Gamma = m'/\mu; \tag{32.23b}$$

it will be independent of $\tilde{t}$.

(d) Specify an initial distribution of circumference $2\pi r$ as function of a, and solve the dynamic equation

$$\left(\frac{\partial r}{\partial \tilde{t}}\right)^2 - \frac{2m(a)}{r} = \Gamma^2(a) - 1 \tag{32.23c}$$

to obtain the subsequent evolution of $r(\tilde{t}, a)$. Notice that this equation has identically the same form as in Newtonian theory!

(e) Calculate the remaining quantities of interest from the algebraic equations

$$ds^2 = -d\tilde{t}^2 + (r'\,da/\Gamma)^2 + r^2\,d\Omega^2, \tag{32.23d}$$

$$\rho = \mu n = m'/(4\pi r^2 r'), \tag{32.23e}$$

$$\tilde{\Phi} = 0, \qquad U = \partial r/\partial \tilde{t}. \tag{32.23f}$$

[Note: In this solution, successive "shells" may pass through each other, producing a surface of infinite density as they do ($r' \longrightarrow 0$ where $m' \neq 0$), since there is no pressure built up to stop shell crossing. When this happens, the coordinate system becomes pathological (a no longer increases monotonically outward), but spacetime remains well-behaved. The surface of infinite density (1) produces negligible tidal forces on neighboring dust particles; and (2) like the surface layers of §21.13, it is an idealization that gets smeared down to finite density by finite pressure.]

Exercise 32.9. COLLAPSE WITH UNIFORM DENSITY

Recover the Friedmann-Schwarzschild solution for collapse with uniform density and zero pressure by specifying appropriate forms for $m(a)$ and $r(a)$ in the prescription of exercise 32.8. In the interior of the star, give the dust particles nonzero rest masses, $\mu =$ constant $\neq 0$; in the exterior give them zero rest masses, $\mu = 0$ ("imaginary dust particles" in vacuum). Reduce the resulting metric (32.23d) to that of Friedmann inside the star, and to that of Novikov for the Schwarzschild geometry outside the star [equations (31.12)].

§32.6. THE FATE OF A MAN WHO FALLS INTO THE SINGULARITY AT $r = 0$

The effect of tidal forces on the body of a man falling into the $r = 0$ singularity:

Consider the plight of an experimental astrophysicist who stands on the surface of a freely falling star as it collapses to $R = 0$.

As the collapse proceeds toward $R = 0$, the various parts of the astrophysicist's body experience different gravitational forces. His feet, which are on the surface of the star, are attracted toward the star's center by an infinitely mounting gravitational force; while his head, which is farther away, is accelerated downward by a somewhat smaller, though ever rising force. The difference between the two accelerations (tidal force) mounts higher and higher as the collapse proceeds, finally becoming infinite as R reaches zero. The astrophysicist's body, which cannot withstand such extreme forces, suffers unlimited stretching between head and foot as R drops to zero.

But this is not all. Simultaneous with this head-to-foot stretching, the astrophysicist is pulled by the gravitational field into regions of spacetime with ever-decreasing circumferential area, $4\pi r^2$. In order to accomplish this, tidal gravitational forces must compress the astrophysicist on all sides as they stretch him from head to foot. The circumferential compression is actually more extreme than the longitudinal stretching; so the astrophysicist, in the limit $R \longrightarrow 0$, is crushed to zero volume and indefinitely extended length.

The above discussion can be put on a mathematical footing as follows.

There are three stages in the killing of the astrophysicist: (1) the early stage, when his body successfully resists the tidal forces; (2) the intermediate stage, when it is gradually succumbing; and (3) the final stage, when it has been completely overwhelmed.

Stage 1: body resists deformation; stresses build up

During the early stage, one can analyze the tidal forces by means of the equation of geodesic deviation, evaluated in the astrophysicist's orthonormal frame $\boldsymbol{\omega}^{\hat{\tau}}$, $\boldsymbol{\omega}^{\hat{\rho}}$, $\boldsymbol{\omega}^{\hat{\theta}}$, $\boldsymbol{\omega}^{\hat{\phi}}$ (see §31.2). In this frame, the nonvanishing components of the Riemann curvature tensor are given by equations (31.6):

$$\begin{aligned} R_{\hat{\tau}\hat{\rho}\hat{\tau}\hat{\rho}} &= -2M/r^3, \qquad & R_{\hat{\tau}\hat{\theta}\hat{\tau}\hat{\theta}} &= R_{\hat{\tau}\hat{\phi}\hat{\tau}\hat{\phi}} = M/r^3, \\ R_{\hat{\theta}\hat{\phi}\hat{\theta}\hat{\phi}} &= 2M/r^3, \qquad & R_{\hat{\rho}\hat{\theta}\hat{\rho}\hat{\theta}} &= R_{\hat{\rho}\hat{\phi}\hat{\rho}\hat{\phi}} = -M/r^3. \end{aligned} \tag{32.24a}$$

The equation of geodesic deviation says that two *freely moving* particles, momentarily at rest in the astrophysicist's local inertial frame, and separated by the 3-vector

$$\boldsymbol{\xi} = \xi^{\hat{j}}\boldsymbol{e}_{\hat{j}},$$

must accelerate apart with a relative acceleration given by

$$\begin{aligned} D^2\xi^{\hat{j}}/d\tau^2 &= -R^{\hat{j}}{}_{\hat{\tau}\hat{k}\hat{\tau}}\xi^{\hat{k}} = -R_{\hat{j}\hat{\tau}\hat{k}\hat{\tau}}\xi^{\hat{k}} \\ &= -R_{\hat{\tau}\hat{j}\hat{\tau}\hat{k}}\xi^{\hat{k}}. \end{aligned}$$

Using the components (32.24a) of the curvature tensor, one sees that

$$\begin{aligned} D^2\xi^{\hat{\rho}}/d\tau^2 &= +(2M/r^3)\xi^{\hat{\rho}}, \\ D^2\xi^{\hat{\theta}}/d\tau^2 &= -(M/r^3)\xi^{\hat{\theta}}, \\ D^2\xi^{\hat{\phi}}/d\tau^2 &= -(M/r^3)\xi^{\hat{\phi}}. \end{aligned} \tag{32.24b}$$

To apply these equations to the astrophysicist's body, idealize it (for simplicity) as a homogeneous rectangular box of mass $\mu \approx 165$ pounds ≈ 75 kg, of length $\ell \approx 70$ inches ≈ 1.8 m in the $\boldsymbol{e}_{\hat{\rho}}$ direction, and of width and depth $w \approx 10$ inches ≈ 0.2 m in the $\boldsymbol{e}_{\hat{\theta}}$ and $\boldsymbol{e}_{\hat{\phi}}$ directions. Then calculate the stresses that must be set up in this idealized body to prevent its particles from moving along diverging (and converging) geodesics.

From the form of equations (32.24), it is evident that the principal directions of the stress will be $\boldsymbol{e}_{\hat{\rho}}$, $\boldsymbol{e}_{\hat{\theta}}$, and $\boldsymbol{e}_{\hat{\phi}}$ (i.e., in the $\boldsymbol{e}_{\hat{\rho}}$, $\boldsymbol{e}_{\hat{\theta}}$, $\boldsymbol{e}_{\hat{\phi}}$ basis, the stress tensor will be diagonal). The longitudinal component of the stress, at the astrophysicist's center of mass, can be evaluated as follows. A volume element of his body with mass $d\mu$, located at a height h above the center of mass (distance h measured along $\boldsymbol{e}_{\hat{\rho}}$ direction) would accelerate with $a = (2M/r^3)h$ away from the center of mass, if it were allowed to move freely. To prevent this acceleration, the astrophysicist's muscles must exert a force

$$dF = a\, d\mu = (2M/r^3)h\, d\mu.$$

This force contributes to the stress across the horizontal plane ($\boldsymbol{e}_{\hat{\theta}} \wedge \boldsymbol{e}_{\hat{\phi}}$ plane) through the center of mass. The total force across that plane is the sum of the forces on all mass elements above it (which is also equal to the sum of the forces on the mass elements below it):

$$\begin{aligned} F &= \int_{\text{(region above plane)}} a\, d\mu = \int_0^{\ell/2} \left(\frac{2Mh}{r^3}\right)\left(\frac{\mu}{\ell w^2}\right)(w^2\, dh) \\ &= \frac{1}{4}\frac{\mu M \ell}{r^3}. \end{aligned}$$

The stress is this force divided by the cross-sectional area w^2, with a minus sign because it is a tension rather than a pressure:

$$T_{\hat{\rho}\hat{\rho}} = -\frac{1}{4}\frac{\mu M \ell}{w^2 r^3} \approx -1.1 \times 10^{15}\frac{M/M_\odot}{(r/1\text{ km})^3}\frac{\text{dynes}}{\text{cm}^2}. \tag{32.25a}$$

The components of the stress in the $\boldsymbol{e}_{\hat{\theta}}$ and $\boldsymbol{e}_{\hat{\phi}}$ directions at the center of mass are, similarly,

$$T_{\hat{\theta}\hat{\theta}} = T_{\hat{\phi}\hat{\phi}} = +\frac{1}{8}\frac{\mu M}{\ell r^3} \approx +0.7 \times 10^{13}\frac{M/M_\odot}{(r/1\text{ km})^3}\frac{\text{dynes}}{\text{cm}^2}. \tag{32.25b}$$

(Recall that one atmosphere of pressure is 1.01×10^6 dynes/cm^2.)

Stage 2: body gives way; man dies

The human body cannot withstand a tension or pressure of $\gtrsim 100$ atmospheres $\approx 10^8$ dynes/cm^2 without breaking. Consequently, an astrophysicist on a freely collapsing star of one solar mass will be killed by tidal forces when the star's radius is $R \sim 200$ km $\gg 2M \approx 3$ km.

Stage 3: body gets crushed and distended

By the time the star is much smaller than its gravitational radius, the baryons of the astrophysicist's body are moving along geodesics; his muscles and bones have completely given way. In this final stage of collapse, the timelike geodesics are curves along which the Schwarzschild "time"-coordinate, t, is almost constant [*cf.* the narrowing down of the light cones near $r = 0$ in Figure 32.1,a; also equation (31.2) in the limit $r \ll 2M$]. The astrophysicist's feet touch the star's surface at one value of t—say $t = t_f$—while his head moves along the curve $t = t_h > t_f$. Consequently, the length of the astrophysicist's body increases according to the formula

$$\begin{aligned} \ell_{\text{astroph}} &= [g_{tt}(R)]^{1/2}[t_h - t_f] = [2M/R]^{1/2}[t_h - t_f] \\ &\propto R^{-1/2} \propto (\tau_{\text{collapse}} - \tau)^{-1/3}. \end{aligned} \tag{32.26a}$$

Here $\tau = [-\int^R |g_{rr}|^{1/2}\, dr + \text{constant}]$ is proper time as it would be measured by the astrophysicist if he were still alive, and τ_{collapse} is the time at which he hits $r = 0$. The gravitational field also constrains the baryons of the astrophysicist's body to fall along world lines of constant θ and ϕ during the final stages of collapse. Consequently, his cross-sectional area decreases according to the law

$$\mathscr{A}_{\text{astroph}} = [g_{\theta\theta}(R) g_{\phi\phi}(R)]^{1/2}\, \Delta\theta\, \Delta\phi \propto R^2 \propto (\tau_{\text{collapse}} - \tau)^{4/3}. \tag{32.26b}$$

By combining equations (32.26a,b), one sees that the volume of the astrophysicist's body decreases, during the last few moments of collapse, according to the law

$$\mathscr{V}_{\text{astroph}} = \ell_{\text{astroph}} \mathscr{A}_{\text{astroph}} \propto R^{3/2} \propto (\tau_{\text{collapse}} - \tau). \tag{32.26c}$$

This crushing of matter to infinite density by infinitely large tidal gravitational forces can occur not only on the surface of the collapsing star, but also at any other point along the $r = 0$ singularity outside the surface of the star. Hence, any foolish rocketeer who ventures below the radius $r = 2M$ of the external gravitational field is doomed to destruction.

For further discussion of spacetime singularities, and of the possibility that quantum gravitational effects might force a reconsideration of the singularities predicted by classical gravitation theory, see Chapter 30, §34.6, and Chapter 44.

§32.7. REALISTIC GRAVITATIONAL COLLAPSE—AN OVERVIEW

Review of spherical collapse

Instability, implosion, horizon, and singularity; these are the key stages in the spherical collapse of any star. *Instability:* The star, having exhausted its nuclear fuel, and having contracted slowly inward, begins to squeeze its pressure-sustaining electrons or photons onto its atomic nuclei; this softens the equation of state, which induces an instability [see, e.g., §§10.15 and 11.4 of Zel'dovich and Novikov (1971) for details]. *Implosion:* Within a fraction of a second the instability develops into a full-scale implosion; for realistic density distributions, the stellar core falls rapidly inward on itself, and the outer envelopes trail along behind [see, e.g., the numerical calculations of Colgate and White (1966), Arnett (1966, 1967), May and White (1966), and Ivanova, Imshennik, and Nadezhin (1969)]. *Horizon:* In the idealized spherical case, the star's surface falls through its gravitational radius ("horizon"; end of communication with the exterior; point of no return). From the star's vantage point this happens after a finite, short lapse of proper time. But from an external vantage point the star requires infinite time to reach the horizon, though it becomes black exponentially rapidly in the process [e-folding time $\sim M \sim 10^{-5}(M/M_{\odot})$ sec]. The result is a "black hole", whose boundary is the horizon (gravitational radius), and whose interior can never communicate with the exterior. *Singularity:* From the star's interior vantage point, within a short proper time interval $\Delta\tau \sim M \sim 10^{-5}(M/M_{\odot})$ sec after passing through the horizon, a singularity is reached (zero radius, infinite density, infinite tidal gravitational forces).

Summary of 1972 knowledge about realistic, nonspherical collapse:

Does this basic picture—instability, implosion, horizon, singularity—have any relevance for real stars? Might complications such as rotation, nonsphericity, magnetic fields, and neutrino fluxes alter the qualitative picture? No, not for small initial perturbations from sphericity. Perturbation theory analyses described in Box 32.2 and exercise 32.10 show that *realistic, almost-spherically symmetric collapse, like idealized collapse, is characterized by instability, implosion, horizon;* and Penrose (1965b; see §34.6) proves that *some type of singularity then follows.*

(1) horizon

(2) black hole

Highly nonspherical collapse is more poorly understood, of course. Nevertheless, a number of detailed calculations and precise theorems point with some confidence to two conclusions: (1) *horizons (probably) form when and only when a mass M gets compacted into a region whose circumference in EVERY direction is* $\mathcal{C} \lesssim 4\pi M$ (Box 32.3); (2) *the external gravitational field of a horizon (black hole), after all the "dust" and gravitational waves have cleared away, is almost certainly the Kerr-Newman generalization of the Schwarzschild geometry* (Chapter 33). *If so, then the external field is determined uniquely by the mass, charge, and angular momentum that went "down the hole."* (This nearly proved theorem carries the colloquial title "A black hole has no hair.")

(3) singularity

The interior of the horizon, and the endpoint (if any) of the collapse are very poorly understood today. The various possibilities will be reviewed in Chapter 34. That a singularity occurs one can state with much certainty, thanks to theorems of Penrose, Hawking, and Geroch. But whether all, only some, or none of the collapsing matter and fields ultimately encounter the singularity one does not know.

EXERCISES

Exercise 32.10. PRICE'S THEOREM FOR A SCALAR FIELD [See Price (1971, 1972a), also Thorne (1972), for more details than are presented here.]

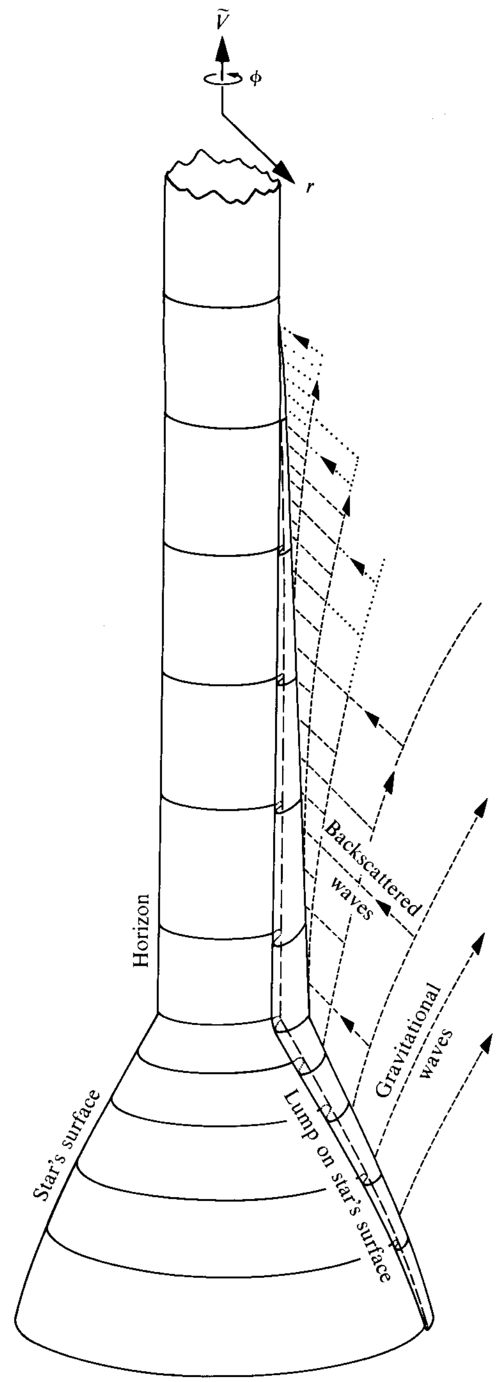
A collapsing spherical star, with an arbitrary nonspherical "scalar charge distribution," generates an external scalar field Φ. The vacuum field equation for Φ is $\Box\Phi = \Phi_{;\alpha}{}^{\alpha} = 0$. Ignore the back-reaction of the field's stress-energy on the geometry of spacetime.

(*continued on page 868*)

Box 32.2 COLLAPSE WITH SMALL NON-SPHERICAL PERTURBATIONS [based on detailed calculations by Richard H. Price (1971, 1972a,b)].

A. Density Perturbations

1. When star begins to collapse, it possesses a small nonspherical "lump" in its density distribution.
2. As collapse proceeds, lump grows larger and larger [instability of collapse against small perturbations—a phenomenon well known in Newtonian theory; see, e.g., Hunter (1967); Lin, Mestel, and Shu (1965)].
3. The growing lump radiates gravitational waves.
4. Waves of short wavelength ($\lambda \ll M$), emitted from near horizon ($r - 2M \lesssim M$), partly propagate to infinity and partly get backscattered by the "background" Schwarzschild curvature of spacetime. Backscattered waves propagate into horizon (surface of black hole; gravitational radius) formed by collapsing star.
5. Waves of long wavelength ($\lambda \gg M$), emitted from near horizon ($r - 2M \lesssim M$), get fully backscattered by spacetime curvature; they never reach out beyond $r \sim 3M$; they end up propagating "down the hole."
6. Is lump on star still there as star plunges through horizon, and does star thereby create a deformed (lumpy) horizon? Yes, according to calculations.
7. *But* external observers can only learn about existence of "final lump" by examining deformation (quadrupole moment) in final gravitational field. That final deformation in field does not and cannot propagate outward with infinite speed (no instantaneous "action at a distance"). It propagates with speed of light, in form of gravitational waves with near-infinite wavelength (infinite redshift from edge of horizon to any external radius). Deformation in final field, like any other wave of long wavelength, gets fully backscattered by curvature of spacetime at $r \lesssim 3M$; it cannot reach external observers. External observers can never learn of existence



Collapse depicted in ingoing Eddington-Finkelstein coordinates of final lump. *Final external field is perfectly spherical, lump-free, Schwarzschild geometry!*

8. Even in region of backscatter ($2M < r \lesssim 3M$), final external field is lump-free. Backscattered waves, carrying information about existence of final lump, interfere destructively with outgoing waves carrying same information. Result is destruction of all deformation in external field and in horizon!

9. Final black hole is a Schwarzschild black hole!

B. Perturbations in Angular Momentum

1. When star begins to collapse, it possesses a small, nonzero intrinsic angular momentum ("spin") $\boldsymbol{S}$.

2. As collapse proceeds, $\boldsymbol{S}$ is conserved (except for a tiny, negligible change due to angular momentum carried off by waves; that change is proportional to square of amplitude of waves, i.e., to square of amplitude of perturbations in star, i.e., to $\boldsymbol{S}^2$).

3. Consequently, external field always and everywhere carries imprint of angular momentum $\boldsymbol{S}$ (on imprints, see Chapter 19). There is no need for that imprint to propagate outward from near horizon. Moreover, it could not so propagate even if it tried, because of the conservation law for $\boldsymbol{S}$ (absence of dipole gravitational waves; see §§36.1 and 36.10).

4. Hence, the final external field is that of an undeformed, slowly rotating black hole:

$$ds^2 = \underbrace{-\left(1 - \frac{2M}{r}\right) dt^2 + \frac{dr^2}{1 - 2M/r} + r^2\, d\Omega^2}_{\text{Schwarzschild geometry}}$$

$$\underbrace{-\left(\frac{4S \sin\theta}{r^2}\right)(r \sin\theta\, d\phi)\, dt.}_{\substack{\text{rotational imprint, see exercise} \\ \text{26.1; also Chapter 19.}}}$$

Here the polar axis has been oriented along $\boldsymbol{S}$.

C. Perturbations in Electromagnetic Field

1. Star possesses a magnetic field generated by currents in its interior, and an electric field due to an arbitrary internal charge distribution; and electromagnetic radiation is emitted by its hot matter. For simplicity, $\boldsymbol{S}$ is assumed zero.

2. Evolution of external electromagnetic field is similar to evolution of perturbations in external gravitational field. Distant observer can never learn "final" values of changeable quantities (magnetic dipole moment, electric dipole moment, quadrupole moments, . . .). Final values try to propagate out from horizon, carried by electromagnetic waves of near-infinite wavelength. But they cannot get out: spacetime curvature reflects them back down the hole; and they superpose destructively with their outgoing counterparts, to produce zero net field.

3. By contrast with all other quantities, which are changeable, the electric monopole moment (total flux of electric field; equal to 4π times total electric charge) is conserved. It never changes from before star collapses, through the collapse stage, into the quiescent black-hole stage.

4. Hence, the final external electromagnetic field is a spherically symmetric coulomb field

$$\left.\begin{array}{l} \boldsymbol{E} = (Q/r^2)\boldsymbol{e}_{\hat{r}} \\ \boldsymbol{B} = 0 \end{array}\right\} \begin{array}{l} \text{as measured by static} \\ \text{observer } (r, \theta, \phi, \text{ constant}); \end{array}$$

and the final spacetime geometry is that of Reissner and Nordstrøm (charged black hole; see exercises 31.8 and 32.1):

$$ds^2 = -\left(1 - \frac{2M}{r} + \frac{Q^2}{r^2}\right) dt^2 + \frac{dr^2}{(1 - 2M/r + Q^2/r^2)} + r^2\, d\Omega^2.$$

Box 32.2 (*continued*)

D. Generalization; Price's Theorem

1. Let the star generate a "zero-rest-mass, integer-spin field." ["Zero rest mass" refers to the quantized particles associated with the classical field. Classically it means the field has a Coulomb-law ($1/r$) fall off at large distances. The spin also is a property of the quantized particles; classically it is most easily visualized as describing the symmetries of a monochromatic plane wave under rotations about the direction of propagation; see §35.6. A scalar field has spin zero; an electromagnetic field has spin one; Einstein's gravitational field has spin two; . . . . Of such fields, only gravitational ($s = 2$) and electromagnetic ($s = 1$) are known to exist in the real universe. See, e.g., Dirac (1936), Gårding (1945), Bargmann and Wigner (1948), Penrose (1965a), for further discussion.]

2. Let the spin-s field be sufficiently weak that its stress-energy perturbs the star's external, Schwarzschild geometry only very slightly.

3. Resolve the external field into spherical harmonics (scalar spherical harmonics for $s = 0$; vector spherical harmonics for $s = 1$; tensor spherical harmonics for $s \geq 2$); and label the spherical harmonics by the usual integer ℓ ($\ell = 0$ for monopole; $\ell = 1$ for dipole; $\ell = 2$ for quadrupole; etc.).

4. All multipole fields with $\ell < s$ are conserved during the collapse (theorem from classical radiation theory). A scalar field ($s = 0$) conserves nothing. The electromagnetic field ($s = 1$) conserves only its monopole parts (electric Coulomb field, and vanishing magnetic Coulomb field). The gravitational field ($s = 2$) conserves its monopole part (with imprint equal to mass), and its dipole parts (with imprints measuring the angular momentum, and the standard gravitational dipole moment—which vanishes if coordinate system is centered on star).

5. For $\ell \geq s$, and only for $\ell \geq s$, radiation is possible (scalar waves can have any multipolarity; electromagnetic waves must be dipole and higher; gravitational waves must be quadrupole and higher; see §36.1).

6. Price's theorem states that, as the nearly spherical star collapses to form a black hole, all things that can be radiated (all multipoles $\ell \geq s$) get radiated completely away—in part "off to infinity"; in part "down the hole" ("what is permitted is compulsory"). The final field is characterized completely by its conserved quantities (multipole moments with $\ell < s$).

7. For proof of Price's theorem in the case of a scalar field, see exercise 32.10.

E. Generalization to Nonclassical Fields

See Hartle (1971, 1972) and Teitelboim (1972b,c) for neutrino fields; Bekenstein (1972a,b) and Teitelboim (1972a) for pion fields.

Box 32.3 COLLAPSE IN ONE AND TWO DIMENSIONS

A. The Question

To produce a black hole (horizon from which nothing can emerge), must one compact matter strongly in all three spatial dimensions, to circumferences $\mathcal{C} \lesssim 4\pi M$ (quasispherical compaction); or is it sufficient to compact only in one or two dimensions?

B. The Answer for One Dimension

Consider, as an example readily generalized, the gravitational collapse of a spheroid of dust (zero pressure). Let the spheroid be highly Newtonian ($r \ggg 2M$) in its initial, momentary state of rest; and let it be slightly flattened (oblate). In Newtonian theory, any homogeneous, nonrotating spheroid of dust remains homogeneous as it collapses; but its deformations grow [see, e.g., Lin, Mestel, and Shu (1965) for details]. Hence, the spheroid of interest implodes to form a pancake of infinite density but finite mass per unit surface area. The final kinetic energy of the dust particles is roughly equal to their final potential energy:

$$\frac{1}{2}v^2 \sim \frac{M}{(\mathcal{C}/2\pi)}$$

$M =$ mass of spheroid,
$\mathcal{C} =$ circumference of final pancake.

Consequently, so long as $\mathcal{C}/2\pi \gg 2M$, the collapse velocities remain much smaller than light, and the gravitational energy remains much smaller than the rest mass-energy. This means that for $\mathcal{C}/2\pi \gg 2M$, the Newtonian analysis is an excellent approximation to general relativity all the way down to the pancake endpoint. Hence, *no horizon can form,* hardly any gravitational waves are emitted, and the whole story is exceedingly simple and fully Newtonian. However, since the pancake endpoint is not a singularity of spacetime (see the remarks at end of exercise 32.8), the evolution can proceed beyond it; and as $\mathcal{C}$ contracts to $\lesssim 4\pi M$, the evolution will become very complicated and highly relativistic (see the "collapse, pursuit, and plunge scenario" of Figure 24.3).

C. The Answer for Two Dimensions

Consider, as an example *not* so readily generalized, the gravitational collapse of a homogeneous prolate spheroid of dust, initially highly Newtonian. Such a spheroid collapses to form a thin "thread" or "spindle" [see Lin, Mestel, and Shu (1965)]. Assume that the spheroid is still Newtonian when its threadlike state is reached. It then has a length ℓ, a mass per unit length $\lambda = M/\ell \ll 1$, and a rapidly contracting equatorial radius $R \ll \ell$. Subsequently, each segment of the thread collapses radially as though it were part of an infinite cylinder. [Ignore the instability of breakup into "beads"; see, e.g., Hunter (1967), Chandrasekhar (1968).] The radial collapse velocity approaches the speed of light and the gravitational energy approaches the rest mass-energy only when the thread has become exceedingly thin, $R \lesssim R_{\rm crit} \sim \ell \exp(-1/4\lambda)$. At this stage, relativistic deviations from Newtonian collapse come into play. Thorne (1972) and Morgan and Thorne (1973) have analyzed the relativistic effects using an idealized infinite-cylinder model. The results are very different from either the spherical case or the pancake case. The collapsing cylinder emits a large flux of gravitational waves; but they are powerless to halt the collapse. *The collapse proceeds inward to a thread-like singularity, without the creation of any horizon* (*no black hole!*).

D. Objection to the Answer, a Reply, and a Conjecture

One can object that the collapses of both pancake and cylinder can be halted short of their endpoints, especially that of the pancake. As the thickness of

Box 32.3 (*continued*)

the pancake approaches zero, the vertical pull of gravity remains finite, but the pressure gradient caused by any finite pressure goes to infinity. Hence, pressure halts the collapse. Subsequently the rim of the pancake contracts toward the relativistic regime $\mathcal{C}/2\pi \lesssim 2M$. In the collapse of a cylinder according to Newtonian theory, with a pressure-density relation $p \propto \rho^{\gamma}$, the gravitational acceleration a_g and pressure-buoyancy acceleration a_p vary as

$$a_g = -2\lambda/R, \qquad a_p \sim \rho^{-1}(p/R) \propto \rho^{\gamma-1}/R.$$

Hence, for $\gamma > 1$ (the most common realistic case) pressure halts the collapse, but for $\gamma < 1$ it does not. Whether this is true also after the relativistic domain is reached, one does not yet know.

Actually, the ability of pressure to halt the collapse is of no importance to the issue of black holes and horizons. The important thing is that *in oblate collapse with final circumference $\mathcal{C} \gg 4\pi M$, and also in prolate collapse with final thread length $\ell \gg 2M$, no horizons are created.* This, coupled with the omnipresent horizons in nearly spherical collapse (Box 32.2) suggests the following conjecture [Thorne (1972)]: *Black holes with horizons form when and only when a mass M gets compacted into a region whose circumference in EVERY direction is $\mathcal{C} \lesssim 4\pi M$.* (Like most conjectures, this one is sufficiently vague to leave room for many different mathematical formulations!)

(a) Resolve the external field into scalar spherical harmonics, using Schwarzschild coordinates for the external Schwarzschild geometry:

$$\Phi = \sum_{\ell} \frac{1}{r} \Psi_\ell(t, r) Y_{\ell m}(\theta, \phi). \tag{32.27a}$$

Show that the vacuum field equation reduces to

$$-\Psi_{\ell,tt} + \Psi_{\ell,r^*r^*} = \left(1 - \frac{2M}{r}\right)\left(\frac{2M}{r^3} + \frac{\ell(\ell+1)}{r^2}\right)\Psi_\ell, \tag{32.27b}$$

where r^* is the "tortoise coordinate" of §25.5 and Figure 25.4:

$$r^* = r + 2M\,\ell n\,(r/2M - 1). \tag{32.27c}$$

Notice that (32.27b) is a flat-space, one-dimensional wave equation with effective potential

$$V_{\text{eff}}(r) = \left(1 - \frac{2M}{r}\right)\left(\frac{2M}{r^3} + \frac{\ell(\ell+1)}{r^2}\right). \tag{32.27d}$$

Part of this effective potential $[\ell(\ell + 1)/r^2]$ is the "centrifugal barrier," and part $[2M/r]$ is due to the curvature of spacetime. Notice the similarity of this effective potential for scalar waves, to the effective potentials for particles and photons moving in the Schwarzschild geometry,

$$(\tilde{V}^2)_{\text{particles}} = (1 - 2M/r)(1 + \tilde{L}^2/r^2),$$
$$(B^{-2})_{\text{photons}} = (1 - 2M/r)r^{-2}$$

(Boxes 25.6 and 25.7). The scalar-wave potential, like the photon potential, is positive for all $r > 2M$. It rises, from 0 at $r = 2M$, to a barrier summit; then falls back to 0 at $r = \infty$.

2

(b) Show that *there exist no physically acceptable, static scalar-wave* perturbations of a Schwarzschild black hole. [More precisely, show that all static solutions to equation (32.27b) become infinite at either the horizon ($r = 2M$, $r^* = -\infty$) or at radial infinity.] This suggests that somehow the black hole formed by collapse must divest itself of the star's external scalar field before it can settle down into a quiescent state.

(c) The general solution to the wave equation (32.27b) can be written in terms of a Fourier transform. For waves that begin near the horizon, propagate outward, and are partially transmitted and partially reflected ("rightward-propagating waves"), show that the general solution is

$$\Psi_\ell(t, r^*) = \int_{-\infty}^{\infty} A(k) R_k^\ell(r^*) e^{-ikt}\, dk, \tag{32.28a}$$

where

$$d^2 R_k^\ell/dr^{*2} = [-k^2 + V_{\text{eff}}(r)] R_k^\ell, \tag{32.28b}$$

$$R_k^\ell = e^{ikr^*} + \Gamma_k^{(R)} e^{-ikr^*} \quad \text{as } r^* \longrightarrow -\infty,$$

$$R_k^\ell = T_k^{(R)} e^{ikr^*} \quad \text{as } r^* \longrightarrow \infty. \tag{32.28c}$$

Show that the "reflection and transmission coefficients for rightward-propagating waves," $\Gamma_k^{(R)}$ and $T_k^{(R)}$, have the following asymptotic forms for $|k| \ll 1/M$ (short wave number; long wavelength):

$$\Gamma_k^{(R)} = -1 + \alpha 2Mik, \qquad T_k^{(R)} = \underbrace{\frac{\beta}{(2\ell - 1)!!}(2Mik)^{\ell+1}} \tag{32.28d}$$

produces complete reflection and complete destructive interference in limit $k \longrightarrow 0$; see Box 32.2 for detailed discussion of consequences

no transmission in limit $k \longrightarrow 0$; see Box 32.2

where α and β are constants of order unity. Give a similar analysis for waves that impinge on a Schwarzschild black hole from outside ("leftward-propagating waves").

(d) Show that, as the star collapses into the horizon, the world line of its surface in (t, r^*) coordinates is

$$r^* = R^*(t) \equiv -t - R_0^* \exp(-t/2M) + \text{const.}, \tag{32.29a}$$

where R_0^* is related to the magnitude a of the surface's 4-acceleration ($a > 0$ for outward 4-acceleration) by

$$R_0^* = (8M/e)\left\{1 + 16Ma\left[Ma + \left(M^2a^2 + \frac{1}{8}\right)^{1/2}\right]\right\}. \tag{32.29b}$$

Thus, the world line of the surface appears to become null near the horizon ($t + r^* \equiv \tilde{V} =$ constant); of course, this is due to pathology of the coordinate system there. Show, further, that the scalar field on the star's surface ($\tilde{V} =$ constant) must vary as

$$\Psi_\ell = Q_{\ell 0} + Q_{\ell 1} e^{-\tilde{U}/4M}, \qquad \tilde{U} \equiv t - r^*. \tag{32.29c}$$

when the star is approaching the horizon ($t \longrightarrow \infty$, $r^* \longrightarrow -\infty$, $\tilde{U} \longrightarrow \infty$), in order that the rate of change of Ψ_ℓ be finite as measured on the star's surface. Notice that $Q_{\ell 0}$ is the "final value" of the scalar field on the star's surface. It can be regarded as an outgoing wave with zero wave number (infinite wavelength); and, consequently, *it gets completely and* *destructively reflected by the effective potential* [see equation (32.28d); also Box 32.2]. Conclusion: All multipoles of the scalar field die out at finite r^* as $t \longrightarrow \infty$. (Price's theorem for a scalar field.) For a more detailed analysis, including the rates at which the multipoles die out, see Price (1971, 1972a) or Thorne (1972).

Exercise 32.11. NEWMAN-PENROSE "CONSTANTS"
[See Press and Bardeen (1971), Bardeen and Press (1972), and Piir (1971) for more details than are presented here.]

Wheeler (1955) showed that Maxwell's equations for an ℓ-pole electromagnetic field residing in the Schwarzschild geometry can be reduced to the wave equation

$$-\Psi_{\ell,tt} + \Psi_{\ell,r^*r^*} = \left(1 - \frac{2M}{r}\right)\frac{\ell(\ell+1)}{r^2}\Psi_\ell \tag{32.30}$$

[electromagnetic analogue of (32.27b)]. After this equation has been solved, the components of the electromagnetic field can be obtained by applying certain differential operators to $\Psi_\ell(t, r^*) Y_{\ell m}(\theta, \phi)$.

(a) Show that the general solution to the electromagnetic wave equation (32.30) for dipole ($\ell = 1$) fields, with outgoing-wave boundary conditions at $r^* \longrightarrow +\infty$, has the form

$$\Psi_1 = f_0(\widetilde{U}) + \frac{f_1(\widetilde{U})}{r} + \frac{f_2(\widetilde{U})}{r^2} + \ldots, \tag{32.31a}$$

where

$$\widetilde{U} = t - r^* \text{ is "retarded time", and}$$

$$f_1' = f_0, \qquad f_2' = 0, \qquad \ldots, \qquad f_n' = -\frac{(n+1)(n-2)}{2n} f_{n-1} + (n-2)Mf_{n-2}. \tag{32.31b}$$

When spacetime is flat ($M = 0$), this solution becomes

$$\Psi_1 = f_1'(\widetilde{U}) + f_1(\widetilde{U})/r. \tag{32.31N}$$

[The $1/r$ fall-off of the radiation field $f_1'(\widetilde{U})$ has been factored out of Ψ_1; see the scalar-wave function (32.27a).] The terms $f_2(\widetilde{U})/r^2 + \ldots$, which are absent in flat spacetime, are attributable to backscatter of the outgoing waves by the curvature of spacetime. They are sometimes called the "tail" of the waves.

(b) Show that the general static dipole field has the form (32.31a) with

$$(f_0)_{\text{static}} = 0; \qquad (f_1)_{\text{static}} \equiv D = \text{dipole moment};$$
$$(f_2)_{\text{static}} = \frac{3}{2}MD; \ldots . \tag{32.32}$$

(c) Consider a star (not a black hole!) with a dipole field that is initially static. At time $t = 0$, let the star suddenly change its dipole moment to a new static value D'. Equations (32.31b) demand that f_2 be conserved ["*Newman-Penrose (1965) constant*"]. Hence, f_2 *will always exhibit a value, $\frac{3}{2}MD$, corresponding to the old dipole moment; it can never change to $\frac{3}{2}MD'$*. This is a manifestation of the tail of the waves that are generated by the sudden change in dipole moment. To understand this tail effect more clearly, and to discover an important flaw in the above result, evaluate the solution (32.31) for retarded time $\widetilde{U} > 0$, using the assumptions

$$\text{(1) field has static form (32.32) for } \widetilde{U} < 0,$$
$$\text{(2) } f_1 = D' \text{ for } \widetilde{U} > 0. \tag{32.33}$$

Put the answer in the form

$$\Psi_1 = \frac{D'}{r} + \frac{\frac{3}{2}MD}{r^2} + \sum_{n=3}^{\infty} \frac{2M(D' - D)(-1)^{n+1}(n+1)\tilde{U}^{n-2}}{(2r)^n} + O\left(\frac{M^2}{r^3}, \frac{M^2\tilde{U}}{r^4}\right). \tag{32.34}$$

(The terms neglected are small compared to those kept for all $\tilde{U}/r$, so long as $r \gg M$.) Evidently, so long as the series converges the Newman-Penrose "constant" (coefficient of $1/r^2$) remembers the old D value and is conserved, as claimed above. Show, however, that the series diverges for $\tilde{U} > 2r$—i.e., it diverges inside a sphere that moves outward with asymptotically $\frac{1}{3}$ the speed of light. Thus, *the Newman-Penrose "constant" is well-defined and conserved only outside the "$\frac{1}{3}$-speed-of-light cone."*

(d) Sum the series in (32.34) to obtain a solution valid even for $\tilde{U} > 2r$:

$$\begin{aligned}\Psi &= \underbrace{\frac{D'}{r} + \frac{3}{2}\frac{MD'}{r^2}}_{\text{new static solution}} - \underbrace{\frac{2M(D'-D)}{r}\frac{(\tilde{U}+3r)}{(\tilde{U}+2r)^2}}_{\text{"tail term"}} + O\left(\frac{M^2}{r^3}\right) \\ &= \text{the series (32.34) for } \tilde{U} < 2r \text{ (domain of convergence of that series)} \\ &= \frac{D'}{r} + \frac{3}{2}\frac{MD'}{r^2} + O\left(\frac{M}{\tilde{U}r}, \frac{M^2}{r^3}\right) \qquad \text{for } \tilde{U} \gg r \gg M.\end{aligned} \tag{32.35}$$

From this result conclude that *at fixed r and late times the electromagnetic field becomes asymptotically static, and the Newman-Penrose "constant" assumes the new value $\frac{3}{2}MD'$ appropriate to the new dipole moment.*

CHAPTER **33**

BLACK HOLES

A luminous star, of the same density as the Earth, and whose diameter should be two hundred and fifty times larger than that of the Sun, would not, in consequence of its attraction, allow any of its rays to arrive at us; it is therefore possible that the largest luminous bodies in the universe may, through this cause, be invisible.

P. S. LAPLACE (1798)

§33.1. WHY "BLACK HOLE"?

A dialog explaining why black holes deserve their name

Sagredus. What is all this talk about "black holes"? When an external observer watches a star collapse, he sees it implode with ever-increasing speed, until the relativistic stage is reached. Then it appears to slow down and become "frozen," just outside its horizon (gravitational radius). However long the observer waits, he never sees the star proceed further. How can one reasonably give the name "black hole" to such a frozen object, which never disappears from sight?

Salvatius. Let us take the name "black hole" apart. Consider first the blackness. Surely nothing can be blacker than a black hole. The very redshift that makes the collapsing star appear to freeze also makes it darken and become black. In the continuum approximation, where one ignores the discreteness of photons, the intensity of the radiation received by distant observers decreases exponentially as time passes, $L \propto \exp(-t/3\sqrt{3}M)$, with an exceedingly short e-folding time

$$\tau = 3\sqrt{3}M = (2.6 \times 10^{-5}\ \mathrm{sec})(M/M_{\odot}).$$

Within a fraction of a second, the star is essentially black. Discreteness of photons makes it even blacker. The number of photons emitted before the star crosses its horizon is finite, so the exponential decay cannot continue

For a more detailed exposition of the foundations of "black-hole physics," see DeWitt and DeWitt (1973).

forever. Eventually—only $10^{-3}(M/M_{\odot})$ seconds after the star begins to dim (see exercise 32.2)—the last photon that will ever get out reaches the distant observers. Thereafter nothing emerges. The star is not merely "essentially black"; it is *"absolutely black."*

Sagredus. Agreed. But it is the word "hole" that concerns me, not "black." How can one possibly regard the name "hole" as appropriate for an object that hovers forever just outside its horizon. True, absence of light makes the object invisible. But couldn't one always see it by shining a flashlight onto its surface? And couldn't one always fly down to its surface in a rocket ship and scoop up a few of the star's baryons? After all, as seen from outside the baryons at its surface will never, never, never manage to fall into the horizon!

Salvatius. Your argument *sounds* persuasive. To test its validity, examine the collapse of a spherically symmetric system, using the ingoing Eddington-Finkelstein diagram of Figure 33.1. Let a family of external observers shine their flashlights onto the star's surface, as you have suggested. Let the surface of the star be carefully silvered so it reflects back all light that reaches it. Initially (low down in the spacetime diagram of Figure 33.1) the ingoing light beams

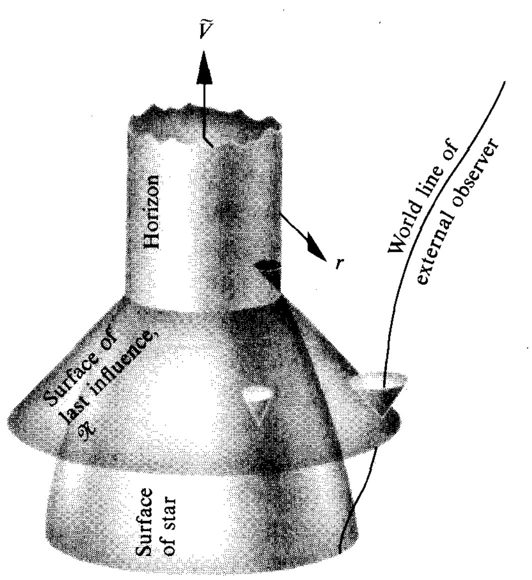


Figure 33.1.
Spherical gravitational collapse of a star to form a black hole, as viewed in ingoing Eddington-Finkelstein coordinates. The "surface of last influence," $\mathcal{R}$, is an ingoing null surface that intersects the horizon in coincidence with the surface of the collapsing star. After an external observer, moving forward in time, has passed through the surface of last influence, he cannot interact with and influence the star before it plunges through the horizon. Thus, one can think of the surface of last influence as the "birthpoint" of the black hole. Before passing through this surface, the external observer can say his flashlight is probing the shape of a collapsing star; afterwards, he can regard his signals as probes of a black hole. For further discussion, see text.

reach the star's surface and are reflected back to the flashlights with no difficulty. But there is a critical point—an ingoing radial null surface $\mathcal{R}$—beyond which reflection is impossible. Photons emitted inward along $\mathcal{R}$ reach the star just as it is passing through its horizon. After reflection these photons fly "outward" along the horizon, remaining forever at $r = 2M$. Other photons, emitted inward after the flashlight has passed through $\mathcal{R}$, reach the surface of the star and are reflected only after the star is inside its horizon. Such photons can never return to the shining flashlights. Once inside the horizon, they can never escape. Thus, the total number of photons returned is finite and is subject to the same blackness decay law as is the intrinsic luminosity of the star. Moreover, *if the observers do not turn on their flashlights until after they pass through the null surface $\mathcal{R}$, they can never receive back any reflected photons!* Evidently, flashlights are of no help in seeing the "frozen star."

Sagredus. I cannot escape the logic of your argument. Nevertheless, seeing is not the only means of interacting with the frozen star. I have already suggested swooping down in a rocket ship and stealing a few baryons from its surface. Similarly, one might let matter fall radially inward onto the frozen star. When the matter hits the star's surface, its great kinetic energy of infall will be converted into heat and into outpouring radiation.

Salvatius. Thus it might seem at first sight. But examine again Figure 33.1. No swooping rocket ship and no infalling matter can move inward faster than a light ray. Thus, if the decision to swoop is made after the ship passes through the surface $\mathcal{R}$, the rocket ship has no possibility of reaching the star before it plunges through the horizon; the rocket and pilot cannot touch the star, sweep up baryons, and return to tell their tale. Similarly, infalling matter to the future of $\mathcal{R}$ can never hit the star's surface before passing through the horizon. The surface $\mathcal{R}$ is, in effect, a "surface of last influence." Once anybody or anything has passed through $\mathcal{R}$, he or it has no possibility whatever of influencing or interacting with the star in any way before it plunges through the horizon. *Thus, from a "causal" or "interaction" standpoint, the collapsing star becomes a hole in space at the surface $\mathcal{R}$.* This hole is not black at first. Radiation from the collapsing star still emerges after $\mathcal{R}$ because of finite light-propagation times, just as radiation still reaches us today from the hot big-bang explosion of the universe. But if an observer at radius $r \gg 2M$ waits for a time $2r$ after passing through $\mathcal{R}$ (time for $\mathcal{R}$ to reach horizon, plus time for rays emitted at $R \sim 3M$ to get back to observer), then he will see the newly formed hole begin to turn black; and within a time $\Delta t \sim (10^{-3}$ seconds$)(M/M_\odot)$ thereafter, it will be completely black.

Sagredus. You have convinced me. For all practical purposes the phrase "black hole" is an excellent description. The alternative phrases "frozen star" and "collapsed star," which I find in the pre-1969 physics literature, emphasize an optical-illusion aspect of the phenomenon. Attention must be directed away from the star that created the black hole, because beyond the surface of last influence one has no means to interact with that star. The star is irrelevant to the subsequent physics and astrophysics. Only the horizon and its external spacetime geometry are relevant for the future. Let us agree to call that horizon the "surface of a black hole," and its external geometry the "gravitational field of the black hole."

Salvatius. Agreed.

§33.2. THE GRAVITATIONAL AND ELECTROMAGNETIC FIELDS OF A BLACK HOLE

The collapse of an electrically neutral star endowed with spherical symmetry produces a spherical black hole with external gravitational field described by the Schwarzschild line element

$$ds^2 = -(1 - 2M/r)\,dt^2 + \frac{dr^2}{1 - 2M/r} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{33.1}$$

The surface of the black hole, i.e., the horizon, is located at $r = 2M =$ (gravitational radius). Only the region on and outside the black hole's surface, $r \geq 2M$, is relevant to external observers. Events inside the horizon can never influence the exterior.

The gravitational collapse of a realistic star (nonspherical, collapse with small but nonzero net charge of one sign or the other) produces a black hole somewhat different from the simple Schwarzschild hole. For collapse with small charge and small asymmetries, perturbation-theory calculations (Box 32.2) predict a final black hole with external field determined entirely by the mass M, charge Q, and intrinsic angular momentum S of the collapsing star. For fully relativistic collapse, with large asymmetries and possibly a large charge, the final black hole (if one forms) is also characterized uniquely by M, Q, and S. This is the conclusion that strongly suggests itself in 1972 from a set of powerful theorems described in Box 33.1.

The structure of a black hole is determined uniquely by its mass M, charge Q, and intrinsic angular momentum, S

Why M, Q, and S should be the complete governors of the final external field of the black hole, one can understand heuristically as follows. Of all quantities intrinsic to any isolated source of gravity and electromagnetism, only M, Q, and S possess (and are defined in terms of) *unique, conserved imprints* in the distant external fields of the source (conserved Gaussian flux integrals; see Box 19.1 and §20.2). When a star collapses to form a black hole, its distant external fields are forced to maintain unchanged the imprints of M, Q, and S. In effect, M, Q, and S provide anchors or constraints on the forms of the fields. Initially other constraints are produced by the distributions of mass, momentum, stress, charge, and current inside the star. But ultimately the star plunges through a horizon, cutting itself off causally from the external universe. (The nonpropagation of long-wavelength waves through curved spacetime plays a key role in this cutoff; see Box 32.2.) Subsequently, the only anchors remaining for the external fields are the conserved imprints of M, Q, and S. Consequently, the external fields quickly settle down into unique shapes corresponding to the given M, Q, and S. Of course, the settling down involves dynamic changes of the fields and an associated outflow of gravitational and electro-

Heuristic explanation of the M-Q-S uniqueness

Box 33.1 A BLACK HOLE HAS NO "HAIR"

The following theorems come close to proving that *the external gravitational and electromagnetic fields of a stationary black hole* (a black hole that has settled down into its "final" state) *are determined uniquely by the hole's mass M, charge Q, and intrinsic angular momentum S*—i.e., the black hole can have no "hair" (no other independent characteristics). For a detailed review, see Carter (1973).

I. Stephen Hawking (1971b, 1972a): A stationary black hole must have a horizon with spherical topology; and it must either be static (zero angular momentum), or axially symmetric, or both.

II. Werner Israel (1967a, 1968): Any *static* black hole with event horizon of spherical topology has external fields determined uniquely by its mass M and charge Q; moreover, those external fields are the Schwarzschild solution if $Q = 0$, and the Reissner-Nordstrøm solution (exercises 31.8 and 32.1) if $Q \neq 0$ (both special cases of Kerr-Newman; see §33.2).

III. Brandon Carter (1970): "All uncharged, stationary, axially symmetric black holes with event horizons of spherical topology fall into disjoint families not deformable into each other. The black holes in each family have external gravitational fields determined uniquely by two parameters: the mass M and the angular momentum S." (Note: the "Kerr solutions"—i.e., "Kerr-Newman" with $Q = 0$ —form one such family; it is very likely that there are no others, but this has not been proved as of December 1972. It is also likely that Carter's theorem can be extended to the case with charge; but this has also not yet been done.)

IV. Conclusions made by combining all three theorems:

(a) All stationary black holes are axially symmetric.

(b) All static (nonrotating) black holes are characterized uniquely by M and Q, and have the Reissner-Nordstrøm form.

(c) All uncharged, rotating black holes fall into distinct and disjoint families, with each black hole in a given family characterized uniquely by M and S. The Kerr solutions form one such family. There may well be no other family.

V. Remarks and Caveats:

(a) The above statements of the theorems are all somewhat heuristic. Each theorem makes several highly technical assumptions, not stated here, about the global properties of spacetime. These assumptions seem physically reasonable and innocuous, but they might not be.

(b) Progress in black-hole physics is so rapid that, by the time this book is published, there may well exist theorems more powerful than the above, which really prove that "a black hole has no hair."

(c) For insight into the techniques of "global geometry" used in proving the above theorems and others like them, see Chapter 34; for greater detail see the forthcoming book by Hawking and Ellis (1973).

(d) For analyses which show that a black hole cannot exert any weak-interaction forces caused by the leptons which have gone down it, see Hartle (1971, 1972) and Teitelboim (1972b,c). For similar analyses which show absence of strong-interaction forces from baryons that have gone down the hole, see Bekenstein (1972a,b) and Teitelboim (1972a).

magnetic waves. And, of course, the outflowing waves carry off mass and angular momentum (but not charge), thereby leaving M and S changed. And, of course, the external fields must then readjust themselves to the new M and S. But the process will quickly converge, producing a black hole with specific final values of M, Q, and S and with external fields determined uniquely by those values.

The problem of calculating the external fields for given M, Q, and S and their given imprints, is analogous to the problem of Plateau—to calculate the shape of a soap film anchored to a wire of given shape.* One calculates the shape of the soap film by seeking a surface of minimum area spanning the bent wire. The condition of minimum area leads to a differential equation describing the soap film, which must be solved subject to the constraint imposed by the shape of the wire.

Variational principle for black-hole structure

To calculate the external fields of a black hole, one can extremize the "action integral" $\int(\mathcal{R} + \mathcal{L})\sqrt{-g}\, d^4x$ for interacting gravitational and electromagnetic fields (see Chapter 21) subject to the anchored-down imprints of M, Q, and S at radial infinity, and subject to the existence of a physically nonsingular horizon (no infinite curvature at horizon!). Extremizing the action is equivalent to solving the coupled Einstein-Maxwell field equations subject to the constraints imprinted by M, Q, and S, and the existence of the horizon. The derivation of the solution and the proof of its uniqueness are much too complex to be given here. (See references cited in Box 33.1.) However, the solution turns out to be the *"Kerr-Newman geometry"* and its associated electromagnetic field.†

Details of black-hole structure:

Written in the t, r, θ, ϕ coordinates of Boyer and Lindquist (1967) (generalization of Schwarzschild coordinates), the Kerr-Newman geometry has the form

(1) metric ("Kerr-Newman geometry")

$$\begin{aligned} ds^2 = &-\frac{\Delta}{\rho^2}[dt - a\sin^2\theta\, d\phi]^2 + \frac{\sin^2\theta}{\rho^2}[(r^2 + a^2)\,d\phi - a\,dt]^2 \\ &+ \frac{\rho^2}{\Delta}dr^2 + \rho^2\, d\theta^2, \end{aligned} \tag{33.2}$$

where

$$\Delta \equiv r^2 - 2Mr + a^2 + Q^2, \tag{33.3a}$$

$$\rho^2 \equiv r^2 + a^2\cos^2\theta, \tag{33.3b}$$

$$a \equiv S/M \equiv \text{angular momentum per unit mass.} \tag{33.4}$$

The corresponding electromagnetic field tensor, written as a 2-form (recall: $\boldsymbol{dx}^\alpha \wedge \boldsymbol{dx}^\beta \equiv \boldsymbol{dx}^\alpha \otimes \boldsymbol{dx}^\beta - \boldsymbol{dx}^\beta \otimes \boldsymbol{dx}^\alpha$) is

(2) electromagnetic field

$$\begin{aligned} \boldsymbol{F} = &\, Q\rho^{-4}(r^2 - a^2\cos^2\theta)\,\boldsymbol{dr} \wedge [\boldsymbol{dt} - a\sin^2\theta\,\boldsymbol{d\phi}] \\ &+ 2Q\rho^{-4}ar\cos\theta\sin\theta\,\boldsymbol{d\theta} \wedge [(r^2 + a^2)\,\boldsymbol{d\phi} - a\,\boldsymbol{dt}]. \end{aligned} \tag{33.5}$$

*On the problem of Plateau see, e.g., Courant (1937), Darboux (1941), or p. 157 of Lipman Bers (1952).

†The uncharged ($Q = 0$) version was first found as a solution to Einstein's vacuum field equations by Kerr (1963). The charged generalization was first found as a solution to the Einstein-Maxwell field equations by Newman, Couch, Chinnapared, Exton, Prakash, and Torrence (1965). Only later was the connection to black holes discovered; see Box 33.1.

Expressions (33.2) for the metric and (33.5) for the electromagnetic field are sufficiently long to be somewhat frightening. Therefore, it is helpful to develop some qualitative insight into them and into their implications before attempting detailed computations with them. Boxes 33.2, 33.3, and 33.4 develop qualitative insight by presenting, without derivation, a summary of the key features of the Kerr-Newman geometry and a summary of the physics and astrophysics of black holes. The remainder of this chapter is a Track-2 justification and derivation of some, but not all, of the results cited in Boxes 33.2–33.4.

(continued on page 891)

Box 33.2 KERR-NEWMAN GEOMETRY AND ELECTROMAGNETIC FIELD

I. Equations for metric and electromagnetic field

A. Parameters appearing in equations:
$M =$ mass, $Q =$ charge, $a \equiv S/M =$ angular momentum per unit mass, all as measured by their standard imprints on the distant fields.

B. Constraint on parameters:
The Kerr-Newman geometry has a horizon, and therefore describes a black hole, if and only if $M^2 \geq Q^2 + a^2$. It seems likely that in any collapsing body which violates this constraint, centrifugal forces and/or electrostatic repulsion will halt the collapse before a size $\sim M$ is reached; see equation (33.56).

C. Limiting cases:

$Q = 0$,	Kerr (1963) geometry;
$S = 0$,	Reissner-Nordstrøm geometry and electromagnetic field (exercises 31.8 and 32.1);
$Q = S = 0$,	Schwarzschild geometry;
$M^2 = Q^2 + a^2$	"Extreme Kerr-Newman geometry."

D. Boyer-Lindquist (1967) coordinates (t, r, θ, ϕ—generalization of Schwarzschild coordinates; black hole rotates in ϕ direction):

$$\begin{aligned} ds^2 = &-(\Delta/\rho^2)[dt - a\sin^2\theta\, d\phi]^2 + (\sin^2\theta/\rho^2)[(r^2 + a^2)\, d\phi - a\, dt]^2 \\ &+ (\rho^2/\Delta)\, dr^2 + \rho^2\, d\theta^2; \end{aligned} \tag{1}$$

$$\Delta \equiv r^2 - 2Mr + a^2 + Q^2, \qquad \rho^2 \equiv r^2 + a^2\cos^2\theta. \tag{2}$$

$$\begin{aligned} \boldsymbol{F} = &\ Q\rho^{-4}(r^2 - a^2\cos^2\theta)\, \boldsymbol{dr} \wedge [\boldsymbol{dt} - a\sin^2\theta\, \boldsymbol{d\phi}] \\ &+ 2Q\rho^{-4}ar\cos\theta\sin\theta\, \boldsymbol{d\theta} \wedge [(r^2 + a^2)\, \boldsymbol{d\phi} - a\, \boldsymbol{dt}]. \end{aligned} \tag{3}$$

E. Kerr coordinates [$\widetilde{V}, r, \theta, \widetilde{\phi}$—generalization of ingoing Eddington-Finkelstein coordinates; $(\widetilde{V}, \theta, \widetilde{\phi}) =$ constant is an ingoing, "radial," null geodesic; black hole rotates in $\widetilde{\phi}$ direction]:
Relationship to Boyer-Lindquist:

$$\boldsymbol{d}\widetilde{V} = \boldsymbol{d}t + (r^2 + a^2)(\boldsymbol{d}r/\Delta),$$
$$\boldsymbol{d}\widetilde{\phi} = \boldsymbol{d}\phi + a(\boldsymbol{d}r/\Delta). \tag{4}$$

$$\begin{aligned} ds^2 = &-[1 - \rho^{-2}(2Mr - Q^2)]\, d\widetilde{V}^2 + 2\, dr\, d\widetilde{V} + \rho^2\, d\theta^2 \\ &+ \rho^{-2}[(r^2 + a^2)^2 - \Delta a^2 \sin^2\theta] \sin^2\theta\, d\widetilde{\phi}^2 - 2a \sin^2\theta\, d\widetilde{\phi}\, dr \\ &- 2a\rho^{-2}(2Mr - Q^2) \sin^2\theta\, d\widetilde{\phi}\, d\widetilde{V}. \end{aligned} \tag{5}$$

$$\begin{aligned} \boldsymbol{F} = Q\rho^{-4}[&(r^2 - a^2 \cos^2\theta)\, \boldsymbol{d}r \wedge \boldsymbol{d}\widetilde{V} - 2a^2 r \cos\theta \sin\theta\, \boldsymbol{d}\theta \wedge \boldsymbol{d}\widetilde{V} \\ &- a \sin^2\theta (r^2 - a^2 \cos^2\theta)\, \boldsymbol{d}r \wedge \boldsymbol{d}\widetilde{\phi} + 2ar(r^2 + a^2) \cos\theta \sin\theta\, \boldsymbol{d}\theta \wedge \boldsymbol{d}\widetilde{\phi}]. \end{aligned} \tag{6}$$

II. Properties of spacetime geometry

A. Symmetries (§33.4):
The metric coefficients in Boyer-Lindquist coordinates are independent of t and ϕ, and in Kerr coordinates are independent of $\widetilde{V}$ and $\widetilde{\phi}$. Thus the spacetime geometry is "time-independent" (stationary) and axially symmetric. The "Killing vectors" (§25.2) associated with these two symmetries are $(\partial/\partial t)_{r,\theta,\phi} = (\partial/\partial \widetilde{V})_{r,\theta,\widetilde{\phi}}$ and $(\partial/\partial\phi)_{t,r,\theta} = (\partial/\partial\widetilde{\phi})_{\widetilde{V},r,\theta}$.

B. Dragging of inertial frames and static limit (§33.4):

1. The "dragging of inertial frames" by the black hole's angular momentum produces a precession of gyroscopes relative to distant stars. By this precession one defines and measures the angular momentum of the black hole (see §§19.2 and 19.3).
2. The dragging becomes more and more extreme the nearer one approaches the horizon of the black hole. Before the horizon is reached, at a surface described by

$$r = r_0(\theta) \equiv M + \sqrt{M^2 - Q^2 - a^2\cos^2\theta}, \tag{7}$$

the dragging becomes so extreme that no observer can possibly remain at rest there (i.e., be "static") relative to the distant stars. At and inside this surface (called the "*static limit*"), all observers with fixed r and θ must orbit the black hole in the same direction in which the hole rotates:

$$\Omega \equiv d\phi/dt > \frac{a\sin\theta - \sqrt{\Delta}}{(r^2 + a^2)\sin\theta - \sqrt{\Delta}\, a\sin^2\theta}$$
$$(\geq 0 \text{ for } a = S/M > 0 \text{ and } r \leq r_0).$$

No matter how hard an observer, at fixed (r, θ) inside the static limit, blasts his rocket engines, he can never halt his angular motion relative to the distant stars.

3. The mathematical foundation for the above statement is this: world lines of the form $(r, \theta, \phi) =$ constant [tangent vector $\propto \partial/\partial t =$ "Killing vector in time direction"] change from being timelike outside the static limit to being spacelike inside it. Therefore, on and inside the static limit, no observer can remain at rest.

C. Horizon (§33.4):

1. The horizon is located at

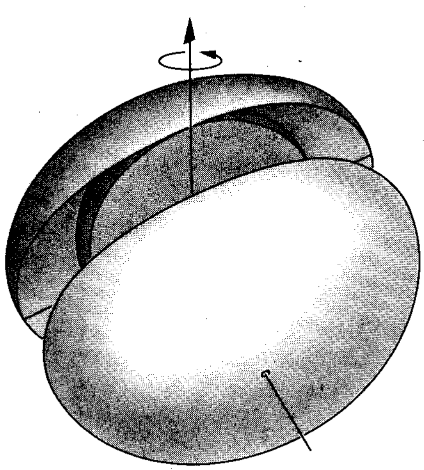
$$r = r_+ \equiv M + \sqrt{M^2 - Q^2 - a^2}. \tag{8}$$

2. As with the Schwarzschild horizon of a nonrotating black hole, so also here, particles and photons can fall inward through the horizon; but no particle or

Box 33.2 (*continued*)

photon can emerge outward through it.
3. The horizon is "generated" by outgoing null geodesics (outgoing photon world lines).

D. Ergosphere (§33.4):
1. The "ergosphere" is the region of spacetime between the horizon and the static limit. It plays a fundamental role in the physics of black holes (Box 33.3; §33.7).
2. The static limit and the horizon touch at the point where they are cut by the axis of rotation of the black hole ($\theta = 0, \pi$); they are well-separated elsewhere with the static limit outside the horizon, unless $a = 0$ (no rotation). When $a = 0$, the static limit and horizon coincide; there is no dragging of inertial frames; there is no ergosphere.



Qualitative representation of horizon, ergosphere, and static limit [adapted from Ruffini and Wheeler (1971b)].

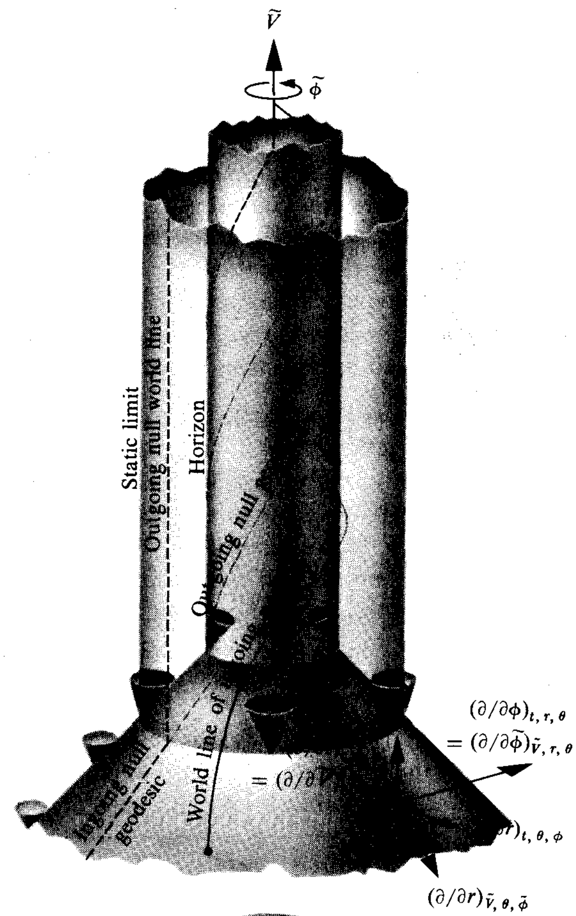
E. Singularity in Boyer-Lindquist coordinates:
1. For a nonrotating black hole, the Schwarzschild coordinates become singular at the horizon. One manifestation of the singularity is the infinite amount of coordinate time required for any particle or photon to fall inward through the horizon, $t \longrightarrow \infty$ as $r \longrightarrow 2M$. One way to remove the singularity (Eddington-Finkelstein way) is to replace t by a null coordinate
$$\widetilde{V} = t + r + 2M \ln|r/2M - 1|$$
attached to infalling photons [so $(\partial/\partial r)_{\widetilde{V},\theta,\phi}$ is vector tangent to photon world lines].
2. For a rotating black hole, the Boyer-Lindquist coordinates, being generalizations of the Schwarzschild coordinates, are also singular at the horizon. It requires an infinite coordinate time for any particle or photon to fall inward through the horizon, $t \longrightarrow \infty$ as $r \longrightarrow r_+$. But that is not all. The dragging of inertial frames forces particles and photons near the horizon to orbit the black hole with $\Omega \equiv d\phi/dt > 0$. Consequently, for a particle falling through the horizon ($r \longrightarrow r_+$), just as $t \longrightarrow \infty$, so also $\phi \longrightarrow \infty$ (infinite twisting of world lines around horizon).
3. To remove the coordinate singularity, one must perform an infinite compression of coordinate time, and an infinite untwisting in the neighborhood of the horizon. Kerr coordinates achieve this by replacing t with a null coordinate $\widetilde{V}$, and ϕ with an untwisted angular coordinate $\widetilde{\phi}$:
$$\boldsymbol{d}\widetilde{V} = \boldsymbol{d}t + (r^2 + a^2)(\boldsymbol{d}r/\Delta),$$
$$\boldsymbol{d}\widetilde{\phi} = \boldsymbol{d}\phi + a(\boldsymbol{d}r/\Delta).$$

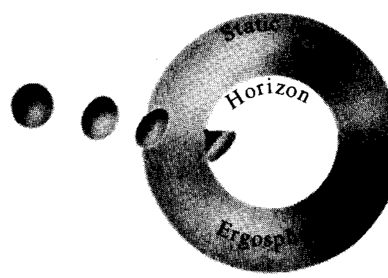
Both of the new coordinates are attached to the world lines of a particular family of infalling photons; $(\partial/\partial r)_{\widetilde{V},\theta,\widetilde{\phi}}$ is the field of vectors tangent to the world lines of this family of photons (ingoing principal null congruence; §33.6).

F. Spacetime diagram:

1. A spacetime diagram in Kerr coordinates looks much like an Eddington-Finkelstein diagram for the Schwarzschild geometry. In both cases, one plots the surfaces of constant $\tilde{V}$ not as horizontal planes, but as "backward light cones" ("45-degree surfaces"), because they are generated by the world lines of ingoing photons. Equivalently, one plots surfaces of constant $\tilde{t} \equiv \tilde{V} - r$ as horizontal planes.
2. The key differences between a Kerr diagram and an Eddington-Finkelstein diagram are: (a) Because the Kerr-Newman geometry is not spherical, a Kerr diagram with one rotational degree of freedom suppressed loses information about the geometry. Kerr diagrams are usually made for the equatorial "plane," $\theta = \pi/2$. (b) Just as the horizon pulls the light cones inward, so the dragging of inertial frames tilts the light cones in the direction of increasing $\tilde{\phi}$, for $a > 0$ and $r =$ constant. (c) The *ingoing* edge of a light cone ($dr/d\tilde{V} = -\infty$) does *not* tilt toward increasing $\tilde{\phi}$; the transformation from Boyer-Lindquist coordinates to Kerr coordinates untwists the tilt with decreasing r, which would otherwise be produced by "frame dragging."
3. The shapes of the light cones reveal the special features of the static limit and horizon. At the static limit, a vertical world line [$r, \theta, \tilde{\phi}$ constant; $(\partial/\partial \tilde{V})_{r,\theta,\tilde{\phi}} = (\partial/\partial t)_{r,\theta,\phi} =$ tangent vector] lies on the light cone. At the horizon the light cones tilts fully inward, except for a single line of tangency to the horizon. Notice that the line of tangency has $d\tilde{\phi}/d\tilde{V} = a/(r_+{}^2 + a^2) \neq 0$. Equivalently, the outgoing null geodesics, which generate the horizon, twist about it ("barber-pole-twist")—yet another manifestation of the dragging of inertial frames.

Kerr diagram for equatorial slice ($\theta = \pi/2$) through the spacetime of an "extreme Kerr" black hole ($Q = 0, a = M$).





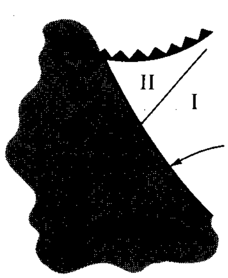
View from above showing the shapes of the light cones as a function of radius

Box 33.2 ***(continued)***

4. The Kerr diagram, like the Eddington-Finkelstein diagram, describes infall through the horizon in a faithful, nonsingular way.
5. [The term "Kerr diagram" is a misnomer. Kerr has not published such diagrams himself, though nowadays others construct such diagrams using his coordinate system. Penrose is the originator and greatest exploiter of such diagrams (see, e.g., Penrose, 1969). But several other types of diagrams bear Penrose's name, so it would be confusing to name them all after him.]

G. Maximal analytic extension of Kerr-Newman geometry:

1. When one abstracts the Schwarzschild geometry away from all sources (Chapter 31), one discovers that it describes an expanding and recontracting bridge, connecting two different universes. But in the context of black holes, only half of the Schwarzschild geometry (regions I and II) is relevant. The other half (regions III and IV) gets fully replaced by the interior of the star that collapsed to form the black hole. Because only a



part of the Schwarzschild geometry comes into play, ingoing Eddington-Finkelstein coordinates—which describe I and II well, but III and IV badly—are well-suited to black-hole physics.

2. Similarly, when one abstracts the Kerr-Newman geometry away from all sources, one discovers that it describes a much larger, and more complex spacetime manifold than one might ever have suspected. This "maximum analytic extension" of the Kerr-Newman geometry has been analyzed in detail by Boyer and Lindquist (1967) and by Carter (1966a, 1968a). But it is totally irrelevant to the subject of black holes, for two reasons. First, as with Schwarzschild, the star that collapsed to form the black hole replaces most of the inward extension of the Kerr-Newman manifold. Second, even outside the star, the Kerr-Newman geometry does not properly represent the true geometry at early times. At early times the star has not got far down the road to collapse. Gravitational moments of the star arise from mountains or prominences or turbulence or other particularities that have not yet gone into the meat grinder. The geometry departs from flatness (1) by a term that varies for large distances as mass divided by distance, and (2) by another term that varies as angular momentum divided by the square of the distance and multiplied by a spherical harmonic of order one, but also (3) by higher-order terms proportional to higher-order mass moments multiplied by higher spherical harmonics. These higher-order terms normally will deviate at early times from the corresponding terms in the mathematical analysis of the Kerr-Newman geometry—though the deviations will die out as time passes. For a system endowed with spherical symmetry, no such higher-order terms do occur or can occur. Therefore the geometry outside is

Schwarzschild in character at all stages of the collapse. However, when the system lacks spherical symmetry, the geometry outside initially departs from Kerr-Newman character. *Only well after the collapse occurs* (asymptotic future), *and in the region at and outside the horizon, is the Kerr-Newman geometry a faithful descriptor of a black hole.* This region is described in a nonsingular manner by Kerr coordinates and Kerr diagrams; and it is the only region that this book will explore.

H. Test-particle orbits
See §§33.5–33.8 and Box 33.5.

III. Properties of electromagnetic field (§33.3):

A. Far from the black hole, where spacetime is nearly flat, in the usual spherical orthonormal frame ($\boldsymbol{\omega}^{\hat{t}} = \boldsymbol{d}t$, $\boldsymbol{\omega}^{\hat{r}} = \boldsymbol{d}r$, $\boldsymbol{\omega}^{\hat{\theta}} = r\,\boldsymbol{d}\theta$, $\boldsymbol{\omega}^{\hat{\phi}} = r \sin\theta\,\boldsymbol{d}\phi$), the electric and magnetic fields have dominant components

$$E_{\hat{r}} = \frac{Q}{r^2};$$

$$B_{\hat{r}} = \frac{2Qa}{r^3}\cos\theta,\ B_{\hat{\theta}} = \frac{Qa}{r^3}\sin\theta.$$

These reveal that

Q = charge of black hole,
$\mathfrak{M} \equiv Qa$ = magnetic dipole moment of black hole.

B. Notice that the gyromagnetic ratio, $\gamma \equiv$ (magnetic moment)/(angular momentum), is equal to Q/M = (charge/mass), just as for an electron!

C. Notice that the value of the magnetic moment, like all other features of the black hole, is determined uniquely by the hole's mass, charge, and angular momentum: $\mathfrak{M} = QS/M$. This illustrates the theorem (Box 33.1) that a black hole has no "hair."

D. Other electric and magnetic moments are nonzero, but are determined uniquely by M, S, and Q.

E. Near the black hole, the curvature of spacetime deforms the electric and magnetic fields produced by the charged, rotating black hole. For a mathematical description of this deformed field, see Cohen and Wald (1971); for a diagrammatic representation, Hanni and Ruffini (1973).

Box 33.3 THE ASTROPHYSICS OF BLACK HOLES

Black holes in nature should participate in astrophysical processes that are as varied as those for stars. By searching for observable phenomena associated with these processes, astronomers have a good chance of discovering the first black hole sometime during the 1970's. This box lists some possible astrophysical processes, and a few relevant references.

I. Mechanisms of Formation

A. "Direct, in isolation": A massive star ($M \gtrsim 3M_\odot$) collapses, almost spherically, producing a collapsed neutron-star core that is too massive to support itself against gravity. Gravity pulls the core on inward, producing a horizon and black hole. [May

Box 33.3 (*continued*)

and White (1966, 1967); Chapter 32 of this book.]

B. "Indirect, in isolation": "Collapse, pursuit, and plunge scenario" depicted in Figure 24.3 [Ruffini and Wheeler (1971b).]

C. "In the thick of things": Stars collected into a dense cluster (e.g., the nucleus of a galaxy) exchange energy. Some acquire energy and move out into a halo. Others lose energy and make a more compact cluster. This process of segregation continues. The cluster becomes so compact that collisions ensue and gas is driven off. The gas moves toward the center of the gravitational potential well. Out of it new stars form. The process continues. Eventually star-star collisions may become sufficiently energetic and inelastic that the centers of the colliding stars coalesce. In this way supermassive objects may be built up and may evolve. Ultimately (1) many "small" stars may collapse to form "small" black holes ($M \sim M_\odot$); (2) one or more supermassive stars may collapse to form huge black holes ($M \sim 10^4 M_\odot$ to $10^9 M_\odot$); (3) the entire conglomerate of stars and gas and holes may become so dense that it collapses to form a single gigantic hole. [Sanders (1970), Spitzer (1971), Lynden-Bell (1967, 1969), Colgate (1967), §§24.5, 24.6, 25.7 of this book.]

D. "Primordially": Perturbations in the initial density distribution of the expanding universe may produce collapse, resulting in "primordial black holes." Those holes would subsequently grow by accretion of radiation and matter. By today all such holes might have grown into enormous objects [$M \sim 10^{17} M_\odot$; Zel'dovich and Novikov (1966)]; but some of them might have avoided such growth and might be as small as 10^{-5} grams [Hawking (1971a)].

II. How many black holes are there in our galaxy today?

Peebles (1972) has given an excellent review of this issue and of prospects for finding black holes in the near future. He says "a good fraction of the mass of the disc of our galaxy was deposited [long ago] in stars capable of collapsing to black holes.... The indication is that the galaxy's disk may contain on the order of 10^9 black holes."

III. "Live" black holes versus "dead" black holes

A. A Schwarzschild black hole is "dead" in the sense that one can never extract from it any of its mass-energy. One aspect of this "deadness"—the fact that a Schwarzschild black hole is stable against small perturbations—is essential (1) to the identification of a black hole with the ultimate "ground state" of a large mass, and (2) to any assertion that general relativity theory predicts the possible existence of black holes. [For a proof of stability see Vishveshwara (1970). The problem was formulated, and most of the necessary techniques developed, by Regge and Wheeler (1957), with essential contributions also by Zerilli (1970a).] Thus a small pulse of gravitational (or other) radiation impinging on a Schwarzschild black hole does not initiate a transition of the black hole into a very different object or state.

B. A Kerr-Newman black hole—which is rotating or charged or both—is not dead. The rotational and electromagnetic contributions to the mass-energy *can* be extracted. (See §§33.7 and 33.8 for mathematical details.) Thus, such black holes are "live"; they can inject energy into their surroundings. By a suitable arrangement of external apparatus, one can trigger an exponentially growing energy release [Press and Teukolsky (1972).] But for a perturbed black hole in isolation, the release is always "controlled" and damped; i.e., Kerr black holes are stable in any classical context [Press and Teukolsky (1973)].

C. Most objects (massive stars; galactic nuclei; . . .) that can collapse to form black holes have so much angular momentum that the holes they produce should be "very live" (a nearly equal to M; S nearly equal to M^2). [Bardeen (1970a).]

D. By contrast, it is quite probable (but far from certain) that no black hole in the universe has substantial charge—i.e., that all black holes have $Q \ll M$. A black hole with $Q \sim M$ (say, $Q > 0$ for concreteness) would exert attractive electrostatic forces on electrons, and repulsive electrostatic forces on protons, that are larger than the hole's gravitational pull by the factor

$$\frac{\text{(electrostatic force)}}{\text{(gravitational force)}} = \frac{eQ}{\mu M} \sim \frac{e}{\mu} \sim 10^{20}.$$

Here e is the electron charge and μ is the electron (or proton) mass. Such huge differential forces are likely to pull in enough charge from outside the hole to neutralize it.

E. But one has learned from the "unipolar induction process" for neutron stars [Goldreich and Julian (1968)] that charge neutralization can sometimes be circumvented. Whether any black-hole process can possibly prevent neutralization one does not know in 1972.

IV. Interaction of a black hole with its environment

A. Gravitational pull: A black hole exerts a gravitational pull on surrounding matter and stars. The pull is indistinguishable, at radii $r \gg M$, from the pull of a star with the same mass.

B. Accretion and emission of x-rays and γ-rays: Gas surrounding a black hole gets pulled inward and is heated by adiabatic compression, by shock waves, by turbulence, by viscosity, etc. Before it reaches the horizon, the gas may become so hot that it emits a large flux of x-rays and perhaps even γ-rays. Thus, accreting matter can convert a black hole into a glowing "white" body [for a review of the literature, see Novikov and Thorne (1973)]. Accretion from a nonrotating gas cloud tends to decrease the angular momentum of a black hole [preferential accretion of particles with "negative" angular momentum; Doroshkevich (1966), Godfrey (1970a)]. But the gas surrounding a hole is likely to be rotating in the same direction as the hole itself, and to maintain $S \sim M^2$ [more precisely, $S \approx 0.998M^2$; Thorne (1973b)].

C. A lump of matter (an "asteroid" or a "planet" or a star) falling into a black hole should emit a burst of gravitational waves as it falls. The total energy radiated is $E \sim 0.01\mu(\mu/M)$, where μ is the mass of the object. [Zerilli (1970b); Davis, Ruffini, Press, and Price (1971); Figure 36.2 of this book.]

D. An object in a stable orbit around a black hole should spiral slowly inward because of loss of energy through gravitational radiation, until it reaches the most tightly bound, stable circular orbit. It should then fall quickly into the hole, emitting a "last-gasp burst" of waves. The total energy radiated during the slow inward spiral is equal to the binding energy of the last stable circular orbit:

$$E_{\text{radiated}} = \mu - E_{\text{last orbit}}$$
$$= \begin{cases} 0.0572\mu \text{ for Schwarzschild hole,} \\ 0.4235\mu \text{ for Kerr hole with } S = M^2,\ Q = 0. \end{cases}$$

Here μ is the rest mass of the captured object. [Box 33.5.] The total energy in the last-gasp burst is $E \sim 0.01\mu(\mu/M)$ if $\mu \ll M$. [Fig. 36.2.]

Box 33.3 (*continued*)

E. When matter falls down a black hole, it can excite the hole's external spacetime geometry into vibration. The vibrations are gradually converted into gravitational waves, some of which escape, others go down the hole. [Press (1971, Goebel (1972).] These vibrations are analogous to an "incipient gravitational geon" [Wheeler (1962); Christodoulou (1971)]—except that for a vibrating black hole the background Kerr geometry holds the vibration energy together (prevents it from propagating away immediately), whereas in a geon it is curvature produced by the "vibration energy" itself that prevents disruption.

F. By a non-Newtonian, induction-zone (i.e., nonradiative) gravitational interaction, a black hole gradually transfers its angular momentum to any non-axially-symmetric, nearby distribution of matter or fields. [Hawking (1972a); Ipser (1971), Press (1972), Hawking and Hartle (1972).]

G. A star or planet falling into a large black hole will get torn apart by tidal gravitational forces. If the tearing occurs near but outside the horizon, it may eject a blob of stellar matter that goes out with relativistic velocity ("tube-of-toothpaste effect"). Moreover, the outgoing jet may extract a substantial amount of rotational energy from the hole's ergosphere—i.e., the hole might throw it off with a rest mass plus kinetic energy in excess of the rest mass of the original infalling object. [Wheeler (1971d); §§33.7 and 33.8.]

H. The magnetic field lines of a charged black hole may be anchored to surrounding plasma, may get wound up as the hole rotates, and may shake, twitch, and excite the plasma.

V. Collisions between black holes

A. Two black holes can collide and coalesce; but there is no way to blast a black hole apart into several black holes [Hawking (1972a); exercise 34.4].

B. When two black holes collide and coalesce, the surface area of the final black hole must exceed the sum of the surface areas of the two initial black holes ("second law of black-hole dynamics"; Hawking (1971a,b); Box 33.4; §34.5). This constraint places an upper limit on the amount of gravitational radiation emitted in the collision. For example, if all three holes are of the Schwarzschild variety and the two initial holes have equal masses $M/2$, then

$$4\pi(2M_{\text{final}})^2 \geq 4\pi[2(M/2)]^2 + 4\pi[2(M/2)]^2,$$
$$M_{\text{final}} \geq M/\sqrt{2},$$

so the energy radiated is

$$E_{\text{radiated}} \leq M - M/\sqrt{2} = 0.293M.$$

VI. Where and how to search for a black hole [For a detailed review, see Peebles (1971)]:

A. When it forms, by the burst or bursts of gravitational radiation given off during formation [Figure 24.3].

B. In a binary star system: black-hole component optically invisible, but may emit x-rays and γ-rays due to accretion; visible component shows telltale Doppler shifts [Hoyle, Fowler, Burbidge, and Burbidge (1964); Zel'dovich and Guseynov (1965); Trimble and Thorne (1969); Pringle and Rees (1972); Shakura and Sunyaev (1973)]. The velocity of the visible component and the period give information on the mass of the invisible component. If mass of this invisible component is four solar masses or more, it cannot be an ordinary star, because an ordinary star of that mass would have $(4)^3 = 64$ times the luminosity of the sun. Neither can it be a white dwarf or a neutron star because either object, so heavy, would instantly collapse to a black hole. Therefore, it is attractive—though not necessarily compelling [see Trimble and Thorne (1969)]—to identify the invisible object as a black hole.

C. [*But* one must not expect to see any noticeable gravitational lens action from a black hole in a binary system: if it taxed the abilities of astronomers for decades to see the black disc of Mercury, 4,800 km in diameter, swim across the great face of the sun, little hope there is to see a black hole with an effective radius of only ~ 3 km, enormously more remote, occult a companion star. Significant lens action requires that the lens (black hole) be separated by a normal interstellar distance from the star it focuses; whence the impact parameter of the focused rays is more than a stellar radius, so the lens action is not more than that of a normal star. Moreover, even with 10^9 black holes in the galaxy, only one per year would pass directly between the Earth and a more distant star, and produce significant lens action (Refsdal, 1964). Chance of watching the right spot on the sky at the right time with a sufficiently strong telescope: nil!]

D. At the center of a globular cluster, where a black hole may settle down, attract normal stars to its vicinity, and thereby produce a cusp in the distribution of light from the cluster. [Cameron and Truran (1971), Peebles (1971).]

E. In the nucleus of a galaxy, including even the Milky Way, where a single huge black hole ($M \sim 10^4$ to $10^8 M_\odot$) might sit as an end-product of earlier activity of the galactic nucleus. Such a hole will emit gravitational waves, light, and radio waves as it accretes matter. Much of the light may be converted into infrared radiation by surrounding dust. The black hole may also produce jets and other nuclear activity. [Lynden-Bell (1969), Lynden-Bell and Rees (1971), Wheeler (1971d), Peebles (1971).]

Box 33.4 THE LAWS OF BLACK-HOLE DYNAMICS

The black-hole processes described in Box 33.3 are governed by the standard laws of physics: general relativity, plus Maxwell electrodynamics, plus hydrodynamic, quantum mechanical, and other laws for the physics of matter and radiation. From these standard laws of physics, one can derive certain "rules" or "constraints," which all black-hole processes must satisfy. Those rules have a power, elegance, and simplicity that rival and resemble the power, elegance, and simplicity of the laws of thermodynamics. Therefore, they have been given the analogous name "the laws of black-hole dynamics" (Israel 1971). This box states two of the laws of black-hole

Box 33.4 *(continued)*

dynamics and some of their ramifications. Two additional laws, not discussed here, have been formulated by Bardeen, Carter, and Hawking (1973).

I. The First and Second Laws of Black-Hole Dynamics.

A. The first law.

1. Like the first law of thermodynamics, the first law of black-hole dynamics is the standard law of conservation of total energy, supplemented by the laws of conservation of total momentum, angular momentum, and charge. For detailed discussions of these conservation laws, see Box 19.1 and Chapter 20.
2. Specialized to the case where matter falls down a black hole and gravitational waves pour out, the first law takes the form depicted and discussed near the end of Box 19.1.
3. Specialized to the case of infalling electric charge, the first law says that the total charge Q of a black hole, as measured by the electric flux emerging from it, changes by an amount equal to the total charge that falls down the hole,

$$\Delta Q = q_{\text{that falls in}}.$$

4. Specialized to the case where two black holes collide and coalesce (example given in Box 33.3), the first law says: (a) Let $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ be the 4-momenta of the two black holes as measured gravitationally, when they are so well-separated that they have negligible influence on each other. ($\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are 4-vectors in the surrounding asymptotically flat spacetime.) Similarly, let $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ be their total angular-momentum tensors (not intrinsic angular-momentum vectors!) relative to some arbitrarily chosen origin of coordinates, $\mathscr{P}_0$, in the surrounding asymptotically flat spacetime ($\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ contain orbital angular momentum, as well as intrinsic angular momentum; see Box 5.6.). (b) Let $\boldsymbol{P}_3$ and $\boldsymbol{J}_3$ be the similar total 4-momentum and angular momentum of the final black hole. (c) Let $\boldsymbol{P}_r$ and $\boldsymbol{J}_r$ be the total 4-momentum and angular momentum radiated as gravitational waves during the collision and coalescence. *Then*

$$\boldsymbol{P}_3 = \boldsymbol{P}_1 + \boldsymbol{P}_2 - \boldsymbol{P}_r, \qquad \boldsymbol{J}_3 = \boldsymbol{J}_1 + \boldsymbol{J}_2 - \boldsymbol{J}_r.$$

[Note: to calculate the mass and intrinsic angular momentum of the final black hole from a knowledge of $\boldsymbol{P}_3$ and $\boldsymbol{J}_3$, follow the prescription of Box 5.6. In that prescription, the world line of the final black hole is that world line, in the distant asymptotically Lorentz coordinates, on which the hole's distant spherical field is centered.

B. *The second law* [expounded and applied by Hawking (1971b, 1972a)]. When anything falls down a black hole, or when several black holes collide and coalesce or collide and scatter, or in any other process whatsoever involving black holes, *the sum of the surface areas* (or squares of "irreducible masses"—see equation 3 below) *of all black holes involved can never decrease.* (See §34.5 for proof.) This is the second law of black-hole dynamics.

II. Reversible and Irreversible Transformations; Irreducible Mass [Christodoulou (1970); Christodoulou and Ruffini (1971)—results derived independently of and simultaneously with Hawking's discovery of the second law.]

A. Consider a single Kerr-Newman black hole interacting with surrounding matter and fields. Its surface area, at any moment of time, is given in terms of its momentary mass M, charge Q, and intrinsic angular momentum per unit mass $a \equiv S/M$ by

$$A = 4\pi[r_+{}^2 + a^2] = 4\pi[(M + \sqrt{M^2 - Q^2 - a^2})^2 + a^2] \tag{1}$$

(exercise 33.12). Interaction with matter and fields may change M, Q, and a in various ways; M can even be decreased—i.e., energy can be extracted from the black hole! [Penrose (1969); §33.7.] But whatever may be the changes, they can never reduce the surface area A. Moreover, if any change in M, Q, and a ever increases the surface area, no future process can ever reduce it back to its initial value.

B. Thus, one can classify black-hole processes into two groups.

1. *Reversible transformations* change M, Q, or a or any set thereof, while leaving the surface area fixed. They can be reversed, bringing the black hole back to its original state.
2. *Irreversible transformations* change M, Q, or a or any set thereof, and increase the surface area in the process. Such a transformation can never be reversed. The black hole can never be brought back to its original state after an irreversible transformation.

C. Examples of reversible transformations and of irreversible transformations induced by infalling particles are presented in §§33.7 and 33.8.

D. The reversible extraction of charge and angular momentum from a black hole (decrease in Q and a holding A fixed) necessarily reduces the black hole's mass (energy extraction!). By the time all charge and angular momentum have been removed, the mass has dropped to a final "irreducible value" of

$$M_{\text{ir}} = (A/16\pi)^{1/2} = \begin{pmatrix}\text{mass of Schwarzschild} \\ \text{black hole of surface area } A\end{pmatrix}. \tag{2}$$

Box 33.4 (*continued*)

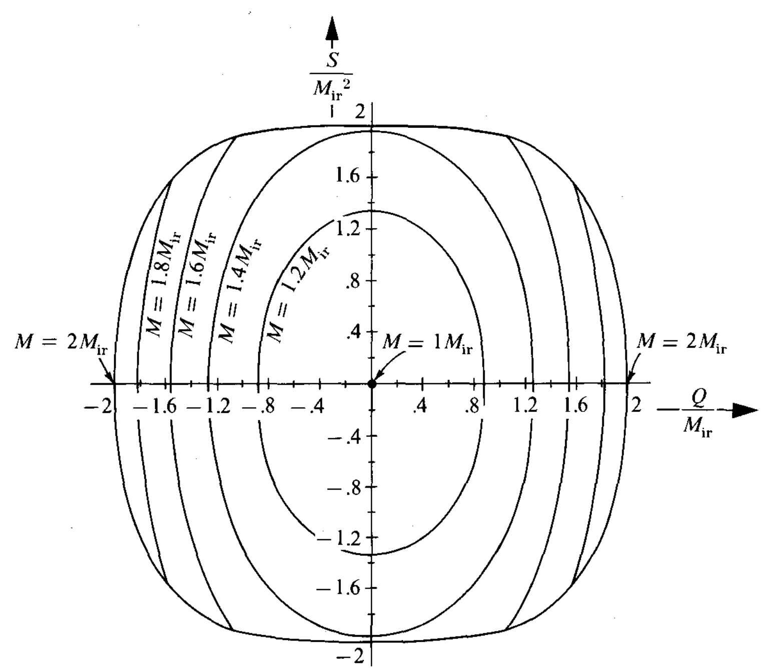
E. Expressed in terms of this final, irreducible mass, the initial mass-energy of the black hole (with charge Q and intrinsic angular momentum S) is

$$M^2 = \left(M_{\mathrm{ir}} + \frac{Q^2}{4M_{\mathrm{ir}}}\right)^2 + \frac{S^2}{4M_{\mathrm{ir}}^{\ 2}} \qquad (3)$$

[This formula, derived by Christodoulou and Ruffini, may be obtained by combining equations (1), (2), and $S = Ma$].

F. Thus, *one can regard the total mass-energy of a black hole as made up of an irreducible mass, an electromagnetic mass-energy, and a rotational energy. But* one must resist the temptation to think of these contributions as adding linearly. On the contrary, they combine in a way [equation (3)] analogous to the way rest mass and linear momentum combine to give energy, $E^2 = m^2 + p^2$.

G. Contours of constant M/M_{ir} are depicted below in the "charge-angular momentum plane." Black holes can exist only in the interior of the region depicted ($Q^2 + a^2 \leq M^2$). [Diagram adapted from Christodoulou (1971).]



H. Since a black hole's irreducible mass is proportional to the square root of its surface area, one can restate the second law of black-hole dynamics as follows:

In black-hole processes the sum of the squares of the irreducible masses of all black holes involved can never decrease.

1

2

§33.3. MASS, ANGULAR MOMENTUM, CHARGE, AND MAGNETIC MOMENT

The rest of this chapter is Track 2. To be prepared for it, one needs to have covered the Track-2 part of Chapter 32 (gravitational collapse). In reading it, one will be helped greatly by Chapter 25 (orbits in Schwarzschild geometry). The rest of this chapter is needed as preparation for Chapter 34 (singularities and global methods).

It is instructive to verify that the constants M, Q, and a, which appear in equations (33.2)–(33.5) for the Kerr-Newman geometry and electromagnetic field, are actually the black hole's mass, charge, and angular momentum per unit mass, as claimed above.

Mass and angular momentum are defined by their imprints on the spacetime geometry far from the black hole. Therefore, to calculate the mass and angular momentum, one can expand the line element (33.2) in powers of $1/r$ and examine the leading terms:

The metric far outside a black hole: imprints of mass and angular momentum

$$\begin{aligned} ds^2 = &-\left[1 - \frac{2M}{r} + O\left(\frac{1}{r^2}\right)\right] dt^2 - \left[\frac{4aM}{r}\sin^2\theta + O\left(\frac{1}{r^2}\right)\right] dt\, d\phi \\ &+ \left[1 + O\left(\frac{1}{r}\right)\right][dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)]. \end{aligned} \tag{33.6}$$

The examination is facilitated by transforming to asymptotically Lorentz coordinates—$x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, $z = r\cos\theta$:

$$\begin{aligned} ds^2 = &-\left[1 - \frac{2M}{r} + O\left(\frac{1}{r^2}\right)\right] dt^2 - \left[\frac{4aM}{r^3} + O\left(\frac{1}{r^4}\right)\right][x\, dy - y\, dx] \\ &+ \left[1 + O\left(\frac{1}{r}\right)\right][dx^2 + dy^2 + dz^2]. \end{aligned} \tag{33.6$'$}$$

Direct comparison with the "standard form" [equation (19.13)] of the metric far from a stationary rotating source reveals that (1) the parameter M is, indeed, the mass of the black hole; and (2) the intrinsic angular momentum vector of the black hole is

$$\boldsymbol{S} = (aM)\,\partial/\partial z = (aM)\cdot\begin{pmatrix}\text{unit vector pointing along polar axis}\\ \text{of Boyer-Lindquist coordinates}\end{pmatrix}. \tag{33.7}$$

The charge is defined for the black hole, as for any source, by a Gaussian flux integral of its electric field over a closed surface surrounding the hole. The electric

2

The electromagnetic field far outside a black hole:

(1) electric field

field in the asymptotic rest frame of the black hole has as its orthonormal components

$$
\begin{aligned}
E_{\hat{r}} &= E_r = F_{rt} = Q/r^2 + O(1/r^3), \\
E_{\hat{\theta}} &= E_\theta/r = F_{\theta t}/r = O(1/r^4), \\
E_{\hat{\phi}} &= E_\phi/r\sin\theta = F_{\theta t}/r\sin\theta = 0.
\end{aligned}
\tag{33.8}
$$

Hence, the electric field is purely radial with a Gaussian flux integral of $4\pi Q$, which reveals Q to be the black hole's charge.

A similar calculation of the dominant components of the magnetic field reveals

(2) magnetic field

$$
\begin{aligned}
B_{\hat{r}} &= F_{\hat{\theta}\hat{\phi}} = \frac{F_{\theta\phi}}{r^2\sin\theta} = 2\frac{Qa}{r^3}\cos\theta + O\left(\frac{1}{r^4}\right), \\
B_{\hat{\theta}} &= F_{\hat{\phi}\hat{r}} = \frac{F_{\phi r}}{r\sin\theta} = \frac{Qa}{r^3}\sin\theta + O\left(\frac{1}{r^4}\right), \\
B_{\hat{\phi}} &= F_{\hat{r}\hat{\theta}} = \frac{F_{r\theta}}{r} = 0.
\end{aligned}
\tag{33.9}
$$

This is a dipole magnetic field, and from it one immediately reads off the value

(3) magnetic dipole moment

$$\mathscr{M} = Qa = (Q/M)S = \underbrace{(\text{charge/mass})}_{[\text{"gyromagnetic ratio"}]} \times (\text{angular momentum}) \tag{33.10}$$

for the magnetic moment of the black hole.

Nonspherical shape of hole's geometry

Just as the rotation of the black hole produces a magnetic field, so it also produces nonspherical deformations in the gravitational field of the black hole [see Hernandez (1967) for quantitative discussion]. But those deformations, like the magnetic moment, are *not* freely specifiable. They are determined uniquely by the mass, charge, and angular momentum of the black hole.

§33.4. SYMMETRIES AND FRAME DRAGGING

The metric components (33.2) of a Kerr-Newman black hole are independent of the Boyer-Lindquist time coordinate t and angular coordinate ϕ. This means (see §25.2) that

Killing vectors for the Kerr-Newman geometry

$$\boldsymbol{\xi}_{(t)} \equiv (\partial/\partial t)_{r,\theta,\phi} \text{ and } \boldsymbol{\xi}_{(\phi)} \equiv (\partial/\partial\phi)_{t,r,\theta} \tag{33.11}$$

are Killing vectors associated with the stationarity (time-translation invariance) and axial symmetry of the black hole. The scalar products of these Killing vectors with themselves and each other are

$$\boldsymbol{\xi}_{(t)} \cdot \boldsymbol{\xi}_{(t)} = g_{tt} = -\left(\frac{\Delta - a^2\sin^2\theta}{\rho^2}\right) = -\left(1 - \frac{2Mr - Q^2}{\rho^2}\right), \quad (33.12a)$$

$$\boldsymbol{\xi}_{(t)} \cdot \boldsymbol{\xi}_{(\phi)} = g_{t\phi} = \frac{a\sin^2\theta\,(\Delta - r^2 - a^2)}{\rho^2} = -\frac{(2Mr - Q^2)a\sin^2\theta}{\rho^2}, \quad (33.12b)$$

$$\boldsymbol{\xi}_{(\phi)} \cdot \boldsymbol{\xi}_{(\phi)} = g_{\phi\phi} = \frac{[(r^2 + a^2)^2 - \Delta a^2\sin^2\theta]\sin^2\theta}{\rho^2}. \quad (33.12c)$$

Since Killing vectors are geometric properties of spacetime, with existence independent of any and all coordinate systems, their scalar products also have coordinate-free meaning. It so happens (not by chance, but by careful choice of coordinates!) that the Boyer-Lindquist metric components g_{tt}, $g_{t\phi}$, and $g_{\phi\phi}$ are equal to these coordinate-independent scalar products. Thus g_{tt}, $g_{t\phi}$, and $g_{\phi\phi}$ can be thought of as three scalar fields which embody information about the symmetries of spacetime. By contrast, the metric coefficients $g_{rr} = \rho^2/\Delta$ and $g_{\theta\theta} = \rho^2$ carry no information at all about the symmetries.* They depend, for their existence and values, on the specific Boyer-Lindquist choice of coordinates.

Stationary observers

Any observer who moves along a world line of constant (r, θ) with uniform angular velocity sees an unchanging spacetime geometry in his neighborhood. Hence, such an observer can be thought of as *"stationary"* relative to the local geometry. If and only if his angular velocity is zero, that is, if and only if he moves along a world line of constant (r, θ, ϕ), will he also be *"static"* relative to the black hole's asymptotic Lorentz frame (i.e., relative to the "distant stars").

Static observers

Angular velocity and 4-velocity of a stationary observer

The precise definition of "angular velocity relative to the asymptotic rest frame" —or simply "angular velocity"—is

$$\Omega \equiv \frac{d\phi}{dt} = \frac{d\phi/d\tau}{dt/d\tau} = \frac{u^\phi}{u^t} \quad (33.13a)$$

(see exercise 33.2). In terms of Ω, the Killing vectors, and the scalar products of Killing vectors, *the 4-velocity of a stationary observer is*

$$\begin{aligned} \boldsymbol{u} = u^t(\partial/\partial t + \Omega\partial/\partial\phi) &= \frac{\boldsymbol{\xi}_{(t)} + \Omega\boldsymbol{\xi}_{(\phi)}}{|\boldsymbol{\xi}_{(t)} + \Omega\boldsymbol{\xi}_{(\phi)}|} \\ &= \frac{\boldsymbol{\xi}_{(t)} + \Omega\boldsymbol{\xi}_{(\phi)}}{(-g_{tt} - 2\Omega g_{t\phi} - \Omega^2 g_{\phi\phi})^{1/2}}. \end{aligned} \quad (33.13b)$$

A stationary observer is static if and only if Ω vanishes.

The stationary observers at given r, θ cannot have any and every angular velocity. Only those values of Ω are allowed for which the 4-velocity $\boldsymbol{u}$ lies inside the future light cone—i.e., for which

$$(\boldsymbol{\xi}_{(t)} + \Omega\boldsymbol{\xi}_{(\phi)})^2 = g_{tt} + 2\Omega g_{t\phi} + \Omega^2 g_{\phi\phi} < 0.$$

*This is not quite true. Kerr-Newman spacetime possesses, in addition to its two Killing vectors, also a "Killing tensor" which is closely linked to the Boyer-Lindquist coordinates r and θ. See Walker and Penrose (1970); also §33.5.

Frame dragging, static limit, and ergosphere

Thus, *the angular velocities of stationary observers are constrained by*

$$\Omega_{\min} < \Omega < \Omega_{\max}, \tag{33.14}$$

where

$$\Omega_{\min} = \omega - \sqrt{\omega^2 - g_{tt}/g_{\phi\phi}}, \tag{33.15a}$$

$$\Omega_{\max} = \omega + \sqrt{\omega^2 - g_{tt}/g_{\phi\phi}}, \tag{33.15b}$$

$$\omega \equiv \frac{1}{2}(\Omega_{\min} + \Omega_{\max}) = -\frac{g_{\phi t}}{g_{\phi\phi}} = \frac{(2Mr - Q^2)a}{(r^2 + a^2)^2 - \Delta a^2 \sin^2\theta}, \tag{33.16}$$

and it is assumed that $S/M = a > 0$. The following features of these limits are noteworthy. (1) Far from the black hole, one has $r\Omega_{\min} = -1$ and $r\Omega_{\max} = +1$, corresponding to the standard limits imposed by the speed of light in flat spacetime. (2) With decreasing radius, $\Omega_{\min}$ increases ("dragging of inertial frames"). Finally, when g_{tt} reaches zero, i.e., at

$$r = r_0(\theta) \equiv M + \sqrt{M^2 - Q^2 - a^2 \cos^2\theta}, \tag{33.17}$$

$\Omega_{\min}$ becomes zero. At and inside this surface, all stationary observers must orbit the black hole with positive angular velocity. Thus, *static observers exist outside and only outside* $r = r_0(\theta)$. For this reason $r = r_0(\theta)$ is called the "static limit"; see Box 33.2. (3) As one moves through the static limit into the "ergosphere," one sees the allowed range of angular velocities become ever more positive (ever more "frame dragging"). At the same time, one sees the allowed range narrow down, until finally, at the horizon

$$r = r_+ \equiv M + \sqrt{M^2 - Q^2 - a^2}, \tag{33.18}$$

the limits $\Omega_{\min}$ and $\Omega_{\max}$ coalesce ($\omega^2 = g_{tt}/g_{\phi\phi}$). Thus, at the horizon there are no stationary observers. All timelike world lines point inward. There is no escape from the black hole's "pull."

Further features of stationary observers and "frame dragging" are explored in the exercises.

EXERCISES

Exercise 33.1. KERR DESCRIPTION OF KILLING VECTORS

(a) Use the transformation law from Boyer-Lindquist coordinates to Kerr coordinates [equation (4) of Box 33.2] to show that

$$\boldsymbol{\xi}_{(t)} \equiv (\partial/\partial t)_{r,\theta,\phi} = (\partial/\partial \tilde{V})_{r,\theta,\tilde{\phi}}, \tag{33.19a}$$

$$\boldsymbol{\xi}_{(\phi)} \equiv (\partial/\partial \phi)_{t,r,\theta} = (\partial/\partial \tilde{\phi})_{\tilde{V},r,\theta}. \tag{33.19b}$$

Verify explicitly by examining metric components that

$$g_{\tilde{V}\tilde{V}} = g_{tt}, \qquad g_{\tilde{V}\tilde{\phi}} = g_{t\phi}, \qquad g_{\tilde{\phi}\tilde{\phi}} = g_{\phi\phi}, \tag{33.19c}$$

in accordance with equations (33.19a,b).

(b) Show that for a stationary observer (world line of constant r, θ), the angular velocity expressed in terms of Kerr coordinates is

$$\Omega \equiv d\phi/dt = d\tilde{\phi}/d\tilde{V} = u^{\tilde{\phi}}/u^{\tilde{V}},$$

so that the entire discussion of stationary observers in terms of Kerr coordinates is identical to the discussion in terms of Boyer-Lindquist coordinates. Differences between the coordinate systems show up only when one moves along world lines of changing r. Reconcile this fact with the fact that both coordinate systems use the *same* coordinates (r, θ) but different time and aximuthal coordinates (t, ϕ versus $\tilde{V}, \tilde{\phi}$).

Exercise 33.2. OBSERVATIONS OF ANGULAR VELOCITY

An observer, far from a black hole and at rest in the hole's asymptotic Lorentz frame, watches (with his eyes) as a particle moves along a stationary (nongeodesic) orbit near the black hole. Let $\Omega = d\phi/dt$ be the particle's angular velocity, as defined and discussed above. The distant observer uses his stopwatch to measure the time required for the particle to make one complete circuit around the black hole (one complete circuit relative to the distant observer himself; i.e., relative to the hole's asymptotic Lorentz frame).

(a) Show that the circuit time measured is $2\pi/\Omega$. Thus, *Ω can be regarded as the particle's "angular velocity as measured from infinity."*

(b) Let the observer moving with the particle measure its circuit time relative to the asymptotic Lorentz frame, using his eyes and a stopwatch he carries. Show that his answer for the circuit time must be

$$\Delta\tau = \frac{2\pi}{\Omega}\underbrace{(-g_{tt} - 2\Omega g_{t\phi} - \Omega^2 g_{\phi\phi})^{1/2}}_{\text{["redshift factor"]}}. \tag{33.20}$$

Exercise 33.3. LOCALLY NONROTATING OBSERVERS (Bardeen 1970b)

(a) Place a rigid, circular mirror ("ring mirror") at fixed (r, θ) around a black hole. Let an observer at (r, θ) with angular velocity Ω emit a flash of light. Some of the photons will get caught by the mirror and will skim along its surface, circumnavigating the black hole in the positive-ϕ direction. Others will get caught and will skim along in the negative-ϕ direction. Show that the observer will receive back the photons from both directions simultaneously only if his angular velocity is

$$\Omega = \omega(r, \theta) = \text{expression (33.16)}.$$

Thus in this case, and only in this case, can the observer regard the $+\phi$ and $-\phi$ directions as equivalent in terms of local geometry. Put differently, in this case and only in this case is the observer "nonrotating relative to the local spacetime geometry." Thus, it is appropriate to use the name *"locally nonrotating observer"* for an observer who moves with the angular velocity $\Omega = \omega(r, \theta)$.

(b) Associated with the axial symmetry of a black hole is a conserved quantity, $\boldsymbol{p}_\phi \equiv \boldsymbol{p} \cdot \boldsymbol{\xi}_{(\phi)}$, for geodesic motion. This quantity for any particle—whether it is moving along a geodesic or not—is called the "component of angular momentum along the black hole's spin axis," or simply the particle's "angular momentum." (See §33.5 below.) Show that of all stationary observers at fixed (r, θ), only the "locally nonrotating observer" has zero angular momentum. [*Note:* Bardeen, Press, and Teukolsky (1972) have shown that the "locally nonrotating observer" can be a powerful tool in the analysis of physical processes near a black hole.]

Exercise 33.4. ORTHONORMAL FRAMES OF LOCALLY NONROTATING OBSERVERS

(a) Let spacetime be filled with world lines of locally nonrotating observers, and let each such observer carry an orthonormal frame with himself. Show that the spatial orientations of these frames can be so chosen that their basis 1-forms are

$$\begin{aligned} \boldsymbol{\omega}^{\hat{t}} &= |g_{tt} - \omega^2 g_{\phi\phi}|^{1/2}\, \boldsymbol{d}t, & \boldsymbol{\omega}^{\hat{\phi}} &= (g_{\phi\phi})^{1/2}(\boldsymbol{d}\phi - \omega\, \boldsymbol{d}t),\\ \boldsymbol{\omega}^{\hat{r}} &= (\rho/\Delta^{1/2})\, \boldsymbol{d}r, & \boldsymbol{\omega}^{\hat{\theta}} &= \rho\, \boldsymbol{d}\theta. \end{aligned} \tag{33.21}$$

More specifically, show that these 1-forms are orthonormal and that the dual basis has

$$\partial/\partial\hat{t} = \boldsymbol{u} \equiv \text{4-velocity of locally nonrotating observer.} \tag{33.22}$$

Show that $\boldsymbol{u} = -\boldsymbol{\omega}^{\hat{t}}$ is a rotation-free field of 1-forms [$\boldsymbol{d\omega}^{\hat{t}} \wedge \boldsymbol{\omega}^{\hat{t}} = 0$; exercise 4.4].

(b) One sometimes meets the mistaken notion that a "locally nonrotating observer" is in some sense locally inertial. To destroy this false impression, verify that: (i) such an observer has nonzero 4-acceleration,

$$\boldsymbol{a} = \Gamma_{\hat{j}\hat{t}\hat{t}}\, \boldsymbol{e}_{\hat{j}} = \frac{1}{2} \boldsymbol{\nabla} \ln |g_{tt} - \omega^2 g_{\phi\phi}|; \tag{33.23}$$

(ii) if such an observer carries gyroscopes with himself, applying the necessary accelerations at the gyroscope centers of mass, he sees the gyroscopes precess relative to his orthonormal frame (33.21) with angular velocity

$$\begin{aligned} \boldsymbol{\Omega}^{(\text{precess})} &= \Gamma_{\hat{\theta}\hat{\phi}\hat{t}}\, \boldsymbol{e}_{\hat{r}} + \Gamma_{\hat{\phi}\hat{r}\hat{t}}\, \boldsymbol{e}_{\hat{\theta}} \\ &= \frac{1}{2} \frac{g_{\phi\phi}{}^{1/2}}{|g_{tt} - \omega^2 g_{\phi\phi}|^{1/2}} \left[\frac{\omega_{,\theta}}{\rho} \boldsymbol{e}_{\hat{r}} - \frac{\Delta^{1/2}\omega_{,r}}{\rho} \boldsymbol{e}_{\hat{\theta}} \right]. \end{aligned} \tag{33.24}$$

[Hints: See exercise 19.2, equation (13.69), and associated discussions. The calculation of the connection coefficients is performed most easily using the methods of differential forms; see §14.6.]

Exercise 33.5. LOCAL LIGHT CONES

Calculate the shapes of the light cones depicted in the Kerr diagram for an uncharged ($Q = 0$) Kerr black hole (part II.F of Box 33.2). In particular, introduce a new time coordinate

$$\tilde{t} \equiv \tilde{V} - r \tag{33.25}$$

for which the slices of constant $\tilde{t}$ are horizontal surfaces in the Kerr diagram. Then the Kerr diagram plots $\tilde{t}$ vertically, r radially, and $\tilde{\phi}$ azimuthally, while holding $\theta = \pi/2$ ("equatorial slice through black hole").

(a) Show that the light cone emanating from given $\tilde{t}$, r, $\tilde{\phi}$ has the form

$$\frac{dr}{d\tilde{t}} = a\left(\frac{d\tilde{\phi}}{d\tilde{t}}\right) - \frac{2M/r}{1 + 2M/r} \pm \sqrt{\frac{1}{(1 + 2M/r)^2} - \frac{r^2(d\tilde{\phi}/d\tilde{t})^2}{1 + 2M/r}}.$$

(b) Show that the light cone slices through the surface of constant radius along the curves

$$dr/d\tilde{t} = 0,\quad d\tilde{\phi}/d\tilde{t} = \Omega_{\text{min}} \text{ and } \Omega_{\text{max}}, \tag{33.26b}$$

where Ω_{min} and Ω_{max} are given by expressions (33.15a,b) (minimum and maximum allowed angular velocities for stationary observers).

(c) Show that at the static limit, $r = r_0(\pi/2)$, the light cone is tangent to a curve of constant r, θ, $\tilde{\phi}$.

(d) Show that the light cone slices the surface of constant $\tilde{\phi}$ along the curves

$$\frac{d\tilde{\phi}}{d\tilde{t}} = 0, \qquad \frac{dr}{d\tilde{t}} = -1 \text{ and } \frac{1 - 2M/r}{1 + 2M/r}. \tag{33.26c}$$

(e) Show that the light cone is tangent to the horizon.
(f) Make pictures of the shapes of the light cone as a function of radius.
(g) Describe qualitatively how the light cone must look near the horizon in Boyer-Lindquist coordinates. (*Note:* it will look "crazy" because the coordinates are singular at the horizon.)

---

§33.5. EQUATIONS OF MOTION FOR TEST PARTICLES [Carter (1968a)]

Let a test particle with electric charge e and rest mass μ move in the external fields of a black hole. Were there no charge down the black hole, the test particle would move along a geodesic (zero 4-acceleration). But the charge produces an electromagnetic field, which in turn produces a Lorentz force on the particle: $\mu \boldsymbol{a} = e\boldsymbol{F} \cdot \boldsymbol{u}$. (Here $\boldsymbol{u}$ is the particle's 4-velocity, and $\boldsymbol{a} \equiv \boldsymbol{\nabla}_{\boldsymbol{u}} \boldsymbol{u}$ is its 4-acceleration.)

The geodesic equation, $\boldsymbol{a} = 0$, for the uncharged case is equivalent to Hamilton's equations

$$dx^\mu/d\lambda = \partial\mathcal{H}/\partial p_\mu, \qquad dp_\mu/d\lambda = -\partial\mathcal{H}/\partial x^\mu, \tag{33.27a}$$

where λ is an affine parameter so normalized that

$$d/d\lambda = \boldsymbol{p} = \text{4-momentum}, \tag{33.27b}$$

and where

$$\mathcal{H} \equiv \text{"super-Hamiltonian"} = \frac{1}{2} g^{\mu\nu} p_\mu p_\nu \tag{33.27c}$$

(see exercise 25.2). Similarly (see exercise 33.6) the Lorentz-force equation, $\mu \boldsymbol{a} = e\boldsymbol{F} \cdot \boldsymbol{u}$, for the charged case is equivalent to Hamilton's equations written in terms of position x^μ and "generalized momentum" π_μ:

$$dx^\mu/d\lambda = \partial\mathcal{H}/\partial \pi_\mu, \qquad d\pi_\mu/d\lambda = -\partial\mathcal{H}/\partial x^\mu. \tag{33.28a}$$

Superhamiltonian for a charged test particle in any electromagnetic field in curved spacetime

The form of the superhamiltonian $\mathcal{H}$, in terms of the metric coefficients at the particle's location, $g^{\mu\nu}(x^\alpha)$, and the particle's charge e and generalized momentum π_μ, is

$$\mathcal{H} = \frac{1}{2} g^{\mu\nu}(\pi_\mu - eA_\mu)(\pi_\nu - eA_\nu). \tag{33.28b}$$

[See §7.3 of Goldstein (1959) for the analogous superhamiltonian in flat spacetime.]

2

The first of Hamilton's equations for this superhamiltonian reduces to

$$p^\mu \equiv (\text{4-momentum}) \equiv dx^\mu/d\lambda = \pi^\mu - eA^\mu \tag{33.29a}$$

(value of π^μ in terms of p^μ, e, and A^μ); the second, when combined with the first, reduces to the Lorentz-force equation

$$dp^\mu/d\lambda + \Gamma^\mu{}_{\alpha\beta}p^\alpha p^\beta = eF^{\mu\nu}p_\nu. \tag{33.29b}$$

$$\left[\begin{matrix} p_\nu, \text{ not } u_\nu \text{ because} \\ \lambda = \tau/\mu \end{matrix}\right]$$

For a Kerr-Newman black hole, the vector potential in Boyer-Lindquist coordinates can be put in the form

Vector potential for a charged black hole

$$\boldsymbol{A} = -\frac{Qr}{\rho^2}(\boldsymbol{d}t - a\sin^2\theta\, \boldsymbol{d}\phi), \tag{33.30}$$

as one verifies by checking that

$$\boldsymbol{dA} = \frac{1}{2}(A_{\beta,\alpha} - A_{\alpha,\beta})\, \boldsymbol{d}x^\alpha \wedge \boldsymbol{d}x^\beta$$

reduces to the Faraday 2-form of equation (33.5).

There is good reason for going through all this formalism, rather than tackling head-on the Lorentz-force equation in its most elementary coordinate version,

$$\frac{d^2x^\alpha}{d\lambda^2} + \Gamma^\alpha{}_{\beta\gamma}\frac{dx^\beta}{d\lambda}\frac{dx^\gamma}{d\lambda} = eF^\alpha{}_\beta\frac{dx^\beta}{d\lambda}.$$

"Constants of motion" for a charged test particle moving around a charged black hole:

The Hamiltonian formalism enables one to discover immediately two constants of the motion; the elementary Lorentz-force equation does not. The key point is that the components A_μ of $\boldsymbol{A}$ [equation (33.30)] and the components $g^{\mu\nu}$ of the metric [inverse of $g_{\mu\nu}$ of equation (33.2); see (33.35)] are independent of t and ϕ (stationarity and axial symmetry of both the electromagnetic field and the spacetime geometry). Consequently, the superhamiltonian is also independent of t and ϕ; and therefore Hamilton's equation

$$d\pi_\alpha/d\lambda = -\partial\mathcal{H}/\partial x^\alpha$$

(1) "energy at infinity" E
(2) "axial component of angular momentum" L_z

guarantees that π_t *and* π_ϕ *are constants of the motion.*

Far from the black hole, where the vector potential vanishes and the metric becomes

$$ds^2 = -dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2),$$

the constants of the motion become

$$\pi_t = p_t = -p^t = -\text{energy},$$

$$\pi_\phi = p_\phi = rp^{\hat\phi} = \begin{pmatrix}\text{projection of angular momentum} \\ \text{along black hole's rotation axis}\end{pmatrix}.$$

Thus it is appropriate to adopt the names and notation

$$E \equiv (\text{"energy at infinity"}) \equiv -\pi_t = -(p_t + eA_t), \tag{33.31a}$$

$$L_z \equiv \begin{pmatrix} \text{"axial component of angular} \\ \text{momentum", or simply} \\ \text{"angular momentum"} \end{pmatrix} \equiv \pi_\phi = p_\phi + eA_\phi \tag{33.31b}$$

for the constants of the motion $-\pi_t$ and π_ϕ.

A third constant of the motion is the particle's rest mass

(3) rest mass μ

$$\mu = |\boldsymbol{p}| = (-g^{\alpha\beta}p_\alpha p_\beta)^{1/2}. \tag{33.31c}$$

In general, four constants of the motion are needed to determine uniquely the orbit of a particle through four-dimensional spacetime. If the black hole were to possess an additional symmetry—e.g., if it were spherical, rather than merely axially symmetric—then automatically there would be a fourth constant of the motion. But in general, black holes are not spherical; so test-particle motion around a black hole possesses only three *obvious* constants. It is rather remarkable, then, that a constant turns out to exist. It was discovered by Carter (1968a), using Hamilton-Jacobi methods. As of 1973, nobody has given a cogent geometric explanation of why this fourth constant should exist—although hints of an explanation may be found in Carter (1968c) and Walker and Penrose (1970).

Carter's "fourth constant" of the motion, as derived in exercise 33.7, is

(4) "$\mathscr{Q}$"

$$\mathscr{Q} = p_\theta{}^2 + \cos^2\theta[a^2(\mu^2 - E^2) + \sin^{-2}\theta L_z{}^2]. \tag{33.31d}$$

The constant of the motion

$$\mathscr{K} \equiv \mathscr{Q} + (L_z - aE)^2, \tag{33.31e}$$

obtained by combining $\mathscr{Q}$, L_z, and E, is often used in place of $\mathscr{Q}$. Whereas $\mathscr{Q}$ can be negative, $\mathscr{K}$ is always nonnegative:

$$\begin{aligned}\mathscr{K} &= p_\theta{}^2 + (L_z - aE\sin^2\theta)^2/\sin^2\theta + a^2\mu^2\cos^2\theta \\ &\geq 0 \text{ everywhere} \\ &= 0 \text{ only for case of photon } (\mu = 0) \text{ moving along polar axis } (\theta = 0, \pi).\end{aligned}$$

The contravariant components of the test particle's 4-momentum, $p^\alpha = dx^\alpha/d\lambda$, are readily expressed in terms of the constants E, L_z, μ, $\mathscr{Q}$, by combining equations (33.31) with the metric coefficients (33.2) and the components of the vector potential (33.30). The result is

Equations of motion for charged test particles

$$\rho^2\, d\theta/d\lambda = \sqrt{\Theta}, \tag{33.32a}$$

$$\rho^2\, dr/d\lambda = \sqrt{R}, \tag{33.32b}$$

$$\rho^2\, d\phi/d\lambda = -(aE - L_z/\sin^2\theta) + (a/\Delta)P, \tag{33.32c}$$

$$\rho^2\, dt/d\lambda = -a(aE\sin^2\theta - L_z) + (r^2 + a^2)\Delta^{-1}P. \tag{33.32d}$$

2

Here $\rho^2 = r^2 + a^2 \cos^2\theta$ as defined in equation (33.3b), and the functions Θ, R, P are defined by

$$\Theta = \mathscr{Q} - \cos^2\theta\,[a^2(\mu^2 - E^2) + L_z{}^2/\sin^2\theta], \tag{33.33a}$$

$$P = E(r^2 + a^2) - L_z a - eQr, \tag{33.33b}$$

$$R = P^2 - \Delta[\mu^2 r^2 + (L_z - aE)^2 + \mathscr{Q}]. \tag{33.33c}$$

When working in Kerr coordinates (to avoid the coordinate singularity at the horizon), one must replace equations (33.32c) and (33.32d) by

$$\rho^2\, d\tilde{V}/d\lambda = -a(aE\sin^2\theta - L_z) + (r^2 + a^2)\Delta^{-1}(\sqrt{R} + P), \tag{33.32c$'$}$$

$$\rho^2\, d\tilde{\phi}/d\lambda = -(aE - L_z/\sin^2\theta) + a\Delta^{-1}(\sqrt{R} + P). \tag{33.32d$'$}$$

[These follow from (33.32) and the transformation between the two coordinate systems—see equations (4) of Box 33.2.] In the above equations, the signs of $\sqrt{R}$ and $\sqrt{\Theta}$ can be chosen independently; but once chosen, they must be used consistently everywhere.

Applications of these equations of motion will play a key role in the rest of this chapter.

EXERCISES

Exercise 33.6. SUPERHAMILTONIAN FOR CHARGED-PARTICLE MOTION

Show that Hamilton's equations (33.28a) for the Hamiltonian (33.28b) reduce to equation (33.29a) for the value of the generalized momentum, and to the Lorentz force equation (33.29b). [*Hint:* Use the relation $(g^{\alpha\beta}g_{\beta\gamma})_{,\mu} = 0$.]

Exercise 33.7. HAMILTON-JACOBI DERIVATION OF EQUATIONS OF MOTION [Based on Carter (1968a)]

Derive the first-order equations of motion (33.32) for a charged particle moving in the external fields of a Kerr-Newman black hole. Use the Hamilton-Jacobi method [Boxes 25.3 and 25.4 of this book; also Chapter 9 of Goldstein (1959)], as follows.

(a) Throughout the superhamiltonian $\mathscr{H}$ of equation (33.28b), replace the generalized momentum π_α by the gradient $\partial S/\partial x^\alpha$ of the Hamilton-Jacobi function.

(b) Write down the Hamilton-Jacobi equation [generalization of equation (2) of Box 25.4] in the form

$$-\frac{\partial S}{\partial \lambda} = \mathscr{H}\left[x^\alpha, \frac{\partial S}{\partial x^\beta}\right] = \frac{1}{2}\, g^{\alpha\beta}\left(\frac{\partial S}{\partial x^\alpha} - eA_\alpha\right)\left(\frac{\partial S}{\partial x^\beta} - eA_\beta\right). \tag{33.34a}$$

(c) Show that the metric components $g^{\alpha\beta}$ for a Kerr-Newman black hole in Boyer-Lindquist coordinates are given by

$$\boldsymbol{g} \equiv g^{\alpha\beta}\frac{\partial}{\partial x^\alpha}\frac{\partial}{\partial x^\beta} = -\frac{1}{\Delta\rho^2}\left[(r^2 + a^2)\frac{\partial}{\partial t} + a\frac{\partial}{\partial \phi}\right]^2 + \frac{1}{\rho^2\sin^2\theta}\left[\frac{\partial}{\partial \phi} + a\sin^2\theta\,\frac{\partial}{\partial t}\right]^2$$
$$+ \frac{\Delta}{\rho^2}\left(\frac{\partial}{\partial r}\right)^2 + \frac{1}{\rho^2}\left(\frac{\partial}{\partial \theta}\right)^2. \tag{33.35}$$

(d) Use these metric components and the components (33.30) of the vector potential to bring the Hamilton-Jácobi equation (33.33) into the concrete form

$$-\frac{\partial S}{\partial \lambda} = -\frac{1}{2}\frac{1}{\Delta\rho^2}\left[(r^2+a^2)\frac{\partial S}{\partial t} + a\frac{\partial S}{\partial \phi} - eQr\right]^2 + \frac{1}{2}\frac{1}{\rho^2\sin^2\theta}\left[\frac{\partial S}{\partial \phi} + a\sin^2\theta\frac{\partial S}{\partial t}\right]^2 + \frac{1}{2}\frac{\Delta}{\rho^2}\left(\frac{\partial S}{\partial r}\right)^2 + \frac{1}{2}\frac{1}{\rho^2}\left(\frac{\partial S}{\partial \theta}\right)^2. \tag{33.34b}$$

(e) Solve this Hamilton-Jacobi equation by separation of variables. [*Hint:* Because the equation has no explicit dependence on λ, ϕ, or t, the solution must take the form

$$S = \frac{1}{2}\mu^2\lambda - Et + L_z\phi + S_r(r) + S_\theta(\theta), \tag{33.36a}$$

where the values of the "integration constants" follow from $\partial S/\partial\lambda = -\mathcal{H}$, $\partial S/\partial t = \pi_t$, $\partial S/\partial\phi = \pi_\phi$. Insert this assumed form into (33.35) and solve for $S_r(r)$ and $S_\theta(\theta)$ to obtain

$$S_r = \int \Delta^{-1}\sqrt{R}\,dr, \qquad S_\theta = \int \sqrt{\Theta}\,d\theta, \tag{33.36b}$$

where $R(r)$ and $\Theta(\theta)$ are the functions defined in equation (33.33). Notice that the constant $\mathcal{Q}$ arises naturally as a "separation-of-variables constant" in this procedure. It was in this way that Carter originally discovered $\mathcal{Q}$, following Misner's suggestion that he seek analogies to a constant in Newtonian dipole fields (Corben and Stehle, 1960, p. 209).]

(f) By successively setting $\partial S/\partial[\mathcal{Q} + (L_z - aE)^2]$, $\partial S/\partial\mu^2$, $\partial S/\partial E$, and $\partial S/\partial L_z$ to zero, obtain the following equations describing the test-particle orbits:

$$\int^\theta \frac{d\theta}{\sqrt{\Theta}} = \int^r \frac{dr}{\sqrt{R}}, \tag{33.37a}$$

$$\lambda = \int^\theta \frac{a^2\cos^2\theta}{\sqrt{\Theta}}\,d\theta + \int^r \frac{r^2}{\sqrt{R}}\,dr, \tag{33.37b}$$

$$t = \int^\theta \frac{-a(aE\sin^2\theta - L_z)}{\sqrt{\Theta}}\,d\theta + \int^r \frac{(r^2+a^2)P}{\Delta\sqrt{R}}\,dr, \tag{33.37c}$$

$$\phi = \int \frac{-(aE\sin^2\theta - L_z)}{\sin^2\theta\,\sqrt{\Theta}}\,d\theta + \int \frac{aP}{\Delta\sqrt{R}}\,dr. \tag{33.37d}$$

(g) By differentiating these equations and combining them, obtain the equations of motion (33.32) cited in the text.

(h) Derive equations (33.31) for E, L_z, μ, and $\mathcal{Q}$ by setting $\partial S/\partial x^\alpha = \pi_\alpha = p_\alpha + eA_\alpha$.

---

§33.6. PRINCIPAL NULL CONGRUENCES

Principal null congruences for the spacetime geometry of a black hole

Two special families of *photon* trajectories "mold themselves into" the Kerr-Newman geometry in an especially harmonious way. They are called the "*principal null congruences*" of the geometry. ("Congruence" is an elegant word that means "space-filling family of curves.") These congruences are the solutions to the test-particle equations of motion (33.32) with

$$\mu = 0 \text{ (zero rest mass; photon)}, \tag{33.38a}$$

$$e = 0 \text{ (zero charge on photon)}, \tag{33.38b}$$

$$L_z = aE\sin^2\theta \qquad \begin{pmatrix}\text{a permissible value for } L_z\\ \text{only because } d\theta/d\lambda \text{ turns}\\ \text{out to be zero}\end{pmatrix}, \tag{33.38c}$$

$$\mathcal{Q} = -(L_z - aE)^2 = -a^2E^2\cos^4\theta. \tag{33.38d}$$

For these values of the constants of motion, the equations of motion (33.32) reduce to

$$k^\theta \equiv d\theta/d\lambda = 0, \tag{33.39a}$$

$$k^r \equiv dr/d\lambda = \pm E \qquad \begin{matrix}\text{("+" for outgoing photons,}\\ \text{"−" for ingoing),}\end{matrix} \tag{33.39b}$$

$$k^\phi \equiv d\phi/d\lambda = aE/\Delta, \tag{33.39c}$$

$$k^t \equiv dt/d\lambda = (r^2 + a^2)E/\Delta. \tag{33.39d}$$

Significance of the principal null congruences

In what sense are these photon trajectories more interesting than others? (1) They mold themselves to the spacetime curvature in such a way that, if $C_{\alpha\beta\gamma\delta}$ is the Weyl conformal tensor (§13.5), and ${}^*C_{\alpha\beta\gamma\delta} = \epsilon_{\alpha\beta\mu\nu}C^{|\mu\nu|}{}_{\gamma\delta}$ is its dual, then

$$C_{\alpha\beta\gamma[\delta}k_{\epsilon]}k^\beta k^\gamma = 0, \qquad {}^*C_{\alpha\beta\gamma[\delta}k_{\epsilon]}k^\beta k^\gamma = 0. \tag{33.40}$$

[This relationship implies that the Kerr-Newman geometry is of "Petrov-Pirani type D" and that these photon trajectories are "doubly degenerate, principal null congruences." For details of the meanings and implications of these terms see, e.g., §8 of Sachs (1964), or Ehlers and Kundt (1962), or the original papers by Petrov (1954, 1969) and Pirani (1957).] (2) By suitable changes of coordinates (exercise 33.8), one can bring the Kerr-Newman metric into the form

$$ds^2 = (\eta_{\alpha\beta} + 2Hk_\alpha k_\beta)\,dx^\alpha\,dx^\beta, \tag{33.41}$$

where H is a scalar field and k_α are the components of the wave vector for one of the principal null congruences (either one; but not both!). [This was the property of the Kerr-Newman metric that led to its original discovery (Kerr, 1963). For further detail on metrics of this form, see Kerr and Schild (1965).] (3) In Kerr coordinates (Box 33.2), the ingoing principal null congruence is

$$\underbrace{r = -E\lambda}_{\text{arbitrary normalization factor; can be removed by redefinition of }\lambda}, \qquad \theta = \text{const}, \qquad \tilde{\phi} = \text{const}, \qquad \tilde{V} = \text{const}. \tag{33.42a}$$

These ingoing photon world lines are the generators of the conical surface $\tilde{V} = \text{const.}$ in the Kerr diagram of Box 33.2. (4) The only kind of particle that can remain forever at the horizon is a photon with world line in the outgoing principal null congruence (exercise 33.9). Such photon world lines are "generators" of the horizon (dotted curves with a "barber-pole twist" in Kerr diagram of Box 33.2). They have angular velocity

$$\Omega = \frac{d\phi}{dt} = \frac{d\tilde{\phi}}{d\tilde{V}} = \frac{a}{r_+{}^2 + a^2} = \frac{a}{2M^2 - Q^2 + 2M(M^2 - a^2 - Q^2)^{1/2}}. \tag{33.42b}$$

EXERCISES

Exercise 33.8. KERR-SCHILD COORDINATES

(a) ~~Show that in Kerr~~ coordinates the ingoing null congruence (33.39) has the form (33.42a). Also show that the covariant components of the wave vector—after changing to a new affine parameter $\lambda_{\text{new}} = \lambda_{\text{old}}E$—are

$$k_r^{(\text{in})} = 0, \qquad k_\theta^{(\text{in})} = 0, \qquad k_{\tilde{\phi}}^{(\text{in})} = a\sin^2\theta, \qquad k_{\tilde{V}}^{(\text{in})} = -1. \tag{33.43}$$

(b) Introduce new coordinates $\tilde{t}$, x, y, z, defined by

$$x + iy = (r + ia)e^{i\tilde{\phi}}\sin\theta, \qquad z = r\cos\theta, \qquad \tilde{t} = \tilde{V} - r; \tag{33.44a}$$

and show that in this "Kerr-Schild coordinate system" the metric takes the form

$$ds^2 = (\eta_{\alpha\beta} + 2Hk_\alpha^{(\text{in})}k_\beta^{(\text{in})})\,dx^\alpha\,dx^\beta, \tag{33.44b}$$

where

$$H = \frac{Mr - \frac{1}{2}Q^2}{r^2 + a^2(z/r)^2}, \tag{33.44c}$$

$$k_\alpha^{(\text{in})}\,dx^\alpha = -\frac{r(x\,dx + y\,dy) - a(x\,dy - y\,dx)}{r^2 + a^2} - \frac{z\,dz}{r} - d\tilde{t}. \tag{33.44d}$$

For the transformation to analogous coordinates in which

$$ds^2 = (\eta_{\alpha\beta} + 2Hk_\alpha^{(\text{out})}k_\beta^{(\text{out})})\,dx^\alpha\,dx^\beta.$$

see, e.g., Boyer and Lindquist (1967).

Exercise 33.9. NULL GENERATORS OF HORIZON

(a) Show that in Kerr coordinates the outgoing principle null congruence is described by the tangent vector

$$\frac{d\theta}{d\lambda} = 0, \qquad \frac{dr}{d\lambda} = E, \qquad \frac{d\tilde{\phi}}{d\lambda} = 2a\frac{E}{\Delta}, \qquad \frac{d\tilde{V}}{d\lambda} = 2(r^2 + a^2)\frac{E}{\Delta}. \tag{33.45}$$

(b) These components of the wave vector become singular at the horizon ($\Delta = 0$), not because of a singularity in the coordinate system—the coordinates are well-behaved!—but because of poor normalization of the affine parameter. For each outgoing geodesic, let Δ_0

be a constant, defined as the value of Δ at the event where the geodesic slices the hypersurface $\tilde{V} = 0$. Then renormalize the affine parameter for each geodesic

$$\lambda_{\text{new}} = (E/\Delta_0)\lambda_{\text{old}}. \tag{33.46}$$

Show that the resulting wave vectors

$$\frac{d\theta}{d\lambda} = 0, \qquad \frac{dr}{d\lambda} = \Delta_0, \qquad \frac{d\tilde{\phi}}{d\lambda} = 2a\frac{\Delta_0}{\Delta}, \qquad \frac{d\tilde{V}}{d\lambda} = 2(r^2 + a^2)\frac{\Delta_0}{\Delta} \tag{33.45$'$}$$

are well-behaved as one approaches the horizon; and show that the geodesics on the horizon have the form

$$\theta = \text{const.}, \qquad r = r_+ = \text{const.}, \qquad \tilde{\phi} = 2a\lambda, \qquad \tilde{V} = 2({r_+}^2 + a^2)\lambda. \tag{33.47}$$

(c) Show that these are the only test-particle trajectories that remain forever on the horizon. [*Hint:* Examine the light cone.]

§33.7. STORAGE AND REMOVAL OF ENERGY FROM BLACK HOLES [Penrose (1969)]

When a small object falls down a large hole:

(1) energy radiated is negligible compared to object's rest mass

When an object falls into a black hole, it changes the hole's mass, charge, and intrinsic angular momentum (first law of black-hole dynamics; Box 33.4). If the infalling object is large, its fall produces much gravitational and electromagnetic radiation. To calculate the radiation emitted, and the energy and angular momentum it carries away—which are prerequisites to any calculation of the final state of the black hole—is an enormously difficult task. But if the object is very small (size of object $\ll$ size of horizon; mass of object $\ll$ mass of hole), and has sufficiently small charge, the radiation it emits in each circuit around the hole is negligible. For example, for gravitational radiation

$$\frac{(\text{energy emitted per circuit})}{(\text{rest mass of object})} \sim \frac{(\text{rest mass of object})}{(\text{mass of hole})} \tag{33.48}$$

[see §36.5; also Bardeen, Press, and Teukolsky (1972)]. Because the energy emitted is negligible, radiation reaction is also negligible, and the object moves very nearly along a test-particle trajectory. In this case, application of the first law of black-hole dynamics is simple and straightforward.

(2) hole's mass, charge, and angular momentum change by $\Delta M = E$, $\Delta Q = e$, $\Delta S = L_z$

Consider, initially, a small object that falls directly into the black hole from far away. According to the first law, it produces the following changes in the mass, charge, and angular momentum of the black hole:

$$\Delta M = E = (\text{“energy at infinity” of infalling object}), \tag{33.49a}$$

$$\Delta Q = e = (\text{charge of infalling object}), \tag{33.49b}$$

$$\Delta S \equiv \Delta|\boldsymbol{S}| = L_z = \begin{pmatrix}\text{component of object's angular momentum}\\ \text{on black hole's rotation axis}\end{pmatrix}. \tag{33.49c}$$

The infalling object will also change the direction of $\boldsymbol{S}$. In the black hole's original asymptotic Lorentz frame, its initial angular momentum vector points in the z-direction,

$$(S_z)_{\text{initial}} = S, \qquad (S_x)_{\text{initial}} = 0, \qquad (S_y)_{\text{initial}} = 0.$$

Consequently, only the z-component of angular momentum of the infalling object can produce any significant change in the magnitude of $\boldsymbol{S}$. But the x- and y-components, L_x and L_y, can change the direction of $\boldsymbol{S}$. If the object has negligible speed at infinity, then it produces the changes (exercise 33.10):

$$\Delta S_x = L_x = -(\sin\phi_\infty)\sqrt{\boldsymbol{\Theta}_\infty} - (\cot\theta_\infty \cos\phi_\infty)L_z, \tag{33.49d}$$

$$\Delta S_y = L_y = (\cos\phi_\infty)\sqrt{\boldsymbol{\Theta}_\infty} - (\cot\theta_\infty \sin\phi_\infty)L_z, \tag{33.49e}$$

$$\Delta(S_x{}^2 + S_y{}^2)^{1/2} = \sqrt{\mathscr{Q}} = (L^2 - L_z{}^2)^{1/2}. \tag{33.49f}$$

Here a subscript "∞" means the value of a quantity at a point on the orbit far from the black hole (at "infinity").

Consider, next, a more complicated process, first conceived of by Penrose (1969): (1) Shoot a small object A into the black hole from outside with energy-at-infinity E_A, charge e_A, and axial component of angular momentum L_{zA}. (2) When the object is deep down near the horizon, let it explode into two parts, B and C, each of which subsequently moves along a new test-particle trajectory, with new constants of the motion e_B and e_C, E_B and E_C, L_{zB} and L_{zC}. (3) So design the explosion that object B falls down the hole and gets captured, but object C escapes back to radial infinity. What will be the changes in mass, charge, and angular momentum of the black hole? According to the first law of black-hole dynamics,

$$\Delta M = \begin{pmatrix} \text{total energy that distant observers see} \\ \text{fall inward past themselves minus} \\ \text{total energy that they see reemerge} \end{pmatrix}$$
$$= E_A - E_C.$$

Similarly, $\Delta Q = e_A - e_C$ and $\Delta S = L_{zA} - L_{zC}$. Not unexpectedly, these changes can be written more simply in terms of the constants of motion for object B, which went down the hole. View the explosion "$A \longrightarrow B + C$" in a local Lorentz frame down near the hole, which is centered on the explosive event. As viewed in that frame, the explosion must satisfy the special relativistic laws of physics (equivalence principle!). In particular, it must obey charge conservation

$$e_A = e_B + e_C \tag{33.50a}$$

and conservation of total 4-momentum

$$(\boldsymbol{p}_A)_{\text{immediately before explosion}} = (\boldsymbol{p}_B + \boldsymbol{p}_C)_{\text{immediately after explosion}}.$$

Moreover, conservation of 4-momentum $\boldsymbol{p}$ and charge e implies also conservation of generalized momentum $\boldsymbol{\pi} \equiv \boldsymbol{p} - e\boldsymbol{A}$,

$$\boldsymbol{\pi}_A = \boldsymbol{p}_A - e_A\boldsymbol{A} = \boldsymbol{p}_B + \boldsymbol{p}_C - (e_B + e_C)\boldsymbol{A} = \boldsymbol{\pi}_B + \boldsymbol{\pi}_C;$$

and hence also conservation of the components of generalized momentum along the vectors $\partial/\partial t$ and $\partial/\partial\phi$,

$$E_A \equiv -\pi_{tA} = -\pi_{tB} - \pi_{tC} = E_B + E_C, \tag{33.50b}$$

$$L_{zA} \equiv \pi_{\phi B} + \pi_{\phi C} = L_{zB} + L_{zC}, \tag{33.50c}$$

(conservation of "energy-at-infinity" and "axial component of angular momentum" in explosion). Combining these conservation laws with the expressions

$$\Delta M = E_A - E_C, \qquad \Delta Q = e_A - e_C, \qquad \text{and } \Delta S = L_{zA} - L_{zC},$$

one obtains

$$\Delta M = E_B, \qquad \Delta Q = e_B, \qquad \Delta S = L_{zB}. \tag{33.51}$$

This result restated in words: the changes in mass, charge, and angular momentum are equal to the "energy-at-infinity," charge, and "axial component of angular momentum" that object B carries inward across the horizon, *even though B* may have ended up on a test-particle orbit that does not extend back to radial infinity!

Changes in M, Q, S for any nonradiative black-hole process

Straightforward extensions of the above thought experiment produce this generalization: *In any complicated black-hole process that involves infalling, colliding, and exploding pieces of matter that emit negligible gravitational radiation, the total changes in mass, charge, and angular momentum of the black hole are*

$$\Delta M = \begin{pmatrix} \text{sum of values of energy-at-infinity, } E, \\ \text{for all objects which cross the horizon—with} \\ E \text{ evaluated for each object at event of crossing} \end{pmatrix}, \tag{33.52a}$$

$$\Delta Q = \begin{pmatrix} \text{similar sum, of charges, } e, \text{ for} \\ \text{all objects crossing horizon} \end{pmatrix}, \tag{33.52b}$$

$$\Delta S = \begin{pmatrix} \text{similar sum of axial components of angular} \\ \text{momentum, } L_z, \text{ for all objects crossing horizon} \end{pmatrix}. \tag{33.52c}$$

Extraction of energy from a black hole by processes in the ergosphere

This result is not at all surprising. It is precisely what one might expect from the most naive of viewpoints. Not so expected, however, is the following consequence [Penrose (1969)]: *By injecting matter into a black hole in a carefully chosen way, one can decrease the total mass-energy of the black hole—i.e., one can extract energy from the hole.*

For *uncharged* infalling objects, the key to energy extraction is the ergosphere [hence its name, coined by Ruffini and Wheeler (1971a) from the Greek word "ϵργον" for "work"]. Outside the ergosphere, the Killing vector $\boldsymbol{\xi}_{(t)} \equiv \partial/\partial t$ is timelike, as is the 4-momentum $\boldsymbol{p}$ of every test particle; and therefore $E = -\boldsymbol{p} \cdot \boldsymbol{\xi}_{(t)}$ is necessarily positive. But inside the ergosphere (between the horizon and the static limit), $\boldsymbol{\xi}_{(t)}$

is spacelike, so for certain choices of timelike momentum vector (certain orbits of uncharged test particles), $E = -\boldsymbol{p}\cdot\boldsymbol{\xi}_{(t)}$ is negative, whereas for others it is positive. The orbits of negative E are confined entirely to the ergosphere. Thus, to inject an uncharged object with negative E into the black hole—and thereby to extract energy from the hole—one must always change its E from positive to negative and therefore also change its orbit, after it penetrates into the ergosphere. Of course, this is not difficult in principle—and perhaps not even in practice; see Figure 33.2.

For a charged object, electromagnetic forces alter the region where there exist orbits of negative energy-at-infinity. If the charges of object and hole have opposite sign, then the hole's electromagnetic field pulls inward on the object, giving it more kinetic energy when near the hole than one would otherwise expect. Thus, $-\boldsymbol{p}\cdot\boldsymbol{\xi}_{(t)}$ becomes an overestimate of E,

$$E = -(\boldsymbol{p} - e\boldsymbol{A})\cdot\boldsymbol{\xi}_{(t)} = -\boldsymbol{p}\cdot\boldsymbol{\xi}_{(t)} + \underbrace{eQr/\rho^2}_{[<0 \text{ if } eQ<0]}; \tag{33.53}$$

and orbits with $E < 0$ exist in a region somewhat larger than the ergosphere. If, on the other hand, e and Q have the same sign, then orbits with $E < 0$ are confined to a region smaller than the ergosphere. For given values e, Q, and rest mass μ, the region where there exist orbits with $E < 0$ is called the *"effective ergosphere."*

The "effective ergosphere" for charged-particle processes

EXERCISE

Exercise 33.10. ANGULAR MOMENTUM VECTOR FOR INFALLING PARTICLE

Derive equations (33.49d,e,f) for the components L_x and L_y of the orbital angular momentum of a particle falling into a black hole. Assume negligible initial speed, $E^2 - \mu^2 \approx 0$.

---

§33.8. REVERSIBLE AND IRREVERSIBLE TRANSFORMATIONS [Christodoulou (1970), Christodoulou and Ruffini (1971)]

Take a black hole of given mass M, charge Q, and angular momentum S. By injection of small objects, make a variety of changes in M, Q, and S. Can one pick an arbitrary desired change, ΔM, ΔQ, and ΔS, and devise a process that achieves it? Or are there limitations?

The second law of black-hole dynamics (nondecreasing surface area of black hole; Box 33.4; proof in §34.5 of next chapter) provides a strict limitation.

Then can all values within that limitation be achieved—and can that limitation be discovered by a direct examination of test-particle orbits?

The answer is yes; and, in fact, the limitation was discovered by Christodoulou (1970) and Christodoulou and Ruffini (1971) from an examination of test-particle orbits, independently of and simultaneously with Hawking's (1971) discovery of the second law of black-hole dynamics.

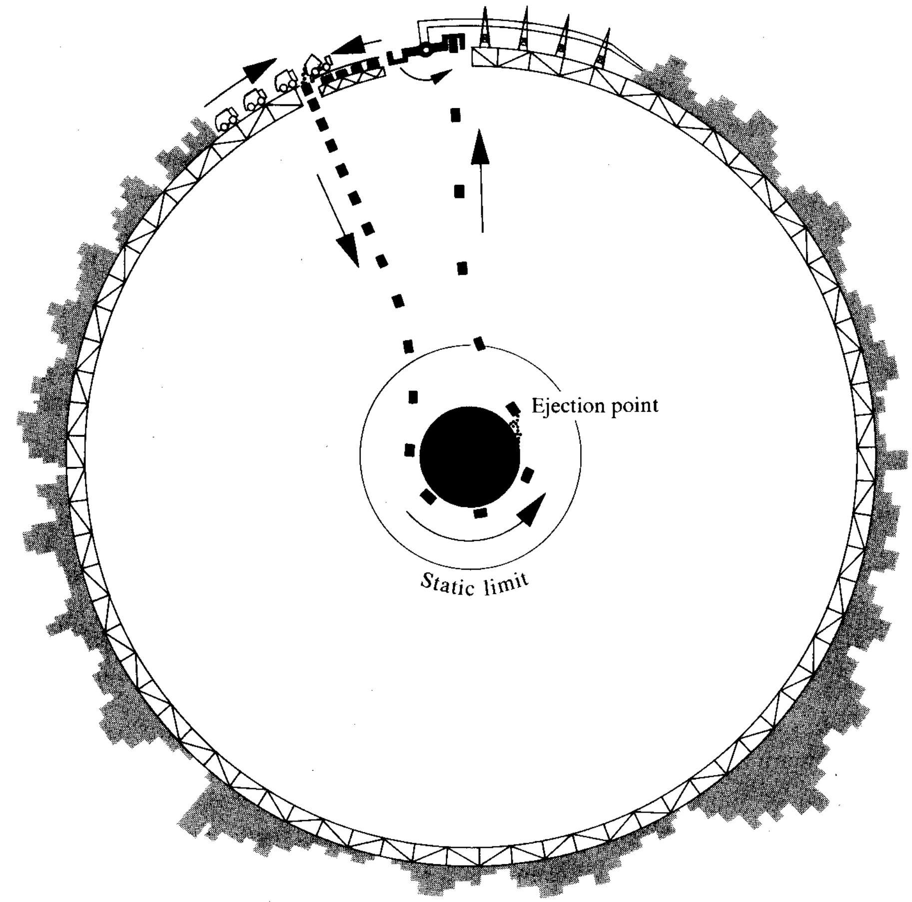


Figure 33.2.
An advanced civilization has constructed a rigid framework around a black hole, and has built a huge city on that framework. Each day trucks carry one million tons of garbage out of the city to the garbage dump. At the dump the garbage is shoveled into shuttle vehicles which are then, one after another, dropped toward the center of the black hole. Dragging of inertial frames whips each shuttle vehicle into a circling, inward-spiraling orbit near the horizon. When it reaches a certain "ejection point," the vehicle ejects its load of garbage into an orbit of negative energy-at-infinity, $E_{\text{garbage}} < 0$. As the garbage flies down the hole, changing the hole's total mass-energy by $\Delta M = E_{\text{garbage ejected}} < 0$, the shuttle vehicle recoils from the ejection and goes flying back out with more energy-at-infinity than it took down

$$E_{\text{vehicle out}} = E_{\text{vehicle + garbage down}} - E_{\text{garbage ejected}}$$
$$> E_{\text{vehicle + garbage down}}.$$

The vehicle deposits its huge kinetic energy in a giant flywheel adjacent to the garbage dump; and the flywheel turns a generator, producing electricity for the city, while the shuttle vehicle goes back for another load of garbage. The total electrical energy generated with each round trip of the shuttle vehicle is

$$\begin{aligned}(\text{Energy per trip}) &= E_{\text{vehicle out}} - (\text{rest mass of vehicle})\\ &= (E_{\text{vehicle + garbage down}} - E_{\text{garbage ejected}}) - (\text{rest mass of vehicle})\\ &= (\text{rest mass of vehicle} + \text{rest mass of garbage} - \Delta M) - (\text{rest mass of vehicle})\\ &= (\text{rest mass of garbage}) + (\text{amount, } -\Delta M, \text{ by which hole's mass decreases}).\end{aligned}$$

Thus, not only can the inhabitants of the city use the black hole to convert the entire rest mass of their garbage into kinetic energy of the vehicle, and thence into electrical power, but they can also convert some of the mass of the black hole into electrical power!

To derive the limitation of nondecreasing surface area from properties of test-particle orbits, one must examine what values of energy-at-infinity, E, are allowed at a given location (r, θ) outside a black hole. Equations (33.32a,b), when combined, yield the value of E in terms of a test particle's location (r, θ), rest mass μ, charge e, axial component of angular momentum L_z, and momenta $p^r = dr/d\lambda, p^\theta = d\theta/d\lambda$ in the r and θ directions:

Properties of test-particle orbits:

$$\alpha E^2 - 2\beta E + \gamma = 0; \qquad E = \frac{\beta + \sqrt{\beta^2 - \alpha\gamma}}{\alpha}, \tag{33.54a}$$

(1) E as function of μ, e, L_z, r, θ, p^r

where

$$\alpha = (r^2 + a^2)^2 - \Delta a^2 \sin^2\theta > 0 \text{ everywhere outside horizon,} \tag{33.54b}$$

$$\beta = (L_z a + eQr)(r^2 + a^2) - L_z a\,\Delta, \tag{33.54c}$$

$$\gamma = (L_z a + eQr)^2 - \Delta(L_z/\sin\theta)^2 - \mu^2\,\Delta\rho^2 - \rho^4[(p^r)^2 + \Delta(p^\theta)^2]. \tag{33.54d}$$

(One must take the positive square root, $+\sqrt{\beta^2 - \alpha\gamma}$, rather than the negative square root; positive square root corresponds to 4-momentum pointing toward future; while negative square root corresponds to past-pointing 4-momentum; see Figure 33.3.)

Several features of the energy equation (33.54) are noteworthy. (1) For orbits in the equatorial "plane," $\theta = \pi/2$ and $p^\theta \equiv 0$, the energy equation yields an effective potential for radial motion (Box 33.5). (2) Orbits of negative E must have $\beta < 0$ and $\gamma > 0$—which can be achieved only if $L_z a < 0$ and/or $eQ < 0$. Thus, *one cannot decrease the mass of a black hole without simultaneously decreasing the magnitude of its charge or angular momentum or both.* (3) For an orbit at given (r, θ), with given e and L_z, E is a minimum if $p^r = p^\theta = \mu = 0$. *Put differently, the rest mass and the r- and θ-components of momentum always contribute positively to E.*

(2) effective potential

(3) negative E requires $L_z a < 0$ and/or $eQ < 0$

By injecting an object into a black hole, produce small changes

$$\delta M = E, \qquad \delta Q = e, \qquad \delta S = L_z,$$

in its mass, charge, and angular momentum. For given changes in Q and S, what range of changes in M is possible? Clearly δM can be made as large as one wishes by making the rest mass μ sufficiently large. But there will be a lower limit on δM. That limit corresponds to the minimum value of E for given e and L_z. The orbit of minimum E crosses the horizon (otherwise no changes in M, Q, S would occur!), so one can evaluate E there. At the horizon, as anywhere, a minimum for E is achieved if $\mu = p^r = p^\theta = 0$. Inserting these values and $r = r_+$ (so $\Delta = 0$) into equations (33.54), one finds

Changes in black-hole properties due to injection of particles:

$$E_{\min} = \frac{L_z a + eQr_+}{r_+{}^2 + a^2}, \tag{33.55}$$

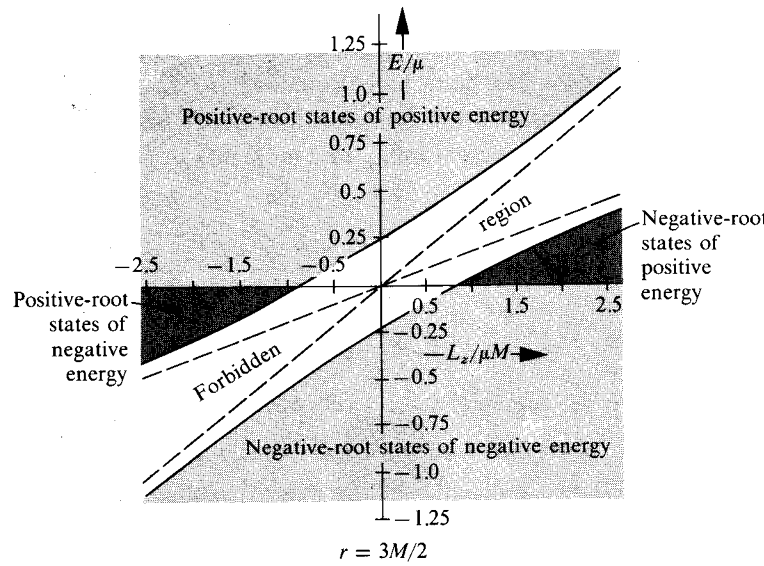


Figure 33.3.
Energy-at-infinity E allowed for a particle of angular momentum L_z and rest mass μ, which is (1) in the "equatorial plane" $\theta = \pi/2$, (2) at radius $r = 3M/2$, (3) of an uncharged ($Q = 0$) extreme-Kerr ($S = M^2$) black hole. E is here plotted against L_z. "Seas" of "positive and negative root" states are shown. The positive root states have energies-at-infinity given by equations (33.54)

$$E = \frac{\beta + \sqrt{\beta^2 - \alpha\gamma}}{\alpha}$$

and have 4-momentum vectors pointing into the future light cone. The negative root states (states of Dirac's "negative energy sea") have energies at infinity given by

$$E = \frac{\beta - \sqrt{\beta^2 - \alpha\gamma}}{\alpha},$$

and have 4-momentum vectors pointing into the past light cone. In the gap between the "seas" no orbits exist (forbidden region). The gap vanishes at the horizon $r = M$ (infinite redshift of *local* energy gap, 2μ, gives zero gap in energy-at-infinity). [Figure adapted from Christodoulou (1971).]

corresponding to changes in the black-hole properties of

(1) limit on δM for given δQ and δS

$$\delta M \geq \frac{a\,\delta S + r_+ Q\,\delta Q}{{r_+}^2 + a^2} \quad \begin{pmatrix}\text{absolute minimum value of}\\ \delta M \text{ for given } \delta S \text{ and } \delta Q\end{pmatrix}. \tag{33.56}$$

Notice an important consequence [Bardeen (1970a)]: if the black hole is initially of the "extreme Kerr-Newman" variety, with $M^2 = a^2 + Q^2$, so that one might fear a change which makes $M^2 < a^2 + Q^2$ and thereby destroys the horizon, one's fears are unfounded. Equation (33.56) then demands (since $r_+ = M$ and $S = Ma$)

$$M\,\delta M \geq a\,\delta a + Q\,\delta Q;$$

(2) preservation of the horizon

so M^2 remains greater than or equal to $a^2 + Q^2$, and the horizon is preserved.

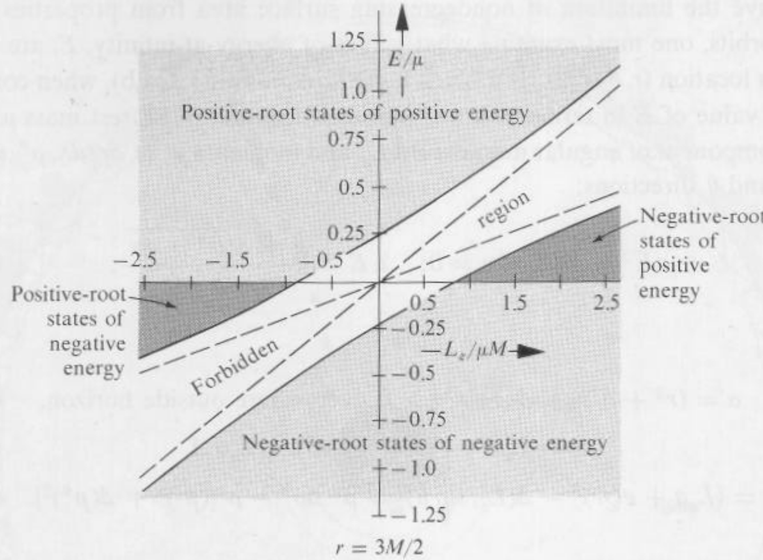


Figure 33.3.
Energy-at-infinity E allowed for a particle of angular momentum L_z and rest mass μ, which is (1) in the "equatorial plane" $\theta = \pi/2$, (2) at radius $r = 3M/2$, (3) of an uncharged ($Q = 0$) extreme-Kerr ($S = M^2$) black hole. E is here plotted against L_z. "Seas" of "positive and negative root" states are shown. The positive root states have energies-at-infinity given by equations (33.54)

$$E = \frac{\beta + \sqrt{\beta^2 - \alpha\gamma}}{\alpha}$$

and have 4-momentum vectors pointing into the future light cone. The negative root states (states of Dirac's "negative energy sea") have energies at infinity given by

$$E = \frac{\beta - \sqrt{\beta^2 - \alpha\gamma}}{\alpha},$$

and have 4-momentum vectors pointing into the past light cone. In the gap between the "seas" no orbits exist (forbidden region). The gap vanishes at the horizon $r = M$ (infinite redshift of *local* energy gap, 2μ, gives zero gap in energy-at-infinity). [Figure adapted from Christodoulou (1971).]

corresponding to changes in the black-hole properties of

(1) limit on δM for given δQ and δS

$$\delta M \geq \frac{a\,\delta S + r_+ Q\,\delta Q}{{r_+}^2 + a^2} \quad \begin{pmatrix}\text{absolute minimum value of}\\ \delta M \text{ for given } \delta S \text{ and } \delta Q\end{pmatrix}. \tag{33.56}$$

Notice an important consequence [Bardeen (1970a)]: if the black hole is initially of the "extreme Kerr-Newman" variety, with $M^2 = a^2 + Q^2$, so that one might fear a change which makes $M^2 < a^2 + Q^2$ and thereby destroys the horizon, one's fears are unfounded. Equation (33.56) then demands (since $r_+ = M$ and $S = Ma$)

$$M\,\delta M \geq a\,\delta a + Q\,\delta Q;$$

(2) preservation of the horizon

so M^2 remains greater than or equal to $a^2 + Q^2$, and the horizon is preserved.

Box 33.5 ORBITS OF TEST PARTICLE IN "EQUATORIAL PLANE" OF KERR-NEWMAN BLACK HOLE

Radial motion is governed by energy equation (33.54) with $\theta = p^\theta = 0$:

$$\alpha E^2 - 2\beta E + \gamma_0 - r^4(p^r)^2 = 0; \qquad E = \frac{\beta + \sqrt{\beta^2 - \alpha\gamma_0 + \alpha r^4(p^r)^2}}{\alpha}; \tag{1}$$

α, β, γ_0 are functions of r and of constants of motion,

$$\alpha = (r^2 + a^2)^2 - \Delta a^2 > 0, \tag{2a}$$

$$\beta = (L_z a + eQr)(r^2 + a^2) - L_z a\Delta, \tag{2b}$$

$$\gamma_0 = (L_z a + eQr)^2 - \Delta L_z{}^2 - \mu^2 r^2 \Delta; \tag{2c}$$

$p^r =$ (radial momentum) is

$$p^r = dr/d\lambda. \tag{3}$$

Thus, equation (1) is an ordinary differential equation for $dr/d\lambda$.

Qualitative features of the radial motion can be read off an effective-potential diagram. The effective potential $V(r)$ is the minimum allowed value of E at radius r:

$$V(r) = \frac{\beta + \sqrt{\beta^2 - \alpha\gamma_0}}{\alpha}.$$

As in the Schwarzschild case (Figure 25.2), the allowed regions for a particle of energy-at-infinity E are the regions with $V(r) \leq E$; and the turning points ($p^r = dr/d\lambda = 0$) occur where $V(r) = E$.

Stable circular orbits occur at the minima of $V(r)$. By examining $V(r)$ closely, one finds that for uncharged black holes the innermost stable circular orbit (most tightly bound orbit) has the characteristics here tabulated [table adapted from Ruffini and Wheeler (1971b)].

Charac-teristic	Newtonian (Figure 25.2)	Schwarzschild ($a = Q = 0$) (Figure 25.2)	Extreme Kerr ($a^2 = M^2$, $Q = 0$) (see figure) [Bardeen (1970a)] if $L_z a > 0$	if $L_z a < 0$
r/M	0	6	1	9
E/μ	$-\infty$	$2\sqrt{2}/3$	$1/\sqrt{3}$	$5/(3\sqrt{3})$
$(\mu - E)/\mu =$ "fractional binding"	$+\infty$	0.0572	0.4226	0.0377
$\lvert L_z\rvert/\mu M$	0	$2\sqrt{3}$	$2/\sqrt{3}$	$22/(3\sqrt{3})$

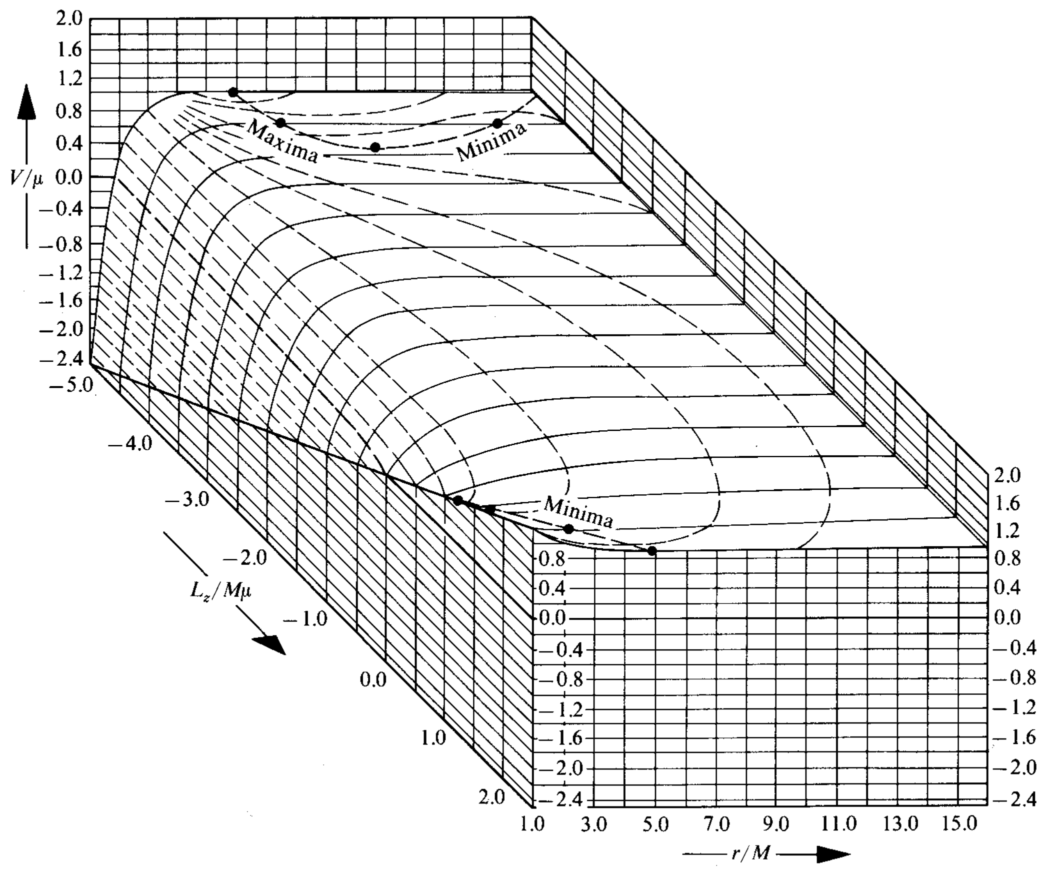
Box 33.5 (*continued*)

For a charged extreme Kerr-Newman black hole ($M^2 = Q^2 + a^2$, $Q \neq 0$ and $a \neq 0$) stable circular orbits with 100 per cent binding ($E = 0$) are achieved in the limit

$$\frac{e}{\mu} \longrightarrow -\infty, \qquad \frac{Q}{M} \longrightarrow 0 \text{ (so } a \longrightarrow M), \qquad \text{and } \left(\frac{e}{\mu}\right)\cdot\left(\frac{Q}{M}\right) \longrightarrow -\infty.$$

[Christodoulou and Ruffini (1971)].

The effective potential for an uncharged, extreme Kerr black hole ($a = M$) is shown in the figure [figure adapted from Ruffini and Wheeler (1971b)]. For detailed diagrams of orbits in the equatorial plane, see de Felice (1968). For many interesting properties of orbits that are not confined to the equatorial plane, see Wilkins (1972).



The general limit (33.56) on the change in mass can be rewritten in an alternative form [Christodoulou (1970), Christodoulou and Ruffini (1971)]:

$$\delta M_{\mathrm{ir}} \geq 0, \tag{33.57}$$

where

(3) irreducible mass

$$M_{\mathrm{ir}} \equiv \frac{1}{2}\sqrt{r_+{}^2 + a^2} = \frac{1}{2}[(M + \sqrt{M^2 - Q^2 - a^2})^2 + a^2]^{1/2} \tag{33.58}$$

is the "irreducible mass" of the black hole. *Equation (33.57) states that no black-hole transformation produced by the injection of small lumps of matter can ever reduce the irreducible mass of a black hole.* This result is actually a special case of the second law of black-hole dynamics, since the surface area of a black hole is

$$A = 16\pi M_{\mathrm{ir}}{}^2 \tag{33.59}$$

(Exercise 33.12).

Equation (33.58) can be combined with $a = S/M$ and inverted to yield

$$M^2 = \Big(\underbrace{M_{\mathrm{ir}}}_{\text{irreducible contribution to mass}} + \underbrace{\frac{Q^2}{4\,M_{\mathrm{ir}}}}_{\text{electromagnetic contribution to mass}}\Big)^2 + \underbrace{\frac{S^2}{4\,M_{\mathrm{ir}}{}^2}}_{\text{rotational contribution to mass}}. \tag{33.60}$$

(4) reversible and irreversible transformations

A black-hole transformation that holds fixed the irreducible mass is *reversible;* one that increases it is *irreversible.* The derivation of equation (33.56) revealed that the only injection processes that actually achieve the minimum possible value for δM (and thus make $\delta M_{\mathrm{ir}} = 0$) are those with $\mu = p^r = p^\theta = 0$ at the horizon, $r = r_+$. Restated in words: To produce a reversible transformation by injecting an object into a black hole, one must (1) give the object a rest mass μ extremely small compared to its charge ϵ or axial component of angular momentum L_z,

$$\mu/e \ll 1 \text{ and/or } \mu^2/L_z{}^2 \ll 1;$$

and (2) set the object down "extremely gently" ($p^r = p^\theta = 0$), extremely close to the horizon ($r = r_+$). This does not sound too difficult until one recalls that objects with $p^r = p^\theta = 0$ at the horizon must be moving outward with the speed of light, and that the nearer one approaches the horizon as one sets down the object, the greater one's danger of "slipping" and getting swallowed!

Clearly, any actual injection process will depart somewhat from irreversibility. Reversibility is an idealized limit, approachable but not attainable.

EXERCISES

Exercise 33.11. IRREDUCIBLE MASS IS IRREDUCIBLE

Show that condition (33.56) is equivalent to $\delta M_{\text{ir}} \geq 0$.

Exercise 33.12. SURFACE AREA OF A BLACK HOLE

Show that the surface area of the horizon of the Kerr-Newman geometry [area of surface $r = r_+$ and $t = \text{const}$ (Boyer-Lindquist coordinates) or $\tilde{V} = \text{const}$ (Kerr coordinates)] is $16\pi M_{\text{ir}}^2$.

Exercise 33.13. ANGULAR VELOCITY OF A BLACK HOLE

A general theorem [Hartle (1970) for relativistic case; Ostriker and Gunn (1969) for nonrelativistic case] says that, if one injects angular momentum into a rotating star while holding fixed all other contributions to its total mass-energy (contributions from entropy and from baryonic rest mass), then the injection produces a change in total mass-energy given by

$$\delta(\text{mass-energy}) = \begin{pmatrix}\text{angular velocity of star}\\ \text{at point of injection}\end{pmatrix}\delta(\text{angular momentum}). \tag{33.61}$$

By analogy, if one injects an angular momentum δS into a rotating black hole while holding fixed all other contributions to its total mass-energy (contributions from irreducible mass and from charge), one identifies the coefficient Ω_h in the equation

$$\delta M = \Omega_h \, \delta S$$

as the angular velocity of the hole:

$$\Omega_h = (\partial M/\partial S)_{Q,M_{\text{ir}}}. \tag{33.62}$$

(a) Show that the angular velocity of a black hole is equal to

$$\Omega_h = \frac{a}{r_+^2 + a^2}. \tag{33.63}$$

Notice that this is precisely the angular velocity of photons that live forever on the horizon [equation (33.42b); "barber-pole twist" of null generators of horizon].

(b) Show that *any* object falling into a black hole acquires an angular velocity (relative to Boyer-Lindquist coordinates) of $\Omega = d\phi/dt = \Omega_h$ in the late stages, as it approaches the horizon. (Recall that the horizon is a singularity of the Boyer-Lindquist coordinates. This is the reason that every object, regardless of its L_z, E, e, μ, $\mathcal{Q}$, can approach and does approach $\Omega = \Omega_h$.)

Exercise 33.14. SEPARATION OF VARIABLES FOR WAVE EQUATIONS

This chapter has studied extensively the motion of small objects in the external fields of black holes. Of almost equal importance, but not so well-understood yet because of its complexity, is the evolution of weak electromagnetic and gravitational perturbations ("waves") in the Kerr-Newman geometry. Just as one had no *à priori* reason to expect a "fourth constant" for test-particle motion in the Kerr-Newman geometry, so one had no reason to expect separability for Maxwell's equations, or for the wave equations describing gravitational perturbations—or even for the scalar wave equation $\Box\psi \equiv -\psi_{,\alpha}{}^{\alpha} = 0$. Thus it came as a great surprise when Carter (1968c) proved separability for the scalar wave equation, and later when Teukolsky (1972, 1973) separated both Maxwell's equations and the wave equations for gravitational perturbations.

Show that separation of variables for the scalar-wave equation in the (uncharged) Kerr geometry yields solutions of the form

$$\psi = (r^2 + a^2)^{-1/2} u_{\ell m}(r) S_{m\ell}(-i\omega a, \cos\theta) e^{i(m\phi - \omega t)}, \tag{33.64a}$$

where m and ℓ are integers with $0 \leq |m| \leq \ell$; $S_{m\ell}$ is a spheroidal harmonic [see Meixner and Schärfke (1954)]; and $u_{\ell m}$ satisfies the differential equation

$$-d^2u/dr^{*2} + Vu = 0. \tag{33.64b}$$

In order to put the equation in this form, define a Regge-Wheeler (1957) "tortoise"-type radial coordinate r^* by

$$dr^* = \Delta^{-1}(r^2 + a^2)\, dr, \tag{33.64c}$$

and find an effective potential $V(r^*)$ given by

$$\begin{aligned} V = -\left(\omega - \frac{ma}{r^2 + a^2}\right)^2 &+ [(m - \omega a)^2 + \mathcal{Q}](r^2 + a^2)^{-2}\, \Delta \\ &+ 2(Mr - a^2)(r^2 + a^2)^{-3}\, \Delta + 3a^2(r^2 + a^2)^{-4}\, \Delta^2. \end{aligned} \tag{33.64d}$$

In this radial equation $\mathcal{Q}$ is a constant (analog of Carter's constant for particle motion), given in terms of m and ℓ by

$$\mathcal{Q} \equiv \lambda_{m\ell} - m^2;\ \lambda_{m\ell} = \begin{bmatrix} \text{eigenfunction of spheroidal harmonic;} \\ \text{see Meixner and Schärfke (1954)} \end{bmatrix}. \tag{33.64e}$$

[These details of the separated solution were derived by Brill *et al.* (1972). For studies of the interaction between fields and Kerr black holes—studies performed using the above solution, and using analogous solutions to the electromagnetic and gravitational wave equations—see Bardeen, Press, and Teukolsky (1972), Misner (1972b), Teukolsky (1972), Ipser (1971), Press and Teukolsky (1973), and Chrzanowski and Misner (1973).]

2

CHAPTER 34

GLOBAL TECHNIQUES, HORIZONS, AND SINGULARITY THEOREMS

§34.1. GLOBAL TECHNIQUES VERSUS LOCAL TECHNIQUES

This chapter is entirely Track 2. §22.5 (geometric optics) and the Track-2 portions of Chapters 32 and 33 (collapse and black holes) are necessary preparation for it. It is not needed as preparation for any later chapter.

Local techniques of analyzing spacetime physics contrasted with global techniques

Until the 1960's, computations in gravitation theory used *local* techniques almost exclusively: the Einstein field equation describes how the stress-energy tensor $\boldsymbol{T}$ at a given event generates curvature $\boldsymbol{G}$ at that same event (*local* physics). When reduced to differential equations for the metric coefficients, $\boldsymbol{G} = 8\pi\boldsymbol{T}$ relates $g_{\alpha\beta}$, $\partial g_{\alpha\beta}/\partial x^{\mu}$, and $\partial^2 g_{\alpha\beta}/\partial x^{\mu}\partial x^{\nu}$ at each given event to $T_{\gamma\delta}$ at that same event (*local* equation). The solution of these differential equations is effected, on a computer or in any initial-value-type analysis, by integrating forward in time from event to event to event (*local* integration). The nongravitational laws of physics are obtained by invoking the equivalence principle in a *local* Lorentz frame at each individual event in spacetime. To build up an understanding of the global structure of spacetime, one performs *local* computations near each event, and then patches the local results together to form a global picture. Why this great reliance on local analyses? Because the laws of gravitation physics take on particularly simple forms when stated locally.

That gravitation physics is also subject to powerful and simple *global* laws, physicists did not realize until the mid 1960's. But since 1963, studies of black holes and of singularities have revealed global laws and global properties of spacetime that rival in their simplicity and elegance even the (local) equivalence principle. An example is the second law of black-hole dynamics: "In an isolated system, the sum of the surface areas of all black holes can never decrease." As a result, there has developed a powerful body of knowledge and techniques for analyzing *directly* the global properties of spacetime.

To give a full treatment of global techniques would require many chapters. Fortunately, a full treatment is being published, almost simultaneously with this book, by Hawking and Ellis (1973). Because Hawking and Ellis are much better qualified to write on this subject than are we (Misner, Thorne, and Wheeler), we have chosen to not write a "competitive" treatment. Instead, we give in this chapter only a brief taste of the subject—enough of a taste to make the reader acquainted with the types of techniques involved and several of the most important results, but not enough to give him a working knowledge of the subject. The topics we have chosen to treat are those that contact most closely the rest of this book: properties of "infinity" in an asymptotically flat spacetime (§34.2); causality and horizons (§§34.3 and 34.4); a proof of the second law of black-hole dynamics (§34.5); and theorems about the evolution of singularities in spacetime (§34.6). For greater detail on global techniques, one can consult not only the book of Hawking and Ellis (1973), but also review articles by Geroch (1971), by Penrose (1968a, 1972), and by Hawking (1973), the thesis of Godfrey (1970b), and the more specialized papers cited in the body of this chapter.

References on global techniques

§34.2. "INFINITY" IN ASYMPTOTICALLY FLAT SPACETIMES

Motivation for studying properties of spacetime near infinity

When performing calculations in asymptotically flat spacetime, one often must examine the asymptotic forms of fields (e.g., the metric, or the curvature tensor, or the electromagnetic field) "at infinity." For example, the mass and angular momentum of an isolated system are determined by the asymptotic form of the metric (Chapter 19). It is rarely sufficient to examine asymptotic forms near "spatial infinity." For example, if one wishes to learn how much mass was carried away by gravitational and electromagnetic waves during a supernova explosion, one must examine the asymptotic form of the metric not just at "spatial infinity," but at "future null infinity" (see Figure 34.1).

Penrose (1964, 1965a) has developed a powerful body of mathematical technique for studying asymptotic properties of spacetime near "infinity." The key to his technique is a "conformal transformation" of spacetime, which brings "infinity" in to a finite radius and thereby converts asymptotic calculations into calculations at "finite points." Penrose's technique also provides rigorous definitions of several types of "infinity" that one encounters in asymptotically flat spacetimes.

The details of Penrose's technique are not of importance to the rest of this chapter. However, this chapter will refer frequently to the various types of "infinity" defined by Penrose. In heuristic terms, they are as follows (see Figure 34.2a).

Specific regions of infinity: I^+, I^0, I^-, $\mathscr{I}^+$, $\mathscr{I}^-$

$I^+ \equiv$ "future timelike infinity":
the region $t \longrightarrow +\infty$ at finite radius r
(region toward which timelike lines extend).

$I^- \equiv$ "past timelike infinity":
the region $t \longrightarrow -\infty$ at finite radius r
(region from which timelike lines come).

$I^0 \equiv$ "spacelike infinity":
the region $r \longrightarrow \infty$ at finite time t
(region toward which spacelike slices extend).

$\mathscr{I}^+ \equiv$ "future null infinity":
the region $t + r \longrightarrow \infty$ at finite $t - r$
(region toward which outgoing null lines extend).

$\mathscr{I}^- \equiv$ "past null infinity":
the region $t - r \longrightarrow -\infty$ at finite $t + r$
(region from which ingoing null lines come).

Note: $\mathscr{I}$ is a script I, and is sometimes given the name "scri."

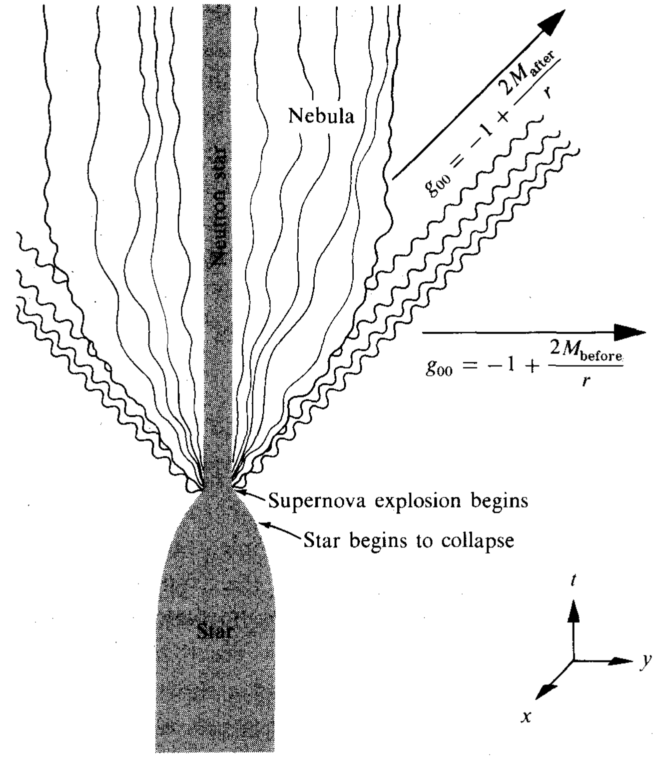


Figure 34.1.
Measurement of the mass-energy radiated as gravitational and electromagnetic waves by a supernova explosion in asymptotically flat spacetime. The mass-energy radiated equals the mass (M_{before}) of the presupernova star, minus the mass (M_{after}) of the neutron star and nebula after the explosion:

$$M_{\text{radiated}} = M_{\text{before}} - M_{\text{after}}.$$

To measure M_{before}, one can examine the asymptotic form (in suitable coordinates) of g_{00} at spatial infinity

$$g_{00} = -1 + \frac{2M_{\text{before}}}{r} + O\left(\frac{1}{r^2}\right) \qquad \text{as } r \longrightarrow \infty,\ t = \text{constant}.$$

But to measure M_{after} in the same way, one must wait, at any fixed r, until the radiation has flowed entirely past that point:

$$g_{00} = -1 + \frac{2M_{\text{after}}}{r} + O\left(\frac{1}{r^2}\right)$$

$$\text{as } r \longrightarrow \infty \text{ with } t - r = \begin{pmatrix}\text{constant value sufficiently large}\\ \text{to be inside the burst of waves}\end{pmatrix}.$$

Put differently, to measure M_{after} one must examine the asymptotic form of g_{00} *not* at "spatial infinity," but rather at "future null infinity."

Coordinate diagrams for exhibiting structure of infinity

It is often useful, in visualizing the asymptotic structure of spacetime, to introduce coordinates that attribute finite coordinate values to infinity. For example, in flat spacetime one can transform from the usual spherical coordinates t, r, θ, ϕ, with

$$ds^2 = -dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2), \tag{34.1}$$

to new spherical coordinates ψ, ξ, θ, ϕ, with

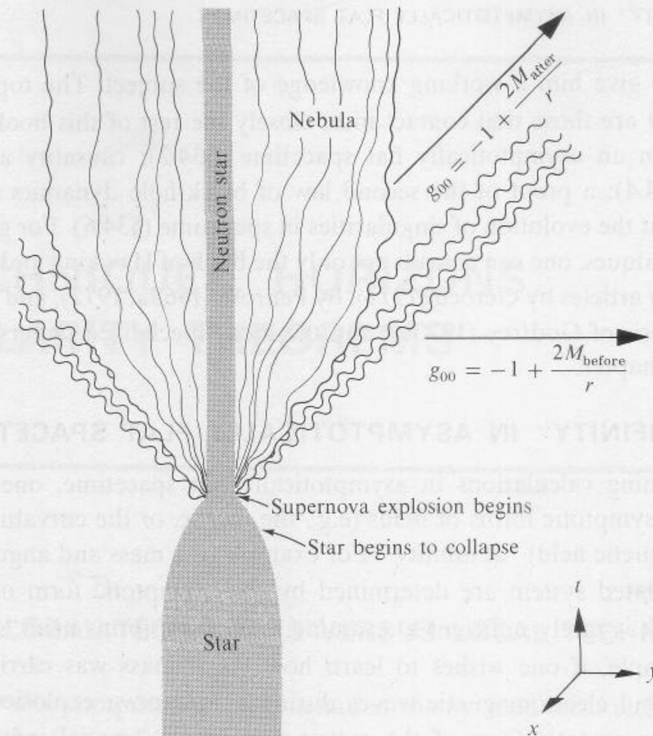


Figure 34.1.
Measurement of the mass-energy radiated as gravitational and electromagnetic waves by a supernova explosion in asymptotically flat spacetime. The mass-energy radiated equals the mass (M_{before}) of the presupernova star, minus the mass (M_{after}) of the neutron star and nebula after the explosion:

$$M_{\text{radiated}} = M_{\text{before}} - M_{\text{after}}.$$

To measure M_{before}, one can examine the asymptotic form (in suitable coordinates) of g_{00} at spatial infinity

$$g_{00} = -1 + \frac{2M_{\text{before}}}{r} + O\left(\frac{1}{r^2}\right) \quad \text{as } r \longrightarrow \infty,\ t = \text{constant}.$$

But to measure M_{after} in the same way, one must wait, at any fixed r, until the radiation has flowed entirely past that point:

$$g_{00} = -1 + \frac{2M_{\text{after}}}{r} + O\left(\frac{1}{r^2}\right)$$

$$\text{as } r \longrightarrow \infty \text{ with } t - r = \begin{pmatrix}\text{constant value sufficiently large}\\ \text{to be inside the burst of waves}\end{pmatrix}.$$

Put differently, to measure M_{after} one must examine the asymptotic form of g_{00} *not* at "spatial infinity," but rather at "future null infinity."

Coordinate diagrams for exhibiting structure of infinity

It is often useful, in visualizing the asymptotic structure of spacetime, to introduce coordinates that attribute finite coordinate values to infinity. For example, in flat spacetime one can transform from the usual spherical coordinates t, r, θ, ϕ, with

$$ds^2 = -dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2), \tag{34.1}$$

to new spherical coordinates ψ, ξ, θ, ϕ, with

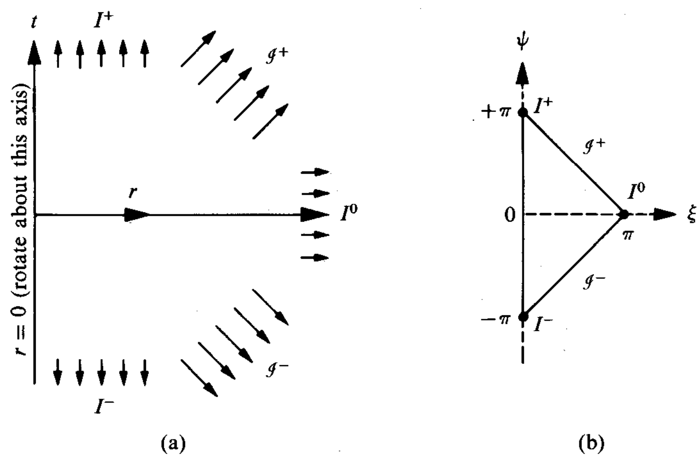


Figure 34.2.
Flat, "Minkowski" spacetime as depicted (a) in the usual spherical coordinates t, r, θ, ϕ of a global Lorentz frame, and (b) in the spherical coordinates of equations (34.2). The five regions of infinity—I^+, I^-, I^0, $\mathscr{I}^+$, $\mathscr{I}^-$—are shown in each coordinate diagram. In both coordinate systems, radial null lines make angles of 45° with the vertical axis, and nonradial null lines make angles less than 45° [see equations (34.1) and (34.2c)]. See exercise 34.1 for further detail.

$$t + r = \tan\frac{1}{2}(\psi + \xi), \tag{34.2a}$$

$$t - r = \tan\frac{1}{2}(\psi - \xi), \tag{34.2b}$$

$$ds^2 = \frac{-d\psi^2 + d\xi^2}{4\cos^2\frac{1}{2}(\psi + \xi)\cos^2\frac{1}{2}(\psi - \xi)} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{34.2c}$$

The resulting ψ, ξ coordinate diagram (Figure 34.2b) depicts $I^+, I^-, I^0, \mathscr{I}^+, \mathscr{I}^-$ more clearly than does the usual t, r, coordinate diagram.

As another example, replace the Kruskal-Szekeres coordinates v, u, θ, ϕ for Schwarzschild spacetime by new coordinates ψ, ξ, θ, ϕ:

$$v + u = \tan\frac{1}{2}(\psi + \xi), \tag{34.3a}$$

$$v - u = \tan\frac{1}{2}(\psi - \xi), \tag{34.3b}$$

$$(1 - r/2M)e^{r/2M} = v^2 - u^2 = \tan\frac{1}{2}(\psi + \xi)\tan\frac{1}{2}(\psi - \xi), \tag{34.3c}$$

$$ds^2 = \frac{32M^3}{r}\frac{e^{-r/2M}(-d\psi^2 + d\xi^2)}{4\cos^2\frac{1}{2}(\psi + \xi)\cos^2\frac{1}{2}(\psi - \xi)} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{34.3d}$$

The resulting coordinate diagram (Figure 34.3) depicts clearly the causal connections between the horizons, the singularities, and the various regions of infinity.

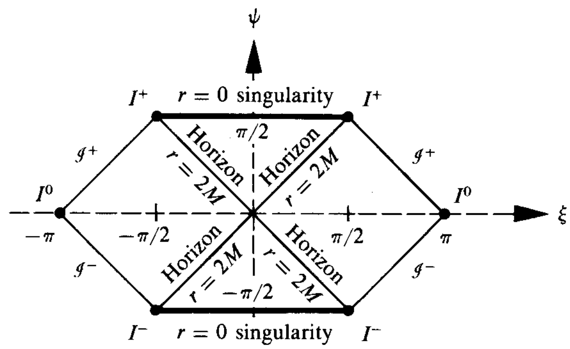


Figure 34.3.
Schwarzschild spacetime as depicted in the ψ, ξ, θ, ϕ coordinates of equations (34.3). This coordinate diagram should be compared with the Kruskal-Szekeres coordinate diagram (Figure 31.3). In both diagrams, radial null geodesics are 45° lines. Each of the asymptotically flat regions (one on each side of the "wormhole" of Figure 31.5a) has its own set of infinities I^+, I^-, I^0, $\mathscr{I}^+$, and $\mathscr{I}^-$. See exercise 34.2 for justification of this diagram.

EXERCISES

Exercise 34.1. FLAT SPACETIME IN ψ, ξ, θ, ϕ COORDINATES

(a) Derive equation (34.2c) from (34.1) and (34.2a,b).

(b) Show that the regions I^+, I^-, I^0, $\mathscr{I}^+$, and $\mathscr{I}^-$ of flat spacetime are located at

$$\begin{aligned}
&I^+\colon \psi = \pi, \xi = 0,\\
&I^-\colon \psi = -\pi, \xi = 0,\\
&I^0\colon \psi = 0, \xi = \pi,\\
&\mathscr{I}^+\colon \psi + \xi = \pi, \quad -\pi < \psi - \xi < \pi,\\
&\mathscr{I}^-\colon \psi - \xi = -\pi, \quad -\pi < \psi + \xi < \pi.
\end{aligned} \tag{34.4}$$

[see equations (34.2)]. These are the regions depicted in Figure 34.2.

(c) Show that in flat spacetime, in a ψ, ξ coordinate diagram (Figure 34.2), radial null lines make angles of 45° with the vertical axis, and nonradial null lines make angles of less than 45°.

Exercise 34.2. SCHWARZSCHILD SPACETIME IN ψ, ξ, θ, ϕ COORDINATES

(a) Derive equations (34.3c,d) from (34.3a,b) and the Kruskal-Szekeres equations (31.14).

(b) Use equations (34.3) to justify the precise form of the coordinate diagram in Figure 34.3.

Exercise 34.3. REISSNER-NORDSTROM SPACETIME

(a) Show that there exists a coordinate system in which the Reissner-Nordstrøm geometry with $0 < |Q| < M$ (exercises 31.8 and 32.1) has the form

$$\begin{aligned}
ds^2 &= F^2(-d\psi^2 + d\xi^2) + r^2(d\theta^2 + \sin^2\theta\, d\phi^2),\\
F &= F(\psi, \xi), \qquad r = r(\psi, \xi),
\end{aligned} \tag{34.5}$$

and in which the horizons and infinities are as shown in Figure 34.4. [*Note:* This is a very difficult exercise unless one has in hand the solution to exercise 31.8(d). For solution, see Carter (1966b).]

(b) Use Figure 34.4 to deduce that the Reissner-Nordstrøm geometry describes a "wormhole" or bridge, connecting two asymptotically flat spacetimes, which: (i) expands to a state of maximum circumference; (ii) recontracts toward a state of minimum circumference, and in the process disconnects its outer regions from the two I^0's (spatial infinity) and reconnects them to a pair of $r = 0$ singularities; (iii) bounces; (iv) reexpands, and in the process

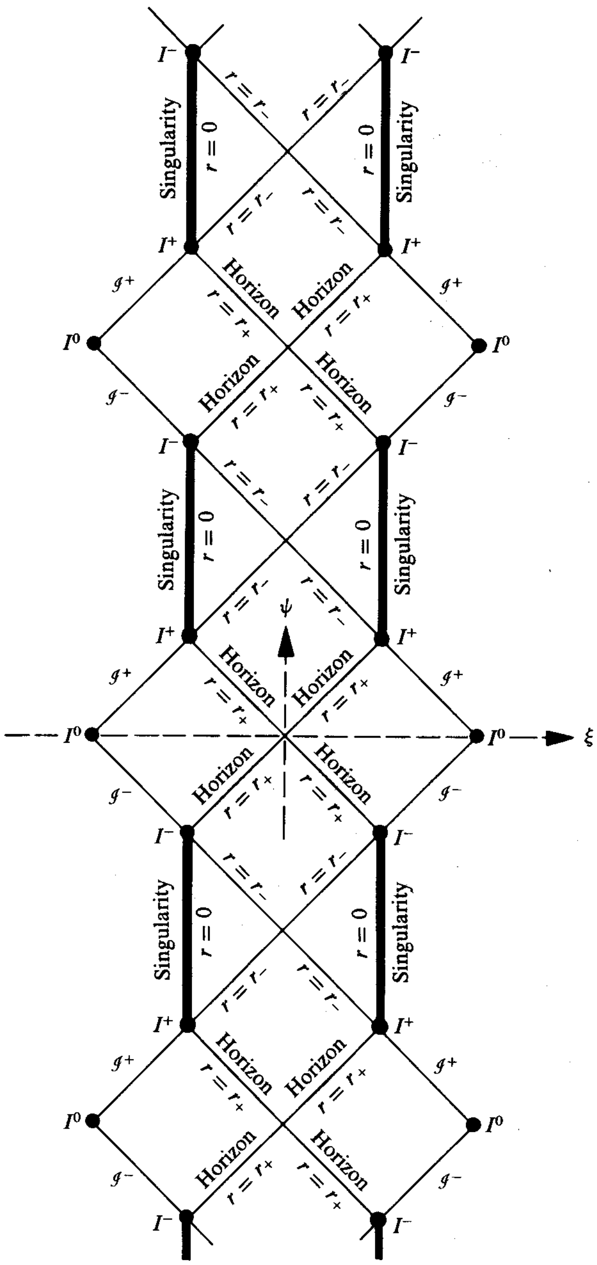


Figure 34.4.
Reissner-Nordstrøm spacetime

$$ds^2 = -\left(1 - \frac{2M}{r} + \frac{Q^2}{r^2}\right) dt^2 + \frac{dr^2}{1 - 2M/r + Q^2/r^2} + r^2(d\theta^2 + \sin^2\theta \, d\phi^2)$$

with $0 < |Q| < M$, as depicted in a new (ψ, ξ, θ, ϕ) coordinate system where the line element has the form

$$ds^2 = F^2(-d\psi^2 + d\xi^2) + r^2(d\theta^2 + \sin^2\theta \, d\phi^2).$$

(see exercise 34.3.) This coordinate diagram reveals the global structure of the geometry, including its singularities at $r = 0$, its horizons at

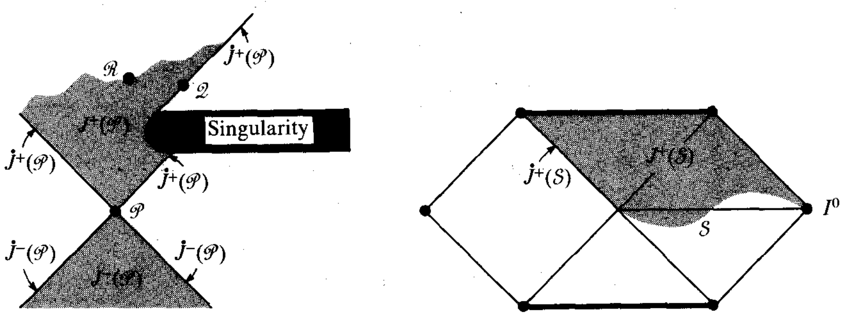
$$r = r_+ = M + \sqrt{M^2 - Q^2}$$

(which limit communication with $\mathscr{I}^+$ and $\mathscr{I}^-$), the null surfaces at

$$r = r_- = M - \sqrt{M^2 - Q^2}$$

(which limit communication with the singularities), and the various asymptotically flat infinities, I^+, I^-, I^0, $\mathscr{I}^+$, and $\mathscr{I}^-$. From this diagram one can read off the "causal structure" of the geometry—i.e., the abilities of various regions to communicate with each other. For detailed discussion of the geometry, see Graves and Brill (1960) and Carter (1966b). For discussions of collapsing charged stars, for which this geometry is the external gravitational field, see Novikov (1966a,b), de la Cruz and Israel (1967), and Bardeen (1968).

disconnects its "outer regions" from the two singularities and reconnects them to a pair of I^0's in two *new* asymptotically flat universes; (v) slows its expansion to a halt; (vi) recontracts toward a state of minimum circumference, and in the process disconnects its outer regions from the two I^0's and reconnects them to a new pair of $r = 0$ singularities; etc. *ad infinitum*.



Note: $\mathscr{P} \prec \mathscr{R}$, $\mathscr{P} \nprec \mathscr{Q}$,
$\mathscr{P} \ll \mathscr{R}$, $\mathscr{P} \not\ll \mathscr{Q}$,
but $\mathscr{Q} \in \dot{J}^+(\mathscr{P})$

Note: S is a spacelike slice which extends from I^0 in to $r = 2M$, but does not include $r = 2M$. $J^+(S)$ does not include the leftmost horizon; but $\dot{J}^+(S)$ is the leftmost horizon.

(a) (b)

Figure 34.5.
Spacetime diagrams illustrating various causal relationships. Diagram (a) is a hypothetical spacetime; diagram (b) is Schwarzschild spacetime (see Figure 34.3). In both diagrams, null lines have slopes of 45°.

§34.3. CAUSALITY AND HORIZONS

Restriction of discussion to asymptotically flat, time-oriented manifolds

Turn now to global* techniques for analyzing black holes. The goals of the discussion will be (1) to define the concept of horizon (this section), (2) to deduce global geometric properties of horizons (next section), and (3) to prove the second law of black-hole dynamics (following section). The entire discussion will be confined to spacetime manifolds that (1) contain at least one asymptotically flat region ("the external universe"; region "outside black holes"), and (2) are "time-oriented." By "*time-oriented*" one means that at each event in spacetime a distinct choice has been made as to which light cone is the future cone and which is the past, and moreover that this choice is continuous from event to event throughout spacetime.

The discussion begins with definitions of a variety of causal relationships between events and regions of spacetime (see Figure 34.5).

Definitions of several causality concepts

Definition: $\mathscr{P} \ll \mathscr{Q}$ or equivalently $\mathscr{Q} \gg \mathscr{P}$ ("the event $\mathscr{P}$ *precedes* the event $\mathscr{Q}$"; "the event $\mathscr{Q}$ *follows* the event $\mathscr{P}$") means that there exists at least one smooth, future-directed timelike curve that extends from $\mathscr{P}$ to $\mathscr{Q}$.

Definition: A *causal curve* $\mathscr{C}(\lambda)$ is any smooth curve that is nowhere spacelike—i.e., that is timelike or null or "zero" [$\mathscr{C}(\lambda) =$ some fixed $\mathscr{P}$, for all λ] or some admixture thereof.

Definition: $\mathscr{P} \prec \mathscr{Q}$ or equivalently $\mathscr{Q} \succ \mathscr{P}$ ("the event $\mathscr{P}$ *causally precedes* the event

*Global, but not fully global; the "universe" of §§34.3–34.5 is asymptotically flat; no account is taken here of possible closure or collapse of the universe or of their consequences.

$\mathscr{Q}$"; the event $\mathscr{Q}$ *causally follows* the event $\mathscr{P}$") means that there exists at least one future-directed causal curve that extends from $\mathscr{P}$ to $\mathscr{Q}$.

Definition: $J^-(\mathscr{P})$, called *the causal past of* $\mathscr{P}$, is the set of all events that causally precede $\mathscr{P}$—i.e., $J^-(\mathscr{P}) = \{\mathscr{Q} \,|\, \mathscr{Q} \prec \mathscr{P}\}$.

Definition: $J^+(\mathscr{P})$, called *the causal future of* $\mathscr{P}$, is the set of all events that causally follow $\mathscr{P}$—i.e., $J^+(\mathscr{P}) = \{\mathscr{Q} \,|\, \mathscr{Q} \succ \mathscr{P}\}$.

Definition: If $\mathcal{S}$ is a region of spacetime—e.g., a segment of a spacelike hypersurface—then $J^-(\mathcal{S})$ (*the causal past of* $\mathcal{S}$) is the set of all events that causally precede at least one event in $\mathcal{S}$—i.e.,

$$J^-(\mathcal{S}) = \{\mathscr{Q} \,|\, \mathscr{Q} \prec \mathscr{P} \text{ for at least one } \mathscr{P} \in \mathcal{S}\}.$$

Definition: Similarly, $J^+(\mathcal{S})$ (*the causal future of* $\mathcal{S}$) is the set of all events that causally follow at least one event in $\mathcal{S}$—i.e.,

$$J^+(\mathcal{S}) = \{\mathscr{Q} \,|\, \mathscr{Q} \succ \mathscr{P} \text{ for at least one } \mathscr{P} \in \mathcal{S}\}.$$

Definition: $\dot{J}^+(\mathcal{S})$ is the *boundary of* $J^+(\mathcal{S})$,
$\dot{J}^-(\mathcal{S})$ is the *boundary of* $J^-(\mathcal{S})$.

Definition: One defines *the future of* $\mathscr{P}$, $I^+(\mathscr{P})$; *the past of* $\mathscr{P}$, $I^-(\mathscr{P})$; *the future of* $\mathcal{S}$, $I^+(\mathcal{S})$; *the past of* $\mathcal{S}$, $I^-(\mathcal{S})$; *the boundary of the future of* $\mathcal{S}$, $\dot{I}^+(\mathcal{S})$; *and the boundary of the past of* $\mathcal{S}$, $\dot{I}^-(\mathcal{S})$ in precisely the same manner as above, except that the phrase "causally precede" is replaced by "precede," and "causally follow" is replaced by "follow." Example:

$$I^+(\mathcal{S}) = \{\mathscr{Q} \,|\, \mathscr{Q} \gg \mathscr{P} \text{ for at least one } \mathscr{P} \in \mathcal{S}\}.$$

Not all these definitions are needed in the following discussion; but the literature on global methods uses these concepts so extensively that the reader should be familiar with them.

Focus attention on a specific spacetime manifold, and in that manifold select out a specific asymptotically flat region. [In the external field of a star, there is but one asymptotically flat region. In the vacuum Schwarzschild geometry without source (Figure 34.3), there are two. In the Reissner-Nordstrøm geometry without source (Figure 34.4), there are infinitely many different asymptotically flat regions.] The selected asymptotically flat region ("external universe") has one future timelike infinity I^+, one past timelike infinity I^-, one spacelike infinity I^0, one future null infinity $\mathscr{I}^+$, and one past null infinity $\mathscr{I}^-$. It may also possess black holes, which form by stellar collapse, and which collide, coalesce, accrete matter, and generally wreak havoc in their immediate vicinities. The surfaces of all black holes ("future horizons") separate the external universe, which can send signals out to $\mathscr{I}^+$, from the black-hole interiors, which cannot. One thus has the definition:

Definition: surfaces of black holes; future horizons—$\dot{J}^-(\mathscr{I}^+)$

Definition: The *totality* (*or* "*union*") *of all future horizons* (surfaces of all black holes) is the region $\dot{J}^-(\mathscr{I}^+)$—i.e., it is the boundary of the domain $J^-(\mathscr{I}^+)$ that can send future-directed causal curves out to future null infinity.

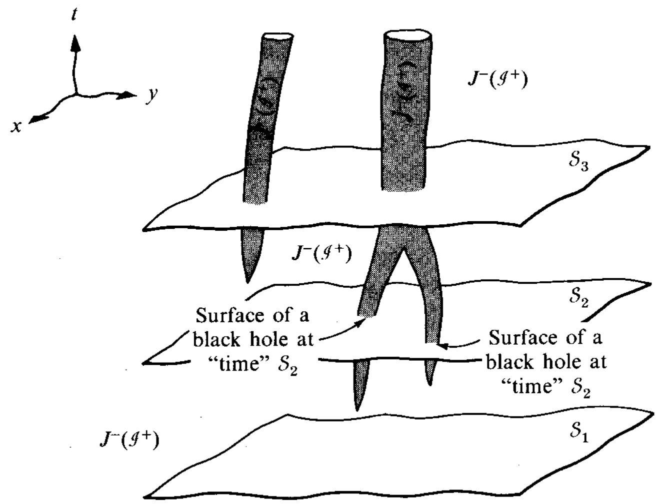


Figure 34.6.
Black holes in an asymptotically flat spacetime (schematic spacetime diagram). $J^-(\mathscr{I}^+)$ is the "external universe"—i.e., the region which can send causal curves to future null infinity. $\dot{J}^-(\mathscr{I}^+)$, the greyish region, is the boundary of the "external universe"—i.e., it is the union of all future horizons. At the "time" of spacelike slice S_1, there are no black holes in the universe. Between S_1 and S_2 two stars collapse to form black holes. The two closed 2-surfaces, in which S_2 intersects $\dot{J}^-(\mathscr{I}^+)$ are the horizons of those black holes at "time" S_2. Between S_2 and S_3, the two original black holes collide and coalesce, while a third black hole is being formed by stellar collapse.

[Similarly, one can define the totality of all past horizons to be $\dot{J}^+(\mathscr{I}^-)$. But past horizons are of little interest for astrophysics. Whereas gravitational collapse produces future horizons in a quite natural manner, past horizons must be primordial in origin—i.e., they must be postulated as initial structure in the origin of the universe [Novikov (1964), Ne'eman (1965)]. There is no good reason to believe that the universe began with or should have begun with such strange initial structure.]

Any given spacelike slice S through spacetime will intersect $\dot{J}^-(\mathscr{I}^+)$ in a number of disjoint, closed, two-dimensional surfaces. Each such 2-surface is the horizon of a single black hole at the "moment of time" S. See Figure 34.6.

§34.4. GLOBAL STRUCTURE OF HORIZONS

The union of all future horizons, $\dot{J}^-(\mathscr{I}^+)$, has an especially simple global geometric structure, as follows.

Penrose's theorem on the structure of $\dot{J}^-(\mathscr{I}^+)$ (future horizons)

THEOREM [Penrose (1968a)]: *$\dot{J}^-(\mathscr{I}^+)$ is generated by null geodesics that have no future end points.* Stated more precisely (see Figure 34.7):

(1) Definition: The "*generators*" of $\dot{J}^-(\mathscr{I}^+)$ are null geodesics which (at least for some finite lapse of affine parameter) lie in $\dot{J}^-(\mathscr{I}^+)$.

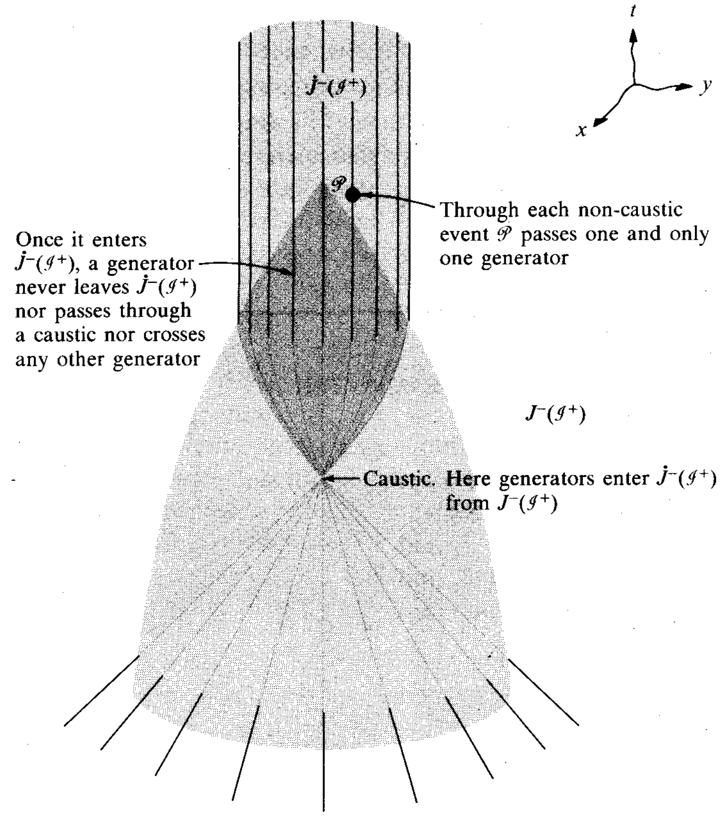


Figure 34.7.
The future horizon $\dot{J}^-(\mathscr{I}^+)$ produced by the spherical gravitational collapse of a star. This horizon illustrates the global geometric structure of $\dot{J}^-(\mathscr{I}^+)$ as spelled out in Penrose's theorem (§34.4 of text). In this special case, there is only one caustic. In general there will be many.

(2) Theorem: When followed into the past, a generator may (but does not have to!) leave $\dot{J}^-(\mathscr{I}^+)$. Each event at which a generator leaves is called a "*caustic*" of $\dot{J}^-(\mathscr{I}^+)$. When a generator leaves, it goes into $J^-(\mathscr{I}^+)$.

(3) Once a generator, being followed into the future, enters $\dot{J}^-(\mathscr{I}^+)$ from $J^-(\mathscr{I}^+)$ at a caustic, it can never thereafter leave $\dot{J}^-(\mathscr{I}^+)$, nor can it ever intersect another generator. [Generators can intersect only at the "caustics," where they enter $\dot{J}^-(\mathscr{I}^+)$.]

(4) Through each noncaustic event of $\dot{J}^-(\mathscr{I}^+)$ there passes one and (aside from normalization of affine parameter) only one generator.

This theorem is proved in Box 34.1.

For a Schwarzschild black hole, the generators of $\dot{J}^-(\mathscr{I}^+)$ are the world lines of radially outgoing photons at the gravitational radius [$r = 2M$, θ and ϕ constant,

$u = +v$; dotted line on horizon in Figure 32.1(c)]. For a Kerr-Newman black hole, the generators of $\dot{J}^-(\mathscr{I}^+)$ are the "barber-pole-twist" null geodesics of Box 33.2(F)—i.e., they are those members of the outgoing principal null congruence that lie on the horizon, $r = r_+$ (§33.6; exercise 33.9). But the theorem is more general. It refers to any black hole—dynamic or static; accreting matter, or coalescing with a neighboring black hole, or existing alone in isolation—in any time-oriented, asymptotically flat spacetime.

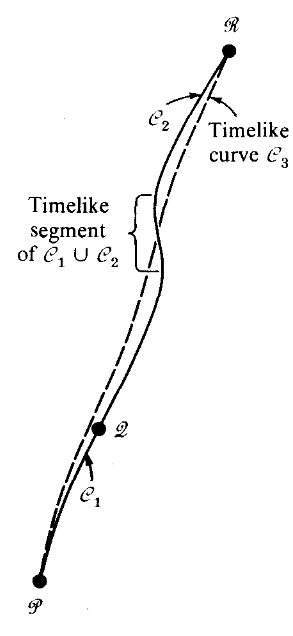
(*continued on page 931*)

Box 34.1 HORIZONS ARE GENERATED BY NONTERMINATING NULL GEODESICS (Penrose 1968a)

A. *Lemma:* If (1) $\mathcal{C}_1(\lambda)$ is a causal, future-directed curve from event $\mathscr{P}$ to event $\mathscr{Q}$, (2) $\mathcal{C}_2(\lambda)$ is a causal, future-directed curve from event $\mathscr{Q}$ to event $\mathscr{R}$, and (3) $\mathscr{P} \not\ll \mathscr{R}$ ($\mathscr{P}$ is not in the past of $\mathscr{R}$), then $\mathcal{C}_1$ and $\mathcal{C}_2$ are null geodesics, and their tangent vectors coincide (aside from normalization) at event $\mathscr{Q}$.

Proof:*

1. Suppose that $\mathcal{C}_1$ were not a null geodesic, or $\mathcal{C}_2$ were not a null geodesic, or both. Then somewhere along $\mathcal{C}_1 \cup \mathcal{C}_2$ there would be a timelike segment, or a nongeodesic null segment, or both.
   a. If $\mathcal{C}_1 \cup \mathcal{C}_2$ contained a timelike segment, then a slight deformation† of $\mathcal{C}_1 \cup \mathcal{C}_2$ would produce a smooth curve $\mathcal{C}_3$ from $\mathscr{P}$ to $\mathscr{R}$ which is everywhere timelike‡—contradicting the assumption $\mathscr{P} \not\ll \mathscr{R}$.

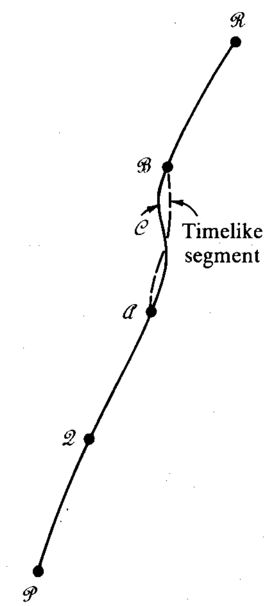


*The proof utilizes some elementary concepts of point-set topology; see, e.g., Wallace (1963) or Kelley (1955).

†One can always deform any curve in any spacetime manifold by a small amount in any direction one wishes, without running into singularities or into other boundaries of the manifold. This is possible because a manifold by definition is *open*. In physical terms, spacetime is open because each event in spacetime must possess a local Lorentz neighborhood which also lies in spacetime.

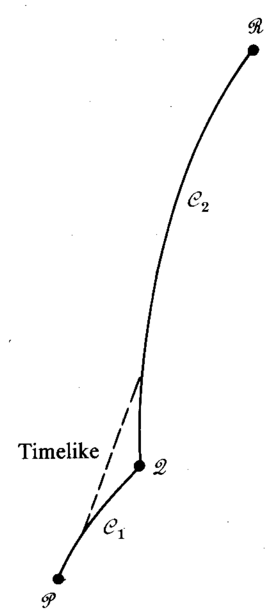
‡One can convince oneself of this, and of similar claims made later in the proof, by arguments using local Lorentz frames. In the literature on global geometry, claims such as this are rarely substantiated—though each author is always convinced that he could do so if forced to by a skeptic. Unfortunately, to substantiate such claims with rigorous arguments would lengthen and complicate the discussion enormously and would tend to obscure the simplicity of the underlying ideas.

b. If $\mathcal{C}_1 \cup \mathcal{C}_2$ contained a nongeodesic null segment $\mathcal{C}$ reaching from event $\mathcal{A}$ to event $\mathcal{B}$, then, when compared to neighboring curves between $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{C}$ would not have stationary length. This means that some curves from $\mathcal{A}$ to $\mathcal{B}$ would have larger squared length—i.e., would be spacelike—while others would have smaller squared length—i.e., would be timelike. Thus, a slight deformation of $\mathcal{C}$ would produce a timelike segment from $\mathcal{A}$ to $\mathcal{B}$. Then a further deformation of $\mathcal{C}_1 \cup \mathcal{C}_2$, as described in (a) above, would produce a smooth timelike curve from $\mathcal{P}$ to $\mathcal{R}$, contradicting $\mathcal{P} \nless \mathcal{R}$.



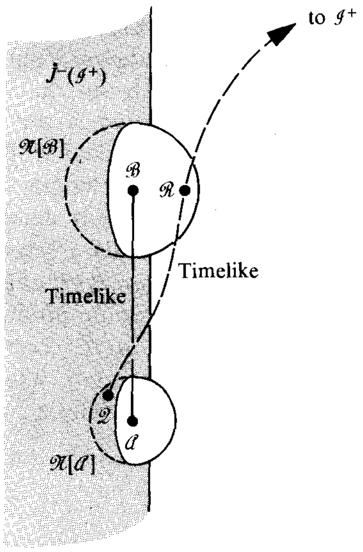
Thus, the supposition is wrong; i.e., both $\mathcal{C}_1$ and $\mathcal{C}_2$ must be null geodesics.

2. Suppose that the tangent vectors of $\mathcal{C}_1$ and $\mathcal{C}_2$ did not agree at their join point, $\mathcal{Q}$. Then one could "round off the corner" at $\mathcal{Q}$, producing a timelike segment there. One could then further deform $\mathcal{C}_1 \cup \mathcal{C}_2$ as in (1a) above, to produce a smooth timelike curve from $\mathcal{P}$ to $\mathcal{R}$—contradicting $\mathcal{P} \nless \mathcal{R}$. Thus, the supposition is wrong; i.e., the tangent vectors must agree at $\mathcal{Q}$. Q.E.D.



Box 34.1 ***(continued)***

B. *Lemma:* If $\mathscr{A} \in \dot{J}^-(\mathscr{I}^+)$ and $\mathscr{B} \in \dot{J}^-(\mathscr{I}^+)$, then $\mathscr{A} \not\ll \mathscr{B}$.

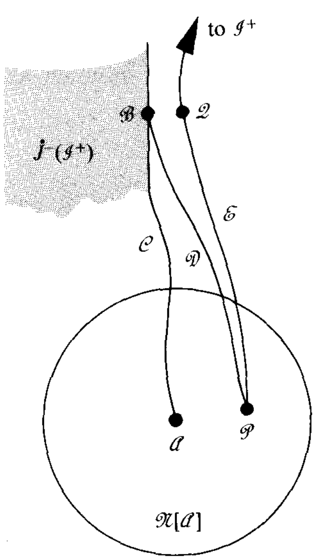


Proof: Assume $\mathscr{A} \ll \mathscr{B}$.

1. Then there exists a timelike curve from $\mathscr{A}$ to $\mathscr{B}$.
2. A slight deformation of that curve which keeps it still timelike will make it link an arbitrary event $\mathscr{Q}$ in some sufficiently small neighborhood $\mathscr{N}[\mathscr{A}]$ to an arbitrary event $\mathscr{R}$ in some small $\mathscr{N}[\mathscr{B}]$.
3. Pick $\mathscr{R}$ to lie in $J^-(\mathscr{I}^+)$. Then join the timelike curve from $\mathscr{Q}$ to $\mathscr{R}$ onto a causal curve from $\mathscr{R}$ to $\mathscr{I}^+$. The resulting curve, when smoothed in a neighborhood of the join, becomes a causal curve from any arbitrary $\mathscr{Q} \in \mathscr{N}[\mathscr{A}]$ to $\mathscr{I}^+$.
4. The existence of such curves implies that $\mathscr{N}[\mathscr{A}] \subset J^-(\mathscr{I}^+)$, and hence that $\mathscr{A} \notin \dot{J}^-(\mathscr{I}^+)$—in contradiction to the original hypotheses.

Conclusion: $\mathscr{A} \not\ll \mathscr{B}$. Q.E.D.

C. *Lemma:* Let $\mathscr{C}(\lambda)$ be a causal curve that intersects $\dot{J}^-(\mathscr{I}^+)$ at some event $\mathscr{B}$. Then when followed into the past from $\mathscr{B}$, $\mathscr{C}(\lambda)$ forever lies in $\dot{J}^-(\mathscr{I}^+) \cup J^-(\mathscr{I}^+)$.



Proof:

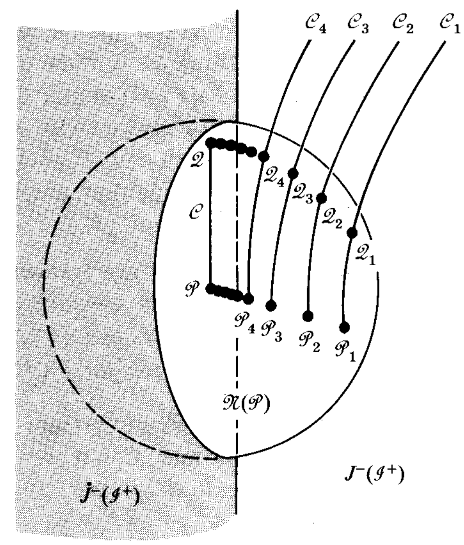
1. Pick an arbitrary event $\mathscr{A}$ on $\mathscr{C}(\lambda)$, with $\mathscr{A} \prec \mathscr{B}$.
2. Construct an arbitrarily small neighborhood $\mathscr{N}[\mathscr{A}]$.
3. A small deformation of $\mathscr{C}$, between $\mathscr{A}$ and $\mathscr{B}$, produces a timelike curve $\mathscr{D}$ from some event $\mathscr{P} \in \mathscr{N}[\mathscr{A}]$ to $\mathscr{B}$.
4. Since $\mathscr{B} \in \dot{J}^-(\mathscr{I}^+)$, a slight deformation of $\mathscr{D}$, keeping it still timelike, produces a curve $\mathscr{E}$ from $\mathscr{P}$ to some event $\mathscr{Q} \in J^-(\mathscr{I}^+)$. $\mathscr{E}$ can then be prolonged, remaining causal, until it reaches $\mathscr{I}^+$. The result is a causal curve from $\mathscr{P}$ to $\mathscr{I}^+$. Hence, $\mathscr{P} \in J^-(\mathscr{I}^+)$.
5. But $\mathscr{P}$ was in an arbitrarily small neighborhood $\mathscr{N}[\mathscr{A}]$. Hence, $\mathscr{A}$ must also be in $J^-(\mathscr{I}^+)$ or else in its boundary, $\dot{J}^-(\mathscr{I}^+)$. Q.E.D.

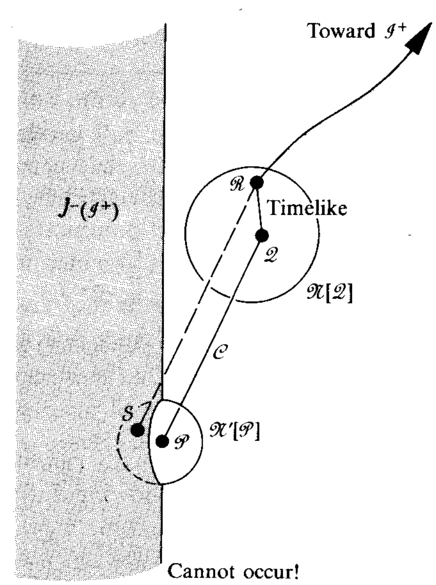
D. *Theorem* [Penrose (1968a)]: $\dot{J}^-(\mathscr{I}^+)$ is generated by null geodesics which have no future endpoints. [See text of §34.4 for more detailed statement of theorem.]

Proof:

1. Pick an arbitrary event $\mathscr{P}$ in $\dot{J}^-(\mathscr{I}^+)$. Prove as follows that through $\mathscr{P}$ there passes a future-directed null geodesic which lies in $\dot{J}^-(\mathscr{I}^+)$:
   a. Construct an arbitrary neighborhood $\mathscr{N}[\mathscr{P}]$. [If $\dot{J}^-(\mathscr{I}^+)$ happens somewhere to encounter a singularity of spacetime, then $\mathscr{N}[\mathscr{P}]$ must be chosen small enough to keep the singularity outside it.]
   b. In $\mathscr{N}[\mathscr{P}] \cap J^-(\mathscr{I}^+)$, construct a sequence of events $\{\mathscr{P}_i\}$ which converges to the event $\mathscr{P}$.
   c. For each i, construct a causal curve $\mathcal{C}_i$ extending from $\mathscr{P}_i$ to $\mathscr{I}^+$.
   d. Let $\mathscr{Q}_i$ be the intersection of $\mathcal{C}_i$ with $\mathscr{N}[\mathscr{P}]$, the boundary of $\mathscr{N}[\mathscr{P}]$. Since $\dot{\mathscr{N}}[\mathscr{P}]$ is a compact set, the sequence $\mathscr{Q}_i$ must have a limit point, $\mathscr{Q}$.
   e. Because there exist causal curves from events $\mathscr{P}_i$ arbitrarily near $\mathscr{P}$ to events $\mathscr{Q}_i$ arbitrarily near $\mathscr{Q}$, there must be a causal curve from $\mathscr{P}$ to $\mathscr{Q}$. Call that curve $\mathcal{C}$.
   f. Since $\mathscr{Q}$ is a limit point of a sequence of events in $J^-(\mathscr{I}^+)$, $\mathscr{Q}$ either lies in $J^-(\mathscr{I}^+)$, or else lies on its boundary $\dot{J}^-(\mathscr{I}^+)$, or both. Suppose $\mathscr{Q} \notin \dot{J}^-(\mathscr{I}^+)$.
      i. Then some small neighborhood $\mathscr{N}[\mathscr{Q}]$ is contained entirely in $J^-(\mathscr{I}^+)$.
      ii. Construct a causal curve from $\mathscr{P}$ to $\mathscr{I}^+$ by going from $\mathscr{P}$ to $\mathscr{Q}$ along the causal curve $\mathcal{C}$, then from $\mathscr{Q}$ along a timelike curve to some event $\mathscr{R} \in \mathscr{N}[\mathscr{Q}]$, and then from $\mathscr{R}$ to $\mathscr{I}^+$ along a causal curve—and by smoothing at the join points $\mathscr{Q}$ and $\mathscr{R}$.
      iii. Since this curve from $\mathscr{P}$ to $\mathscr{I}^+$ has a timelike segment, it can be deformed smoothly, while being kept causal, so that it reaches any desired event $\mathcal{S}$ in some small neighborhood $\mathscr{N}'[\mathscr{P}]$. But this means that $\mathscr{N}'[\mathscr{P}] \subset J^-(\mathscr{I}^+)$, hence that $\mathscr{P} \notin \dot{J}^-(\mathscr{I}^+)$—which contradicts the original definition of $\mathscr{P}$.

Conclusion: $\mathscr{Q} \in \dot{J}^-(\mathscr{I}^+)$.



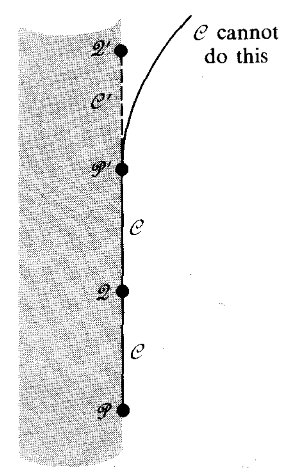


Box 34.1 ***(continued)***

g. By Lemma B, since $\mathscr{P} \in \dot{J}^-(\mathscr{I}^+)$ and $\mathscr{Q} \in \dot{J}^-(\mathscr{I}^+)$, then $\mathscr{P} \not< \mathscr{Q}$. But $\mathscr{C}$ is a future-directed causal curve from $\mathscr{P}$ to $\mathscr{Q}$. Consequently, by Lemma A, $\mathscr{C}$ is a null geodesic.
h. Since the curve $\mathscr{C}$ intersects $\dot{J}^-(\mathscr{I}^+)$ at $\mathscr{Q}$, between $\mathscr{P}$ and $\mathscr{Q}$ it must everywhere lie in $\dot{J}^-(\mathscr{I}^+) \cup J^-(\mathscr{I}^+)$ [Lemma C]. Apply the reasoning of (f) above, with $\mathscr{Q}$ replaced by an arbitrary point on $\mathscr{C}$ between $\mathscr{P}$ and $\mathscr{Q}$. Thereby conclude that, everywhere between $\mathscr{P}$ and $\mathscr{Q}$, $\mathscr{C}$ lies in $\dot{J}^-(\mathscr{I}^+)$.

Summary: Through every event $\mathscr{P} \in \dot{J}^-(\mathscr{I}^+)$ there passes a null geodesic $\mathscr{C}$ which, when followed into the future from $\mathscr{P}$, lies in $\dot{J}^-(\mathscr{I}^+)$. This null geodesic is called a "generator" of $\dot{J}^-(\mathscr{I}^+)$.

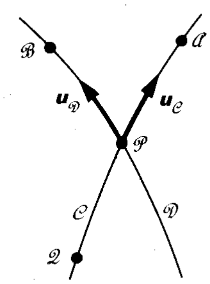
2. Follow the generator $\mathscr{C}$ from $\mathscr{P}$ to $\mathscr{Q}$ and then onward still further. Can it *ever* leave $\dot{J}^-(\mathscr{I}^+)$? No! For suppose it did leave, at some event $\mathscr{P}' \in \dot{J}^-(\mathscr{I}^+)$.
   a. Repeat the entire construction of step 1, with $\mathscr{P}'$ replacing $\mathscr{P}$, to conclude that there is a null geodesic $\mathscr{C}' \subset \dot{J}^-(\mathscr{I}^+)$ extending into the causal future from $\mathscr{P}'$ to some event $\mathscr{Q}'$.
   b. By Lemma B, since $\mathscr{P} \in \dot{J}^-(\mathscr{I}^+)$ and $\mathscr{Q}' \in \dot{J}^-(\mathscr{I}^+)$, $\mathscr{P} \not< \mathscr{Q}'$.
   c. Then by Lemma A the null geodesic $\mathscr{C}$ from $\mathscr{P}$ to $\mathscr{P}'$ and the null geodesic $\mathscr{C}'$ from $\mathscr{P}'$ to $\mathscr{Q}'$ have tangents that coincide at $\mathscr{P}'$ (aside from normalization). Thus, with a renormalization of affine parameter, $\mathscr{C}'$ becomes the prolongation of $\mathscr{C}$—which means that $\mathscr{C}$ does not leave $\dot{J}^-(\mathscr{I}^+)$ at $\mathscr{P}'$.



Conclusion: Once a generator, being followed into the future, enters $\dot{J}^-(\mathscr{I}^+)$, it can never thereafter leave $\dot{J}^-(\mathscr{I}^+)$.

3. Figure 34.7 provides an example of how a null geodesic, being following into the future, can enter $\dot{J}^-(\mathscr{I}^+)$ and become a generator. Lemma C guarantees that, *when a null geodesic enters $\dot{J}^-(\mathscr{I}^+)$, it enters from $J^-(\mathscr{I}^+)$.*

4. As indicated by the example of Fig. 34.7, at a "caustic" [entry point of generators onto $\dot{J}^-(\mathscr{I}^+)$] generators can cross each other. Follow a generator $\mathscr{C}$ to the causal future from its entry point onto $\dot{J}^-(\mathscr{I}^+)$. Can it ever again cross another generator? No. For suppose that at an event $\mathscr{P}$ the generator $\mathscr{C}$ were to cross another generator $\mathscr{D}$.
   a. To the causal future of $\mathscr{P}$, both generators always lie in $\dot{J}^-(\mathscr{I}^+)$. Thus, events $\mathscr{A}$ and $\mathscr{B}$ of the picture are in $\dot{J}^-(\mathscr{I}^+)$.
   b. Since $\mathscr{P}$ is to the causal future of the caustic where $\mathscr{C}$ enters $\dot{J}^-(\mathscr{I}^+)$, there exists an event $\mathscr{Q} \in \dot{J}^-(\mathscr{I}^+) \cap \mathscr{C}$ to the causal past of $\mathscr{P}$.
   c. Since $\mathscr{Q} \in \dot{J}^-(\mathscr{I}^+)$ and $\mathscr{B} \in \dot{J}^-(\mathscr{I}^+)$, $\mathscr{Q} \not\prec \mathscr{B}$ [Lemma B].
   d. Lemma A, applied to the curves $\mathscr{C}$ from $\mathscr{Q}$ to $\mathscr{P}$, and $\mathscr{D}$ from $\mathscr{P}$ to $\mathscr{B}$, then guarantees that the tangent vectors $\boldsymbol{u}_{\mathscr{C}}$ and $\boldsymbol{u}_{\mathscr{D}}$ coincide at $\mathscr{P}$ (aside from normalization), and that therefore (aside from normalization) the geodesics $\mathscr{C}$ and $\mathscr{D}$ are identical. This contradicts the supposition that $\mathscr{C}$ and $\mathscr{D}$ are different generators which cross at $\mathscr{P}$.



Conclusion: Once a generator has entered $\dot{J}^-(\mathscr{I}^+)$, it can never thereafter cross any other generator.

§34.5. PROOF OF SECOND LAW OF BLACK-HOLE DYNAMICS [Hawking (1971b, 1972a, 1973)]

All the tools are now in hand for a proof of the second law of black-hole dynamics.

Proof of second law of black-hole dynamics:

Consider the union of all future horizons, $\dot{J}^-(\mathscr{I}^+)$, in an asymptotically flat spacetime, as depicted in Figure 34.8. Divide up the null-geodesic generators of $\dot{J}^-(\mathscr{I}^+)$ into a large number of infinitesimal bundles, and give each bundle an identifying number, K. As one moves from past toward future along $\dot{J}^-(\mathscr{I}^+)$, one occasionally sees new bundles of generators created in "caustics" of the 3-surface $\dot{J}^-(\mathscr{I}^+)$. The caustic sources of new generators are created by such processes as the infall of matter through the horizon (example: bundle #42 in Figure 34.8), and the collision and coalescence of two black holes (example: bundle #29). But each bundle, once created, can never be destroyed (no termination of null generators as one moves from past toward future).

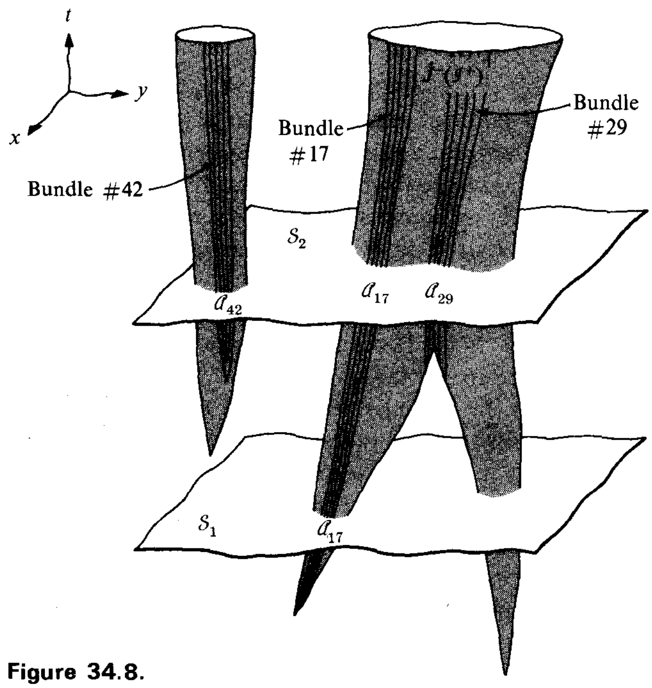


Figure 34.8.
Schematic spacetime diagram used in proving the second law of black-hole dynamics. See text for details of the proof, and see Figure 34.6 for physical interpretation of the diagram.

Focus attention on a specific bundle of generators—bundle $\# K$. At a specific event $\mathscr{P}$ along that bundle, let various observers, moving with various velocities, measure its (two-dimensional) cross-sectional area $\mathscr{A}_K(\mathscr{P})$. As shown in Figure 22.1, exercise 22.13, and exercise 22.14: (1) the cross-sectional area $\mathscr{A}_K(\mathscr{P})$ is independent of the velocity of the observer who measures it—i.e., $\mathscr{A}_K(\mathscr{P})$ depends only on location $\mathscr{P}$ along the bundle; and (2) $\mathscr{A}_K$ changes from event to event along the bundle in a manner governed by the "focusing theorem"

$$\frac{d^2\mathscr{A}_K{}^{1/2}}{d\lambda_K{}^2} \leq 0 \qquad \text{if the energy density } T_{\hat{0}\hat{0}}, \text{ as measured by all observers along the bundle, is nonnegative.} \tag{34.6}$$

Proof assumes nonnegative energy density

Here λ_K is affine parameter along the bundle. *Assume*—in accord with all physical experience and the best assessments of modern physics—*that energy density $T_{\hat{0}\hat{0}}$ can never be negative.* (*This assumption underlies the second law of black-hole dynamics. If it were ever found to be invalid, then one would have to abandon the second law.*)

Suppose that $d\mathscr{A}_K{}^{1/2}/d\lambda_K$ were negative at some event $\mathscr{P}$ along the bundle. Then, according to the focusing theorem, $d\mathscr{A}_K{}^{1/2}/d\lambda_K$ would always remain at least as negative as its value at $\mathcal{P}$—and, hence, after a lapse of affine parameter given by

$$\Delta\lambda_K \leq \left(\frac{\mathcal{A}_K{}^{1/2}}{-d\mathcal{A}_K{}^{1/2}/d\lambda_K}\right)_{\text{at } \mathcal{P}}, \tag{34.7}$$

$\mathcal{A}_K{}^{1/2}$ would go to zero. At the point where $\mathcal{A}_K{}^{1/2}$ reaches zero, adjacent null geodesics in the bundle cross each other, giving rise to events in $\dot{J}^-(\mathscr{I}^+)$ through which pass more than one null geodesic generator. But this violates Penrose's theorem on the global structure of horizons (§34.4).

Thus either the supposition of negative $d\mathcal{A}_K{}^{1/2}/d\lambda_K$ is wrong; or else $d\mathcal{A}_K{}^{1/2}/d\lambda_K$ goes negative, but then, before the generators get a chance to cross [before the finite lapse (34.7) of affine parameter], the generators hit a singularity and cease to exist. *To prove the second law of black-hole dynamics, one must assume that no singularity is hit by the horizon, and thereby conclude that $d\mathcal{A}_K{}^{1/2}/d\lambda_K$ never goes negative. Hawking (1971b, 1972a) makes an alternative assumption which implies $d\mathcal{A}_K{}^{1/2}/d\lambda_K \geq 0$: Hawking assumes that spacetime is "future asymptotically predictable." In essence this means that spacetime possesses no "naked singularities"—i.e., no singularities visible from $\mathscr{I}^+$.* (Naked singularities could influence the evolution of the external universe; and, therefore, unless one knew the laws of physics governing singularities—which one does not—they would prevent one from predicting the future in the external universe.)

Proof assumes that horizon never hits a singularity (no naked singularities)

Under either assumption (no naked singularities; or horizon never hits a singularity), one concludes that

$$d\mathcal{A}_K{}^{1/2}/d\lambda_K \text{ is nonnegative everywhere along bundle } K. \tag{34.8}$$

This result says that the cross-sectional area $\mathcal{A}_K$ of each bundle can never decrease as one moves toward the future along $\dot{J}^-(\mathscr{I}^+)$. Since new bundles can be created, but old ones can never be destroyed as one moves toward the future, *the total cross-sectional area of $\dot{J}^-(\mathscr{I}^+)$ cannot decrease toward the future.* Equivalently, (see Figure 34.8), if $\mathcal{S}_1$ and $\mathcal{S}_2$ are spacelike hypersurfaces with $\mathcal{S}_2$ everywhere to the future of $\mathcal{S}_1$, then the cross-sectional area of $\dot{J}^-(\mathscr{I}^+)$ at its intersection with $\mathcal{S}_2$, $\mathcal{A}(\mathcal{S}_2)$, cannot be less than the cross-sectional area at $\mathcal{S}_1$, $\mathcal{A}(\mathcal{S}_1)$. This is the second law of black-hole dynamics, reformulated in more precise language than that of Chapter 33, and finally proved.

Precise formulation of second law

EXERCISE

Exercise 34.4. A BLACK HOLE CAN NEVER BIFURCATE [Hawking (1972a)]

Make plausible the theorem that no matter how hard one "zaps" a black hole, and no matter what one "zaps" it with, one can never make it bifurcate into two black holes. [*Hint:* By drawing pictures, make it plausible that, at any bifurcation point, some null geodesic generators of $\dot{J}^-(\mathscr{I}^+)$ must leave $\dot{J}^-(\mathscr{I}^+)$ as one follows them into the *future*—in violation of Penrose's theorem (§34.4). Assume that the surface of each hole is topologically a 2-sphere. *Note:* The same argument, time-reversed, shows that if two black holes coalesce, generators enter $\dot{J}^-(\mathscr{I}^+)$ from $J^-(\mathscr{I}^+)$ at the coalescence point; and the surface area of the horizon increases.]

2

§34.6. THEOREMS ON SINGULARITIES, AND THE "ISSUE OF THE FINAL STATE"

Overview of theorems on singularities

Just as global techniques are powerful tools in the analysis of horizons, so they also are powerful in the analysis of spacetime singularities. In fact, it was the proof of Penrose's (1965b) pioneering theorem on singularities that gave birth to global techniques for studying spacetime.

For a detailed introduction to the global analysis of singularities, one can read the book of Hawking and Ellis (1973). Now that the reader has had a taste of global techniques, attention here will focus on a qualitative description of results:

How does gravitational collapse terminate? Is the singularity at the end point of spherical collapse typical, or can asymmetries remove it? That singularities are very general phenomena, and cannot be wished away, has been known since 1965, thanks to theorems on singularities proved by Penrose, Hawking, and Geroch. [For a full list of references, see Hawking and Penrose (1969) or Hawking and Ellis (1973).]

Singularity defined

Before examining the theorems on singularities, one must make precise the concept of a *singularity*. This is not easy, as Geroch (1968) has emphasized in a long treatise on the wide variety of pathologies that can occur in spacetime manifolds. However, after vigorous efforts by many people, Schmidt (1970) finally produced a definition that appears to be satisfactory. Put in heuristic terms, Schmidt's highly technical definition goes something like this. In a spacetime manifold, consider all spacelike geodesics (paths of "tachyons"), all null geodesics (paths of photons), all timelike geodesics (paths of freely falling observers), and all timelike curves with bounded acceleration (paths along which observers are able, in principle, to move). Suppose that one of these curves terminates after the lapse of finite proper length (or finite affine parameter in the null-geodesic case). Suppose, further, that it is impossible to extend the spacetime manifold beyond that termination point—e.g., because of infinite curvature there. Then that termination point, together with all adjacent termination points, is called a "singularity." (What could be more singular than the cessation of existence for the poor tachyon, photon, or observer who moves along the terminated curve?)

Trapped surface defined

Another concept needed in the singularity theorems is that of a *trapped surface*. This concept, devised by Penrose (1965b), is motivated by a close examination of the two-dimensional, spherical surfaces $(r, t) =$ const. inside the horizon of the Schwarzschild geometry. These surfaces signal the nearness of a singularity ($r = 0$) by this property: light rays emitted from one of these surfaces in the perpendicular outward direction (i.e., outgoing, orthogonal, null geodesics) converge toward each other as they propagate; and inward light rays perpendicular to the 2-surface also converge. Penrose gives the name "trapped surface" to any closed 2-surface, spherical or not, that has this property. In Schwarzschild spacetime, the convergence of light rays, both outgoing and ingoing, can be attributed to the "intense pull of gravity," which sucks the photons into the singularity. That this might also be true in asymmetric spacetimes is suggested by the Hawking-Penrose (1969) theorem [the most powerful of a wide class; see Hawking and Penrose (1969) for references to others; and see Boxes 34.2 and 34.3 for introductions to Hawking and Penrose]:

The Hawking-Penrose theorem on singularities

A spacetime M necessarily contains incomplete, inextendable timelike or null geodesics (and is thus singular in the Schmidt sense) if, in addition to Einstein's equations, the following four conditions hold: (1) M contains no closed timelike curves (reasonable causality condition); (2) at each event in M and for each unit timelike vector $\boldsymbol{u}$, the stress-energy tensor satisfies

$$\left(T_{\alpha\beta} - \frac{1}{2}g_{\alpha\beta}T\right)u^\alpha u^\beta \geq 0$$

(reasonable energy condition); (3) the manifold is "general" (i.e., not too highly symmetric) in the sense that every timelike or null geodesic with unit tangent $\boldsymbol{u}$ passes through at least one event where the curvature is not lined up with it in a specific way:

$$u_{[\alpha}R_{\beta]\gamma\delta[\epsilon}u_{\rho]}u^\gamma u^\delta \neq 0 \text{ at some point on the geodesic.}$$

(4) the manifold contains a trapped surface.

All these conditions, except the trapped surface, seem eminently reasonable for any physically realistic spacetime! Note, especially, that the energy condition can be violated only if, as measured by some local observer in his proper frame, the total energy density E is negative or the principal pressures (eigenvalues of stress tensor) P_i are so negative that

$$\sum_i P_i < -E.$$

Relevance of the Hawking-Penrose theorem for gravitational collapse

The relevance of the Hawking-Penrose theorem for collapse follows from the general expectation that, in the real universe, trapped surfaces will always exist just below all future horizons, $\dot{J}^-(\mathscr{I}^+)$. (Exceptions, such as the Kerr metric with $a = M$, are probably a "set of measure zero.") Since horizons and accompanying trapped surfaces are necessarily produced by slightly nonspherical collapse (Box 32.2), and since they probably also result from moderately deformed collapse (§32.7), such collapse presumably produces singularities—or a violation of causality, which is also a rather singular occurrence!

The nature of the singularity at the endpoint of realistic collapse: 4 possibilities

If the singularities are really such a general feature of collapse, then the exact nature of the singularity is of life-and-death importance to anyone who falls through a horizon! Here one is on very shaky ground. Although the main results and conjectures described up to now in this section will probably survive all future research, opinions about the nature of the singularities are likely to change several times more before the whole story is in. Hence, it is safe only to describe the possibilities, not to attempt to judge them.

Possibility 1

The singularity at the endpoint of a realistic collapse is a region of infinite tidal gravitational forces (infinite curvature), which crushes the collapsing matter to infinite density. Examples: the very special, homogeneous crushing of the Oppenheimer-Snyder (1939) spherical collapse (§32.4); also the very special inhomogeneous but spherical crushing described by Podurets (1966); also the special inhomogeneous,

(*continued on page 940*)

Box 34.2 ROGER PENROSE: Born August 8, 1931, Colchester, Essex, England

Roger Penrose started out as an algebraic geometer. However, while at Cambridge from 1952–55 and again from 1957–60, his interest in general relativity was aroused by Hermann Bondi and Dennis Sciama. Because of his pure mathematical background, his approach to the subject was different from those which had been adopted hitherto. He was particularly interested in the global light-cone structure of spacetime and in the equations of zero rest-mass fields, both of which are preserved under conformal transformations. He exploited this conformal invariance to give an elegant and powerful treatment of gravitational radiation in terms of a null surface $\mathscr{I}^+$ at infinity. More recently this interest has led him to develop the theory of twistors, which are the spinors corresponding to the conformal group of Minkowski space. These offer a new and very promising approach to the quantization of spacetime.

His interest in conformal geometry also led him to study the properties of the causality relationships between points of spacetime. These in turn led him to the theorems on the occurrence of singularities in spacetime, which are probably the most important predictions of general relativity, since they seem to imply that spacetime has a beginning or an end.

"If spacetime is considered from the point of view of its conformal structure only, points at infinity can be treated on the same basis as finite points"

[PENROSE, IN INFELD (1964)]

"The argument will be to show that the existence of a trapped surface implies—irrespective of symmetry—that singularities necessarily develop"

[PENROSE (1965b)]

"While the quantum effects of gravitation are normally thought to be significant only when curvatures approach 10^{33} cm^{-1}, all our local physics is based on the Poincare group being a good approximation of a local symmetry group at dimensions greater than 10^{-13} cm. Thus, if curvatures ever even approach 10^{13} cm^{-1}, there can be little doubt but that extraordinary local effects are likely to take place"

[HAWKING AND PENROSE (1969)]

"We are thus presented with what is perhaps the most fundamental unanswered question of general-relativistic collapse theory, namely: does there exist a "cosmic censor" who forbids the appearance of naked singularities, clothing each one in an absolute event horizon?"

[PENROSE (1969)]

"Under normal circumstances, general relativity can, for practical purposes, remain remarkably apart—almost aloof—from the rest of physics. At a space-time singularity, the very reverse must surely be the case!"

"I do not believe that a real understanding of the nature of elementary particles can ever be achieved without a simultaneous deeper understanding of the nature of spacetime itself. But if we are concerned with a level of phenomena for which such an understanding is not *necessary—and this will cover almost all of present-day physics—then the smooth manifold picture presents an (unreasonably!) excellent framework for the discussion of phenomena."*

"The most important single lesson of relativity theory is, perhaps, that space and time are not concepts that can be considered independently of one another but must be combined together to give a four*-dimensional picture of phenomena: the description in terms of* spacetime*"*

[PENROSE (1968a)]

"If a formalism enables one to treat myriads of non-existent types of universe, then (effectively) it contains 'arbitrary parameters,' only special values of which will correspond to the world as it actually is. In the ordinary approach to spacetime as a pseudo-Riemannian differentiable manifold, the dimension of the manifold and the signature of the metric are two such arbitrary parameters."

"As we localize the position of a particle, it jumps essentially along the null cone. Other particles are produced, which leap backward and forward essentially along null directions, without apparent regard for continuity, heeding only the positions of the null cones themselves and "topology" only in the respect in which this term is applied to the structure of graphs"

[PENROSE (1966)]

"My own view is that ultimately physical laws should find their most natural expression in terms of essentially combinatorial *principles, that is to say, in terms of finite processes such as counting or other basically simple manipulative procedures. Thus, in accordance with such a view, should emerge some form of discrete or combinatorial spacetime"*

[PENROSE, IN KLAUDER (1972)]

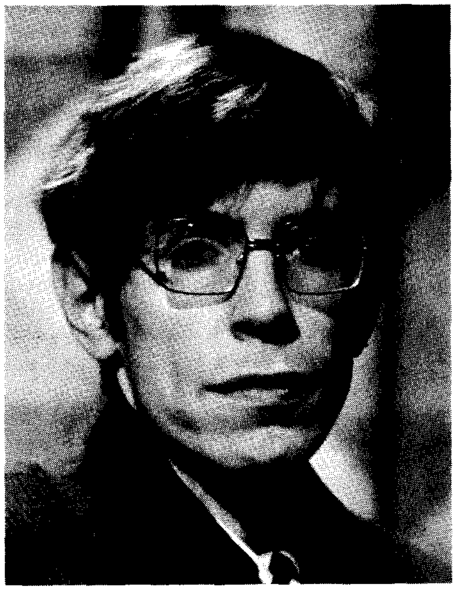
"Complex numbers are . . . a very important constituent of the structure of physical laws. The twistor theory carries this further in suggesting that complex numbers may also be very basically involved in defining the nature of spacetime itself."

[PENROSE AND MACCALLUM (1973)]

"It is thus very tempting to believe that a link between spacetime curvature and quantum processes may be supplied by the use of twistors. Then, roughly speaking, it is the continual slight 'shifting' of the interpretations of the quantum (twistor) operators which results in the curvature of spacetime"

[PENROSE (1968b)]

Box 34.3 STEPHEN W. HAWKING: Born January 8, 1942, Oxford, England



As a research student of Dennis Sciama's in Cambridge, Stephen Hawking's early interest in relativity theory centered mainly on the question of spacetime singularities. With Ellis, he showed that a large class of homogeneous cosmological models must be singular. Then, encouraged by work of Penrose on the singularities arising in gravitational collapse, he developed new techniques which, in a series of papers in the Royal Society of London during 1966–67, established the important result that any plausible general-relativistic cosmology must be singular.

The major portion of his later research has been concerned with black holes. He devised a series of arguments of great ingenuity which, together with the work of Israel and Carter, established to all intents and purposes the result that (vacuum) black holes in general relativity are described by Kerr metrics, that topologies other than spherical cannot occur, and that a certain limit on the energy emitted when two black holes congeal into one must be satisfied.

Some of this work has had substantial pure mathematical interest (e.g., singularity theorems), some of it is concerned with astrophysics (e.g., work with Taylor on helium production in the big bang), some with observations (work with Gibbons on the possibility of black holes in binary star systems) and even experimental developments (with Gibbons on gravitational-wave detectors). In such scope is exhibited not only a considerable insight, depth, and versatility, but also the gift of an extraordinary determination to overcome severe physical handicaps, to seek out and comprehend the truth.

"The observed isotropy of the microwave background indicates that the universe is rotating very little if at all. . . . This could possibly be regarded as an experimental verification of Mach's Principle"

[HAWKING (1969)]

"Undoubtedly, the most important results are the theorems . . . on the occurrence of singularities. These seem to imply either that the general theory of relativity breaks down or that there could be particles whose histories did not exist before (or after) a certain time. The author's own opinion is that the theory probably does break down but only when quantum gravitational effects become important."

"Although we have omitted the singular points from the definition of spacetime, we can still recognize the 'holes' left where they have been cut out by the existence of incomplete geodesics."

"A good physical theory should not only correctly describe the currently experimental knowledge, but should also predict new results which can be tested by experiment, the further the predictions from the original experiments, the greater the credit to the theory if they are found to be correct. Thus observations of whether or not singularities actually occurred, would provide a powerful test of the general theory of relativity in strong fields"

[HAWKING (1966a)]

"The construction of gravitational radiation detectors may open up a whole new field of 'gravitational astronomy' which could be as fruitful as radio astronomy has been in the last two decades. . . . Black hole collisions . . . would be much more effective in converting rest-mass energy into radiation than nuclear reactions, which can release only about 1 per cent of the rest-mass energy. In addition, black holes formed by collisions of smaller black holes can undergo further collisions, releasing more energy, whereas matter that has been fully processed by nuclear reactions cannot yield any more energy by the same means. . . . we are witnessing something really cataclysmic at the centre of our galaxy"

[HAWKING (1972b)]

"One might suggest that prior to the present expansion there was a collapsing phase. In this, local inhomogeneities grew large and isolated singularities occurred. Most of the matter avoided the singularities and reexpanded to give the present observed universe."

"It seems that we should draw a surface around regions where the radius of curvature is less than, say, 10^{-16} cm. On our side of this surface, a manifold picture of spacetime would be appropriate, but we have no idea what structure spacetime would have on the other side"

[HAWKING AND ELLIS (1968)]

"Presumably it would be necessary to consider quantum effects in very strong fields. However, these would not become important until the radius of curvature became of the order of 10^{-14} cm, which for practical purposes is pretty singular."

"The view has been expressed that singularities are so objectionable that if the Einstein equations were to predict their occurrence, this would be a compulsive reason for modifying them. However, the real test of a physical theory is not whether its predicted results are aesthetically attractive but whether they agree with observation. So far there are no observations which would show that singularities do not occur"

[HAWKING (1966b)]

"It is shown that a stationary black hole must have a topologically spherical boundary and must be axisymmetric if it is rotating. These results, together with those of Israel and Carter, go most of the way toward establishing the conjecture that any stationary black hole is a Kerr solution"

[HAWKING (1972a)]

"The fact that we have observed the universe to be isotropic is only a consequence of our existence."

[COLLINS AND HAWKING (1973)]

"Kasner-like" crushing of Lifschitz and Khalatnikov (1963a,b); also, most importantly, the very general "mixmaster" crushing (Chapter 30), discovered in the homogeneous case by Misner (1969b) and by Belinsky and Khalatnikov (1969a), and analyzed in the inhomogeneous case by Belinsky and Khalatnikov (1969b, 1970) and by Khalatnikov and Lifschitz (1970). The mixmaster singularities—and only they among all explicitly known singularities—appear to be generic in this sense: if one perturbs slightly but arbitrarily the initial conditions of a spacetime that evolves a mixmaster singularity, then the resultant perturbed spacetime will also evolve a mixmaster singularity. Because of this, the prevalent opinion today (1973) is that realistic collapse probably produces, inside the horizon, a mixmaster singularity. But that opinion might change tomorrow.

Possibility 2

The singularity is a region of spacetime in which timelike or null geodesics terminate, not because of infinite tidal gravitational forces or infinite crushing, but because of other, more subtle pathologies. Example: "Taub-NUT space" [see Misner and Taub (1968)]. For other examples created specially to exhibit various pathologies, see Geroch (1968).

Possibility 3

The singularity may be sufficiently limited in "size" and influence that all or most of the collapsing matter successfully avoids it. The matter cannot then explode back outward through the horizon that it went down; the horizon is a one-way membrane and will not let anything back out. Instead, the matter may reach a stage of maximum but finite contraction, and then reexplode into some other region of spacetime (multiply connected spacetime topology; "wormhole"). Analytical solutions for collapsing, charged spheres do reexplode in this manner [Novikov (1966); de la Cruz and Israel (1967); Bardeen (1968); see Figure 34.4]. Such a process requires that the "exploding" end of the wormhole be built into the initial conditions of the universe, with mass and angular momentum (as measured by Keplerian orbits and frame dragging) precisely equal to those that go down the black-hole end. This seems physically implausible. So does the "explosion."

Other Possibilities

Various combinations of the above.

Will quantization of spacetime save the universe from singularities?

If, as one suspects today, the singularities are of a very physical, infinite-curvature type, then one must face up to John Wheeler's (1964a) "issue of the final state" in its most raw and disturbing form. Wheeler, when faced with the issue, argues that infinite-curvature singularities signal a breakdown in classical general relativity—a breakdown forced by the onset of quantum gravitational phenomena (see Chapter 44). Whether quantization of gravity will actually save spacetime from such singularities one cannot know until the "fiery marriage of general relativity with quantum physics has been consummated" [Wheeler (1964a); see also Misner (1969c), and the last section of Box 30.1].

PART **VIII**

GRAVITATIONAL WAVES

Wherein the reader voyages on stormy seas of curvature ripples, searching for the ripple-generating storm gods, and battles through an electromagnetic and thermal fog that allows only uncertain visibility upon those seas.

1

CHAPTER 35

PROPAGATION OF GRAVITATIONAL WAVES

Born: *"I should like to put to Herr Einstein a question, namely, how quickly the action of gravitation is propagated in your theory. That it happens with the speed of light does not elucidate it to me. There must be a very complicated connection between these ideas."*

Einstein: *"It is extremely simple to write down the equations for the case when the perturbations that one introduces in the field are infinitely small. Then the g's differ only infinitesimally from those that would be present without the perturbation. The perturbations then propagate with the same velocity as light."*

Born: *"But for great perturbations things are surely very complicated?"*

Einstein: *"Yes, it is a mathematically complicated problem. It is especially difficult to find exact solutions of the equations, as the equations are nonlinear."*

Excerpts from discussion after Einstein's Fall 1913 lecture in Vienna on "The present position of the problem of gravitation," already two years before he had the final field equations [EINSTEIN, 1913a]

§35.1. VIEWPOINTS

Study one idealization after another. Build a catalog of idealizations, of their properties, of techniques for analyzing them. This is the only way to come to grips with so complicated a subject as general relativity!

Spherical symmetry is the idealization that has dominated most of the last 12 chapters. Together with the idealization of matter as a perfect fluid, and of the universe as homogeneous, it has yielded insight into stars, into cosmology, into gravitational collapse.

Turn attention now to an idealization of an entirely different type, one independent of any symmetry considerations at all: the idealization of a "gravitational wave."

Just as one identifies as "water waves" small ripples rolling across the ocean, so one gives the name "gravitational waves" to small ripples rolling across spacetime.

Gravitational waves compared to water waves on ocean:

We are deeply indebted to Mr. James M. Nester, who found and corrected many errors in the equations of this chapter and of a dozen others throughout the book.

(1) approximate nature of a wave

Ripples of what? Ripples in the shape of the ocean's surface; ripples in the shape (i.e., curvature) of spacetime. Both types of waves are idealizations. One cannot, with infinite accuracy, delineate at any moment which drops of water are in the waves and which are in the underlying ocean: Similarly, one cannot delineate precisely which parts of the spacetime curvature are in the ripples and which are in the cosmological background. But one can almost do so; otherwise one would not speak of "waves"!

(2) local viewpoint vs. large-scale viewpoint

Look at the ocean from a rowboat. Waves dominate the seascape. Changes in angle and level of the surface occur every 30 feet or less. These changes propagate. They obey a simple wave equation

$$\left(\frac{1}{g^2}\frac{\partial^4}{\partial t^4}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial x^2}\right)(\text{height of surface})=0.$$

Now get more sophisticated. Notice from a spaceship the large-scale curvature of the ocean's surface—curvature because the Earth is round, curvature because the sun and the moon pull on the water. As waves propagate long distances, this curvature bends their fronts and changes slightly their simple wave equation. Also important over large distance are nonlinear interactions between waves, interaction with the wind, Coriolis forces due to the Earth's rotation, etc.

Spacetime is similar. Propagating through the universe, according to Einstein's theory, must be a complex pattern of small-scale ripples in the spacetime curvature, ripples produced by binary stars, by supernovae, by gravitational collapse, by explosions in galactic nuclei. Locally ("rowboat viewpoint") one can ignore the interaction of these ripples with the large-scale curvature of spacetime and their nonlinear interaction with each other. One can pretend the waves propagate in flat spacetime; and one can write down a simple wave equation for them. But globally one cannot. The large-scale curvature due to quiescent stars and galaxies will produce redshifts and will deform wave fronts; and the energy carried by the waves themselves will help to produce the large-scale curvature. This chapter treats both viewpoints, the local (§§35.2–6), and the global (§§35.7–15).

§35.2. REVIEW OF "LINEARIZED THEORY" IN VACUUM

Linearized theory of gravitational waves:

Idealize, for awhile, the gravitational waves of our universe as propagating through flat, empty spacetime (local viewpoint). Then they can be analyzed using the "linearized theory of gravity," which was introduced in Chapter 18.

Linearized theory, one recalls, is a weak-field approximation to general relativity. The equations of linearized theory are written and solved as though spacetime were flat (special-relativity viewpoint); but the connection to experiment is made through the curved-space formalism of general relativity.

More specifically, linearized theory describes gravity by a symmetric, second-rank tensor field $\bar{h}_{\mu\nu}$. In the standard ("Lorentz," or Hilbert) gauge, this field satisfies the "gauge" or "subsidiary" conditions (coordinate conditions)

(1) Lorentz gauge condition

$$\bar{h}^{\mu\alpha}{}_{,\alpha}=0. \tag{35.1a}$$

(Here, and throughout linearized theory, indices of $\bar{h}_{\mu\nu}$ are raised and lowered with the Minkowski metric $\eta_{\alpha\beta}$.) In this gauge the *propagation equations* for vacuum gravitational fields are the familiar wave equations

$$\Box\bar{h}_{\mu\nu} \equiv \bar{h}_{\mu\nu,\alpha}{}^{\alpha} = 0. \tag{35.1b}$$

(2) propagation equation

Spacetime is really curved in linearized theory, although equations (35.1) are written and solved as though it were not. The global inertial frames of equations (35.1) are only *almost* inertial. In them the metric components are actually

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} + O([h_{\mu\nu}]^2); \tag{35.2a}^*$$

(3) metric

and the "metric perturbation" $h_{\mu\nu}$ is related to the "gravitational field" $\bar{h}_{\mu\nu}$ by

$$h_{\mu\nu} = \bar{h}_{\mu\nu} - \frac{1}{2}\bar{h}\eta_{\mu\nu}, \qquad \bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2}h\eta_{\mu\nu},$$
$$h \equiv h_{\alpha}{}^{\alpha} = -\bar{h} = -\bar{h}_{\alpha}{}^{\alpha}. \tag{35.2b}$$

The metric (35.2a) governs the motion of test particles, the propagation of light, etc., in the usual general-relativistic manner.

Recall the origin of the equations (35.1) that govern $\bar{h}_{\mu\nu}$. The subsidiary conditions $\bar{h}_{\mu}{}^{\alpha}{}_{,\alpha} = 0$ were imposed by specializing the coordinate system; and the Einstein field equations in vacuum then reduced to $\Box\bar{h}_{\mu\nu} = 0$.

Actually, as was shown in Box 18.2, the coordinates of linearized theory are not fully fixed by the conditions $\bar{h}_{\mu}{}^{\alpha}{}_{,\alpha} = 0$. There remains an ambiguity embodied in further "gauge changes" (infinitesimal coordinate transformations), ξ_{μ}, which satisfy a restrictive condition

(4) residual gauge freedom

$$\xi_{\mu,\alpha}{}^{\alpha} = 0 \tag{35.3a}$$

in order to preserve conditions (35.1a). Then

$$x^{\mu}{}_{\text{new}} = x^{\mu}{}_{\text{old}} + \xi^{\mu} \tag{35.3b}$$

is the coordinate transformation and

$$\bar{h}_{\mu\nu\,\text{new}} = \bar{h}_{\mu\nu\,\text{old}} - \xi_{\mu,\nu} - \xi_{\nu,\mu} + \eta_{\mu\nu}\xi^{\alpha}{}_{,\alpha} \tag{35.3c}$$

is the gauge change. All this was derived and discussed in Chapter 18.

§35.3. PLANE-WAVE SOLUTIONS IN LINEARIZED THEORY

The simplest of all solutions to the linearized equations $\bar{h}_{\mu\nu,\alpha}{}^{\alpha} = \bar{h}_{\mu}{}^{\alpha}{}_{,\alpha} = 0$ is the monochromatic, plane-wave solution,

Monochromatic, plane wave

$$\bar{h}_{\mu\nu} = \Re[A_{\mu\nu}\exp(ik_{\alpha}x^{\alpha})]. \tag{35.4a}$$

*A more nearly rigorous treatment defines $h_{\mu\nu} \equiv g_{\mu\nu} - \eta_{\mu\nu}$, and puts the small corrections $0([h_{\mu\nu}]^2)$ into the field equations:

$$\bar{h}_{\mu}{}^{\alpha}{}_{,\alpha} = O([h_{\mu\nu}]^2{}_{,\alpha}), \qquad \bar{h}_{\mu\nu,\alpha}{}^{\alpha} = O([h_{\mu\nu}]^2{}_{,\alpha\beta}).$$

Here $\Re[\ldots\ldots]$ means that one must take the real part of the expression in brackets; while $A_{\mu\nu}$ (*amplitude*) and k_μ (*wave vector*) are constants satisfying

$$k_\alpha k^\alpha = 0 \qquad (\boldsymbol{k} \text{ a null vector}), \tag{35.4b}$$

$$A_{\mu\alpha} k^\alpha = 0 \qquad (\boldsymbol{A} \text{ orthogonal to } \boldsymbol{k}) \tag{35.4c}$$

[consequences of $\bar{h}_{\mu\nu,\alpha}{}^{\alpha} = 0$ and $\bar{h}_{\mu,\alpha}{}^{\alpha} = 0$, respectively; see (35.10) below for the true physics associated with this wave, the curvature tensor]. Clearly, this solution describes a wave with frequency

$$\omega \equiv k^0 = (k_x{}^2 + k_y{}^2 + k_z{}^2)^{1/2}, \tag{35.5}$$

which propagates with the speed of light in the direction $(1/k^0)(k_x, k_y, k_z)$.

Plane wave has two degrees of freedom in amplitude (two polarizations)

At first sight the amplitude $A_{\mu\nu}$ of this plane wave appears to have six independent components (ten, less the four orthogonality constraints $A_{\mu\alpha}k^\alpha = 0$). But this cannot be right! As Track-2 readers have learned in Chapter 21, the gravitational field in general relativity has two dynamic degrees of freedom, not six. Where has the analysis gone astray?

One went astray by forgetting the arbitrariness tied up in the gauge freedom (35.3). The plane-wave vector

$$\xi^\mu \equiv -iC^\mu \exp(ik_\alpha x^\alpha), \tag{35.6}$$

with four arbitrary constants C^μ, generates a gauge transformation that can change arbitrarily four of the six independent components of $A_{\mu\nu}$. One gets rid of this arbitrariness by choosing a specific gauge.

§35.4. THE TRANSVERSE TRACELESS (TT) GAUGE

Transverse-traceless (TT) gauge:

(1) for plane wave

Select a 4-velocity $\boldsymbol{u}$—not at just one event, but the same $\boldsymbol{u}$ throughout all of spacetime (special-relativistic viewpoint!). By a specific gauge transformation (exercise 35.1), impose the conditions

$$A_{\mu\nu} u^\nu = 0. \tag{35.7a}$$

These are only three constraints on $A_{\mu\nu}$, not four, because one of them—$k^\mu(A_{\mu\nu}u^\nu) = 0$—is already satisfied (35.4c). As a fourth constraint, use a gauge transformation (exercise 35.1) to set

$$A^\mu{}_\mu = 0. \tag{35.7b}$$

One now has eight constraints in all, $A_{\mu\alpha}u^\alpha = A_{\mu\alpha}k^\alpha = A_\alpha{}^\alpha = 0$, on the ten components of the amplitude; and the coordinate system (gauge) is now fixed rigidly. Thus, the two remaining free components of $A_{\mu\nu}$ represent the two degrees of freedom (two polarizations) in the plane gravitational wave.

It is useful to restate the eight constraints $A_{\mu\alpha}u^\alpha = A_{\mu\alpha}k^\alpha = A^\mu{}_\mu = 0$ in a Lorentz frame where $u^0 = 1$, $u^j = 0$, and in a form where k^α does not appear explicitly:

$$h_{\mu 0} = 0, \quad \text{i.e., only the spatial components } h_{jk} \text{ are nonzero;} \tag{35.8a}$$

$$h_{kj,j} = 0, \quad \text{i.e., the spatial components are divergence-free;} \tag{35.8b}$$

$$h_{kk} = 0, \quad \text{i.e., the spatial components are trace-free.} \tag{35.8c}$$

(Here and henceforth repeated spatial indices are to be summed, even if both are down; e.g., $h_{kk} \equiv \Sigma_{k=1}^{3} h_{kk}$.) Notice that, since $h = h_\mu{}^\mu = h_{kk} = 0$, *there is no distinction between $h_{\mu\nu}$ and $\bar{h}_{\mu\nu}$ in this gauge.*

(2) for any wave

Turn attention now away from plane waves to arbitrary gravitational waves in linearized theory. Any electromagnetic wave can be resolved into a superposition of plane waves, and so can any gravitational wave. For each plane wave in the superposition, introduce the special gauge (35.8). Note that the gauge conditions are all linear in $h_{\mu\nu}$. Therefore the arbitrary wave will also satisfy conditions (35.8). Thus arises the theorem: *Pick a specific global Lorentz frame of linearized theory (i.e., pick a specific 4-velocity* $\boldsymbol{u}$*). In that frame (where $u^\alpha = \delta^\alpha{}_0$), examine a specific gravitational wave of arbitrary form. One can always find a gauge in which $h_{\mu\nu}$ satisfies the constraints* (35.8). Moreover, in this gauge only the h_{jk} are nonzero. Therefore, one need only impose the six wave equations

$$\Box h_{jk} = h_{jk,\alpha}{}^{,\alpha} = 0. \tag{35.9}$$

Decomposition of spatial tensors

Any symmetric tensor satisfying constraints (35.8) [but not necessarily the wave equations (35.9)] is called a *transverse-traceless* (*TT*) *tensor*—transverse because it is purely spatial ($h_{0\mu} = 0$) and, if thought of as a wave, is transverse to its own direction of propagation ($h_{ij,j} = h_{ij}k_j = 0$); traceless because $h_{kk} = 0$. The most general purely spatial tensor S_{ij} can be decomposed [see Arnowitt, Deser, and Misner (1962) or Box 35.1] into a part S_{ij}^{TT}, which is "transverse and traceless"; a part $S_{ij}^{T} = \frac{1}{2}(\delta_{ij}f_{,kk} - f_{,ij})$, which is "transverse" ($S_{ij,j}^{T} = 0$) but is determined entirely by one function f giving the trace of S ($S_{kk}^{T} = \nabla^2 f$); and a part $S_{ij}^{L} = S_{i,j}^{L} + S_{j,i}^{L}$, which is "longitudinal" and is determined by the vector field S_i^L. In linearized theory h_{ij}^L is a purely gauge part of $h_{\mu\nu}$, whereas h_{ij}^T and h_{ij}^{TT} are gauge-invariant parts of $h_{\mu\nu}$. The special gauge in which $h_{\mu\nu}$ reduces to its transverse-traceless part is called the *TT* or transverse-traceless gauge. The conditions (35.8) defining this gauge can be summarized as

$$h_{\mu\nu} = h_{\mu\nu}^{TT}. \tag{35.8d}$$

As exercise 35.2 illustrates, only pure waves (and not more general solutions of the linearized field equations with source, $\Box h_{\mu\nu} = -16\pi T_{\mu\nu}$) can be reduced to *TT* gauge.

Curvature tensor in TT gauge

In the TT gauge, the time-space components

$$R_{j0k0} = R_{0j0k} = -R_{j00k} = -R_{0jk0}$$

of the Riemann curvature tensor have an especially simple form [see equation (18.9) and exercise 18.4]:

$$R_{j0k0} = -\frac{1}{2} h^{TT}_{jk,00}. \tag{35.10}$$

Recall that the curvature tensor is gauge-invariant (exercise 18.1). It follows that $h_{\mu\nu}$ cannot be reduced to still fewer components than it has in the TT gauge.

Box 35.1 describes methods to calculate $h^{TT}_{\mu\nu}$ from a knowledge of $h_{\mu\nu}$ in some other gauge.

Box 35.1 METHODS TO CALCULATE "TRANSVERSE-TRACELESS PART" OF A WAVE

Problem: Let a gravitational wave $h_{\mu\nu}(t, x^j)$ in an arbitrary gauge of linearized theory be known. How can one calculate the metric perturbation $h^{TT}_{\mu\nu}(t, x^j)$ for this wave in the transverse-traceless gauge?

Solution 1 (valid only for waves; i.e., when $\Box\bar{h}_{\mu\nu} = 0$). Calculate the components R_{j0k0} of ***Riemann*** in the initial gauge; then integrate equation (35.10)

$$h^{TT}_{jk,00} = -2R_{j0k0} \tag{1}$$

to obtain h^{TT}_{jk}. When the wave is monochromatic, $h_{\mu\nu} = h_{\mu\nu}(x^l)e^{-i\omega t}$; then the solution of (1) has the simple form

$$h^{TT}_{jk} = 2\omega^{-2}R_{j0k0}. \tag{2}$$

Solution 2 (valid only for plane waves). "Project out" the TT components in an algebraic manner using the operator

$$P_{jk} = \delta_{jk} - n_j n_k. \tag{3}$$

Here

$$n_k = k_k/|\boldsymbol{k}|$$

is the unit vector in the direction of propagation. Verify that P_{jk} is a projection operator onto the transverse plane:

$$P_{jl}P_{lk} = P_{jk}, \qquad P_{jk}n_k = 0, \qquad P_{kk} = 2.$$

Then the transverse part of h_{jk} is $P_{jl}h_{lm}P_{mk}$ (or in matrix notation, PhP); and the TT part is this quantity diminished by its trace:

$$h^{TT}_{jk} = P_{jl}P_{mk}h_{lm} - \frac{1}{2}P_{jk}(P_{ml}h_{lm}) \tag{4}$$

(index notation),

$$h^{TT} = PhP - \frac{1}{2}P\,\mathrm{Tr}\,(Ph) \tag{4'}$$

(matrix notation).

The sequence of operations that gives h^{TT}_{ij} cuts two parts out of h_{ij}. The first part cut out is

$$h^{T}_{jk} = \frac{1}{2}P_{jk}(P_{lm}h_{lm}), \tag{5}$$

which is transverse but is built from its own trace,

$$h^{T} \equiv \mathrm{Tr}\,(PhP) = \mathrm{Tr}\,(Ph) = P_{lm}h_{ml}.$$

EXERCISES

Exercise 35.1. TRANSFORMATION OF PLANE WAVE TO TT GAUGE

Let a plane wave of the form (35.4) be given, in some arbitrary gauge of linearized theory. Exhibit explicitly the transformation that puts it into the TT gauge. [*Hint:* Work in a Lorentz frame where the 4-velocity u^μ of the TT gauge is $u^0 = 1$, $u^j = 0$. Solve for the four constants C^μ of the generating function (35.6) by demanding that $\bar{h}_{\mu\nu}$ satisfy the TT constraints (35.7).]

Exercise 35.2. LIMITATION ON EXISTENCE OF TT GAUGE

Although the metric perturbation $h_{\mu\nu}$ for any *gravitational wave* in linearized theory can be put into the TT form (35.8), nonradiative $h_{\mu\nu}$'s cannot. Consider, for example, the external field of a rotating, spherical star, which cannot be written as a superposition of plane waves:

The second part cut out of h_{ij} is the longitudinal part

$$\begin{aligned} h^L{}_{jk} &= h_{jk} - P_{jl}P_{mk}h_{lm} \\ &= n_l n_k h_{jl} + n_j n_l h_{lk} - n_j n_k(n_l n_m h_{lm}); \end{aligned} \tag{6}$$

or

$$h^L = h - PhP \tag{6'}$$

Solution 3 (general case). Fourier analyze any symmetric array $h_{ij} = \int h_{ij}(k, t) \exp(ik_m x^m)\, d^3k$, and apply the formulas (4) from solution 2 to each Fourier component individually. But note that in this case one can write the projection operator in the direction-independent form

$$P_{jk} = \delta_{jk} - \frac{1}{\nabla^2}\partial_j\,\partial_k \tag{7}$$

or

$$n_l n_m = \frac{1}{\nabla^2}\partial_l\,\partial_m \tag{8}$$

(provided the formulas are written with all h's standing on the right), since $\partial_l = ik_l$ under the Fourier integral. Of course the operation $1/\nabla^2$ can be evaluated by other methods, e.g., by Green's functions, as well as by Fourier analysis. [The quantity $\psi \equiv \nabla^{-2}f$ stands for the solution ψ of the Poisson equation $\nabla^2\psi = f$.] The advantage of this method is its power in certain analytic computations (see, e.g., below).

Gauge Transformations. The change in $h_{\mu\nu}$ due to a gauge transformation is

$$\delta h_{\mu\nu} = -(\partial_\nu\xi_\mu + \partial_\mu\xi_\nu). \tag{9}$$

The transverse part of this change is

$$P_{jl}P_{km}(\delta h_{lm}) = -P_{jl}P_{km}(\partial_l\xi_m + \partial_m\xi_l) = 0. \tag{10}$$

To verify this formula for a plane wave (solution 2), note that $\partial_l = i|\boldsymbol{k}|n_l$ and $P_{jl}n_l = 0$. To verify the same result in general, use equation (7) to give the result

$$P_{jl}\,\partial_l = 0. \tag{11}$$

Thus both h_{ij}^{TT} of equation (4), and h_{ij}^T of equation (5) are gauge-invariant:

$$\delta h_{ij}^{TT} = \delta h_{ij}^T = 0. \tag{12}$$

In empty space ($T_{\mu\nu} = 0$), both h_{ij}^T and another gauge-invariant quantity $\tilde{h}_{0k}$ (discussed in exercise 35.4) vanish, by virtue of the field equations.

$$h_{00} = \frac{2M}{r}, \qquad h_{jk} = \frac{2M}{r}\,\delta_{jk}, \qquad h_{0k} = -2\epsilon_{k\ell m}\frac{S^\ell x^m}{r^3},$$
$$r = (x^2 + y^2 + z^2)^{1/2}$$

[see equation (19.5)]. Here M is the star's mass and $\boldsymbol{S}$ is its angular momentum. Show that this *cannot* be put into TT gauge. [Hint: Calculate R_{j0k0} and from it, by means of (35.10), infer h_{jk}^{TT}. Then calculate R_{0xyz} in both the original gauge and the new gauge, and discover that they disagree—not only by virtue of the mass term, but also by virtue of the angular-momentum term.]

Exercise 35.3. A CYLINDRICAL GRAVITATIONAL WAVE

To restore one's faith, which may have been shaken by exercise 35.2, one can consider the radiative solution whose only nonvanishing component $h_{\mu\nu}$ is

$$\bar{h}_{zz} = 4A\cos(\omega t)J_0(\omega\sqrt{x^2 + y^2}),$$

where J_0 is the Bessel function. This solution represents a superposition of ingoing and outgoing cylindrical gravitational waves. For this gravitational field calculate R_{j0k0}, and from it infer h_{jk}^{TT}. Then calculate several other components of $R_{\alpha\beta\gamma\delta}$ (e.g., R_{xyxy}) in the original gauge and in TT gauge, and verify that the answers are the same.

Exercise 35.4. NON-TT PARTS OF METRIC PERTURBATION [Track 2]

From Box 35.1 establish the formula $h^T = \nabla^{-2}(h_{kk,\ell\ell} - h_{k\ell,k\ell})$; then verify the gauge invariance of h^T directly, by showing that $h_{kk,\ell\ell} - h_{k\ell,k\ell}$ is gauge-invariant. Use $\delta h_{ij} = \xi_{i,j} + \xi_{j,i}$. Show similarly that the quantities $\tilde{h}_{0k}$ defined by

$$\tilde{h}_{0k} = \bar{h}_{0k} - \nabla^{-2}(\bar{h}^{\mu}_{0,\mu k} + \bar{h}_{k\ell,\ell 0})$$

are gauge-invariant. Show from the gauge-invariant linearized field equations (18.5) that

$$\nabla^2 h^T = -16\pi T^{00},$$
$$\nabla^2 \tilde{h}_{0k} = -16\pi T_{0k},$$

so h^T and $\tilde{h}_{0k}$ must vanish for waves in empty space.

§35.5. GEODESIC DEVIATION IN A LINEARIZED GRAVITATIONAL WAVE

Action of a gravitational wave on separation of two test particles

The oscillating curvature tensor of a gravitational wave produces oscillations in the separation between two neighboring test particles, A and B. Examine the oscillations from the viewpoint of A. Use a coordinate system ("*proper reference frame of A*"), with spatial origin $x^{\hat{j}} = 0$, attached to A's world line (comoving coordinates); with coordinate time equal to A's proper time ($x^{\hat{0}} = \tau$ on world line $x^{\hat{j}} = 0$); and with orthonormal spatial axes attached to gyroscopes carried by A ("nonrotating frame"). This coordinate system, appropriately specialized, is a local Lorentz frame not just at one event $\mathscr{P}_0$ on A's geodesic world line, but all along A's world line:

$$ds^2 = -dx^{\hat{0}^2} + \delta_{\hat{j}\hat{k}}\,dx^{\hat{j}}\,dx^{\hat{k}} + O(|x^{\hat{j}}|^2)\,dx^{\hat{\alpha}}\,dx^{\hat{\beta}}. \tag{35.11}$$

[Proof: such a "proper reference frame" was set up for accelerated particles in Track 2's §13.6. The line element (13.71) derived there, when specialized to particle A ($a_{\hat{j}} = 0$ because A falls freely; $\omega^{\hat{i}} = 0$ because the spatial axes are attached to gyroscopes) reduces to the above form, as in equation (13.73).]

As the gravitational wave passes, it produces an oscillating curvature tensor, which wiggles the separation vector $\boldsymbol{n}$ reaching from particle A to particle B:

$$D^2 n^{\hat{j}}/d\tau^2 = -R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}} n^{\hat{k}} = -R_{\hat{j}\hat{0}\hat{k}\hat{0}} n^{\hat{k}}. \tag{35.12}$$

The components of the separation vector are nothing but the coordinates of particle B, since particle A is at the origin of its own proper reference frame; thus,

$$n^{\hat{j}} = x_B{}^{\hat{j}} - x_A{}^{\hat{j}} = x_B{}^{\hat{j}}.$$

Moreover, at $x^{\hat{j}} = 0$ [where the calculation (35.12) is being performed], the $\Gamma^{\hat{\mu}}{}_{\hat{\alpha}\hat{\beta}}$ vanish for all $x^{\hat{0}}$; so $d\Gamma^{\hat{\mu}}{}_{\hat{\alpha}\hat{\beta}}/d\tau$ also vanish. This eliminates all Christoffel-symbol corrections in $D^2 n^{\hat{j}}/D\tau^2$. Hence, equation (35.12) reduces to

$$d^2 x_B{}^{\hat{j}}/d\tau^2 = -R_{\hat{j}\hat{0}\hat{k}\hat{0}} x_B{}^{\hat{k}}. \tag{35.13}$$

There is a TT coordinate system that, to first order in the metric perturbation h_{jk}^{TT}, moves with particle A and with its proper reference frame. To first order in h_{jk}^{TT}, the TT coordinate time t is the same as proper time τ, and $R_{j0k0}^{TT} = R_{\hat{j}\hat{0}\hat{k}\hat{0}}$. Hence, equation (35.13) can be rewritten

$$d^2 x_B{}^{\hat{j}}/dt^2 = -R_{j0k0}^{TT} x_B{}^{\hat{k}} = \frac{1}{2}(\partial^2 h_{jk}^{TT}/\partial t^2) x_B{}^{\hat{k}}. \tag{35.14}$$

Suppose, for concreteness, that the particles are at rest relative to each other before the wave arrives ($x_B{}^{\hat{j}} = x_{B(0)}{}^{\hat{j}}$ when $h_{jk}^{TT} = 0$). Then the equation of motion (35.14) can be integrated to yield

$$x_B{}^{\hat{j}}(\tau) = x_{B(0)}{}^{\hat{k}} \left[\delta_{jk} + \frac{1}{2} h_{jk}^{TT} \right]_{\text{at position of } A}. \tag{35.15}$$

This equation describes the wave-induced oscillations of B's location, as measured in the proper reference frame of A.

Transverse character of relative accelerations

Turn to the special case of a plane wave. Suppose the test-particle separation lies in the direction of propagation of the wave. Then the wave cannot affect the separation; there is no oscillation:

$$h_{jk}^{TT} x_{B(0)}{}^{\hat{k}} \propto h_{jk}^{TT} k_k = 0.$$

Only separations in the transverse direction oscillate; *the wave is transverse not only in its mathematical description (h_{jk}^{TT}), but also in its physical effects (geodesic deviation)*!

EXERCISE

Exercise 35.5. ALTERNATIVE CALCULATION OF RELATIVE OSCILLATIONS

Introduce a *TT* coordinate system in which, at time $t = 0$, the two particles are both at rest. Use the geodesic equation to *show* that subsequently they both always remain at rest in the *TT* coordinates, despite the action of the wave. This means that the contravariant components of the separation vector are always constant in the *TT* coordinate frame:

$$n^j = x_B{}^j - x_A{}^j = \text{const.}$$

Call this constant $x_{B(0)}{}^j$. *Transform* these components to the comoving orthonormal frame; the answer should be equation (35.15).

§35.6. POLARIZATION OF A PLANE WAVE

Polarization of gravitational waves:

(1) States of linear polarization, "+" and "×"

Geodesic deviation in the transverse direction provides a means for studying and characterizing the polarizations of plane waves.

Consider a plane, monochromatic wave propagating in the z direction. In the *TT* gauge the constraints $h^{TT}_{0\mu} = 0$, $h^{TT}_{ij,j} \equiv ik_j h^{TT}_{ij} = 0$, and $h^{TT}_{kk} = 0$ reveal that the only nonvanishing components of $h^{TT}_{\mu\nu}$ are

$$\begin{aligned} h^{TT}_{xx} &= -h^{TT}_{yy} = \Re\{A_+ e^{-i\omega(t-z)}\}, \\ h^{TT}_{xy} &= h^{TT}_{yx} = \Re\{A_\times e^{-i\omega(t-z)}\}. \end{aligned} \tag{35.16}$$

The amplitudes A_+ and $A_\times$ represent two independent modes of polarization.

As for electromagnetic plane waves (Figure 35.1), so also for gravitational plane waves (Figure 35.2), one can resolve a given wave into two linearly polarized components, or, alternatively, into two circularly polarized components.

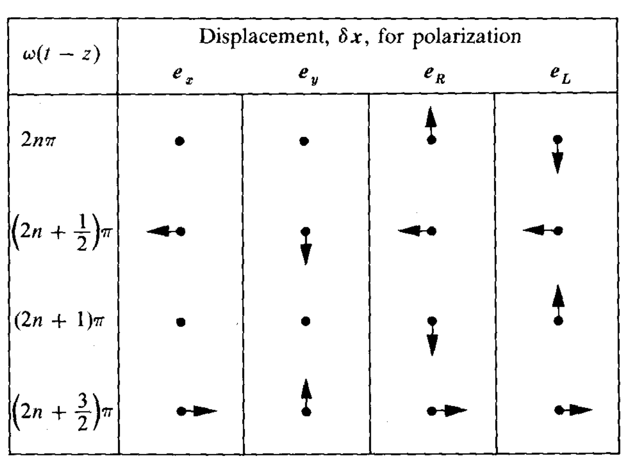


Figure 35.1.
Plane Electromagnetic Waves.
Polarization vector: $\boldsymbol{e}_P$
Vector Potential

$$\boldsymbol{A} = \Re[A_0 e^{-i\omega(t-z)}\boldsymbol{e}_P]$$

Acceleration of a test charge:

$$\begin{aligned} \boldsymbol{a} &= (q/m)\boldsymbol{E} = (q/m)(-\partial \boldsymbol{A}/\partial t) \\ &= \Re[i\omega(q/m)A_0 e^{-i\omega(t-z)}\boldsymbol{e}_P] \end{aligned}$$

Displacement of charge relative to inertial frame:

$$\delta \boldsymbol{x} = \Re\left[\frac{q/m}{i\omega} A_0 e^{-i\omega(t-z)}\boldsymbol{e}_P\right]$$

For *linearly polarized waves,* the unit polarization vectors of electromagnetic theory are $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$. A test charge hit by a plane wave with polarization vector $\boldsymbol{e}_x$ oscillates in the x-direction relative to an inertial frame; and similarly for $\boldsymbol{e}_y$. By analogy, the *unit linear-polarization tensors* for gravitational waves are

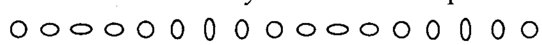
$$\boldsymbol{e}_+ \equiv \boldsymbol{e}_x \otimes \boldsymbol{e}_x - \boldsymbol{e}_y \otimes \boldsymbol{e}_y, \tag{35.17a}$$

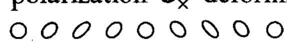
$$\boldsymbol{e}_\times \equiv \boldsymbol{e}_x \otimes \boldsymbol{e}_y + \boldsymbol{e}_y \otimes \boldsymbol{e}_x. \tag{35.17b}$$

The plane wave (35.16), when $A_\times = 0$, has polarization $\boldsymbol{e}_+$ and can be rewritten

$$h_{jk} = \Re\{A_+ e^{-i\omega(t-z)} e_{+jk}\}. \tag{35.18}$$

Its effect in altering the geodesic separation between two test particles depends on the direction of their separation. To see the effect in all directions at once, consider a circular ring of test particles in the transverse (x, y) plane, surrounding a central particle (Figure 35.2). As the plane wave (35.18) (polarization $\boldsymbol{e}_+$) passes, it deforms what was a ring as measured in the proper reference frame of the central particle into an ellipse with axes in the x and y directions that pulsate in and out:



etc. By contrast (Figure 35.2), a wave of polarization $\boldsymbol{e}_\times$ deforms the ring at a 45-degree angle to the x and y directions:  etc.

(2) States of circular polarization

For circularly polarized waves, the unit polarization vectors of electromagnetic theory are

$$e_R = \frac{1}{\sqrt{2}}(e_x + ie_y), \qquad e_L = \frac{1}{\sqrt{2}}(e_x - ie_y) \tag{35.19}$$

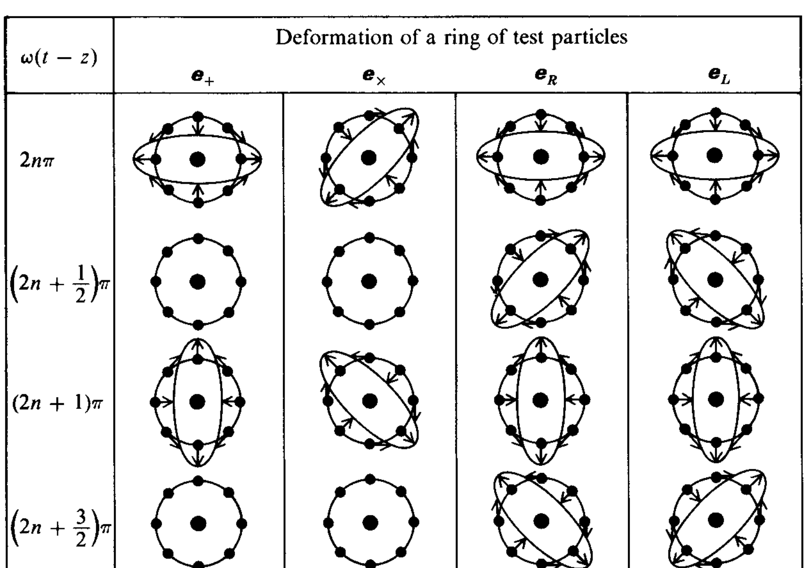


Figure 35.2.
Plane Gravitational Waves. Polarization tensor:

$$\boldsymbol{e}_P$$

Metric perturbation:

$$h_{jk} = \Re\,[A_0 e^{-i\omega(t-z)} e_{Pjk}]$$

Tidal acceleration between two test particles:

$$\frac{D^2 n_{\hat{j}}}{D\tau^2} = -R_{\hat{j}0\hat{k}0} n_{\hat{k}} = \frac{1}{2}\frac{\partial^2 h_{jk}}{\partial t^2} n_k = \Re\left[-\frac{1}{2}\omega^2 A_0 e^{-i\omega(t-z)} e_{Pjk} n_{\hat{k}}\right]$$

Separation between two test particles:

$$n_{\hat{j}} = n_{\hat{j}}^{(0)} + \Re\left[\frac{1}{2} A_0 e^{-i\omega(t-z)} e_{Pjk} n_{\hat{k}}^{(0)}\right]$$

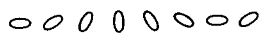
Position of test particle B in proper reference frame of test particle A. (In drawing, A is the central particle and B is any peripheral particle):

$$x_B{}^{\hat{j}} = x_{B(0)}{}^{\hat{j}} + \Re\left[\frac{1}{2} A_0 e^{-i\omega(t-z)} e_{Pjk} x_{B(0)}{}^{\hat{k}}\right]$$

Similarly, the *unit circular polarization tensors* of gravitation theory are

$$\boldsymbol{e}_R = \frac{1}{\sqrt{2}}(\boldsymbol{e}_+ + i\boldsymbol{e}_\times), \qquad \boldsymbol{e}_L = \frac{1}{\sqrt{2}}(\boldsymbol{e}_+ - i\boldsymbol{e}_\times). \tag{35.20}$$

A test charge hit by an electromagnetic wave of polarization $\boldsymbol{e}_R$ moves around and around in a circle in the righthanded direction (counterclockwise for a wave propagating toward the reader); for $\boldsymbol{e}_L$ it circles in the lefthanded (clockwise) direction (see Figure 35.1). Similarly (Figure 35.2), a gravitational wave of polarization $\boldsymbol{e}_R$ rotates the deformation of a test-particle ring in the righthanded direction,



while a wave of $\boldsymbol{e}_L$ rotates it in the lefthanded direction. The individual test particles in the ring rotate in small circles relative to the central particle. However, just as the drops in an ocean wave do not move along with the wave, so the particles on the ring do not move *around* the central particle with the rotating ellipse.

Spin-2 character of gravitational field and its relation to symmetries of waves

Notice from Figure 35.2 that, at any moment of time, a gravitational wave is invariant under a rotation of 180° about its direction of propagation. The analogous angle for electromagnetic waves (Figure 35.1) is 360°, and for neutrino waves it is 720°. This behavior is intimately related to the spin of the zero-mass particles associated with the quantum-mechanical versions of these waves: gravitons have spin 2, photons spin 1, and neutrinos spin 1/2. The classical radiation field of a spin-S particle is always invariant under a rotation of $360°/S$ about its propagation direction.

A radiation field of any spin S has precisely two orthogonal states of linear polarization. They are inclined to each other at an angle of $90°/S$; thus, for a neutrino field, with $S = \frac{1}{2}$, the two states are distinguished as $|\uparrow\rangle$ and $|\downarrow\rangle$ (spin up and spin down; 180° angle). For an electromagnetic wave $S = 1$ and two orthogonal states of polarization are $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$ (90° angle). For a gravitational wave $S = 2$, and two orthogonal states are $\boldsymbol{e}_+$ and $\boldsymbol{e}_\times$ (45° angle).

EXERCISES

Exercise 35.6. ROTATIONAL TRANSFORMATIONS FOR POLARIZATION STATES

Consider two Lorentz coordinate systems, one rotated by an angle θ about the z direction relative to the other:

$$t' = t, \qquad x' = x\cos\theta + y\sin\theta, \qquad y' = y\cos\theta - x\sin\theta, \qquad z' = z. \tag{35.21}$$

Let $|\uparrow\rangle$ and $|\downarrow\rangle$ be quantum-mechanical states of a neutrino with spin-up and spin-down relative to the x direction; and similarly for $|\uparrow'\rangle$ and $|\downarrow'\rangle$. Let $\boldsymbol{e}_x$, $\boldsymbol{e}_y$, $\boldsymbol{e}_{x'}$, $\boldsymbol{e}_{y'}$ be the unit polarization vectors in the two coordinate systems for an electromagnetic wave traveling in the z-direction; and similarly $\boldsymbol{e}_+$, $\boldsymbol{e}_\times$, $\boldsymbol{e}_{+'}$, $\boldsymbol{e}_{\times'}$ for a gravitational wave in linearized theory. Derive the following transformation laws:

$$\begin{aligned}
|\uparrow'\rangle &= |\uparrow\rangle \cos\tfrac{1}{2}\theta + |\downarrow\rangle \sin\tfrac{1}{2}\theta; & |\downarrow'\rangle &= -|\uparrow\rangle \sin\tfrac{1}{2}\theta + |\downarrow\rangle \cos\tfrac{1}{2}\theta;\\
\boldsymbol{e}_{x'} &= \boldsymbol{e}_x\cos\theta + \boldsymbol{e}_y\sin\theta; & \boldsymbol{e}_{y'} &= -\boldsymbol{e}_x\sin\theta + \boldsymbol{e}_y\cos\theta; \qquad (35.22)\\
\boldsymbol{e}_{+'} &= \boldsymbol{e}_+\cos 2\theta + \boldsymbol{e}_\times\sin 2\theta; & \boldsymbol{e}_{\times'} &= -\boldsymbol{e}_+\sin 2\theta + \boldsymbol{e}_\times\cos 2\theta.
\end{aligned}$$

What is the generalization to the linear-polarization basis states for a radiation field of arbitrary spin S?

Exercise 35.7. ELLIPTICAL POLARIZATION

Discuss elliptically polarized gravitational waves in a manner analogous to the discussion of linearly and circularly polarized waves in Figure 35.2.

---

§35.7. THE STRESS-ENERGY CARRIED BY A GRAVITATIONAL WAVE

Exercise 18.5 showed that in principle one can build detectors which extract energy from gravitational waves. Hence, it is clear that the waves must carry energy.

Unfortunately, to derive and justify an expression for their energy requires a somewhat more sophisticated viewpoint than linearized theory. Such a viewpoint will be developed later in this chapter (§§35.13 and 35.15). But for the benefit of Track-1 readers, the key result is stated here.

Approximate localization of energy in a gravitational wave

In accordance with the discussions in §§19.4 and 20.4, the stress-energy carried by gravitational waves cannot be localized inside a wavelength. One cannot say whether the energy is carried by the crest of a wave, by its trough, or by its "walls." However, one *can* say that a certain amount of stress-energy is contained in a given "macroscopic" region (region of several wavelengths' size), and one can thus talk about a tensor for an *effective* smeared-out stress-energy of gravitational waves, $T^{(\mathrm{GW})}_{\mu\nu}$. In a (nearly) inertial frame of linearized theory, $T^{(\mathrm{GW})}_{\mu\nu}$ is given by

Effective stress-energy tensor for gravitational waves:

(1) expressed in terms of metric perturbations

$$T^{(\mathrm{GW})}_{\mu\nu} = \frac{1}{32\pi}\langle h^{TT}_{jk,\mu} h^{TT}_{jk,\nu}\rangle, \tag{35.23}$$

where $\langle\ \rangle$ denotes an average over several wavelengths and h^{TT}_{jk} means the (gauge-invariant) transverse-traceless part of $h_{\mu\nu}$, which is simply h_{jk} in the TT gauge. Another formula for $T^{(\mathrm{GW})}_{\mu\nu}$, valid in any arbitrary gauge, with $\bar{h} \neq 0$, $\bar{h}_{\mu}{}^{\alpha}{}_{,\alpha} \neq 0$, and $\bar{h}_{0\mu} \neq 0$ is

$$T^{(\mathrm{GW})}_{\mu\nu} = \frac{1}{32\pi}\left\langle \bar{h}_{\alpha\beta,\mu}\bar{h}^{\alpha\beta}{}_{,\nu} - \frac{1}{2}\,\bar{h}_{,\mu}\bar{h}_{,\nu} - \bar{h}^{\alpha\beta}{}_{,\beta}\bar{h}_{\alpha\mu,\nu} - \bar{h}^{\alpha\beta}{}_{,\beta}\bar{h}_{\alpha\nu,\mu}\right\rangle \tag{35.23$'$}$$

This stress-energy tensor, like any other, is divergence-free in vacuum

(2) subject to conservation law

$$T^{(\mathrm{GW})}{}_{\mu}{}^{\nu}{}_{,\nu} = 0; \tag{35.24}$$

and it contributes to the large-scale background curvature (which linearized theory ignores) just as any other stress-energy does:

(3) role as source of background curvature

$$G^{(\mathrm{B})}_{\mu\nu} = 8\pi(T^{(\mathrm{GW})}_{\mu\nu} + T^{(\mathrm{matter})}_{\mu\nu} + T^{(\mathrm{other\ fields})}_{\mu\nu}). \tag{35.25}$$

In writing here the term $T^{(\mathrm{GW})}_{\mu\nu}$ for the effective smeared-out energy density of the gravitational wave, one is foregoing any further insertion of the gravitational wave into the Einstein equation. Otherwise one might end up counting twice over the contribution of the same wave to the background curvature of space, even though expressed in very different formalisms.

(4) for a plane, monochromatic wave

According to equation (35.23), the stress-energy tensor for the plane wave,

$$h_{\mu\nu} = \Re\{(A_+ e_{+\mu\nu} + A_\times e_{\times\mu\nu})e^{-i\omega(t-z)}\}, \tag{35.26}$$

is

$$T_{tt}^{(\mathrm{GW})} = T_{zz}^{(\mathrm{GW})} = -T_{tz}^{(\mathrm{GW})} = \frac{1}{32\pi}\omega^2(|A_+|^2 + |A_\times|^2). \tag{35.27}$$

Notice that the background radius of curvature $\mathcal{R}$ (ignored by linearized theory), and the mean reduced wavelength $\lambda\!\!\!^{-}$ (= wavelength/2π) and amplitude $\mathcal{A}$ of the gravitational waves satisfy

$$\begin{aligned} \mathcal{R}^{-2} &\sim \text{typical magnitude of components of } R^{(\mathrm{B})}_{\alpha\beta\gamma\delta} \\ &\sim T_{\mu\nu}^{(\mathrm{GW})} \sim \mathcal{A}^2/\lambda\!\!\!^{-2} \text{ if } T_{\mu\nu}^{(\mathrm{GW})} \text{ is chief source of background curvature,} \\ &\gg T_{\mu\nu}^{(\mathrm{GW})} \sim \mathcal{A}^2/\lambda\!\!\!^{-2} \text{ if } T_{\mu\nu}^{(\mathrm{GW})} \text{ is not chief source.} \end{aligned}$$

Consequently, the dimensionless numbers $\mathcal{A}$ and $\lambda\!\!\!^{-}/\mathcal{R}$ are related by

$$\mathcal{A} \lesssim \lambda\!\!\!^{-}/\mathcal{R}. \tag{35.28}$$

Conditions for validity of gravitational-wave formalism

Thus, *the whole concept of a small-scale ripple propagating in a background of large-scale curvature breaks down, and the whole formalism of this chapter becomes meaningless, if the dimensionless amplitude of the wave approaches unity. One must always have $\mathcal{A} \ll 1$ as well as $\lambda\!\!\!^{-} \ll \mathcal{R}$ if the concept of a gravitational wave is to make any sense!*

§35.8. GRAVITATIONAL WAVES IN THE FULL THEORY OF GENERAL RELATIVITY

Nonlinear effects in gravitational waves:

Curving up of the background spacetime by the energy of the waves is but one of many new effects that enter, when one passes from linearized theory to the full, nonlinear general theory of relativity.

(1) radiation damping

In linearized theory one can consider a localized source of gravitational waves (e.g., a vibrating bar) in steady oscillation, radiating a strictly periodic wave. But the exact theory insists that the energy of the source decrease secularly, to counterbalance the energy carried off by the radiation (energy conservation; gravitational radiation damping; see §§36.8 and 36.11). This makes an exactly periodic wave impossible, though a very nearly periodic one can certainly be emitted [Papapetrou (1958); Arnowitt, Deser, and Misner as reported by Misner (1964b)].

(2) refraction

In the real universe there are spacetime curvatures due not only to the energy of gravitational waves, but also, and more importantly, to the material content of the universe (planets, stars, gas, galaxies). As a gravitational wave propagates through these curvatures, its wave fronts change shape ("refraction"), its wavelength changes (gravitational redshift), and it backscatters off the curvatures to some extent. If the wave is a pulse, the backscatter will cause its shape and polarization to keep changing and will produce "tails" that spread out behind the moving pulse, traveling slower than light [see exercise 32.10; also Riesz (1949), DeWitt and Brehme (1960), DeWitt and DeWitt (1964a), Kundt and Newman (1968), Couch *et. al.* (1968)]. However, so long as $\mathcal{A} \ll 1$ and $\lambdabar/\mathcal{R} \ll 1$, these effects will be extremely small locally. They can only build up over distances of the order of $\mathcal{R}$—and sometimes not even then. Thus, locally, linearized theory will remain highly accurate.

(3) redshift

(4) backscatter

(5) tails

Even in an idealized universe containing nothing but gravitational waves, backscatter and tails are produced by the interaction of the waves with the background curvature that they themselves produce.

If the reduced wavelength $\lambdabar = \lambda/2\pi$ and the mass-energy m of a pulse of waves satisfy $\lambdabar \ll m$, it is possible (in principle) to focus the pulse into a region of size $r < m$, whereupon a part of the energy of the pulse will undergo gravitational collapse to a singularity, leaving behind a black hole [see Ruffini and Wheeler (1970), and pp. 7–24 of Christodoulou (1971)]. Short of a certain critical strength, no part of the pulse undergoes such a collapse. But it does undergo a time delay before reexpanding. This time delay is definable and measurable in the asymptotically flat space, far from the domain where the energy a little earlier underwent temporary focusing into dimensions of order $\lambdabar$.

(6) self-gravitational attraction

All these effects can be analyzed in general relativity theory using approximation schemes which, in first order, are similar to or identical to linearized theory. Later in this chapter (§§35.13–35.15), one such approximation scheme will be developed. But first it is helpful to study an exact solution that exhibits some of these effects.

§35.9. AN EXACT PLANE-WAVE SOLUTION

The rest of this chapter is Track 2. No earlier Track-2 material is needed as preparation for it, but Chapter 20 (conservation laws) and §22.5 (geometric optics) will be found to be helpful. It is not needed as preparation for any later chapter.

Any exact gravitational-wave solution that can be given in closed mathematical form must be highly idealized; otherwise it could not begin to cope with the complexities outlined above. Consequently, mathematically exact solutions are useful for pedagogical purposes only. However, pedogogy should not be condemned: it is needed not only by students, but also by veteran workers in the field of relativity, who even today are only beginning to develop intuition into the nonlinear regime of geometrodynamics!

From the extensive literature on exact solutions, we have chosen, as a compromise between realism and complexity, the following plane wave [Bondi *et. al.* (1959), Ehlers and Kundt (1962)]:

Exact plane-wave solution of vacuum field equation:

(1) form of metric

$$\begin{aligned} ds^2 &= L^2(e^{2\beta}\,dx^2 + e^{-2\beta}\,dy^2) + dz^2 - dt^2 \\ &= L^2(e^{2\beta}\,dx^2 + e^{-2\beta}\,dy^2) - du\,dv. \end{aligned} \tag{35.29a}$$

Here

$$u = t - z, \qquad v = t + z, \qquad L = L(u), \qquad \beta = \beta(u). \tag{35.29b}$$

2

The forms that the functions $L(u)$ ("background factor") and $\beta(u)$ ("wave factor") can take are determined by the vacuum field equations. In the null coordinate system u, v, x, y, the only component of the Ricci tensor that does not vanish identically is (see Box 14.4, allowing for the difference in coordinates, $2v_{\text{there}} = v_{\text{here}}$)

$$R_{uu} = -2L^{-1}[L'' + (\beta')^2 L], \tag{35.30}$$

where the prime denotes d/du. Thus, Einstein's equations in vacuum read

(2) generation of "background factor" L by "wave factor" β

$$L'' + (\beta')^2 L = 0. \tag{35.31}$$

("effect of wave factor on background factor")

(3) linearized limit

The linearized version of this equation is $L'' = 0$, since $(\beta')^2$ is a second-order quantity. Therefore the solution corresponding to linearized theory is

$$L = 1, \quad \beta(u) \text{ arbitrary but small.}$$

The corresponding metric is

$$ds^2 = (1 + 2\beta)\, dx^2 + (1 - 2\beta)\, dy^2 + dz^2 - dt^2, \qquad \beta = \beta(t - z). \tag{35.32}$$

Notice that this is a plane wave of polarization $\boldsymbol{e}_+$ propagating in the z-direction. (See exercise 35.10 at end of §35.12 for the extension to a wave possessing both polarizations, $\boldsymbol{e}_+$ and $\boldsymbol{e}_\times$.)

(4) special case: a plane-wave pulse

Return attention to the exact plane wave, and focus on the case where the "wave factor" $\beta(u)$ is a pulse of duration $2T$, and $|\beta'| \ll 1/T$ throughout the pulse. Then the exact solution (Figure 35.3) is: (1) for $u < -T$ (flat spacetime; pulse has not yet arrived),

$$\beta = 0, \qquad L = 1; \tag{35.33a}$$

(2) for $-T < u < +T$ (interior of pulse),

$$\beta = \beta(u) \text{ is arbitrary, except that } |\beta'| \ll 1/T,$$

$$L(u) = 1 - \int_{-T}^{u} \left\{ \int_{-T}^{\bar{u}} [\beta'(\bar{\bar{u}})]^2 \, d\bar{\bar{u}} \right\} d\bar{u} + O([\beta' T]^4); \tag{35.33b}$$

(3) for $u > T$ (after the pulse has passed),

$$\beta = 0, \qquad L = 1 - \frac{u}{a}, \qquad a \equiv \frac{1}{\int_{-T}^{T} (\beta')^2\, du} + \frac{O([\beta' T]^2)}{\int_{-T}^{T} (\beta')^2\, du}. \tag{35.33c}$$

Before discussing the physical interpretation of this exact solution, one must come to grips with the singularity in the metric coefficients at $u = a \gg T$. (There $L = 0$, so $g_{xx} = g_{yy} = 0$.) Is this a physical singularity like the region $r = 0$ of the Schwarzschild geometry, or is it merely a coordinate singularity as $r = 2M$ is in Schwarzschild coordinates (Chapters 31, 32, and 33)? The only nonzero components of the Riemann tensor for the metric (35.29) are (see Box 14.4)

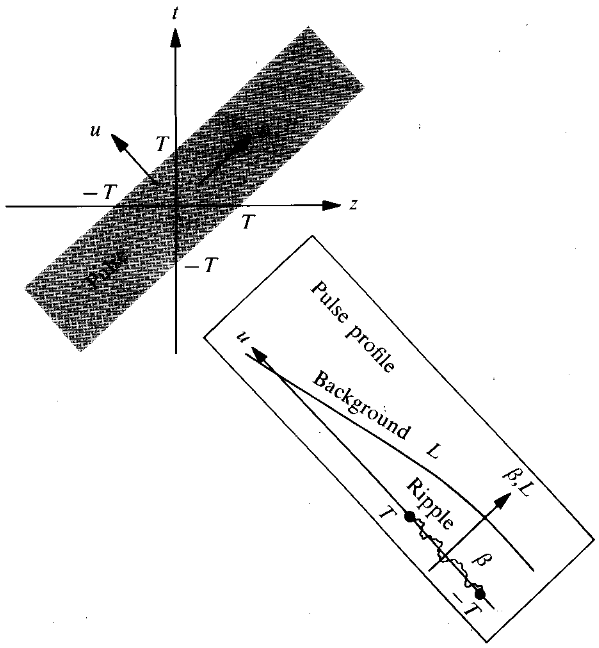


Figure 35.3.
Spacetime diagram and pulse profile for an exact plane-wave solution to Einstein's vacuum field equations. The metric has the form

$$ds^2 = L^2(e^{2\beta}\, dx^2 + e^{-2\beta}\, dy^2) + dz^2 - dt^2.$$

The "wave factor" $\beta(u) \equiv \beta(t - z)$ (short-scale ripples) and the "background factor" $L(u) \equiv L(t - z)$ (large-scale bending of the background geometry by the effective mass-energy of the "ripply" gravitational wave) are shown in the drawing and are given analytically by equations (35.33).

$$R^{x}{}_{uxu} = \frac{1}{2} R_{uu} - \beta'' - 2(L'/L)\beta', \tag{35.34}$$

$$R^{y}{}_{uyu} = \frac{1}{2} R_{uu} + \beta'' + 2(L'/L)\beta'.$$

Moreover, these components both vanish in any extended region where $\beta = 0$. Thus, *spacetime is completely flat in regions where the "wave factor" vanishes—which is everywhere outside the pulse!* In particular, spacetime is flat near $u = a$, so the singularity there must be a coordinate singularity, not a physical singularity. To eliminate this singularity, one can perform the coordinate transformation

(5) spacetime is flat outside the pulse

$$x = \frac{X}{1 - U/a}, \quad y = \frac{Y}{1 - U/a}, \quad u = U, \quad v = V + \frac{X^2 + Y^2}{a - U} \tag{35.35}$$

throughout the region to the future of the pulse ($u > T$), where

$$ds^2 = (1 - u/a)^2(dx^2 + dy^2) - du\, dv. \tag{35.36a}$$

In the new X, Y, U, V, coordinates the metric has the explicitly flat form

$$ds^2 = dX^2 + dY^2 - dU\,dV \qquad \text{for } U = u > T. \tag{35.36b}$$

EXERCISES

Exercise 35.8. GLOBALLY WELL-BEHAVED COORDINATES FOR PLANE WAVE [based on Ehlers and Kundt (1962)]

Find a coordinate transformation similar to (35.35), which puts the exact plane-wave solution (35.29a), (35.31), into the form

$$ds^2 = dX^2 + dY^2 - dU\,dV + (X^2 - Y^2)F\,dU^2, \tag{35.37}$$
$$F = F(U) \text{ completely arbitrary.}$$

This coordinate system has the advantage of no coordinate singularities anywhere; while the original coordinate system has the advantages of an easy transition to linearized theory, and easy interpretation of the action of the wave on test particles.

Exercise 35.9. GEODESIC COMPLETENESS FOR PLANE-WAVE MANIFOLD [based on Ehlers and Kundt (1962)]

Prove that the coordinate system (X, Y, U, V) of exercise 35.8 completely covers its spacetime manifold. More specifically, show that every geodesic can be extended in both directions for an arbitrarily large affine-parameter length without leaving the X, Y, U, V coordinate system. (This property is called *geodesic completeness.*) [*Hint:* Choose an arbitrary event and an arbitrary tangent vector $d/d\lambda$ there. They determine an arbitrary geodesic. Perform a coordinate transformation that leaves the form of the metric unchanged and puts $d/d\lambda$ either in the $(\widetilde{U}, \widetilde{V})$ = constant 2-surface, or in the $\widetilde{X}, \widetilde{Y})$ = constant 2-surface. Verify that the two coordinate systems cover the same region of spacetime. Then analyze completeness of $d/d\lambda$'s geodesic in $(\widetilde{X}, \widetilde{Y}, \widetilde{U}, \widetilde{V})$ coordinates.]

§35.10. PHYSICAL PROPERTIES OF THE EXACT PLANE WAVE

Spacetime is completely flat both before the arrival of the plane-wave pulse ($u < -T$) and after it has passed ($u > T$). This is the message of the last paragraph.

Flatness outside gravitational-wave pulses is unusual

Complete flatness outside the pulse is very atypical for gravitational waves in the full, nonlinear general theory of relativity. It occurs in this example only because the wave fronts (surfaces of constant u and v, i.e., constant z and t) are perfectly flat 2-surfaces. If the wave fronts were bent (e.g., spherical), the energy carried by the pulse would produce spacetime curvature outside it.

Action of exact gravitational-wave pulse on test particles:

To see nonlinear effects in action, turn from the flat geometry outside the pulse to the action of the pulse on freely falling test particles. Consider a family of particles that are all at rest in the original t, x, y, z coordinate system (world lines: $[x, y, z]$ = constant) before the pulse arrives. Then even while the pulse is passing, and after it has gone, the particles remain at rest in the coordinate system. (This is true for any metric, such as (35.29a), with $g_{0\mu} = -\delta^0{}_\mu$, for then $\Gamma^\mu{}_{00} = 0$, so $x^\mu = \delta^\mu{}_0\tau +$ const. satisfies the geodesic equation.)

Two particles whose separation is in the direction of propagation of the pulse (z-direction) have not only constant coordinate separation, $\Delta x = \Delta y = 0$ and $\Delta z \neq 0$; they also have constant proper separation, $\Delta s = g_{zz}{}^{1/2}\, \Delta z = \Delta z$. Hence, the exact plane wave is completely transverse, like a plane wave of linearized theory.

(1) transverse character of relative accelerations

Neighboring particles transverse to the propagation direction, ($\Delta x \neq 0$, $\Delta y \neq 0$, $\Delta z = 0$) have a proper separation that wiggles as the pulse passes:

$$\begin{aligned}\Delta s &= L(t - z)[e^{2\beta(t-z)}(\Delta x)^2 + e^{-2\beta(t-z)}(\Delta y)^2]^{1/2}\\ &\approx L[(1 + 2\beta)(\Delta x)^2 + (1 - 2\beta)(\Delta y)^2]^{1/2}.\end{aligned} \tag{35.38}$$

Superimposed on the usual linearized-theory type of wiggling, due to the "wave factor" β, is a very small net acceleration of the particles toward each other, due to the "background factor" L [note the form of $L(u)$ in Figure 35.3]. This is an acceleration of almost Newtonian type, caused by the gravitational attraction of the energy that the gravitational wave carries between the two particles. The total effect of all the energy that passes is to convert the particles from an initial state of relative rest, to a final state of relative motion with speed

(2) gravitational attraction due to energy in pulse

$$v_{\text{final}} = d\,\Delta s/dt = d(L\,\Delta s_i)/dt = -\Delta s_i/a, \tag{35.39}$$

where

$$\Delta s_i = [(\Delta x)^2 + (\Delta y)^2]^{1/2} = (\text{initial particle separation}).$$

[Recall: $L_{\text{initial}} = 1$, $L_{\text{final}} = 1 - u/a = 1 - (t - z)/a$; equation (35.33).]

Precisely the same effect can be produced by a pulse of electromagnetic waves (§35.11).

§35.11. COMPARISON OF AN EXACT ELECTROMAGNETIC PLANE WAVE WITH THE GRAVITATIONAL PLANE WAVE

Consider the metric

$$ds^2 = L^2(u)(dx^2 + dy^2) - du\,dv, \qquad \begin{Bmatrix} u = t - z \\ v = t + z \end{Bmatrix}, \tag{35.40}$$

which is always flat if it satisfies the vacuum Einstein equations ($R_{\mu\nu} = 0$ or $L'' = 0$), and therefore cannot represent a gravitational wave. In this metric the electromagnetic potential

An electromagnetic plane-wave pulse

$$\boldsymbol{A} = A_\mu\, \boldsymbol{d}x^\mu = \mathcal{A}(u)\, \boldsymbol{d}x \tag{35.41}$$

satisfies Maxwell's equations for arbitrary $\mathcal{A}(u)$. It represents an electromagnetic plane wave analogous to the gravitational plane wave of the last few sections. The only nonzero field components of this wave are

$$F_{ux} = \mathcal{A}', \text{ i.e., } F_{tx} = -F_{zx} = \mathcal{A}'; \tag{35.42}$$

so the electric vector oscillates back and forth in the x-direction, the magnetic vector oscillates in the y-direction, and the wave propagates in the z-direction. The stress-energy tensor in x, y, u, v, coordinates has only

$$T_{uu} = (4\pi L^2)^{-1}(\mathcal{A}')^2 \tag{35.43}$$

nonzero.

The Maxwell equations are already satisfied by the potential (35.41) in the background metric (35.40), as the reader can verify. In order to make that metric itself equally acceptable, one need only impose the Einstein equations $G_{\mu\nu} = 8\pi T_{\mu\nu}$. They read [see equation (35.30) with $\beta = 0$]

$$L'' + (4\pi T_{uu})L = 0. \tag{35.44}$$

Electromagnetic plane wave and gravitational plane wave produce same gravitational attractions

This has exactly the form of the equation $L'' + (\beta')^2 L = 0$ for the gravitational plane wave. Consequently, the relative motions of uncharged test particles produced by the "background factor" $L(u)$ are the same whether $L(u) \neq 1$ is produced by the stress-energy of an electromagnetic wave, or by a corresponding gravitational wave with

$$[(\beta')^2/4\pi]_{\text{grav wave}} = [T_{uu}]_{\text{em wave}} = (\mathcal{A}')^2/4\pi L^2. \tag{35.45}$$

The analogy can be made even closer. Decrease the wavelength of the waves, while holding $(\beta')^2/4\pi$ and $(\mathcal{A}')^2/4\pi L^2$ fixed:

$$\langle(\beta')^2/4\pi\rangle = \langle(\mathcal{A}')^2/4\pi L^2\rangle = \text{const}; \qquad \lambdabar \longrightarrow 0.$$

In the limit of very small wavelength (i.e., from a viewpoint whose characteristic length is $\gg \lambdabar$), the two solutions are completely indistinguishable. Their metrics are identical ($\lambdabar \longrightarrow 0$ and $\langle(\beta')^2\rangle = \text{const.}$ imply $\beta \longrightarrow 0$), and their jigglings of test particles are too small to be seen. Only their curving up of spacetime ($L \neq 1$) and the associated gravitational pull of their energy are detectable.

§35.12. A NEW VIEWPOINT ON THE EXACT PLANE WAVE

Exact gravitational plane waves reexamined in the language of "short-wave approximation":

(1) ripples vs. background

The above comparison suggests a viewpoint that was sketched briefly in the introduction to this chapter and in §35.8. Think of the exact gravitational plane-wave solution [Figure 35.3; equations (35.29) and (35.33)] as ripples in the spacetime curvature, described by $\beta(u)$, propagating on a very slightly curved background spacetime, characterized by $L(u)$. The most striking difference between the background and the ripples is not in the magnitude of their spacetime curvatures, but in their characteristic lengths. The ripples have characteristic length

$$\lambdabar \equiv (\text{typical reduced wavelength, } \lambda/2\pi, \text{ of waves}); \tag{35.46}$$

the background has characteristic length ("radius of curvature of background geometry")

$$\mathcal{R} \sim |L/L''|^{1/2}{}_{\text{inside wave}} \sim 1/|\beta'|. \tag{35.47}$$

Recall that $\lambdabar$ is somewhat smaller than the pulse length, $2T$. Recall also that $|\beta' T| \ll 1$. Conclude that the characteristic lengths of the "wave factor" and the "background factor" differ greatly:

$$\lambdabar \ll \mathcal{R}. \tag{35.48}$$

This difference in scales enables one to separate out the background from the ripples.

The ripples are very much smaller in scale ($\lambdabar \ll \mathcal{R}$) than the background. Nevertheless the local curvature in a ripple is correspondingly larger than the background curvature [equations (35.30), (35.34)]; thus,

$$\begin{aligned}(R^x{}_{uxu})_{\text{background}} &= (R^y{}_{uyu})_{\text{background}} = -L''/L \sim 1/\mathcal{R}^2,\\ (R^x{}_{uxu})_{\text{waves}} &= -(R^y{}_{uyu})_{\text{waves}} = -\beta'' \sim |\beta'|/\lambdabar \sim 1/(\lambdabar\mathcal{R})\\ &\sim (\mathcal{R}/\lambdabar)(R^x{}_{uxu})_{\text{background}}.\end{aligned} \tag{35.49}$$

One is reminded of the mottled surface of an orange!

The metric for the background of the gravitational plane wave is the same as for the electromagnetic one [equation (35.40)]:

$$ds^2 = g^{(\mathrm{B})}_{\mu\nu}\, dx^\mu\, dx^\nu = L^2(dx^2 + dy^2) - du\, dv. \tag{35.50}$$

By comparison with equation (35.29a), one sees that the metric for the full spacetime (background plus ripple) is

$$ds^2 = (g^{(\mathrm{B})}_{\mu\nu} + h_{\mu\nu})\, dx^\mu\, dx^\nu, \tag{35.51}$$

$$h_{xx} = -h_{yy} = 2\beta, \text{ all other } h_{\mu\nu} = 0. \tag{35.52}$$

(2) propagation of ripples in background

(Recall, in the region where $\beta \neq 0$, L is very nearly 1.) One can think of the ripples as a transverse, traceless, symmetric tensor field $h_{\mu\nu}$ analogous to the electromagnetic field A_μ, propagating in the background geometry. Just as the electromagnetic field produces the background curvature via

$$G_{uu} = -2L''/L = 8\pi T_{uu},$$

so the gravitational-wave ripples $h_{\mu\nu}$ produce the background curvature via equation (35.31), which one can rewrite as

$$G^{(\mathrm{B})}_{uu} = -2L''/L = 8\pi T^{(\mathrm{EFF})}_{uu}. \tag{35.53}$$

Here

(3) effective stress-energy tensor for ripples

$$T^{(\mathrm{EFF})}_{uu} \equiv \frac{1}{4\pi}(\beta')^2 = \frac{1}{32\pi} h_{jk,u} h_{jk,u} \tag{35.54}$$

is the "effective stress-energy tensor" for the gravitational waves. Notice that it agrees, except for averaging, with the expression (35.23) that was written down without justification in §35.7.

EXERCISE

Exercise 35.10. PLANE WAVE WITH TWO POLARIZATIONS PRESENT

The exact plane-wave solution (35.29) has polarization $\boldsymbol{e}_+$. Construct a similar solution, containing two arbitrary amplitudes, $\beta(u)$ and $\gamma(u)$, for polarizations $\boldsymbol{e}_+$ and $\boldsymbol{e}_\times$. Extend the discussions of §§35.9–35.12 to this solution.

§35.13. THE SHORTWAVE APPROXIMATION

The remainder of this chapter extends the above viewpoint in a rigorous manner to very general gravitational-wave solutions. This extension is called the "shortwave formalism"; it was largely devised by Isaacson (1968a,b), though it was built on foundations laid by Wheeler (1964a) and by Brill and Hartle (1964). Versions that are even more rigorous have been given in the W.K.B. or geometric-optics limit by Choquet-Bruhat (1969), and by MacCallum and Taub (1973).

Foundations for shortwave formalism:

(1) $\mathcal{R}$, $\lambdabar$, and $\mathcal{A}$ defined

Consider gravitational waves propagating through a *vacuum* background spacetime. As in §35.7, let $\mathcal{R}$ be the typical radius of curvature of the background; let $\lambdabar$ and $\mathcal{A}$ be the typical reduced wavelength $(\lambda/2\pi)$ and amplitude of the waves; and demand both $\mathcal{A} \ll 1$ and $\lambdabar/\mathcal{R} \ll 1$. The background curvature might be due entirely to the waves, or partly to waves and partly to nearby matter and nongravitational fields.

(2) demand that $\mathcal{A} \ll 1$ and $\lambdabar/\mathcal{R} \ll 1$

(3) split of metric into background plus perturbation; "steady coordinates"

The analysis uses a coordinate system closely "tuned" to spacetime in the sense that the metric coefficients can be split into "background" coefficients plus perturbations

$$g_{\mu\nu} = g^{(\mathrm{B})}_{\mu\nu} + h_{\mu\nu} \tag{35.55}$$

with these properties: (1) the amplitude of the perturbation is $\mathcal{A}$

$$h_{\mu\nu} \lesssim (\text{typical value of } g^{(\mathrm{B})}_{\mu\nu}) \cdot \mathcal{A}; \tag{35.56a}$$

(2) the scale on which $g^{(\mathrm{B})}_{\mu\nu}$ varies is $\gtrsim \mathcal{R}$

$$g^{(\mathrm{B})}_{\mu\nu,\alpha} \lesssim (\text{typical value of } g^{(\mathrm{B})}_{\mu\nu})/\mathcal{R}; \tag{35.56b}$$

(3) the scale on which $h_{\mu\nu}$ varies is $\sim \lambdabar$

$$h_{\mu\nu,\alpha} \sim (\text{typical value of } h_{\mu\nu})/\lambdabar. \tag{35.56c}$$

Such coordinates are called "*steady*."

(4) Split of Ricci curvature tensor

A rather long computation (exercise 35.11) shows that the Ricci tensor for an expanded metric of the form (35.55) is

$$R_{\mu\nu} = \underset{?}{R^{(\mathrm{B})}_{\mu\nu}} + \underset{\mathcal{A}/\lambdabar^2}{R^{(1)}_{\mu\nu}(h)} + \underset{\mathcal{A}^2/\lambdabar^2}{R^{(2)}_{\mu\nu}(h)} + \underset{\mathcal{A}^3/\lambdabar^2}{\text{error}}. \tag{35.57}$$

Here a marker ($\mathcal{A}/\bar{\lambda}^2$, etc.) has been placed under each term to show its typical order of magnitude; $R^{(\mathrm{B})}_{\mu\nu}$ is the Ricci tensor for the background metric $g^{(\mathrm{B})}_{\mu\nu}$; and $R^{(1)}_{\mu\nu}$ and $R^{(2)}_{\mu\nu}$ are expressions defined by

$$R^{(1)}_{\mu\nu}(h) \equiv \frac{1}{2}(-h_{|\mu\nu} - h_{\mu\nu|\alpha}{}^{\alpha} + h_{\alpha\mu|\nu}{}^{\alpha} + h_{\alpha\nu|\mu}{}^{\alpha}), \tag{35.58a}$$

$$\begin{aligned} R^{(2)}_{\mu\nu}(h) \equiv \frac{1}{2}\bigg[&\frac{1}{2}h_{\alpha\beta|\mu}h^{\alpha\beta}{}_{|\nu} + h^{\alpha\beta}(h_{\alpha\beta|\mu\nu} + h_{\mu\nu|\alpha\beta} - h_{\alpha\mu|\nu\beta} - h_{\alpha\nu|\mu\beta}) \\ &+ h_{\nu}{}^{\alpha|\beta}(h_{\alpha\mu|\beta} - h_{\beta\mu|\alpha}) - \left(h^{\alpha\beta}{}_{|\beta} - \frac{1}{2}h^{|\alpha}\right)(h_{\alpha\mu|\nu} + h_{\alpha\nu|\mu} - h_{\mu\nu|\alpha})\bigg]. \end{aligned} \tag{35.58b}$$

In these expressions and everywhere below, indices are raised and lowered with $g^{(\mathrm{B})}_{\mu\nu}$, and an upright line denotes a covariant derivative with respect to $g^{(\mathrm{B})}_{\mu\nu}$ (whereas in Chapter 21 it denoted covariant derivative with respect to 3-geometry).

Split of vacuum field equations into "wave part" ($\propto\mathcal{A}$) plus "coarse-grain part" ($\propto\mathcal{A}^2$ and smooth on scale $\bar{\lambda}$) plus "fluctuational corrections" ($\propto\mathcal{A}^2$ and ripply on scale $\bar{\lambda}$)

At the heart of the shortwave formalism is its method for solving the vacuum field equations $R_{\mu\nu} = 0$. One begins by selecting out of expression (35.57) the part linear in the amplitude of the wave $\mathcal{A}$, and setting it equal to zero. The action of the waves to curve up the background is a nonlinear phenomenon (linearized theory shows no sign of it); so $R^{(\mathrm{B})}_{\mu\nu}$ cannot be linear in $\mathcal{A}$. Hence, in expression (35.57), $R^{(1)}_{\mu\nu}(h)$ is the only linear term, and it must vanish by itself

$$R^{(1)}_{\mu\nu}(h) = 0. \tag{35.59a}$$

[Of course $h_{\mu\nu}$ may contain nonlinear correction terms—call them $j_{\mu\nu}$—of order $\mathcal{A}^2$, which must not be constrained by this linear equation. They will be determined by (35.59c), below.]

One next splits the remainder of $R_{\mu\nu}$ into a part that is free of ripples—i.e., that varies only on scales far larger than $\bar{\lambda}$ ("coarse-grain viewpoint"), and a second part that contains the fluctuations. This split can be accomplished by averaging over several wavelengths (see exercise 35.14 for a precise treatment of the averaging process, also see Choquet-Bruhat (1969) for a class of solutions where such averaging is not required):

$$\underset{?}{R^{(\mathrm{B})}_{\mu\nu}} + \underset{\mathcal{A}^2/\bar{\lambda}^2}{\langle R^{(2)}_{\mu\nu}(h)\rangle} + \underset{\mathcal{A}^3/\bar{\lambda}^2}{\text{error}} = 0 \quad \begin{bmatrix}\text{smooth}\\ \text{part}\end{bmatrix} \tag{35.59b}$$

$$\underset{\substack{\mathcal{A}^2/\bar{\lambda}^2 \\ \begin{bmatrix}\text{nonlinear cor-}\\ \text{rection to } h\end{bmatrix}}}{R^{(1)}_{\mu\nu}(j)} + \underset{\mathcal{A}^2/\bar{\lambda}^2}{R^{(2)}_{\mu\nu}(h)} - \underset{\mathcal{A}^2/\bar{\lambda}^2}{\langle R^{(2)}_{\mu\nu}(h)\rangle} + \underset{\mathcal{A}^3/\bar{\lambda}^2}{\text{error}} = 0 \quad \begin{bmatrix}\text{fluctuating}\\ \text{part}\end{bmatrix} \tag{35.59c}$$

That's all there is to it!—except for reducing the equations to manageable form, and a fuller interpretation of the physics.

Physical interpretation of the three parts of field equations:

Begin with the interpretation.

2

(1) propagation of waves

Equation (35.59a) is an equation for the propagation of the gravitational waves $h_{\mu\nu}$.

(2) production of background curvature by energy of waves; $T^{(\mathrm{GW})}_{\mu\nu}$ defined

Equation (35.59b) shows how the stress-energy in the waves creates the background curvature. It can be rewritten in the more suggestive form

$$G^{(\mathrm{B})}_{\mu\nu} \equiv R^{(\mathrm{B})}_{\mu\nu} - \frac{1}{2} R^{(\mathrm{B})} g^{(\mathrm{B})}_{\mu\nu} = 8\pi T^{(\mathrm{GW})}_{\mu\nu} \text{ in vacuum,} \tag{35.60}$$

where

$$T^{(\mathrm{GW})}_{\mu\nu} \equiv -\frac{1}{8\pi}\left\{\langle R^{(2)}_{\mu\nu}(h)\rangle - \frac{1}{2} g^{(\mathrm{B})}_{\mu\nu}\langle R^{(2)}(h)\rangle\right\} \tag{35.61}$$

is the stress-energy tensor for the gravitational waves. Now one sees the origin of the statement in §35.7, that the stress-energy of gravitational waves is well-defined only in a smeared-out sense.

(3) nonlinear self-interaction of waves

Finally, equation (35.59c) shows how the gravitational waves h generate nonlinear corrections j to themselves (wave-wave scattering, harmonics of the fundamental frequency, etc.). These higher-order effects will not be investigated in this chapter.

EXERCISE

Exercise 35.11. CONNECTION COEFFICIENTS AND CURVATURE TENSORS FOR A PERTURBED METRIC

In a specific coordinate frame of an arbitrary spacetime, write the metric coefficients in covariant representation in the form

$$g_{\mu\nu} = g^{(\mathrm{B})}_{\mu\nu} + h_{\mu\nu}. \tag{35.62a}$$

(At the end of the calculation, one can split $h_{\mu\nu}$ into two parts, $h_{\mu\nu} \longrightarrow h_{\mu\nu} + j_{\mu\nu}$; and out of this split obtain the formulas used in the text.) Assume that the typical components of $h_{\mu\nu}$ are much less than those of $g^{(\mathrm{B})}_{\mu\nu}$; so one can expand Christoffel symbols and curvature tensors in $h_{\mu\nu}$. Raise and lower indices of $h_{\mu\nu}$ with $g^{(\mathrm{B})}_{\mu\nu}$; and denote by a "$_|$" covariant derivatives relative to $g^{(\mathrm{B})}_{\mu\nu}$ and by a "$;$" covariant derivatives relative to $g_{\mu\nu}$.

(a) Here $g_{\mu\nu}$ and $g^{(\mathrm{B})}_{\mu\nu}$ can be thought of as two different metrics coexisting in the spacetime manifold. Show that the difference between the corresponding covariant derivatives, $\boldsymbol{\nabla} - \boldsymbol{\nabla}^{(\mathrm{B})} \equiv \boldsymbol{S}$—indeed, the difference between any two covariant derivatives!—is a tensor with components

$$S^{\mu}{}_{\beta\gamma} = \Gamma^{\mu}{}_{\beta\gamma} - \Gamma^{(\mathrm{B})\mu}{}_{\beta\gamma} \tag{35.62b}$$

[Hint: See part B of Box 10.3.]

(b) Show that

$$g^{\mu\nu} = g^{(\mathrm{B})\mu\nu} - h^{\mu\nu} + h^{\mu\alpha}h_{\alpha}{}^{\nu} - h^{\mu\alpha}h_{\alpha}{}^{\beta}h_{\beta}{}^{\nu} + \cdots, \tag{35.62c}$$

and also that

$$g^{\mu\nu} = g^{(\mathrm{B})\mu\nu} - h^{\mu\nu} + h^{\mu\alpha}h_{\alpha}{}^{\nu} - h^{\mu\alpha}h_{\alpha}{}^{\beta}h_{\beta\gamma}g^{\gamma\nu}. \tag{35.62c'}$$

(c) By calculating in a local Lorentz frame of $g^{(B)}_{\mu\nu}$ and then transforming back to the original frame, show that

$$S^{\mu}{}_{\beta\gamma} = \frac{1}{2} g^{\mu\alpha}(h_{\alpha\beta|\gamma} + h_{\alpha\gamma|\beta} - h_{\beta\gamma|\alpha}), \tag{35.62d}$$

$$R^{\alpha}{}_{\beta\gamma\delta} - R^{(B)\alpha}{}_{\beta\gamma\delta} = S^{\alpha}{}_{\beta\delta|\gamma} - S^{\alpha}{}_{\beta\gamma|\delta} + S^{\alpha}{}_{\mu\gamma}S^{\mu}{}_{\beta\delta} - S^{\alpha}{}_{\mu\delta}S^{\mu}{}_{\beta\gamma}, \tag{35.62e}$$

$$R_{\beta\delta} - R^{(B)}{}_{\beta\delta} = S^{\alpha}{}_{\beta\delta|\alpha} - S^{\alpha}{}_{\beta\alpha|\delta} + S^{\alpha}{}_{\mu\alpha}S^{\mu}{}_{\beta\delta} - S^{\alpha}{}_{\mu\delta}S^{\mu}{}_{\beta\alpha}. \tag{35.62f}$$

(d) Show that expression (35.62f) reduces to

$$R_{\beta\delta} = R^{(B)}{}_{\beta\delta} + R^{(1)}_{\beta\delta}(h) + R^{(2)}_{\beta\delta}(h) + \cdots \tag{35.62g}$$

where $R^{(1)}$ and $R^{(2)}$ are defined by equations (35.58).

---

§35.14. EFFECT OF BACKGROUND CURVATURE ON WAVE PROPAGATION

Focus attention on the propagation equation $R^{(1)}_{\mu\nu}(h) = 0$. As in linearized theory, so also here, the propagation is studied more simply in terms of

$$\bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \frac{1}{2} h g^{(B)}_{\mu\nu}, \tag{35.63}$$

$\bar{h}_{\mu\nu}$ defined

than in terms of $h_{\mu\nu}$. Rewritten in terms of $\bar{h}_{\mu\nu}$, $R^{(1)}_{\mu\nu}(h) = 0$ says

$$\bar{h}_{\mu\nu|\alpha}{}^{\alpha} + g^{(B)}_{\mu\nu}\bar{h}^{\alpha\beta}{}_{|\beta\alpha} - 2\bar{h}_{\alpha(\mu|}{}^{\alpha}{}_{\nu)} + 2R^{(B)}_{\alpha\mu\beta\nu}\bar{h}^{\alpha\beta} - 2R^{(B)}_{\alpha(\mu}\bar{h}_{\nu)}{}^{\alpha} = 0. \tag{35.64}$$

Propagation equation for waves on curved background

[To obtain this, invert equation (35.63) obtaining $h_{\mu\nu} = \bar{h}_{\mu\nu} - \frac{1}{2}g^{(B)}_{\mu\nu}\bar{h}$; insert this into (35.58a) and equate to zero; then commute covariant derivatives using the identity (16.6b); finally contract to obtain an expression for $\bar{h}_{|\alpha}{}^{\alpha}$ and substitute that back in.]

The propagation equation (35.64) can be simplified by a special choice of gauge. An infinitesimal coordinate transformation

Specialization to "Lorentz gauge"

$$x^{\mu}{}_{\text{new}}(\mathscr{P}) = x^{\mu}{}_{\text{old}}(\mathscr{P}) + \xi^{u}(\mathscr{P}) \tag{35.65a}$$

induces a first-order change in the functional forms of the metric coefficients given by

$$h_{\mu\nu\,\text{new}}(x^{\alpha}{}_{\text{new}}) = h_{\mu\nu\,\text{old}}(x^{\alpha}{}_{\text{new}}) - 2\xi_{(\mu|\nu)} \tag{35.65b}$$

[analog of the gauge transformation of linearized theory, equation (35.3c); see exercise 35.12]. By an appropriate choice of the four functions ξ^{μ}, one can enforce the four "Lorentz gauge conditions"

$$\bar{h}_{\mu}{}^{\alpha}{}_{|\alpha} = 0 \tag{35.66}$$

in the new coordinate system (exercise 35.13). This choice of gauge is analogous to that of linearized theory. It makes the second and third terms in the propagation equation vanish. (For additional gauge conditions of the "TT" type, see exercise 35.13.)

Coupling of waves to Ricci tensor can be ignored

The last term of the propagation equation, $-2R^{(B)}_{\alpha(\mu}\bar{h}_{\nu)}{}^{\alpha}$, vanishes to within the precision of the analysis, for this reason: attention has been confined to vacuum; so the only source of a nonvanishing Ricci tensor is the stress-energy carried by the gravitational waves themselves [equation (35.60)]; hence $R^{(B)}_{\alpha\beta} \sim \mathcal{a}^2/\lambda\!\!\!^{-2}$ and

$$R^{(B)}{}_{\alpha(\mu}\bar{h}_{\nu)}{}^{\alpha} \sim \mathcal{a}^3/\lambda\!\!\!^{-2}. \tag{35.67}$$

This is of the same order as $R^{(3)}_{\mu\nu}(h)$, the third-order correction to the Ricci tensor, which is far below the precision of the analysis. For consistency in the analysis it will therefore be neglected.

Summary of this section thus far: by choosing a gauge where $\bar{h}_{\mu}{}^{\alpha}{}_{|\alpha} = 0$, and by discarding terms of higher order than the precision of the analysis, one obtains the vacuum propagation equation

Propagation equation in Lorentz gauge and its realm of validity

$$\bar{h}_{\mu\nu|\alpha}{}^{\alpha} + 2R^{(B)}_{\alpha\mu\beta\nu}\bar{h}^{\alpha\beta} = 0, \tag{35.68}$$

subject to the Lorentz gauge condition

$$\bar{h}_{\mu\alpha}{}^{|\alpha} = 0.$$

Equation (35.68) is accurate to first order in the amplitude [corrections $\propto \mathcal{a}^2$ are embodied in equation (35.59c)]; and its accuracy is independent of the ratio $\lambda\!\!\!^{-}/\mathcal{R}$, as one sees from equations (35.59). Thus, *it can be applied whenever the waves are weak, even if the wavelength is large!*

Lists of effects absent from and contained in propagation equation

All nonlinear interactions of the wave with itself are neglected in this first-order propagation equation. Absent is the mechanism for waves to scatter off each other and off the background curvature that they themselves produce. Also absent are any hints of a change in shape of pulse due to self-interaction as a pulse of waves propagates. There are no signs of the gravitational collapse that one knows must occur when a mass-energy m of gravitational waves gets compressed into a region of size $\lesssim m$. To see all these effects, one must turn to corrections of second order in $\mathcal{a}$ and higher [e.g., equations (35.59c) and (35.60)].

Actually contained in the propagation equation are all effects due to the linear action of the background curvature on the propagating wave. These effects are explored, for short wavelengths ($\lambda\!\!\!^{-}/\mathcal{R} \ll 1$) and for nearly flat wave fronts, in exercises 35.15–35.17 at the end of the chapter. The effects considered include a gravitational redshift of gravitational radiation and gravitational deflection of the direction of propagation of gravitational radiation, identical to those for light; and also a rotation of the polarization tensor. When the wavelength is not small ($\lambda\!\!\!^{-}/\mathcal{R}$ not $\ll 1$), the propagation equation includes a back-scatter of the gravitational waves off the background curvature and a resultant pattern of wave "tails" analogous to that explored in exercise 32.10 [see, e.g., Couch *et al.* (1968), Price (1971a), Bardeen and Press (1972), Unt and Keres (1972)].

EXERCISES

Exercise 35.12. GAUGE TRANSFORMATIONS IN A CURVED BACKGROUND

(a) Show that the infinitesimal coordinate transformation (35.65a) induces the change (35.65b) in the functional form of the metric perturbation.

(b) Discuss the relationship between this gauge transformation and the concept of a Killing vector (§25.2).

Exercise 35.13. TRANSVERSE-TRACELESS GAUGE FOR GRAVITATIONAL WAVES PROPAGATING IN A CURVED BACKGROUND

(a) Show that, in vacuum in a curved background spacetime, the gauge condition $\bar{h}_\mu{}^\alpha{}_{|\alpha} = 0$ is preserved by transformations whose generator satisfies the wave equation $\xi_\mu{}^{|\alpha}{}_\alpha = 0$.

(b) Locally (over distances much smaller than $\mathcal{R}$) linearized theory is applicable; so there exists such a transformation which makes [see equations (35.7b) and (35.8a)]

$$\bar{h} = 0 + \text{error}, \qquad \bar{h}_{\mu\alpha}u^\alpha = 0 + \text{error}. \tag{35.69}$$

Here u^α is a vector field that is as nearly covariantly constant as possible ($u^\alpha{}_{|\beta} = 0$); i.e., it is a constant vector in the inertial coordinates of linearized theory; and the errors are small over distances much less than $\mathcal{R}$. Show that $\bar{h} = 0$ *can be imposed globally along with* $\bar{h}_{\mu\alpha}{}^{|\alpha} = 0$; i.e., show that, if it is imposed on an initial hypersurface, the propagation equation (35.68) preserves it.

(c) Show that in general, the background curvature prevents any vector field from being covariantly constant ($u^{\hat{\alpha}}{}_{|\beta} \sim u^{\hat{\alpha}}/\mathcal{R}$ at best); and from this *show that* $\bar{h}_{\mu\alpha}u^\alpha = 0$ *cannot be imposed globally along with* $\bar{h}_\mu{}^\alpha{}_{|\alpha} = 0$.

§35.15. STRESS-ENERGY TENSOR FOR GRAVITATIONAL WAVES

Turn now to an evaluation of the effective stress-energy tensor $T^{(\mathrm{GW})}_{\mu\nu}$ of equation (35.61). The evaluation requires averaging various quantities over several wavelengths. Useful rules for manipulating quantities inside the averaging brackets $\langle\ \ \rangle$ are the following (see exercise 35.14 for justification).

The averaging process involved in "coarse-grain" viewpoint

(1) Covariant derivatives commute; e.g., $\langle h\ h_{\mu\nu|\alpha\beta}\rangle = \langle h\ h_{\mu\nu|\beta\alpha}\rangle$. The fractional errors made by freely commuting are $\sim(\lambda\!\!\!^-/\mathcal{R})^2$, well below the inaccuracy of the computation.

(2) Gradients average out to zero; e.g., $\langle(h_{|\alpha}h_{\mu\nu})_{|\beta}\rangle = 0$. Fractional errors made here are $\lesssim \lambda\!\!\!^-/\mathcal{R}$.

(3) As a corollary, one can freely integrate by parts, flipping derivatives from one h to the other; e.g., $\langle h\ h_{\mu\nu|\alpha\beta}\rangle = \langle -h_{|\beta}h_{\mu\nu|\alpha}\rangle$.

A straightforward but long calculation using these rules, using equation (35.58b) for $R^{(2)}_{\mu\nu}(h)$, using definition (35.63) of $\bar{h}_{\mu\nu}$, using the propagation equation (35.64), and using the definition (35.61) of $T^{(\mathrm{GW})}_{\mu\nu}$, yields $\langle R^{(2)}(h)\rangle = 0$, and

Evaluation of effective stress-energy tensor for gravitational waves, $T^{(\mathrm{GW})}_{\mu\nu}$

$$T^{(\mathrm{GW})}_{\mu\nu} = \frac{1}{32\pi}\langle \bar{h}_{\alpha\beta|\mu}\bar{h}^{\alpha\beta}{}_{|\nu} - \frac{1}{2}\bar{h}_{|\mu}\bar{h}_{|\nu} - 2\bar{h}^{\alpha\beta}{}_{|\beta}\bar{h}_{\alpha(\mu|\nu)}\rangle. \tag{35.70}$$

This is the result quoted in equation (35.23′), except that there one used an inertial coordinate system of linearized theory, where covariant derivatives and ordinary derivatives are the same. In a gauge where $\bar{h}_{\mu}{}^{\alpha}{}_{|\alpha} = 0$, the last term vanishes. When, in addition, $\bar{h}_{\mu\nu}$ is traceless (see exercise 35.13), the second term vanishes; and there remains only

$$T^{(\mathrm{GW})}_{\mu\nu} = \frac{1}{32\pi}\langle \bar{h}_{\alpha\beta|\mu}\bar{h}^{\alpha\beta}{}_{|\nu}\rangle \qquad \text{if } \bar{h}_{\mu}{}^{\alpha}{}_{|\alpha} = \bar{h} = 0. \tag{35.70$'$}$$

Accuracy of expression for $T^{(\mathrm{GW})}_{\mu\nu}$

These expressions for the effective stress-energy of a gravitational wave have fractional errors of order $\mathcal{A}$, due to the neglect of second-order corrections to $h_{\mu\nu}$; they also have fractional errors of order $\lambda\!\!\!^{-}/\mathcal{R}$, due to the averaging process, which makes no sense when $\lambda\!\!\!^{-}$ approaches $\mathcal{R}$ in magnitude. Since $\mathcal{A} \lesssim \lambda\!\!\!^{-}/\mathcal{R}$ (35.28), the dominant errors in $T^{(\mathrm{GW})}_{\mu\nu}$ are $\sim\lambda\!\!\!^{-}/\mathcal{R}$.

Properties of $T^{(\mathrm{GW})}_{\mu\nu}$

To this accuracy, the stress-energy tensor for gravitational waves is on an equal footing with any other stress-energy tensor. It plays the same role in producing background curvature; and it enters into conservation laws in the same way. For example, one can show, either by direct calculation or from the identity $G^{(\mathrm{B})\mu\nu}{}_{|\nu} = 0$, that

$$T^{(\mathrm{GW})\mu\nu}{}_{|\nu} = 0 + \text{error}, \tag{35.71}$$

where the error $\sim(\lambda\!\!\!^{-}/\mathcal{R})(T^{(\mathrm{GW})\mu\nu}/\mathcal{R})$ is negligible in the shortwave approximation.

Some of the properties of $T^{(\mathrm{GW})}_{\mu\nu}$ have already been explored in §35.7. Further properties are explored in exercises 35.18 and 35.19.

EXERCISES

Exercise 35.14. BRILL-HARTLE AVERAGE

Isaacson (1968b) introduces the following averaging scheme, which he names "Brill-Hartle averaging."

(a) In the small region, of size several times $\lambda\!\!\!^{-}$, where the averaging occurs, there will be a unique geodesic of $g^{(\mathrm{B})}_{\mu\nu}$ connecting any two points $\mathcal{P}'$ and $\mathcal{P}$; so given a tensor $\boldsymbol{E}(\mathcal{P}')$ at $\mathcal{P}'$, one can parallel transport it along this geodesic to $\mathcal{P}$, getting there a tensor $\boldsymbol{E}(\mathcal{P}')_{\to\mathcal{P}}$.

(b) Let $f(\mathcal{P}', \mathcal{P})$ be a weighting function that falls smoothly to zero when $\mathcal{P}'$ and $\mathcal{P}$ are separated by many wavelengths, and such that

$$\int f(\mathcal{P}', \mathcal{P})\sqrt{-g^{(\mathrm{B})}(\mathcal{P}')}\, d^4x' = 1. \tag{35.72}$$

(c) Then the average of the tensor field $\boldsymbol{E}(\mathcal{P}')$ over several wavelengths about the point $\mathcal{P}$ is

$$\langle \boldsymbol{E}\rangle_{\mathcal{P}} \equiv \int \boldsymbol{E}(\mathcal{P}')_{\to\mathcal{P}} f(\mathcal{P}', \mathcal{P})\sqrt{-g^{(\mathrm{B})}(\mathcal{P}')}\, d^4x'. \tag{35.73}$$

(i) Show that there exists an entity $g^{(\mathrm{B})\alpha'}_{\mu}(\mathcal{P}, \mathcal{P}')$, whose primed index transforms as a tensor at $\mathcal{P}'$ and whose unprimed index transforms as a tensor at $\mathcal{P}$, such that (for $\boldsymbol{E}$ second rank)

$$E_{\alpha\beta}(\mathcal{P}')_{\to\mathcal{P}} = g^{(\mathrm{B})\mu'}_{\alpha} g^{(\mathrm{B})\nu'}_{\beta} E_{\mu'\nu'}(\mathcal{P}'). \tag{35.74}$$

This entity is called the "bivector of geodesic parallel displacement"; see DeWitt and Brehme (1960) or Synge (1960a).

(ii) Rewriting expression (35.73) in coordinate language as

$$\langle E_{\alpha\beta}(x)\rangle = \int g_{\alpha}^{(\mathrm{B})\mu'}(x,x')g_{\beta}^{(\mathrm{B})\nu'}(x,x')E_{\mu'\nu'}(x')f(x,x')\sqrt{-g^{(\mathrm{B})}(x')}\,d^4x', \qquad (35.73')$$

derive the three averaging rules cited at the beginning of §35.15. [For solution, see Appendix of Isaacson (1968b).]

Exercise 35.15. GEOMETRIC OPTICS

Develop geometric optics for gravitational waves of small amplitude propagating in a curved background. Pattern the analysis after geometric optics for electromagnetic waves (§22.5). In particular, let $\bar{h}_{\mu\nu}$ have an amplitude that varies slowly (on a scale $\mathcal{L} \lesssim \mathcal{R}$) and a phase θ that varies rapidly ($\theta_{,\hat{\alpha}} \sim 1/\lambda\!\!\!^{-}$). Expand the amplitude in powers of $\lambda\!\!\!^{-}/\mathcal{L}$, so that

$$\bar{h}_{\mu\nu} = \Re\{A_{\mu\nu} + \epsilon B_{\mu\nu} + \epsilon^2 C_{\mu\nu} + \cdots\}e^{i\theta/\epsilon}. \qquad (35.75)$$

Here ϵ is a formal expansion parameter, actually equal to unity, which reminds one that the terms attached to ϵ^n are proportional to $(\lambda\!\!\!^{-}/\mathcal{R})^n$. Define the following quantities (with $A^*_{\mu\nu}$ denoting the complex conjugate of $A_{\mu\nu}$):

$$\text{“wave vector”: } k_\alpha \equiv \theta_{,\alpha} \qquad (35.76a)$$

$$\text{“scalar amplitude”: } \mathcal{A} \equiv \left(\frac{1}{2}A^*_{\mu\nu}A^{\mu\nu}\right)^{1/2} \qquad (35.76b)$$

$$\text{“polarization”: } e_{\mu\nu} \equiv A_{\mu\nu}/\mathcal{A}. \qquad (35.76c)$$

By inserting expression (35.75) into the gauge condition (35.66) and the propagation equation (35.68), derive the fundamental equations of geometrical optics as follows.

(a) The rays (curves perpendicular to surfaces of constant phase) are null geodesics; i.e.

$$k_\alpha k^\alpha = 0, \qquad (35.77a)$$

$$k_{\alpha|\beta}k^\beta = 0. \qquad (35.77b)$$

(b) The polarization is orthogonal to the rays and is parallel transported along them;

$$e_{\mu\alpha}k^\alpha = 0, \qquad (35.77c)$$

$$e_{\mu\nu|\alpha}k^\alpha = 0. \qquad (35.77d)$$

(c) The scalar amplitude decreases as the rays diverge away from each other in accordance with

$$\mathcal{A}_{,\alpha}k^\alpha = -\frac{1}{2}k^\alpha{}_{|\alpha}\mathcal{A}. \qquad (35.77e)$$

i.e.,

$$(\mathcal{A}^2k^\alpha)_{|\alpha} = 0 \text{ (“conservation of gravitons”)}. \qquad (35.77f)$$

(d) The correction $B_{\mu\nu}$ to the amplitude obeys

$$B_{\mu\alpha}k^\alpha = iA_{\mu\alpha}{}^{|\alpha}, \qquad (35.77g)$$

$$B_{\mu\nu|\alpha}k^\alpha = -\frac{1}{2}k^\alpha{}_{|\alpha}B_{\mu\nu} + \frac{1}{2}iA_{\mu\nu|\alpha}{}^\alpha + iR^{(\mathrm{B})}_{\alpha\mu\beta\nu}A^{\alpha\beta}. \qquad (35.77h)$$

In accordance with exercise 35.13, specialize the gauge so that $\bar{h} = 0$, i.e.,

$$e_\alpha{}^\alpha = 0. \qquad (35.77i)$$

Then show that the stress-energy tensor (35.70′) for the waves is

$$T_{\mu\nu}^{(\mathrm{GW})} = \frac{1}{32\pi}\mathcal{A}^2 k_\mu k_\nu. \tag{35.77j}$$

This has the same form as the stress-energy tensor for a beam of particles with zero rest mass (see §5.4). Show explicitly that $T^{(\mathrm{GW})\mu\nu}{}_{|\nu} = 0$.

Exercise 35.16. GRAVITONS

Show that geometric optics, as developed in the preceding exercise, is equivalent to the following: "A graviton is postulated to be a particle of zero rest mass and 4-momentum $\boldsymbol{p}$, which moves along a null geodesic ($\nabla_{\boldsymbol{p}}\boldsymbol{p} = 0$). It parallel transports with itself ($\nabla_{\boldsymbol{p}}\boldsymbol{e} = 0$) a transverse ($\boldsymbol{e}\cdot\boldsymbol{p} = 0$) traceless ($e_\alpha{}^\alpha = 0$) polarization tensor $\boldsymbol{e}$. Geometric optics is the theory of a stream of such gravitons moving through spacetime." Exhibit the relationship between the quantities in this version of geometric optics and the quantities in the preceding version (e.g., $\boldsymbol{p} = \hbar\boldsymbol{k}$, where $\hbar$ is Planck's reduced constant $h/2\pi$).

Exercise 35.17. GRAVITATIONAL DEFLECTION OF GRAVITATIONAL WAVES

Show that gravitational waves of short wavelength passing through the solar system experience the same redshift and gravitational deflection as does light. (One should be able to infer this directly from exercise 35.15.)

Exercise 35.18. GAUGE INVARIANCE OF $T_{\mu\nu}^{(\mathrm{GW})}$

Show that the stress-energy tensor $T_{\mu\nu}^{(\mathrm{GW})}$ of equation (35.70) is invariant under gauge transformations of the form (35.65).

Exercise 35.19. $T_{\mu\nu}^{(\mathrm{GW})}$ EXPRESSED AS THE AVERAGE OF A STRESS-ENERGY PSEUDOTENSOR

Calculate the average over several wavelengths of the Landau-Lifshitz stress-energy pseudotensor [equation (20.22)] for gravitational waves with $\lambda\!\!\!^{-}/\mathcal{R} \ll 1$. The result should be equal to $T_{\mu\nu}^{(\mathrm{GW})}$. [Hint: Work in a gauge where $\bar{h}_\mu{}^\alpha{}_{|\alpha} = \bar{h} = 0$, to simplify the calculation.]

Exercise 35.20. SHORTWAVE APPROXIMATION FROM A VARIATIONAL VIEWPOINT

Readers who have studied the variational approach to gravitation theory in Chapter 21 may find attractive the following derivation of the basic equations of the shortwave approximation. It was devised, independently, by Sàndor Kovàcs and Bernard Schutz, and by Bryce DeWitt (unpublished, 1971). MacCallum and Taub (1973) give a "non-Palatini" version.

(a) Define

$$g_{\mu\nu} \equiv g_{\mu\nu}^{(\mathrm{B})} + h_{\mu\nu}, \qquad \bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \frac{1}{2}g_{\mu\nu}^{(\mathrm{B})}h, \tag{35.78a}$$

$$W^\mu{}_{\beta\gamma} \equiv \frac{1}{2}g^{\mu\alpha}_{(\mathrm{B})}(h_{\alpha\beta|\gamma} + h_{\alpha\gamma|\beta} - h_{\beta\gamma|\alpha}). \tag{35.78b}$$

Raise and lower indices on $h_{\mu\nu}$ and $W^\mu{}_{\beta\gamma}$ with the background metric. Using the results of exercise 35.11, derive the following expression for the Lagrangian of the gravitational field:

$$\mathcal{L} \equiv \frac{1}{16\pi}(-g)^{1/2}R = \mathcal{L}' + \begin{pmatrix}\text{perfect divergence}\\ \text{of form } \partial\mathcal{Q}^\alpha/\partial x^\alpha\end{pmatrix} + \begin{pmatrix}\text{corrections of order}\\ \mathcal{a}^3/\lambda\!\!\!^{-2},\ R_{\mu\nu}^{(\mathrm{B})}\mathcal{a}, \text{ and smaller}\end{pmatrix}, \tag{35.78c}$$

where

$$\mathcal{L}' \equiv \frac{1}{16\pi}(-g^{(B)})^{1/2}[R^{(B)} - \bar{h}^{\mu\nu}(W^{\alpha}{}_{\mu\nu|\alpha} - W^{\alpha}{}_{\mu\alpha|\nu}) \\ + g^{(B)\mu\nu}(W^{\alpha}{}_{\beta\alpha}W^{\beta}{}_{\mu\nu} - W^{\alpha}{}_{\beta\nu}W^{\beta}{}_{\mu\alpha})]. \tag{35.78d}$$

[*Hint:* recall that

$$(-g^{(B)})^{1/2}B^{\alpha}{}_{|\alpha} = \partial[(-g^{(B)})^{1/2}B^{\alpha}]/\partial x^{\alpha}$$

for any B^{α}.] Drop the corrections of order $\mathcal{A}^3/\bar{\lambda}^2$ from $\mathcal{L}$; and, knowing in advance that the field equations will demand $R^{(B)}_{\mu\nu} \sim \mathcal{A}^2/\bar{\lambda}^2$, drop also the corrections of order $R^{(B)}_{\mu\nu}\mathcal{A}$. Knowing that a perfect divergence contributes nothing in an extremization calculation, drop the divergence term from $\mathcal{L}$. Then $\mathcal{L}'$ is the only remaining part of $\mathcal{L}$.

(b) Extremize $I \equiv \int \mathcal{L}'\, d^4x$ by the Palatini method (§21.2); i.e., abandon (temporarily) definition (35.78b) of $W^{\mu}{}_{\beta\gamma}$, and extremize I with respect to independent variations of $W^{\mu}{}_{\beta\gamma} = W^{\mu}{}_{\gamma\beta}$, $\bar{h}^{\mu\nu} = \bar{h}^{\nu\mu}$, and $g^{\mu\nu}_{(B)} = g^{\nu\mu}_{(B)}$. Show that extremization with respect to $W^{\mu}{}_{\beta\gamma}$ leads back to equation (35.78b) for $W^{\mu}{}_{\beta\gamma}$ in terms of $h_{\mu\nu}$. Show that extremization with respect to $\bar{h}^{\mu\nu}$, when combined with equations (35.78a,b), leads to the propagation equation for gravitational waves (35.64). Show that extremization with respect to $g^{(B)\mu\nu}$, when combined with equations (35.78a,b) and with the propagation equation (35.64), and when averaged over several wavelengths, leads to

$$G^{(B)}_{\mu\nu} = 8\pi T^{(GW)}_{\mu\nu},$$

where $T^{(GW)}_{\mu\nu}$ is given by equation (35.70). [*Warning:* The amount of algebra in this exercise is enormous, unless one chooses to impose the gauge conditions $\bar{h} = \bar{h}_{\alpha}{}^{\beta}{}_{|\beta} = 0$ from the outset.]

1

CHAPTER 36

GENERATION OF GRAVITATIONAL WAVES

Matter is represented by curvature, but not every curvature does represent matter; there may be curvature "in vacuo."

G. LEMAITRE in Schilpp (1949), p. 440

§36.1. THE QUADRUPOLE NATURE OF GRAVITATIONAL WAVES

Masses in an isolated, nearly Newtonian system move about each other. How much gravitational radiation do they emit?

Generation of gravitational waves analyzed by electromagnetic analog

For an order-of-magnitude estimate, one can apply the familiar radiation formulas of electromagnetic theory, with the replacement $e^2 \longrightarrow -m^2$, which converts the static coulomb law into Newton's law of attraction. This procedure treats gravity as though it were a spin-one (vector) field, rather than a spin-two (tensor) field; consequently, it introduces moderate errors in numerical factors and changes angular distributions. But it gives an adequate estimate of the total power radiated.

In electromagnetic theory, electric-dipole radiation dominates, with a power output or "luminosity," L, given (see §4.4 and Figure 4.6) by

$$L_{\text{electric dipole}} = (2/3)e^2\boldsymbol{a}^2$$

for a single particle with acceleration $\boldsymbol{a}$ and dipole moment changing as $\ddot{\boldsymbol{d}} = e\ddot{\boldsymbol{x}} = e\boldsymbol{a}$;

$$L_{\text{electric dipole}} = (2/3)\,\ddot{\boldsymbol{d}}^2$$

for a general system with dipole moment $\boldsymbol{d}$. [Geometric units: luminosity in cm of mass-energy per cm of light travel time; charge in cm, $e = (G^{1/2}/c^2)e_{\text{conv}} = (2.87 \times 10^{-25}\,\text{cm/esu}) \times (4.8 \times 10^{-10}\,\text{esu}) = 1.38 \times 10^{-34}$ cm, acceleration in cm of distance per cm of time per cm of time. For conventional units, with e in esu or $(\text{g cm}^3/\text{sec}^2)^{1/2}$, insert a factor c^{-3} on the right and get L in erg/sec]. The gravitational analog of the electric dipole moment is the mass dipole moment

Why gravitational waves cannot be dipolar

$$\boldsymbol{d} = \sum_{\text{particles } A} m_A \boldsymbol{x}_A.$$

Its first time-rate of change is the total momentum of the system,

$$\dot{\boldsymbol{d}} = \sum_{\text{particles } A} m_A \dot{\boldsymbol{x}}_A = \boldsymbol{p}.$$

The second time-rate of change of the mass dipole moment has to vanish because of the law of conservation of momentum, $\ddot{\boldsymbol{d}} = \dot{\boldsymbol{p}} = 0$. Therefore *there can be no mass dipole radiation in gravitation physics.*

The next strongest types of electromagnetic radiation are magnetic-dipole and electric-quadrupole. Magnetic-dipole radiation is generated by the second time-derivative of the magnetic moment, $\ddot{\boldsymbol{\mu}}$. Here again the gravitational analog is a constant of the motion, the angular momentum,

$$\boldsymbol{\mu} = \sum_A (\text{position of } A) \times (\text{current due to } A) = \sum_A \boldsymbol{r}_A \times (m\boldsymbol{v}_A) = \boldsymbol{J};$$

so it cannot radiate. *Thus, there can be no gravitational dipole radiation of any sort.*

When one turns to quadrupole radiation, one finally gets a nonzero result (see Figure 36.1). The power output predicted by electromagnetic theory,

$$L_{\text{electric quadrupole}} = \frac{1}{20}\dddot{Q}^2 \equiv \frac{1}{20}\dddot{Q}_{jk}\dddot{Q}_{jk},$$

$$Q_{jk} \equiv \sum_A e_A \left(x_{Aj}x_{Ak} - \frac{1}{3}\delta_{jk}{r_A}^2\right)$$

$(Q_{jk\text{ here}} = Q_{jk\text{ in much other literature}})$, has as its gravitational counterpart

$$L_{\text{mass quadrupole}} = \frac{1}{5}\langle \dddot{I}^2\rangle \equiv \frac{1}{5}\langle \dddot{I}_{jk}\dddot{I}_{jk}\rangle, \tag{36.1}$$

Gravitational-wave power output expressed in terms of "reduced quadrupole moment" of source

$$I\!\!\!-_{jk} \equiv \sum_A m_A \left(x_{Aj}x_{Ak} - \frac{1}{3}\delta_{jk}{r_A}^2\right) = \int \rho\left(x_j x_k - \frac{1}{3}\delta_{jk}r^2\right)d^3x. \tag{36.2}$$

Formula (36.1) contains the correct factor of 1/5, which comes from tensor calculations (see §36.10), instead of the incorrect factor 1/20 suggested by the electromagnetic analog; and the righthand side of (36.1) has been averaged ("$\langle\ \ \rangle$") over several characteristic periods of the source to accord with one's inability to localize the energy of gravitational radiation inside a wavelength.

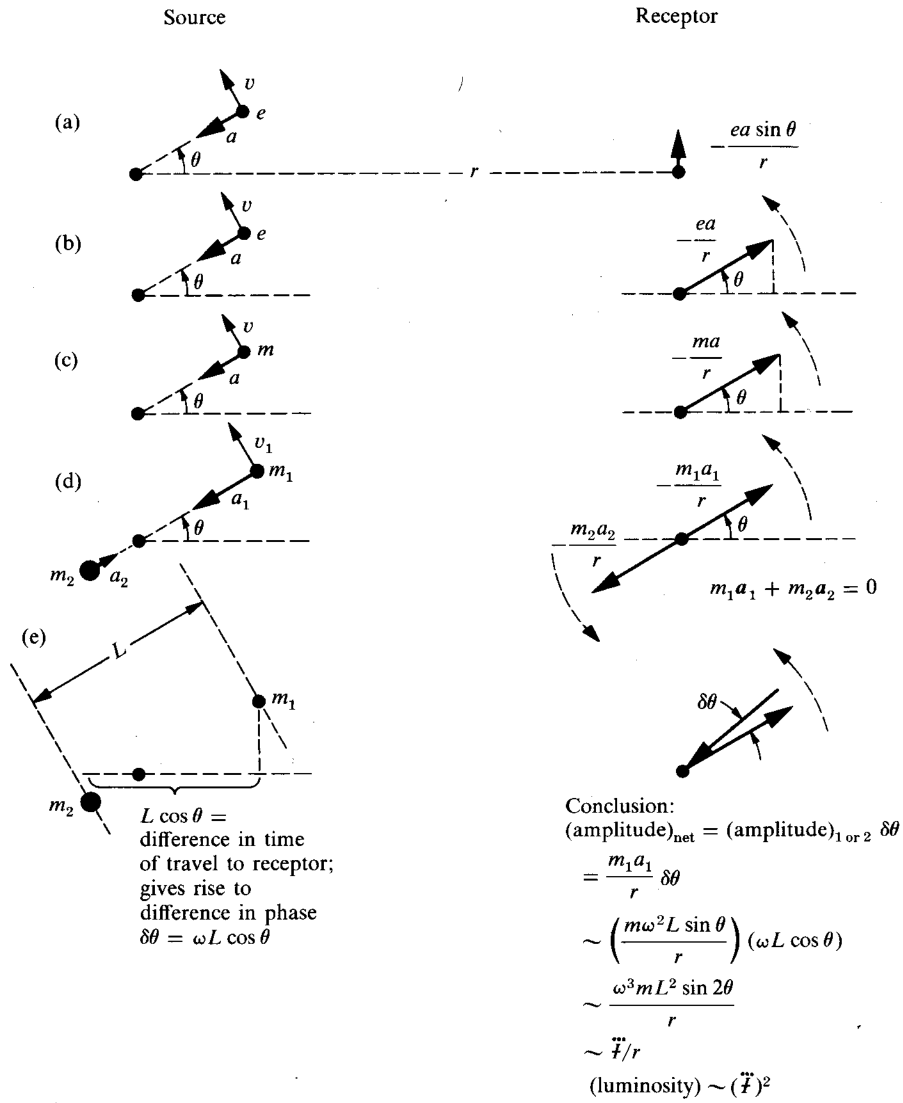


Figure 36.1.

Why gravitational radiation is ordinarily weak. In brief, contributions to the amplitude of the outgoing wave from the mass dipole moments of the separate masses cancel, $(m_1\boldsymbol{a}_1 + m_2\boldsymbol{a}_2)/r = 0$ (principle that action equals reaction).

(a) Radiation from an accelerated charge (see §4.4 and Figure 4.6).

(b) Representation of the field at the great distance r in terms of the typical rotating-vector diagram of electrical engineering; however, here, for ease of visualization, the *vertical* projection of the rotating vector gives the observed field (usual dipole-radiation field produced by a charge in circular orbit).

(c) Corresponding rotating-vector diagram for gravitational radiation, based on the simplified model of the gravitation field as a spin-one or vector field (to be contrasted with its true tensor character; hence details of angular distribution and total radiation as given by this simple diagram are not correct; but order of magnitude of luminosity is correct).

(d) The two masses m_1 and m_2 that hold each other in orbit give equal and opposite contributions to the amplitude of the outgoing wave because of the principle that action equals reaction. (In electromagnetic radiation from a hydrogen atom, the corresponding radiation amplitudes do not cancel: $e_{\text{elec}}\boldsymbol{a}_{\text{elec}} + e_{\text{prot}}\boldsymbol{a}_{\text{prot}} \sim e_{\text{elec}}\boldsymbol{a}_{\text{elec}} \neq 0$).

(e) In a better approximation, one has to allow for the difference in time of arrival at the receptor of the effects from the two masses. The two vectors that formerly opposed each other exactly are now drawn inclined, at the phase angle $\delta\theta$. The amplitude of the resulting field goes as $\ddot{\not{I}}$, where $\not{I}$ is the reduced quadrupole moment; and the luminosity is proportional to $\dddot{\not{I}}^2$.

Notation: There is no ambiguity about the definition of the *"second moment of the mass distribution"* as it appears throughout the physics and mathematics literature

$$I_{jk} \equiv \int \rho x_j x_k \, d^3x.$$

Nor is there any ambiguity about how one constructs the moment of inertia tensor $\mathcal{I}_{jk}$ from this second moment of the mass distribution:

$$\mathcal{I}_{jk} = \delta_{jk} \text{ trace } (I_{ab}) - I_{jk} = \int \rho(r^2 \delta_{jk} - x_j x_k) \, d^3x.$$

The moments that characterize a source radiating quadrupole gravitational radiation are here taken, equally unambiguously, to be the *"trace-free part of the second moment of the mass distribution":*

$$I\!\!\!-_{jk} = I_{jk} - \frac{1}{3}\delta_{jk} \text{ trace } (I_{ab}) = I_{jk} - \frac{1}{3}\delta_{jk} I = \int \rho(x_j x_k - \frac{1}{3}\delta_{jk} r^2) \, d^3x. \quad (36.3)$$

This notation is adopted because it simplifies formulas, it simplifies calculations, it meshes well with much of the literature of gravitational-wave theory [e.g. Peters (1964), Peres and Rosen (1964)], and it is easy to remember. Another name for the quantities $I\!\!\!-_{jk}$ is *reduced quadrupole moment.* This terminology makes clear the distinction between the quantities used here and the three-times-larger quantities that are called quadrupole moments in the standard text of Landau and Lifshitz (1962) and in the literature on nuclear quadrupole moments, and the 3/2-times-larger quantities used in the theory of spherical harmonics:

$$Q_{zz} \begin{pmatrix} \text{Landau and Lifshitz; also} \\ \text{nuclear quadrupole moments} \end{pmatrix} = \int \rho(3z^2 - r^2) \, d^3x,$$

$$Q_{zz} \text{ (theory of spherical harmonics)} = \int \rho\left(\frac{3}{2}z^2 - \frac{1}{2}r^2\right) d^3x,$$

$$I\!\!\!-_{zz} \begin{pmatrix} \text{reduced quadrupole moment;} \\ \text{unambiguous measure of} \\ \text{source strength adopted here} \end{pmatrix} = \int \rho\left(z^2 - \frac{1}{3}r^2\right) d^3x.$$

Thus the $I\!\!\!-_{jk}$ notation has the merit of circumventing the existing ambiguity in the literature.

Why gravitational waves are ordinarily quadrupolar

That electromagnetic radiation is predominantly dipolar (spherical-harmonic index $l = 1$), and gravitational radiation is quadrupolar ($l = 2$) are consequences of a general theorem. Consider a classical radiation field, whose associated quantum mechanical particles have integer spin S, and zero rest mass. Resolve that radiation field into spherical harmonics—i.e., into multipole moments. All components with $l < S$ will vanish; in general those with $l \geq S$ will not; and this is independent of the nature of the source! [See, e.g., Couch and Newman (1972).] Since the lowest nonvanishing multipoles generally dominate for a slowly moving source (speeds $\ll c$), electromagnetic radiation ($S = 1$) is ordinarily dipolar ($l = S = 1$), while gravitational radiation ($S = 2$) is ordinarily quadrupolar ($l = S = 2$). Closely connected with this theorem is the "topological fixed-point theorem" [e.g., Lifshitz (1949)], which distinguishes between scalar, vector, and tensor fields. For a scalar disturbance, such as a pressure wave, there is no difficulty in having a spherically symmetric source. Thus, over a sphere of a great radius r, there is no difficulty in having a pressure field that everywhere, at any one time, takes on the same value p. In contrast, there is no way to lay down on the surface of a 2-sphere a continuous vector field, the magnitude of which is non-zero and everywhere the same ("no way to comb smooth the hair on the surface of a billiard ball"). Likewise, there is no way to lay down on the surface of a 2-sphere a continuous non-zero transverse-traceless 2×2 matrix field that differs from one point to another at most by a rotation. Topology thus excludes the possibility of any spherically symmetric source of gravitational radiation whatsoever.

§36.2. POWER RADIATED IN TERMS OF INTERNAL POWER FLOW

Expression (36.1) for the power output can be rewritten in a form that is easier to use in order-of-magnitude estimates. Notice that the reduced quadrupole moment is

$$\dddot{I\!\!\!-}_{jk} \sim \frac{\begin{pmatrix}\text{mass of that part of}\\ \text{system which moves}\end{pmatrix} \times \begin{pmatrix}\text{size of}\\ \text{system}\end{pmatrix}^2}{\begin{pmatrix}\text{time for masses to move from}\\ \text{one side of system to other}\end{pmatrix}^3} = \frac{MR^2}{T^3}$$

$$\sim \frac{M(R/T)^2}{T} \sim \frac{\begin{pmatrix}\text{nonspherical part}\\ \text{of kinetic energy}\end{pmatrix}}{T};$$

$$\dddot{I\!\!\!-}_{jk} \sim L_{\text{internal}} \equiv \begin{pmatrix}\text{power flowing from one}\\ \text{side of system to other}\end{pmatrix}: \tag{36.4}$$

Gravitational-wave power output in terms of internal power flow of source

Consequently, equation (36.1) says that *the power output in gravitational waves ("luminosity") is roughly the square of the internal power flow*

$$L_{\text{GW}} \sim (L_{\text{internal}})^2. \tag{36.5}$$

If this equation seems crazy (who but a fool would equate a power to the square of a power?), recall that in geometrized units power is dimensionless. The conversion factor to conventional units is

$$L_o \equiv c^5/G = 3.63 \times 10^{59}\ \text{erg/sec} = 2.03 \times 10^5 M_\odot c^2/\text{sec}. \tag{36.6}$$

One may freely insert this factor of $L_o = 1$ wherever one wishes in order to feel more comfortable with the appearance of the equations. For example, one can rewrite equation (36.5) in the form

$$L_{\text{GW}}/L_{\text{internal}} \sim L_{\text{internal}}/L_o. \tag{36.7}$$

In applying the equation $L_{GW} \sim (L_{internal})^2$, one must be careful to ignore those internal power flows that cannot radiate at all, i.e., those that do not accompany a time-changing quadrupole moment. For example, in a star the internal power flows associated with spherical pulsation and axially symmetric rotation must be ignored.

Conservation of energy guarantees that radiation reaction forces will pull down the internal energy of the system at the same rate as gravitational waves carry energy away (see Box 19.1). The characteristic time-scale for radiation reaction to change the system markedly is

Characteristic time-scale for radiation-reaction effects

$$\begin{aligned}\tau_{react} &\sim [1/(\text{rate at which energy is lost})] \times [\text{energy in motions that radiate}]\\ &\sim [1/L_{GW}] \times [(L_{internal}) \times (\text{characteristic period } T \text{ of internal motions})]\\ &\sim (L_{internal}/L_{GW})T \sim (L_0/L_{internal})T.\end{aligned} \tag{36.8}$$

Consequently, *radiation reaction is important in one characteristic period only if the system achieves the enormous internal power flow*

$$L_{internal} \gtrsim L_o = 3.63 \times 10^{59} \text{ ergs/sec} = 1!$$

§36.3. LABORATORY GENERATORS OF GRAVITATIONAL WAVES

As a laboratory generator of gravitational waves, consider a massive steel beam of radius $r = 1$ meter, length $l = 20$ meters, density $\rho = 7.8$ g/cm^3, mass $M = 4.9 \times 10^8$ g (490 tons), and tensile strength $t = 40{,}000$ pounds per square inch or 3×10^9 dyne/cm^2. Let the beam rotate about its middle (so it rotates end over end), with an angular velocity ω limited by the balance between centrifugal force and tensile strength

Power output from a rotating steel beam

$$\omega = (8t/\rho l^2)^{1/2} = 28 \text{ radians/sec.}$$

The internal power flow is

$$\begin{aligned}L_{internal} &= \left(\frac{1}{2} I\omega^2\right)\omega = \frac{1}{24} Ml^2\omega^3\\ &\approx 2 \times 10^{18} \text{ erg/sec} \approx 10^{-41} L_o.\end{aligned}$$

The order of magnitude of the power radiated is

$$L_{GW} \sim (10^{-41})^2 L_o \sim 10^{-23} \text{ erg/sec.} \tag{36.9}$$

(An exact calculation using equation (36.1) gives 2.2×10^{-22} erg/sec; see Exercise 36.1.) Evidently the construction of a laboratory generator of gravitational radiation is an unattractive enterprise in the absence of new engineering or a new idea or both.

To rely on an astrophysical source and to build a laboratory or solar-system detector is a more natural policy to consider. Detection will be discussed in the next chapter. Here attention focuses on astrophysical sources.

EXERCISE

Exercise 36.1. GRAVITATIONAL WAVES FROM ROTATING BEAM

A long steel beam of length l and mass M rotates end over end with angular velocity ω. Show that the power it radiates as gravitational waves is

$$L_{\mathrm{GW}} = \frac{2}{45} M^2 l^4 \omega^6. \tag{36.10}$$

Use this formula to verify that the rod described in the text radiates 2.2×10^{-22} ergs/sec.

§36.4. ASTROPHYSICAL SOURCES OF GRAVITATIONAL WAVES: GENERAL DISCUSSION

Consider a highly dynamic astrophysical system (a star pulsating and rotating wildly, or a collapsing star, or an exploding star, or a chaotic system of many stars). If its mass is M and its size is R, then according to the virial theorem (exercise 39.6) its kinetic energy is $\sim M^2/R$. The characteristic time-scale for mass to move from one side of the system to the other, T, is

$$T \sim \frac{R}{(\text{mean velocity}} \sim \frac{R}{(M/R)^{1/2}} = \left(\frac{R^3}{M}\right)^{1/2} \tag{36.11a}$$

($\sim$ time of free fall; $\sim$ time to turn one radian in Kepler orbit; see Chapter 25). Consequently, the internal power flow is

$$L_{\text{internal}} \sim \frac{(\text{kinetic energy})}{T} \sim \left(\frac{M^2}{R}\right)\left(\frac{M}{R^3}\right)^{1/2} \sim \left(\frac{M}{R}\right)^{5/2}. \tag{36.11b}$$

Power output from violent astrophysical sources, in terms of mass and radius

The gravitational-wave output or "luminosity" is the square of this quantity, or

$$L_{\mathrm{GW}} \sim (M/R)^5 L_o. \tag{36.11c}$$

(If the system is rather symmetric, or if only a small portion of its mass is in motion, then its quadrupole moment does not change much, and the estimate of L_{GW} must be reduced accordingly. The wave amplitude goes down in proportion to the fraction of the mass in motion, and the power is reduced in proportion to the square of that fraction.)

Clearly, *the maximum power output occurs when the system is near its gravitational radius;* and because nothing, not even gravitational waves, can escape from inside the gravitational radius, *the maximum value of the output is $\sim L_o = 3.63 \times 10^{59}$ ergs/second, regardless of the nature of the system!*

Upper limit on power output

Actually, the above derivation of this limit and of equation (36.11c) uses approximations to general relativity that break down near the gravitational radius. [Velocities small compared to light are required in deriving the standard formula (36.1) for

L_{GW} (see §36.7); nearly Newtonian fields are required for the virial theorem arguments of (36.11a), as well as for the L_{GW} formula.] Nevertheless, in rough order of magnitude, equation (36.11c) is valid to quite near the Schwarzschild radius, say, $R \sim 3M$; and inside that point gravity is so strong that no system can resist collapse for an effective length of time much longer than $T \sim M$.

Radiation reaction in astrophysical sources

The time required for radiation-reaction forces to affect a system substantially [equation (36.8)] is of the order

$$\tau_{\text{react}} \sim (L_o/L_{\text{internal}})T \sim (R/M)^{5/2}T, \tag{36.11d}$$

where T is the characteristic time (36.11a) of rotation or free fall. (Note how one inserts and removes the factor $L_o = 1$ at will!) Consequently, *the effect of radiation reaction, as integrated over one period, is unimportant except when the system is near its gravitational radius.*

When a system such as a pulsating star is settling down into an equilibrium state, the radiation reaction will damp its internal motions. On the other hand, when the system, like a binary star system, is far from any state of equilibrium, then loss of energy (and angular momentum) to radiation under certain circumstances may speed up the angular velocity or speed up the internal motions and augment the radiation.

§36.5. GRAVITATIONAL COLLAPSE, BLACK HOLES, SUPERNOVAE, AND PULSARS AS SOURCES

Gravitational waves from:
(1) stellar collapse and formation of a black hole

Since $L_{GW} \sim (M/R)^5 L_o$, the most intense gravitational waves reaching Earth must come from a dynamic, deformed system near its gravitational radius (L_{GW} drops by a factor 100,000 with each increase by 10 of R!). The scenario of Figure 24.3 gives an impression of some of the dynamic processes that not only may happen but probably must happen. The sequence of events sketched out there includes pulses of gravitational radiation interspersed with intervals of continuous radiation of gradually increasing frequency. Pulse number one comes at the time of the original collapse of a star with white-dwarf core to a pancake-shaped neutron star. The details of what then goes on will differ enormously depending on the original mass and angular momentum of this "pancake." In the illustration, this pancake fragments into a constellation of corevolving neutron stars, which then one by one undergo "pursuit and plunge."

Whether in this kind of scenario or otherwise, perhaps the most favorable source of gravitational radiation is a star (the original very temporary "pancake" or one of the fragments therefrom) collapsing through its gravitational radius in a highly nonspherical manner. Such a star should terminate life with a last blast of gravitational waves, which carry off a sizeable fraction of its rest mass. Thus an order-of-magnitude estimate gives

$$\begin{aligned}(\text{energy radiated}) = \int L_{GW}\,dt \sim L_o \cdot \begin{pmatrix}\text{time during which}\\ \text{peak luminosity occurs}\end{pmatrix} \\ \sim L_o M = M.\end{aligned} \tag{36.12}$$

(Whether the energy radiated is $0.9M$, or $0.1M$, or $0.01M$ is not known for certain today; but it must lie in this range of orders of magnitude.) The radiation should be weak at low frequencies; it should rise to a peak at a frequency a little smaller than $1/M$; and it should cut off sharply for circular frequencies above $\omega \sim 1/M$.

(2) the fall of debris into a black hole

Matter ("debris"; see Figure 24.3) falling into a black hole can also be a significant source of gravitational waves. The infalling matter will radiate only weakly when it is far from the gravitational radius; but as it falls through the gravitational radius (between $r \sim 4M$ and $r = 2M$), it should emit a strong burst. If m is the mass of an infalling lump of matter and M is the total mass of the black hole, then the total energy in the final burst is

$$E_{\text{radiated}} \sim m^2/M, \tag{36.13}$$

and it comes off in a time $\sim M$ with a power output of $L_{\text{GW}} \sim (m/M)^2 L_o$. (See exercise 36.2.) Actually, this is an extremely rough estimate of the energy output. In the limit where the infalling lump is small in both size and mass [(size of lump) $\ll$ (gravitational radius of black hole); $m \ll M$; "delta-function lump"], one can perform an exact calculation of the spectrum and energy radiated by treating the lump and the waves as small perturbations on the Schwarzschild geometry of the black hole. The foundations for such a treatment were given by Zerilli (1970b). Zerilli's formula was corrected and applied to the case of head-on impact by Davis, Ruffini, Press, and Price (1971). They predict the spectrum of Figure 36.2 and the total energy output

$$E_{\text{radiated}} = 0.0104 m^2/M \tag{36.14}$$

for $m \ll M$ and (size of lump) $\ll M$.

(3) collisions of black holes

A collision between black holes should also produce a strong burst of gravitational waves—through such collisions are probably very rare!

(4) supernova explosions

Not quite so rare, but still not common, are supernova explosions (about one per galaxy per 100 years). According to current theory as verified by pulsar observations, a supernova is triggered by the collapse of the core of a highly evolved star (see §24.3). The collapse itself and the subsequent wild gyrations of the collapsed core (neutron star) should produce a short, powerful burst of gravitational waves. The characteristics of the burst, as estimated with formulas (36.11), and assuming large departures from sphericity, are

$$\begin{aligned}
&(\text{energy radiated}) \sim (\text{neutron-star binding energy}) \\
&\qquad \sim M^2/R \sim 0.1M \sim 10^{53}\ \text{ergs}, \\
&(\text{mean frequency}) \sim 1/T \sim (M/R^3)^{1/2} \sim 0.03M^{-1} \sim 3000\ \text{Hz}, \\
&(\text{power output}) \sim (M/R)^5 L_o \sim 10^{-5} L_o \sim 3 \times 10^{54}\ \text{ergs/sec},
\end{aligned} \tag{36.15}$$

$$\begin{pmatrix} \text{time for gravitational} \\ \text{radiation to damp the} \\ \text{motion if turbulence,} \\ \text{heat conduction, and other} \\ \text{effects do not damp it} \\ \text{sooner} \end{pmatrix} = \tau \sim M(M/R)^{-4} \sim 0.1\ \text{sec} \sim 300\ \text{periods}.$$

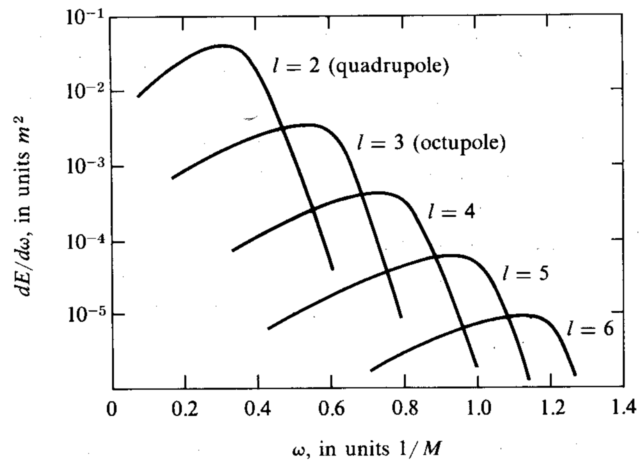


Figure 36.2.
Spectrum of the gravitational waves emitted by a "delta-function" lump of matter of mass m, falling head-on into a nonrotating (Schwarzschild) black hole of mass $M \gg m$. The total energy radiated is distributed among multipoles according to the empirical law

$$\text{(energy in } l\text{-pole waves)} \approx (0.44\ m^2/M)e^{-2l},$$

and the total spectrum peaks at angular frequency

$$\omega_{\max} = 0.32/M.$$

These results were calculated by treating the infalling lump and the gravitational waves as small perturbations on the Schwarzschild geometry of the black hole. The relevant perturbation-theory equations were derived by Zerilli (1970), and were solved numerically to give these results by Davis, Ruffini, Press, and Price (1971).

In the last stages of the stellar pulsations, when the amplitude $\xi = \delta r$ has dropped to $\delta r/r \ll 1$, one can calculate the pulsation frequencies and damping times exactly by treating the fluid motions and gravitational waves as small perturbations of an equilibrium stellar model. The results of such a calculation, which are in good agreement with the above rough estimates, are shown in Box 36.1.

(5) young pulsars

Long after the pulsations of the neutron star have been damped out by gravitational radiation reaction and by other forces, the star will continue to rotate; and as it rotates, carrying along with its rotation an off-axis-pointing magnetic moment, it will beam out the radio waves, light, and x-rays that astronomers identify as "pulsar radiation." In this pulsar phase, gravitational radiation is important only if the star is somewhat deformed from axial symmetry (axial symmetry $\Longrightarrow$ constant quadrupole moment $\Longrightarrow$ no gravitational waves). According to estimates in exercise 36.3, a deformation that contains only 0.001 of the star's mass could radiate 10^{38} ergs per second for the youngest known pulsar (Crab nebula); and the accompanying radiation reaction could be a significant source of the pulsar's slowdown. However, it is not at all clear today (1973)—indeed, it seems unlikely—that the neutron star could support even so small a deformation.

(continued on page 986)

Box 36.1 GRAVITATIONAL WAVES FROM PULSATING NEUTRON STARS

The table given here, taken from Thorne (1969a), shows various characteristics of the quadrupole oscillations of several typical neutron-star models. Note that the gravitational waves emitted by the most massive models (1) have frequencies $\nu = 1/T_n \sim 3{,}000$ Hz, (2) last for a time of $\sim\frac{1}{3}$ second, (3) damp out the stellar vibrations after only $\sim$1,000 oscillations, and (4) carry off a total energy of $\sim(10^{54}\text{ ergs}) \times (\delta R/R)^2$, where $\delta R/R$ is the initial fractional amplitude of vibration of the star's surface.

These results are *not* based on the nearly Newtonian slow-motion formalism of this chapter [equation (36.1), §§36.7 and 36.8], because that formalism is invalid here: the reduced wavelength of the radiation, $\lambdabar \sim 15$ km for waves from the most massive star, is not large compared to the star's gravitational radius, $2M \sim 6$ km; and the star's internal gravitational field is not weak (M/R as large as 0.29). Consequently, these results were derived using an alternative technique, which is valid for rapid motions and strong internal fields, but which assumes small perturbations away from the equilibrium stellar model. See Thorne (1969a) and papers cited therein for details.

QUADRUPOLE PULSATIONS OF NEUTRON STARS

Equation of state	ρ_c ($g\ cm^{-3}$)	$M/M_\odot$	$2M/R$	n	T_n (msec)	τ_n (sec)	τ_n/T_n	$\frac{E_M}{\langle(\delta R/R)^2\rangle}$ (ergs)	$\frac{\text{Power}}{\langle(\delta R/R)^2\rangle}$ (ergs sec^{-1})	$\frac{(\delta R/R)}{(\delta r/r)_c}$	$\delta\theta_s/\delta\theta_c$
H-W	3×10^{14}	0.405	0.0574	0	1.197	13.	11000	7.8×10^{50}	1.2×10^{50}	+ 7.4	+ 3.1
H-W	6×10^{15}	0.682	0.240	0	0.3109	0.19	610	2.8×10^{52}	2.9×10^{53}	+ 5.2	+ 3.7
				1	0.1713	0.28	1600	3.6×10^{51}	2.6×10^{52}	− 14.	− 3.3
				2	0.1179	1.3	11000	2.6×10^{50}	3.9×10^{50}	+ 55.	+ 5.9
				3	0.0938	24.	250000	8.9×10^{48}	$7. \times 10^{47}$	−350.	−24.
V_γ	5.15×10^{14}	0.677	0.159	0	0.6991	1.7	2400	5.7×10^{52}	$7. \times 10^{52}$	+ 1.4	+ 1.3
				1	0.2358	11.	47000	6.0×10^{50}	1.1×10^{50}	− 38.	− 4.7
V_γ	3×10^{15}	1.954	0.580	0	0.3777	0.22	600	1.7×10^{54}	1.6×10^{55}	+ 1.9	+ 3.1
				1	0.1556	1.6	10000	1.5×10^{54}	1.9×10^{54}	− 2.1	− 0.66
				2	0.1026	2.6	25000	5.2×10^{53}	4.0×10^{53}	+ 2.9	+ 0.40

2	0.1026	2.6	25000	5.2×10^{53}	4.0×10^{53}	+ 2.9	+ 0.40

The columns in the table have the following meanings.

Equation of state: the equation of state $p(\rho)$ used in constructing the equilibrium stellar model and in calculating the adiabatic index from $\gamma = [(\rho + p)/p]\, dp/d\rho$; H-W is the Harrison-Wheeler equation of state in the tabular form given by Hartle and Thorne (1968), Table 1; V_γ is the Levinger-Simmons-Tsuruta-Cameron V_γ equation of state in the tabular form given by Hartle and Thorne (1968), Table 2.

ρ_c: central density of total mass-energy for the equilibrium stellar model.

$M/M_\odot$: total mass-energy of the equilibrium model (i.e., the mass that governs distant Keplerian orbits), in units of the sun's mass.

$2M/R = 2GM/Rc^2$: ratio of the gravitational radius of the equilibrium model to its actual radius (radii are defined by $4\pi R^2$ = surface area).

n: the "order" of the pulsational normal mode under study (for all models given here, n is also the number of nodes in the radial relative eigenfunction, $\delta r/r$.). Note: $n = 0$ is the fundamental (quadrupole) mode.

$T_n = 2\pi/\omega_n$: the pulsation period of the quasinormal mode measured in milliseconds.

τ_n: the damping time for the amplitude of the normal mode measured in seconds.

$\tau_n/T_n = \omega_n \tau_n / 2\pi$: the number of pulsation periods required for the amplitude to drop by a factor of $1/e$.

$E_{\mathfrak{M}}/\langle(\delta R/R)^2\rangle$: energy of pulsation of the star, divided by the square of the relative amplitude of radial motion of the star's surface averaged over its surface.

Power$/\langle(\delta R/R)^2\rangle$: the power radiated as gravitational waves, divided by the averaged square of the relative amplitude at the star's surface.

$(\delta R/R)(\delta r/r)_c^{-1}$: relative amplitude of radial motion at the star's surface divided by relative amplitude at the star's center.

$\delta\theta_s/\delta\theta_c = \delta\phi_s/\delta\phi_c$: amplitude of the angular displacement of the star's fluid at its surface divided by the same amplitude at its center.

Of the sources discussed in this section, most are "impulsive" rather than continuous (star collapsing through gravitational radius; debris falling into a black hole; collision between black holes; supernova explosion). They give rise to bursts of gravitational waves. An order-of-magnitude method of analyzing such bursts is spelled out in Box 36.2.

It is difficult and risky to pass from the above description of processes that should generate gravitational waves to an estimate of the characteristics of the waves that actually bathe the earth. For such an estimate, made in 1972 and subject to extensive revision as one's understanding of the universe improves, see Press and Thorne (1972).

EXERCISES

Exercise 36.2. GRAVITATIONAL WAVES FROM MATTER FALLING INTO A BLACK HOLE

A lump of matter with mass m falls into a black hole of mass M. Show that a burst of gravitational waves is emitted with duration $\sim M$ and power $L_{GW} \sim (m/M)^2 L_0$, so that the total energy radiated is given in crude order of magnitude by equation (36.13).

Exercise 36.3. GRAVITATIONAL WAVES FROM A VIBRATING NEUTRON STAR

Idealize a neutron star as a sphere of incompressible fluid of mass M and radius R, with structure governed by Newton's laws of gravity. Let the star pulsate in its fundamental quadrupole mode. Using Newtonian theory, calculate: the angular frequency of pulsation, ω; the energy of pulsation E_{puls}; the quantity $\frac{1}{5}\langle \dddot{I}^{\,2}\rangle$, which, according to equation (36.1), is the power radiated in gravitational waves, L_{GW}; and the e-folding time, $\tau = E_{\text{puls}}/L_{GW}$, for radiating away the energy of the pulsations. Compare the answers with equations (36.15)—which are based on a much cruder approximation—and with the results in Box 36.1, which are based on much better approximations. [For solution, see Table 13 of Wheeler (1966).]

Exercise 36.4. PULSAR SLOWDOWN

The pulsar NPO532 in the Crab Nebula has a period of 0.033 seconds and is slowing down at the rate $dP/dt = 1.35 \times 10^{-5}$ sec/yr. Assuming the pulsar is a typical neutron star, calculate the rate at which it is losing rotational energy. If this energy loss is due primarily to gravitational radiation reaction, what is the magnitude of the star's nonaxial deformation? [For solution, see Ferrari and Ruffini (1969); for a rigorous strong-field analysis, see Ipser (1970).]

§36.6. BINARY STARS AS SOURCES

Binary stars as sources of gravitational waves:

The most numerous sources of weak gravitational waves are binary star systems. Moreover, roughly half of all stars are in binary or multiple systems [see, for example, the compilation of Allen (1962)]. According to Kepler's laws, two stars of masses m_1 and m_2 that circle each other have angular frequency ω and separation a coupled to each other by the formula

$$\omega^2 a^3 = m_1 + m_2 \equiv M.$$

Box 36.2 ANALYSIS OF BURSTS OF RADIATION FROM IMPULSE EVENTS*

	Electromagnetism	*Gravitation*
Typical moment relevant for radiation	$d_x(t)$	$\rlap{-}I_{xz}(t)$
Its Fourier transform	$(2\pi)^{-1/2}\int d_x \exp[i\omega t]\,dt$	$(2\pi)^{-1/2}\int \rlap{-}I_{xz} \exp[i\omega t]\,dt$
Name for this quantity	$d_x(\omega)$	$\rlap{-}I_{xz}(\omega)$
Time decomposition of total radiative energy loss ΔE	$c^{-3}\int \ddot{d}^2(t)\,dt$	$Gc^{-5}\int \dddot{\rlap{-}I}^2(t)\,dt$
Decomposition of ΔE according to circular frequency	$c^{-3}\int \ddot{d}^2(\omega)\,d\omega$	$Gc^{-5}\int \dddot{\rlap{-}I}^2(\omega)\,d\omega$
Integrand nearly constant with respect to ω from $\omega = 0$ up to a critical value of ω, beyond which radiation falls off very fast	$\omega_{\text{crit}} \sim 1/\Delta t$	$\omega_{\text{crit}} \sim 1/\Delta t$
$-d\,\Delta E/d\omega$ for $\omega < \omega_{\text{crit}}$	$\sim c^{-3}\ddot{d}^2(0)$	$\sim Gc^{-5}\dddot{\rlap{-}I}^2(0)$
Zero frequency moment that enters this formula	$\sim(e_1\,\Delta v_{x1} + e_2\,\Delta v_{x2})$	$\Delta(\text{«Kinetic Energy»})_{xz}$
Rewrite of $-d\,\Delta E/d\omega$	$\sim(e\,\Delta v)^2/c^3$	$\sim G[\Delta(\text{«K.E.»})_{xz}]^2/c^5$
Total energy of pulse	$\sim$This$/\Delta t$	$\sim$This$/\Delta t$

*Box adapted from pp. 113 and 114 of Wheeler (1962).

As sample applications of this analysis, Wheeler (1962) cites the following:

Parameter	*One atomic-nucleus fission of 180 MeV*	*Fission bomb yield 17 kilotons at 10% efficiency*	*Meteorite striking earth at escape velocity*	*Explosion of star when 10^{-4} of mass is released*
Mass	4×10^{-22} g	10^4 g	10^9 g	2×10^{33} g
Velocity	1.2×10^9 cm/s	4×10^8 cm/s	11×10^5 cm/s	4×10^8 cm/s
Energy	2.9×10^{-4} erg	7×10^{20} erg	6×10^{20} erg	1.8×10^{50} erg
Fraction assumed relevant to radiative moment	1	0.1	1	0.1
Time integral of this moment $=$ «K.E.»$_{xz}$	2.9×10^{-4} erg	7×10^{19} erg	6×10^{20} erg	1.8×10^{49} erg
«K.E.»$_{xz}/c^2$	3.2×10^{-25} g	0.08 g	0.67 g	2×10^{28} g
$\dfrac{dE}{d\omega} \sim \dfrac{G}{c}\left(\dfrac{\text{«K.E.»}_{xz}}{c^2}\right)^2$	$2.3 \times 10^{-67}\dfrac{\text{erg}}{\text{rad/s}}$	$1.4 \times 10^{-20}\dfrac{\text{erg}}{\text{rad/s}}$	$1.0 \times 10^{-18}\dfrac{\text{erg}}{\text{rad/s}}$	$9 \times 10^{38}\dfrac{\text{erg}}{\text{rad/s}}$
Δt	10^{-21} s	10^{-8} s	10^{-3} s	10^4 s
$\Delta\omega \sim 1/\Delta t$	10^{21} rad/s	10^8 rad/s	10^3 rad/s	10^{-4} rad/s
$\Delta E_{\text{radiated}}$	10^{-46} erg	10^{-12} erg	10^{-15} erg	10^{35} erg
Assumed distance to detector	10^3 cm	10^3 cm	10^9 cm	10^{23} cm
$\Delta E/4\pi r^2$	10^{-53} erg/cm^2	10^{-19} erg/cm^2	10^{-34} erg/cm^2	10^{-12} erg/cm^2

The reader might find it informative to extend this table to the bursts of waves emitted by (1) debris falling into a black hole, (2) collisions between two black holes, and (3) a supernova explosion in which a star of two solar masses collapses to nuclear densities, ejecting half its mass in the process.

In this motion the kinetic energy is

$$(\text{kinetic energy}) = -\frac{1}{2}(\text{potential energy}) = \frac{1}{2}\frac{m_1 m_2}{a}.$$

The power that they radiate as gravitational waves can be estimated roughly as the square of the circulating power, $L \sim \omega \times$ (kinetic energy); thus,

$$L_{\text{GW}} \sim \frac{\mu^2 M^3}{4a^5} L_o,$$

where $\mu = m_1 m_2/M$ is the familiar reduced mass, and $M = m_1 + m_2$ is the total mass of this binary system.

(1) power output

An *exact* calculation based on equation (36.1) gives a result larger than this by a factor $\sim$30: for a binary system of semimajor axis a and eccentricity ε, the power output averaged over an orbital period is

$$L_{\text{GW}} = \frac{32}{5}\frac{\mu^2 M^3}{a^5} f(\varepsilon) L_o, \tag{36.16a}$$

where $f(\varepsilon)$ is the dimensionless "correction function,"

$$f(\varepsilon) = \left[1 + \frac{73}{24}\varepsilon^2 + \frac{37}{96}\varepsilon^4\right][1 - \varepsilon^2]^{-7/2}. \tag{36.16b}$$

[See exercise 36.6 at end of §36.8; also Peters and Mathews (1963).]

(2) effects of radiation reaction

As the binary system loses energy by gravitational radiation, the stars spiral in toward each other (decrease of energy; tightening of gravitational binding). For circular orbits the energy, $E = -\frac{1}{2}m_1 m_2/a = -\frac{1}{2}\mu M/a$, decreases as

$$\begin{aligned} dE/dt &= 1/2(\mu M/a^2)(da/dt) \\ &= -L_{\text{GW}} = -\frac{32}{5}\frac{\mu^2 M^3}{a^5}. \end{aligned}$$

Consequently, the evolution of the orbital radius is given by the formula

$$a = a_o(1 - t/\tau_o)^{1/4}, \tag{36.17a}$$

where $a_o = a_{\text{today}}$ and

$$\tau_o = \frac{1}{4}\left(\frac{-E}{L_{\text{GW}}}\right)_{\text{today}} = \frac{5}{256}\frac{a_o{}^4}{\mu M^2}. \tag{36.17b}$$

Thus, unless nongravitational forces intervene, the two stars will spiral together in a time τ_o (*spiral time*). For an elliptical orbit, the eccentricity also evolves. Radiation is emitted primarily at periastron. Therefore the braking forces of radiation reaction act there with greatest force. This effect deprives the stars of some of the kinetic energy of the excursions in their separation ("radial kinetic energy"). In consequence, the orbit becomes more nearly circular. [See Peters and Mathews (1963) for detailed calculations.]

The calculated power output, flux at Earth, and damping times are shown in Box 36.3 for several known binary stars and several interesting hypothetical cases. Notice that in the most favorable known cases the period is a few hours; the damping time is the age of the universe (could the absence of better cases be due to radiation reaction's having destroyed them?); the output of power in the form of gravitational waves is $\sim 10^{30}$ to 10^{32} ergs/sec (approaching the light output of the sun, 3.9×10^{33} ergs/sec); and the calculated flux at the Earth is $\sim 10^{-10}$ to 10^{-12} ergs/sec (too small to detect in 1973, but perhaps not too small several decades hence; see Chapter 37).

(3) particular binaries observed by astronomers

The hypothetical cases in Box 36.3 illustrate the general relations for astrophysical systems that were derived in §36.4—namely, that only as the system approaches its gravitational radius can $L_{\rm GW}$ approach L_o, and only then can damping remove nearly the whole energy in a single period.

§36.7. FORMULAS FOR RADIATION FROM NEARLY NEWTONIAN SLOW-MOTION SOURCES

Turn now from illustrative astrophysical sources to rigorous formulas valid for a wide variety of sources. One such formula has already been written down,

$$L_{\rm GW} = \frac{1}{5}\langle \dddot{I\!\!\!-}_{jk}\dddot{I\!\!\!-}_{jk}\rangle, \tag{36.1}$$

but it has not yet been derived, nor has its realm of validity been discussed.

This formula for the power output is actually valid for any "nearly Newtonian, slow-motion source"—more particularly, for any source in which

The "nearly Newtonian, slow-motion approximation" for analyzing sources of gravitational waves

$$\text{(size of source)/(reduced wavelength of waves)} \ll 1, \tag{36.18a}$$

$$|\text{Newtonian potential}| \ll \text{(size of source)/(reduced wavelength)}, \tag{36.18b}$$

$$\frac{|\text{typical stresses}|}{\text{(mass density)}} \ll \frac{\text{(size of source)}}{\text{(reduced wavelength)}}. \tag{36.18c}$$

It is not valid, except perhaps approximately, for fast-motion or strong-field sources. Moreover, there is no formalism available today which can handle effectively and *in general* the fast-motion case or the strong-field case.

The rest of this chapter is devoted to a detailed analysis of gravitational waves from nearly Newtonian, slow-motion sources. But the analysis (Track 2; §§36.9–36.11) will be preceded by a Track-1 summary in this section and the next.

For any source of size R and mean internal velocity v, the characteristic reduced wavelength ($\lambda\!\!\!^- = \lambda/2\pi$) of the radiation emitted is $\lambda\!\!\!^- \sim$ (amplitude of motions)$/v \lesssim R/v$. Consequently the demand (36.18a) that $R/\lambda\!\!\!^-$ be $\ll 1$ [i.e., that the source be confined to a small region deep inside the near (nonradiation) zone] enforces the slow-motion constraint

$$v \ll 1.$$

Box 36.3 GRAVITATIONAL RADIATION FROM SEVERAL BINARY STAR SYSTEMS[a]

Type of system	*Name*	*Period*	$m_1/M_\odot$	$m_2/M_\odot$	*Distance from earth (pc)*	*Spiral time*[b]	L_{GW} *(ergs/sec)*	*Flux at earth (erg/sec cm²)*
Solar System (Sun + Jupiter)	Solar System	11.86 yr.	1.0	9.56×10^{-4}	Earth is in near zone	2.5×10^{23} yr	5.2×10^{10}	—
Typical resolved binaries from compilation of Van de Kamp (1958)	η Cas	480 yr.	0.94	0.58	5.9	9.5×10^{24}	5.6×10^{10}	1.4×10^{-29}
	ξ Boo	149.95 yr.	0.85	0.75	6.7	3.8×10^{23}	3.6×10^{12}	6.7×10^{-28}
	Sirius	49.94 yr.	2.28	0.98	2.6	7.2×10^{21}	1.1×10^{15}	1.3×10^{-24}
	Fu 46	13.12 yr.	0.31	0.25	6.5	3.2×10^{21}	3.6×10^{14}	7.1×10^{-26}
Typical eclipsing binaries from compilation of Gaposhkin (1958)	β Lyr	12.925 day	19.48	9.74	330	7.0×10^{11}	0.057×10^{30}	0.0004×10^{-11}
	UWCMa	4.395 day	40.0	31.0	1470	8.2×10^{9}	$49. \times 10^{30}$	0.019×10^{-11}
	β Per	2.867 day	4.70	0.94	30	3.2×10^{11}	0.014×10^{30}	0.013×10^{-11}
	WUMa	0.33 day	0.76	0.57	110	6.2×10^{9}	0.47×10^{30}	0.032×10^{-11}
Favorable cases from compilation of Braginsky (1965)	UV Leo	0.6 day	1.36	1.25	68	1.0×10^{10}	0.63×10^{30}	0.012×10^{-11}
	V Pup	1.45 day	16.6	9.8	390	2.3×10^{9}	$65. \times 10^{30}$	0.36×10^{-11}
	i Boo	0.268 day	1.35	0.68	12	2.0×10^{9}	3.2×10^{30}	$18. \times 10^{-11}$
	YY Eri	0.321 day	0.76	0.50	42	6.6×10^{9}	0.42×10^{30}	0.20×10^{-11}
	SW Lac	0.321 day	0.97	0.83	75	3.5×10^{9}	1.5×10^{30}	0.21×10^{-11}
	WZ Sge	81 min	0.6	0.03	100	1.1×10^{9} yr	0.5×10^{30}	0.04×10^{-11}
Hypothetical binaries (neutron stars or black holes)	10^4 km	12.2 sec	1.0	1.0	1000	3.2 yr	3.25×10^{41}	2.7×10^{-3}
	10^3 km	0.39 sec	1.0	1.0	1000	2.8 hr	3.25×10^{46}	2.7×10^{2}
	10^2 km	12.2 msec	1.0	1.0	1000	1.0 sec	3.25×10^{51}	2.7×10^{7}
	10 km	0.39 msec	1.0	1.0	1000	0.10 msec	3.25×10^{56}	2.7×10^{12}

[a] Based on tables by Braginsky (1965) and by Ruffini and Wheeler (1971b).
[b] The spiral time, τ_0, as given by equation (36.17b) is the time for the two stars to spiral into each other if no nongravitational forces intervene.

These related conditions, $v \ll 1$ and $R \ll \lambda\!\!\!^{-}$, are satisfied by all presently conceived laboratory generators of gravitational waves. No one has seen how to bring a macroscopic mass up to a speed $v \sim 1$. These conditions are also satisfied by every gravitationally bound, nearly Newtonian system. Thus, for such a system of mass M, the condition for gravitational binding, $\frac{1}{2}Mv^2 \lesssim M^2/R$ guarantees that $v \lesssim (M/R)^{1/2} \ll 1$.

The conditions $M/R \ll R/\lambda\!\!\!^{-}$ and $|T^{jk}|/T^{00} \ll R/\lambda\!\!\!^{-}$ are satisfied by all nearly Newtonian sources of conceivable interest. Typical sources (e.g. binary stars) have

$$\frac{M}{R} \sim \frac{|T^{jk}|}{T^{00}} \sim \left(\frac{R}{\lambda\!\!\!^{-}}\right)^2 \ll \frac{R}{\lambda\!\!\!^{-}}$$

(virial theorem). In those rare cases where $(M/R$ or $|T^{jk}|/T^{00}) \gtrsim R/\lambda\!\!\!^{-}$ (e.g., a marginally stable, slowly vibrating star), the motion is so very slow that the radiation will be too weak to be interesting.

For any nearly Newtonian slow-motion system, there is a spacetime region deep inside the near zone ($r \ll \lambda\!\!\!^{-}$), but outside the boundary of the source ($r > R$), in which vacuum Newtonian gravitation theory is nearly valid. An observer in this Newtonian region can measure the Newtonian potential Φ and can expand it in powers of $1/r$:

Definitions of mass, dipole moment, and reduced quadrupole moment for a slow-motion source

$$\Phi = -\left(\frac{M}{r} + \frac{d_j n^j}{r^2} + \frac{3 I\!\!\!-_{jk} n^j n^k}{2r^3} + \cdots\right), \qquad \text{where } n^j = x^j/r. \qquad (36.19a)$$

He can then give names to the coefficients in this expansion:

$$\begin{aligned} M &\equiv \text{“total mass-energy”} = \text{“active gravitational mass”};\\ d_j &\equiv \text{“dipole moment” [if he chooses the origin of coordinates carefully, he can make } d_j = 0];\\ I\!\!\!-_{jk} &\equiv \text{“reduced quadrupole moment” [because the system is nearly Newtonian, } I\!\!\!-_{jk} \text{ is given by expression (36.3)].}\end{aligned} \qquad (36.19b)$$

As this Newtonian potential reaches out into the radiation zone, the static portions of it $(-M/r - d_j n^j/r^2)$ maintain their Newtonian form, unchanged. But the dynamic part $(-\frac{3}{2} I\!\!\!-_{jk} n^j n^k/r^3)$ ceases to be describable in Newtonian terms. As retardation effects become noticeable (at increasing r values), it gradually changes over into outgoing gravitational waves, which must be described in the full general theory of relativity, or in linearized theory, or in the "shortwave" approximation of §35.13.

If one chooses to use linearized theory in the radiation zone, and if one imposes the transverse-traceless gauge there ($h^{TT}_{0\mu} = 0$, $h^{TT}_{jj} = 0$, $h^{TT}_{jk,k} = 0$), then the gravitational waves take the form [derived later as equation (36.47)]

Properties of gravitational waves in terms of reduced quadrupole moment:

$$h^{TT}_{jk} = \frac{2}{r}\ddot{I\!\!\!-}^{TT}_{jk}(t - r) + \text{corrections of order}\left[\frac{1}{r^2}\dot{I\!\!\!-}^{TT}_{jk}(t - r)\right]. \qquad (36.20)$$

(1) the wave field h^{TT}_{jk}

Here $\ddot{I}^{TT}_{jk}$ is the second time-derivative of the transverse-traceless part of the quadrupole moment (transverse to the radial direction; see §35.4); thus,

$$\ddot{I}^{TT}_{jk} = P_{ja}\ddot{I}_{ab}P_{bk} - \frac{1}{2}P_{jk}P_{ab}\ddot{I}_{ab},$$

$$P_{ab} = (\delta_{ab} - n_a n_b) \qquad \text{(projection operator)}, \tag{36.21}$$

$$n_a = x^a/r \qquad \text{(unit radial vector)}.$$

The effective stress-energy tensor for these outgoing waves (§35.7) has the same form as for a swarm of zero-mass particles traveling radially outward with the speed of light; at large distances its components of lowest nonvanishing order are

(2) effective stress-energy tensor

$$T^{(\mathrm{GW})}_{00} = -T^{(\mathrm{GW})}_{0r} = T^{(\mathrm{GW})}_{rr} = \frac{1}{32\pi}\langle h^{TT}_{jk,0}h^{TT}_{jk,0}\rangle = \frac{1}{8\pi r^2}\langle \dddot{I}^{TT}_{jk}\dddot{I}^{TT}_{jk}\rangle$$

$$= \frac{1}{8\pi r^2}\left\langle \dddot{I}_{jk}\dddot{I}_{jk} - 2n_i\dddot{I}_{ij}\dddot{I}_{jk}n_k + \frac{1}{2}(n_j\dddot{I}_{jk}n_k)^2\right\rangle, \tag{36.22}$$

where $\langle\ \ \rangle$ denotes an average over several wavelengths. (Recall that one cannot localize the energy more closely than a wavelength!) The total power crossing a sphere of radius r at time t is

(3) total power radiated

$$L_{\mathrm{GW}}(t, r) = \int T^{(\mathrm{GW})0r}r^2\, d\Omega = \frac{1}{5}\langle \dddot{I}_{jk}(t-r)\dddot{I}_{jk}(t-r)\rangle. \tag{36.23}$$

(See exercise 36.9.) This is the formula with which this chapter began: equation (36.1).

The wave fronts are not precisely spherical. For example, for a binary star system the wave fronts in the equatorial plane must be spirals. This means that there is a tiny nonradial component of the momentum flux, which decreases in strength as $1/r^3$. Associated with this nonradial momentum is an angular momentum density (angular momentum relative to the system's center, $r = 0$), which drops off as $1/r^2$ [Peters (1964), as corrected by DeWitt (1971), p. 286]:

(4) density of angular momentum

$$\mathcal{J}^i = \frac{1}{8\pi r^2}\epsilon^{ijk}\langle -6n_j\ddot{I}_{km}\dddot{I}_{mp}n_p + 9n_j\ddot{I}_{km}n_m n_p\dddot{I}_{pq}n_q\rangle. \tag{36.24}$$

The integral of this quantity over a sphere is the total angular momentum being transported outward per unit time,

(5) total angular momentum radiated

$$-dJ_j/dt = \int \mathcal{J}_j r^2\, d\Omega = \frac{2}{5}\epsilon^{jkl}\langle \ddot{I}_{ka}\dddot{I}_{al}\rangle. \tag{36.25}$$

(See exercise 36.9.)

§36.8. RADIATION REACTION IN SLOW-MOTION SOURCES*

The conservation laws discussed in Box 19.1 and derived in §20.5 guarantee that the source must lose energy and angular momentum at the same rate as the gravitational waves carry them off. The agent that produces these losses is a tiny component of the spacetime curvature inside the source, which reverses sign if one changes from a (realistic) outgoing-wave boundary condition at infinity to the opposite (unrealistic) ingoing-wave condition. These "radiation-reaction" pieces of the curvature can be described in Newtonian language when the source obeys the nearly Newtonian, slow-motion conditions (36.18).

Outgoing-wave boundary condition gives rise to a Newtonian-type radiation-reaction potential

The dynamical part of the Newtonian potential, in its "standard form"

$$\Phi = -\frac{3}{2}{I\!\!\!-}_{jk}(t)n_j n_k/r^3 + O(1/r^4); \text{ equation (36.18),}$$

has no retardation in it. (Newtonian theory demands action at a distance!) Consequently, there is no way whatsoever for the standard potential to decide, at large radii, whether to join onto outgoing waves or onto ingoing waves. Being undecided, it takes the middle track of joining onto standing waves (half outgoing, plus half ingoing). But this is not what one wants. It turns out (see §36.11) that the join can be made to purely outgoing waves if and only if Φ is augmented by a tiny "radiation-reaction" potential

Form and magnitude of the radiation-reaction potential

$$\Phi = \Phi_{\text{standard Newtonian theory}} + \Phi^{(\text{react})}, \tag{36.26a}$$

$$\Phi^{(\text{react})} = \frac{1}{5}\frac{d^5{I\!\!\!-}_{jk}}{dt^5}x^j x^k. \tag{36.26b}$$

If, instead, one sets $\Phi = \Phi_{\text{standard}} - \Phi^{(\text{react})}$, the potential will join onto purely ingoing waves.

In order of magnitude, the radiation-reaction potential is

$$\Phi^{(\text{react})} \sim \frac{1}{{\lambda\!\!\!^{-}}^5}(MR^2)r^2 \sim \frac{MR^2}{r^3}\left(\frac{r}{\lambda\!\!\!^{-}}\right)^5. \tag{36.27}$$

Consequently, near the source it is tiny compared to the standard Newtonian potential [a factor $(R/\lambda\!\!\!^{-})^5 \sim v^5$ smaller!]. However, at the inner boundary of the radiation zone ($r \sim \lambda\!\!\!^{-}$), it is of the same order of magnitude as the dynamic, quadrupole part of the standard potential.

Effects of the potential:

The radiation-reaction part of the Newtonian potential plays the same role as a producer of accelerations that any other part of the Newtonian potential does. Any particle in the Newtonian region experiences a gravitational acceleration given by

(1) radiation-reaction accelerations

$$a_j = -\Phi_{,j} = -\Phi_{\text{standard},j} - \Phi^{(\text{react})}_{,j}. \tag{36.28}$$

*The ideas and formalism described in this section were devised by Burke (1970), Thorne (1969b), and Chandrasekhar and Esposito (1970). Among the forerunners of these ideas were the papers of Peters (1964), and Peres and Rosen (1964).

Inside the source, this acceleration leads to energy and angular momentum losses given by

$$dE/dt = \int \rho a_j v_j \, d^3x \tag{36.29a}$$

and

$$dJ_j/dt = \int \varepsilon_{jk\ell} x_k \rho a_\ell \, d^3x. \tag{36.29b}$$

(Here ρ is the density, v_j is the velocity, and a_j as above is the acceleration of the matter in the source.) Standard Newtonian theory conserves the energy and angular momentum. Therefore only the reaction part of the potential can produce losses:

(2) loss of energy and angular momentum

$$\begin{aligned} dE/dt &= -\int \rho \Phi^{(\mathrm{react})}_{,j} v_j \, d^3x, \\ dJ_j/dt &= -\int \varepsilon_{jk\ell} x_k \rho \Phi^{(\mathrm{react})}_{,\ell} \, d^3x. \end{aligned} \tag{36.30}$$

A straightforward calculation (exercise 36.5) using expression (36.26b) for the reaction potential yields, for the time-averaged losses,

$$\begin{aligned} dE/dt &= -\frac{1}{5}\langle \dddot{I\!\!\!-}_{jk} \dddot{I\!\!\!-}_{jk} \rangle, \\ dJ_j/dt &= -\frac{2}{5}\varepsilon_{jk\ell}\langle \ddot{I\!\!\!-}_{ka} \dddot{I\!\!\!-}_{a\ell} \rangle. \end{aligned} \tag{36.31}$$

Notice that these results agree with the energy and angular momentum carried off by the radiation as given by equations (36.1) and (36.25). The agreement is an absolute imperative. The laws of conservation of total energy and angular momentum demand it.

A slow-motion electromagnetic system emitting electric dipole radiation has a radiation-reaction potential

Radiation-reaction potential for electromagnetic waves

$$A_j^{(\mathrm{react})} = 0, \qquad A_0^{(\mathrm{react})} = -\Phi^{(\mathrm{react})} = \frac{2}{3}\dddot{d}_j x^j, \tag{36.32}$$

which is completely analogous to $\Phi^{(\mathrm{react})}$ of gravitation theory [see, e.g., Burke (1971)]. However, attention does not usually focus on this potential and the reaction forces it produces. Instead, it focuses on the reaction force in a special case: that of an isolated charge being accelerated by nonelectromagnetic forces. For such a charge, the reaction force is

$$\boldsymbol{F}^{(\mathrm{React})} = \frac{2}{3}e^2 \dddot{\boldsymbol{x}}. \tag{36.33}$$

No such formula is relevant to gravitation theory, because there is no such thing as a gravitationally isolated, radiating particle (i.e., one accelerated by forces that have no coupling to gravity).

EXERCISES

Exercise 36.5. ENERGY AND ANGULAR MOMENTUM LOSSES DUE TO RADIATION REACTION

Derive equations (36.31) for the rate at which gravitational radiation damping saps energy and angular momentum from a slow-motion source. Base the derivation on equations (36.26b) and (36.30).

Exercise 36.6. GRAVITATIONAL WAVES FROM BINARY STAR SYSTEMS

Apply the full formalism of §§36.7 and 36.8 to a binary star system with circular orbits. Calculate the angular distribution of the gravitational waves; the total power radiated; the total angular momentum radiated; the radiation-reaction forces; and the loss of energy and angular momentum due to radiation reaction. Compare the answers with the results quoted in §36.6. [For further details of the solution, see Peters and Mathews (1963).]

1

2

§36.9. FOUNDATIONS FOR DERIVATION OF RADIATION FORMULAS

The rest of this chapter is Track 2. Chapter 20 (conservation laws) is needed as preparation for it. It will be helpful in Chapter 39 (post-Newtonian formalism), but is not needed as preparation for any other chapters.

Turn now from the formulas for radiation from a nearly Newtonian system in slow motion to a derivation of these formulas. Initially (this section) work in the full general theory of relativity without any approximations—not even that of slow motion. Impose only the constraint that the source be isolated, and that spacetime become asymptotically flat far away from it.

Derivation of formula for the gravitational-wave field produced by a slow-motion source:

Use a coordinate system that becomes asymptotically Lorentz as rapidly as spacetime curvature permits, when one moves radially outward from the source toward infinity. Everywhere in this coordinate system, even inside the source, which may be relativistic, define

$$h_{\mu\nu} \equiv g_{\mu\nu} - \eta_{\mu\nu}. \tag{36.34}$$

The $h_{\mu\nu}$ are clearly not the components of a tensor. Neither is $\eta_{\mu\nu}$ the true metric tensor. Nevertheless, one is free to raise and lower indices on $h_{\mu\nu}$ with $\eta_{\mu\nu}$ and to *define*

(1) definition of $\bar{h}_{\mu\nu}$

$$\bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h; \qquad h = h_\alpha{}^\alpha = h_{\alpha\beta}\eta^{\alpha\beta}. \tag{36.35}$$

Moreover, one can always specialize the coordinates so that the four conditions

$$\bar{h}_\mu{}^\alpha{}_{,\alpha} = 0 \tag{36.36}$$

are exactly satisfied everywhere, including the interior of the source.

With these definitions and conventions, $\bar{h}_{\mu\nu}$ becomes the gravitational field of linearized theory far from the source, and also inside the source if gravity is weak there. But if the interior gravity is strong ($|\bar{h}_{\mu\nu}|$ not $\ll 1$), $\bar{h}_{\mu\nu}$ in the interior has no connection whatsoever to linearized theory.

The exact Einstein field equations can be written in terms of $\bar{h}^{\mu\nu}$ as [cf. §20.3; in particular, combine equations (20.14), (20.18), and (20.3); and impose the coordinate condition (36.36)]

(2) field equations in terms of $\bar{h}_{\mu\nu}$

$$\bar{h}^{\mu\nu}{}_{,\alpha\beta}\eta^{\alpha\beta} = -16\pi(T^{\mu\nu} + t^{\mu\nu}), \tag{36.37}$$

(3) philosophy of controlled ignorance

where $T^{\mu\nu}$ are the components of stress-energy tensor, and $t^{\mu\nu}$ are quantities (components of the "stress-energy pseudotensor for the gravitational field") that are of quadratic order and higher in $\bar{h}^{\mu\nu}$. Recall the "philosophy of controlled ignorance" expounded in §19.3. One is so ignorant that nowhere does one ever write down an explicit expression for $t^{\mu\nu}$ in terms of $\bar{h}^{\alpha\beta}$; and this ignorance is so controlled that one will never need such an expression in the calculations to follow! More specifically, the strength of the outgoing wave is proportional to the integral of a complicated expression over the interior of a system where "gravitational stresses" may be comparable to material stresses, $|t^{jk}| \sim |T^{jk}|$. No matter. All that will count for the radiation is the quadrupole part of the field. Moreover, that quadrupole moment is empirically definable by purely Newtonian measurements in the Newtonian region (1) well inside the wave zone, but (2) well outside the surface of the source. One does not have to know the inner workings of a star to define its mass (influence on Kepler orbits outside) nor does one have to know those inner workings to define its quadrupole moment *as sensed externally*.

Einstein's equations (36.37), augmented by an outgoing-wave boundary condition, are equivalent to the integral equations

(4) integral formulation of field equations

$$\bar{h}^{\mu\nu}(t, x^j) = 4\int_{\text{all space}} \frac{[T^{\mu\nu} + t^{\mu\nu}]_{\text{ret}}}{|\boldsymbol{x} - \boldsymbol{x}'|}\, d^3x', \tag{36.38}$$

where

$$|\boldsymbol{x} - \boldsymbol{x}'| \equiv \left[\sum_j (x^j - x^{j'})^2\right]^{1/2}, \qquad d^3x' \equiv dx^{1'}\, dx^{2'}\, dx^{3'},$$

and the subscript "ret" means the quantity is to be evaluated at the retarded spacetime point

$$(t' = t - |\boldsymbol{x} - \boldsymbol{x}'|, x^{j'}).$$

These are integral equations because the unknowns, $\bar{h}^{\mu\nu}$, appear both outside and inside the integral (inside they are contained in $t^{\mu\nu}$). Notice that in passing from the wave equations (36.37) to the integral equations (36.38), one has cavalierly behaved as though $\bar{h}^{\mu\nu}$ were fields in flat spacetime. This is certainly not true; but the mathematical manipulations are valid nevertheless!—and the integral equations (36.38) are valid for any field point (t, x^j), even inside the source.

§36.10. EVALUATION OF THE RADIATION FIELD IN THE SLOW-MOTION APPROXIMATION

(5) specialization to slow motion

Thus far the analysis has been exact. Now it is necessary to introduce the slow-motion assumption of §36.7: $R \ll \lambda\!\!\!^{-}$.

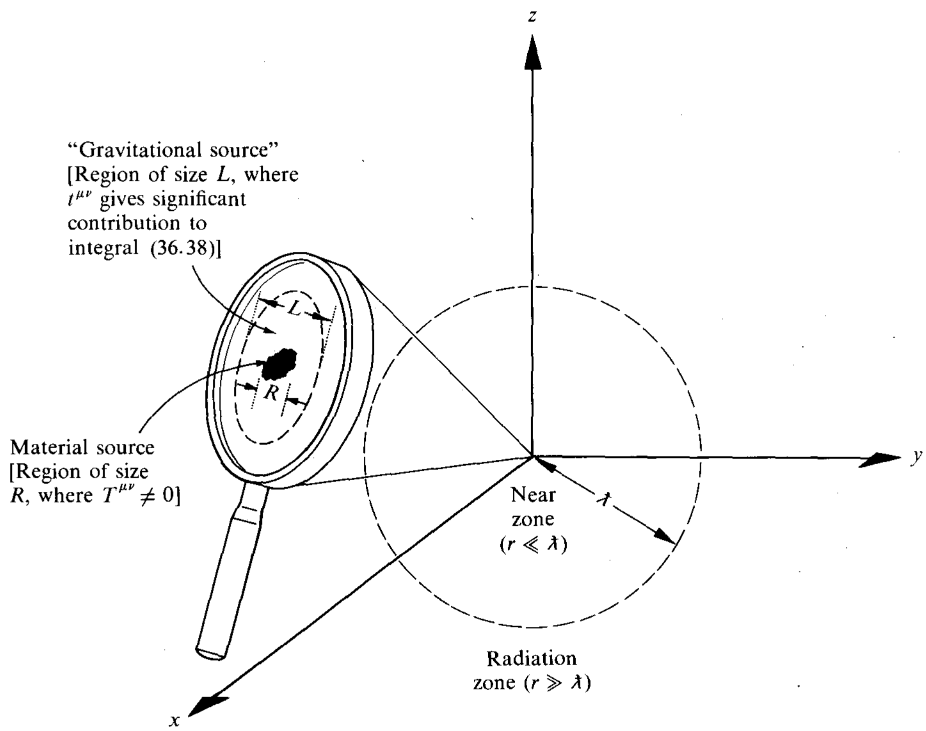


Figure 36.3.
A slow-motion source radiating gravitational waves. The origin of spatial coordinates is located inside the source. The size of the source, R, is very small compared to a reduced wavelength, $R \ll \lambda\!\!\!^{-}$. Significant contributions to the retarded integral (36.38) for $\bar{h}^{\mu\nu}$ come only from a region of size $L \sim R \ll \lambda\!\!\!^{-}$ surrounding the source, because outside the source—but in the near zone ($R \ll r \ll \lambda\!\!\!^{-}$)—the "stress-energy pseudotensor" $t^{\mu\nu}$ dies out as $1/r^4$ (see exercise 36.7).

In the radiation zone, $t^{\mu\nu}$ ceases to die out as $1/r^4$, and begins to die out as $1/r^2$; it is trying to describe (but cannot, really, without appropriate averaging) the stress-energy carried by the gravitational waves. If the source has been emitting waves long enough, contributions from the radiation zone to the retarded integral (36.38) may be nonnegligible:

$$\underbrace{[t^{\mu\nu}]_{\rm ret} \sim \frac{1}{r'^2}}_{[\text{for } r > \lambda\!\!\!^{-}]} \Longrightarrow \int [t^{\mu\nu}]_{\rm ret}\, d^3x' \sim \underbrace{\int \frac{1}{r'^2}\, r'^2\, d\Omega'\, dr'}_{\left[\substack{\text{may have significant contri-}\\ \text{butions from large } r'}\right]}.$$

Such contributions are ignored in the text, in calculations of the radiated waves, because they have nothing whatsoever to do with the emission process itself. Rather, they are part of the propagation process treated in the last chapter. They include the background curvature produced by the stress-energy of the waves, scattering of waves off the background curvature, wave-wave scattering, etc.; and they are totally negligible in the neighborhood of the source itself ($r \lesssim 1{,}000\, \lambda\!\!\!^{-}$, for example) because a slow-motion source radiates so very weakly.

Place the origin of spatial coordinates inside the source, as shown in Figure 36.3. For slow-motion systems, the only significant contributions to the retarded integrals (36.38) come from deep inside the near zone (from a region of size $L \sim R \ll \lambda\!\!\!^{-}$; see Figure 36.3). Confine attention to "field points" (points of observation) x^j far outside this "source region,"

$$|\boldsymbol{x}| \equiv r \gg L \gtrsim |\boldsymbol{x}'|, \tag{36.39a}$$

and expand the retarded integral (36.38) in powers of x'/r—in just the same manner as was done in §19.1. (Such an expansion is justified by and requires the slow-motion assumption, $\lambda\!\!\!^{-}/R \sim \lambda\!\!\!^{-}/L \ll 1$.) The result is

$$\begin{aligned} \bar{h}^{\mu\nu}(t, \boldsymbol{x}) = \frac{4}{r} \int [T^{\mu\nu}(\boldsymbol{x}', t - r) + t^{\mu\nu}(\boldsymbol{x}', t - r)]\, d^3x' \\ + O\left\{\frac{x^j}{r^2 \lambda\!\!\!^{-}} \int x^{j'}[T^{\mu\nu}(\boldsymbol{x}', t - r) + t^{\mu\nu}(\boldsymbol{x}', t - r)]\, d^3x'\right\}. \end{aligned} \tag{36.40}$$

(6) calculation of $\bar{h}^{jk}$ in radiation zone

Of the ten components of $\bar{h}^{\mu\nu}$, only the six spatial ones, $\bar{h}^{jk}$, are of interest, since only they are needed in projecting out the transverse-traceless radiation field $\bar{h}^{TT}_{jk}$. The spatial components are expressed by equations (36.40) in terms of integrals over the "stress distribution" $T^{jk} + t^{jk}$. It will be convenient, in making comparisons with Newtonian theory, to reexpress $\bar{h}^{jk}$ in terms of integrals over the "energy distribution" $T^{00} + t^{00}$. One can make the conversion with the help of the exact equations of motion $T^{\mu\nu}{}_{;\nu} = 0$, which have the special form

$$(T^{\mu\nu} + t^{\mu\nu})_{,\nu} = 0 \tag{36.41}$$

in the coordinate system being used [see equations (36.36) and (36.37); also the discussion in §20.3]. Applying these relations twice in succession, one obtains the identity

$$\begin{aligned} (T^{00} + t^{00})_{,00} &= -(T^{0l} + t^{0l})_{,l0} = -(T^{l0} + t^{l0})_{,0l} \\ &= +(T^{lm} + t^{lm})_{,ml}. \end{aligned}$$

From this and the elementary chain rule for differentiation, it follows that

$$\begin{aligned} [(T^{00} + t^{00})x^j x^k]_{,00} &= (T^{lm} + t^{lm})_{,ml} x^j x^k \\ &= [(T^{lm} + t^{lm})x^j x^k]_{,ml} - 2[(T^{lj} + t^{lj})x^k + (T^{lk} + t^{lk})x^j]_{,l} \\ &\quad + 2(T^{jk} + t^{jk}), \end{aligned}$$

whence

$$\int (T^{jk} + t^{jk})\, d^3x = \frac{1}{2}(d^2 I_{jk}/dt^2), \tag{36.42a}$$

where

$$I_{jk}(t) \equiv \int [T^{00}(t, \boldsymbol{x}) + t^{00}(t, \boldsymbol{x})] x^j x^k\, d^3x. \tag{36.42b}$$

(7) specialization to nearly Newtonian sources

Now introduce the nearly Newtonian assumption. It guarantees that gravitation contributes only a small fraction of the total energy:

$$t^{00} \sim (\Phi_{,j})^2 \sim M^2/R^4 \sim (M/R)T^{00} \ll T^{00};$$

hence

$$I_{jk}(t) = \int T^{00}(t, \boldsymbol{x}) x^j x^k\, d^3x. \tag{36.42b$'$}$$

The quantity I_{jk} thus represents the second moment of the mass distribution.

By combining equations (36.42) and (36.40), and by noting that inside the source $|t^{jk}| \sim |\Phi_{,j}\Phi_{,k}| \sim T^{00}|\Phi|$, one obtains

$$\begin{aligned}\bar{h}^{jk}(t, x) &= \frac{2}{r}\frac{d^2 I_{jk}(t-r)}{dt^2} + O\left[\frac{1}{r}\left(\frac{|T^{jk}|}{T^{00}} + |\Phi|\right)\frac{R}{\lambdabar}M\right] \\ &= \frac{2}{r}\frac{d^2 I_{jk}(t-r)}{dt^2}\left\{1 + \underbrace{O\left[\frac{|T^{jk}|}{T^{00}} + \frac{M}{R}\right]\frac{\lambdabar}{R}}_{\text{[negligible by assumptions (36.18)]}}\right\}. \qquad (36.43)\end{aligned}$$

(8) conversion, by projection, to h_{jk}^{TT}

Actually, what one wants are h_{jk}^{TT}, not $\bar{h}^{jk}$. They can be obtained by first lowering indices, using $\eta_{\ell m} = \delta_{\ell m}$, and then projecting out the TT part using the projection operator for radially traveling waves:

$$P_{\ell m} = \delta_{\ell m} - n_\ell n_m; \qquad n_\ell = x^\ell / r \qquad (36.44)$$

(see Box 35.1). (Because $\bar{h}_{jk}$ and h_{jk} differ only in the trace, they have the same TT parts). The result is

$$h_{jk}^{TT}(t, x) = \frac{2}{r}\frac{d^2 I_{jk}^{TT}(t-r)}{dt^2}, \qquad (36.45\text{a})$$

where

$$I_{jk}^{TT} = P_{j\ell} I_{\ell m} P_{mk} - \frac{1}{2} P_{jk}(P_{\ell m} I_{m\ell}). \qquad (36.45\text{b})$$

(9) reexpression of h_{jk}^{TT} in terms of reduced quadrupole moment

This is not the best form in which to write the answer, because an external observer cannot measure directly the second moment of the mass distribution, I_{jk}. Fortunately, one can replace I_{jk} by the reduced quadrupole moment,

$$\text{\it{I}}\!\!\!-_{jk} \equiv I_{jk} - \frac{1}{3}\delta_{jk} I = \int (T^{00} + t^{00})\left(x^j x^k - \frac{1}{3}\delta_{jk} r^2\right) d^3x, \qquad (36.46)$$

and write

$$h_{jk}^{TT}(t, x) = \frac{2}{r}\frac{d^2 \text{\it{I}}\!\!\!-_{jk}^{TT}(t-r)}{dt^2}. \qquad (36.47)$$

This is allowed because the TT parts of I_{jk} and $\text{\it{I}}\!\!\!-_{jk}$ are identical (exercise 36.8).

The reduced quadrupole moment $\text{\it{I}}\!\!\!-_{jk}$ has a well-defined, elementary physical significance for an observer confined to the exterior of the source. In the near zone ($r \ll \lambdabar$), but outside the source so that vacuum Newtonian theory is very nearly valid, the Newtonian potential is

$$\begin{aligned}\Phi &= -\frac{1}{2}h_{00} = -\frac{1}{2}h^{00} = -\frac{1}{2}\left(\bar{h}^{00} + \frac{1}{2}\bar{h}\right) = -\frac{1}{4}(\bar{h}^{00} + \bar{h}^{jj}) \\ &= -\int_{\text{all space}} \frac{[T^{00} + t^{00} + T^{jj} + t^{jj}]_{\text{ret}}}{|x - x'|} d^3x'\end{aligned}$$

[see equation (36.38)]. Any nearly Newtonian, slow-motion source satisfies

$$|t^{00} + T^{jj} + t^{jj}| \ll T^{00}$$

[recall: $t^{\alpha\beta} \sim (\Phi_{,j})^2 \sim T^{00}|\Phi|$]. Hence, one can write

$$\Phi(\boldsymbol{x}, t) = -\int \frac{[T^{00}(\boldsymbol{x}', t)]}{|\boldsymbol{x} - \boldsymbol{x}'|} d^3x'. \tag{36.48}$$

Expanding $|\boldsymbol{x} - \boldsymbol{x}'|^{-1}$ in powers of $1/r$, one obtains

$$\Phi = -\left(\frac{M}{r} + \frac{d_j x^j}{r^3} + \frac{3 \mathcal{I}_{jk} x^j x^k}{2r^5} + \cdots\right) \text{for} \begin{pmatrix} r \ll \lambda\!\!\!^{-}, \text{ but } r \text{ neverthe-} \\ \text{less large enough that} \\ \text{vacuum Newtonian} \\ \text{theory is valid} \end{pmatrix}, \tag{36.49a}$$

where

$$M = (\text{total mass-energy of source}) = \int T^{00}\, d^3x,$$

$$d_j \equiv (\text{dipole moment of source}) = \int T^{00} x^j\, dx^3, \tag{36.49b}$$

$$\mathcal{I}_{jk} \equiv (\text{reduced quadrupole moment of source}) = \text{expression (36.46)}.$$

Thus, *the quantities $\mathcal{I}_{jk}$, whose second time-derivatives determine the radiation field by equation (36.47), are precisely the components of the star's reduced quadrupole moment, as measured by an observer who explores its Newtonian potential Φ deep inside the near zone* ($r \ll \lambda\!\!\!^{-}$) ("empirical quadrupole moment.").

The final answer (36.47) for the radiation field in terms of $\mathcal{I}_{jk}^{TT}$ was quoted in the summary of results given in §36.7. Also quoted there were expressions for the effective stress-energy tensor of the radiation and for the energy and angular momentum radiated [equations (36.22) to (36.25)]. Those expressions can be derived using the formalism of the shortwave approximation. (See exercise 36.9.)

EXERCISES

Exercise 36.7. MAGNITUDE OF $t^{\mu\nu}$

Consider a slow-motion source of gravitational waves. Show that far from the source, but in the near zone ($R \ll r \ll \lambda\!\!\!^{-}$) the components of the "stress-energy pseudotensor" $t^{\mu\nu}$ die out as $1/r^4$, but in the radiation zone ($r \gg \lambda\!\!\!^{-}$) they die out only as $1/r^2$.

Exercise 36.8. PROOF THAT THE TRANVERSE TRACELESS PARTS OF $\mathcal{I}_{jk}$ AND I_{jk} ARE IDENTICAL

Prove by direct computation that the TT parts of I_{jk} (36.42b) and $\mathcal{I}_{jk}$ (36.46) are identical, no matter where the observer is who does the TT projection (i.e., no matter what the unit vector $\boldsymbol{n}$ in the projection operator may be).

Exercise 36.9. ENERGY AND ANGULAR MOMENTUM RADIATED

(a) For the gravitational waves in asymptotically flat spacetime described by equation (36.47), calculate the smeared-out stress-energy tensor $T_{\mu\nu}^{(\mathrm{GW})}$ of equation (35.23). [*Answer:* equation (36.22).]

(b) Perform the integrals of equations (36.23) and (36.25) to obtain the total power and angular momentum radiated. [*Hint:* Derive and use the following averages over a sphere

$$\frac{1}{4\pi}\int n_i \, d\Omega = 0, \qquad \frac{1}{4\pi}\int n_i n_j \, d\Omega = \frac{1}{3}\delta_{ij}, \qquad \frac{1}{4\pi}\int n_i n_j n_k \, d\Omega = 0,$$

$$\frac{1}{4\pi}\int n_i n_j n_k n_l \, d\Omega = \frac{1}{15}(\delta_{ij}\,\delta_{kl} + \delta_{ik}\,\delta_{jl} + \delta_{il}\,\delta_{jk}).$$

Here $\boldsymbol{n} \equiv \boldsymbol{x}/|\boldsymbol{x}|$ is the unit radial vector.]

---

§36.11. DERIVATION OF THE RADIATION-REACTION POTENTIAL

Derivation of formula for the radiation-reaction potential:

Turn, finally, to a derivation of the radiation-reaction results quoted in §36.8. The analysis starts with the solution (36.43) for the spatial part of the radiation field in the original (i.e., not TT) gauge:

$$\bar{h}^{jk}(t, \boldsymbol{x}) = \frac{2}{r}\ddot{I}_{jk}(t - r). \tag{36.50}$$

Although this solution was originally derived by discarding all terms that die out faster than $1/r$, it is in fact an exact solution to the vacuum field equations $\bar{h}^{jk}{}_{,\alpha}{}^{\alpha} = 0$ of linearized theory. This means that it is valid in the intermediate and near zones ($r \lesssim \lambda\!\!\!^{-}$, but $r > R$) as well as in the radiation zone.

Were one to replace the outgoing-wave condition by an ingoing-wave condition at infinity, the exact solution (36.50) for $\bar{h}^{jk}$ would get replaced by

$$\bar{h}^{jk}(t, \boldsymbol{x}) = \frac{2}{r}\ddot{I}_{jk}(t + r).$$

Thus, in order to delineate the effects of the outgoing-wave boundary condition, one can write the exact solution in the form

(1) formula for $\bar{h}_{jk}$ anywhere outside source, with either outgoing or ingoing waves

$$\bar{h}_{jk}(t, \boldsymbol{x}) = \bar{h}^{jk}(t, \boldsymbol{x}) = \frac{2}{r}\ddot{I}_{jk}(t - \epsilon r), \qquad \epsilon = \pm 1, \tag{36.51}$$

and then focus attention on the effects of the sign of ϵ.

In the near zone ($r \ll \lambda\!\!\!^{-}$), but outside the nearly Newtonian source, this solution for $\bar{h}_{jk}$, as expanded in powers of r, becomes

(2) $\bar{h}_{jk}$ specialized to near zone

$$\bar{h}_{jk} = 2\left[\frac{I^{(2)}_{jk}}{r} - \epsilon I^{(3)}_{jk} + \frac{I^{(4)}_{jk} r}{2!} - \epsilon\frac{I^{(5)}_{jk} r^2}{3!} + \cdots\right], \tag{36.52a}$$

where

$$I^{(n)}_{jk} \equiv d^n I_{jk}(t)/dt^n.$$

2

The corresponding forms of $\bar{h}_{0j}$ and $\bar{h}_{00}$ can be generated from this by the gauge conditions $\bar{h}_{\alpha}{}^{\beta}{}_{,\beta} = 0$; i.e., by $\bar{h}_{j0,0} = \bar{h}_{jk,k}$ and $\bar{h}_{00,0} = \bar{h}_{0j,j}$. The results are:

(3) $\bar{h}_{00}$ and $\bar{h}_{0j}$ in near zone calculated by gauge conditions

$$\begin{aligned}\bar{h}_{0j} = 2\bigg[&-\frac{I^{(1)}_{jk}x^k}{r^3} + \frac{I^{(3)}_{jk}x^k}{2!r} - \epsilon\frac{2I^{(4)}_{jk}x^k}{3!} + \frac{3I^{(5)}_{jk}x^k r}{4!} - \epsilon\frac{4I^{(6)}_{jk}x^k r^2}{5!}\bigg] \\ &+ (\text{static terms not associated with radiation});\end{aligned} \tag{36.52b}$$

$$\begin{aligned}\bar{h}_{00} = 2\bigg[&\frac{(3x^jx^k - r^2\,\delta^{jk})}{r^5}I_{jk} - \frac{(x^jx^k - r^2\,\delta^{jk})}{2!r^3}I^{(2)}_{jk} - \epsilon\frac{2}{3!}I^{(3)}_{jj} \\ &+ \frac{3(x^jx^k + r^2\,\delta^{jk})}{4!r}I^{(4)}_{jk} - \epsilon\frac{4(2x^jx^k + r^2\,\delta^{jk})}{5!}I^{(5)}_{jk} + \cdots\bigg] \\ &+ (\text{static and time-linear terms not associated with radiation}).\end{aligned} \tag{36.52c}$$

The leading term in these expressions rises as $1/r^3$ when one approaches the source:

$$\bar{h}_{00} \approx \frac{2(3x^jx^k - r^2\,\delta^{jk})}{r^5}I_{jk} = \frac{6\not{I}_{jk}n^jn^k}{r^3}.$$

It is precisely the leading term in the dynamic, quadrupole part of the Newtonian potential, $\Phi = -\frac{1}{2}h_{00} = -\frac{1}{4}\bar{h}_{00}$. All other terms without ϵ's in front of them are corrections to the Newtonian potential. They produce effects like the perihelion shift of Mercury that in no way deplete the energy and angular momentum of the system.

(4) plucking out the radiation-reaction potentials from $\bar{h}_{\alpha\beta}$

The terms with ϵ's are associated with radiation reaction. Pluck the leading ones out and call them "reaction potentials":

$$\begin{aligned}\bar{h}^{(\text{react})}_{jk} &= -2I^{(3)}_{jk} - \frac{1}{3}I^{(5)}_{jk}r^2, \\ \bar{h}^{(\text{react})}_{0j} &= -\frac{2}{3}I^{(4)}_{jk}x^k - \frac{1}{15}I^{(6)}_{jk}x^kr^2, \\ \bar{h}^{(\text{react})}_{00} &= -\frac{2}{3}I^{(3)}_{jj} - \frac{1}{15}(2x^jx^k + r^2\,\delta^{jk})I^{(5)}_{jk}.\end{aligned} \tag{36.53}$$

The corresponding metric perturbations $h_{\alpha\beta} = \bar{h}_{\alpha\beta} - \frac{1}{2}\bar{h}\eta_{\alpha\beta}$ are

$$\begin{aligned}h^{(\text{react})}_{jk} &= -2I^{(3)}_{jk} + \frac{2}{3}I^{(3)}_{\ell\ell}\,\delta_{jk} + 0(I^{(5)}_{jk}r^2), \\ h^{(\text{react})}_{0j} &= -\frac{2}{3}I^{(4)}_{jk}x^k + 0(I^{(6)}_{jk}r^3), \\ h^{(\text{react})}_{00} &= -\frac{4}{3}I^{(3)}_{\ell\ell} - \frac{1}{15}(x^jx^k + 3r^2\,\delta_{jk})I^{(5)}_{jk}.\end{aligned} \tag{36.54}$$

These reaction potentials in the near zone are understood most clearly by a change of gauge that brings them into Newtonian form. Set

(5) conversion of radiation-reaction potentials to Newtonian gauge

$$x^{\mu}{}_{\text{new}} = x^{\mu}{}_{\text{old}} + \xi^{\mu}(x), \qquad h_{\mu\nu_{\text{new}}} = h_{\mu\nu_{\text{old}}} - \xi_{\mu,\nu} - \xi_{\nu,\mu}$$

with

$$\begin{aligned} \xi_j &= -I^{(3)}_{jk}x^k + \frac{1}{3}I^{(3)}_{ll}x^j, \\ \xi_0 &= -\frac{2}{3}I^{(2)}_{ll} + \frac{1}{6}I^{(4)}_{jk}x^jx^k - \frac{1}{6}I^{(4)}_{ll}r^2. \end{aligned} \tag{36.55}$$

Then in the new gauge

$$\begin{aligned} h^{(\text{react})}_{jk} &= O(I^{(5)}_{jk}r^2), \qquad h^{(\text{react})}_{0j} = O(I^{(6)}_{jk}r^3), \\ h^{(\text{react})}_{00} &= -\frac{2}{5}\not{I}^{(5)}_{jk}x^jx^k. \end{aligned} \tag{36.56}$$

This gauge is ideally suited to a Newtonian interpretation, since in it the geodesic equation for slowly moving particles has the form

$$d^2x^j/dt^2 = -\Phi^{(\text{react})}_{,j} + \begin{pmatrix}\text{terms not sensitive to} \\ \text{outgoing-wave condition}\end{pmatrix}, \tag{36.57}$$

with

$$\Phi^{(\text{react})} = -\frac{1}{2}h^{(\text{react})}_{00} = \frac{1}{5}\not{I}^{(5)}_{jk}x^jx^k. \tag{36.58}$$

Thus, the leading radiation-reaction effects (with fractional errors $\sim [\lambda\!\!\!^{-}/r]^2$) can be described in the near zone of a nearly Newtonian source by appending the term $\frac{1}{5}\not{I}^{(5)}_{jk}x^jx^k$ to the Newtonian potential. The resulting formalism and a qualitative version of the above derivation were presented in §36.8.

CHAPTER 37

DETECTION OF GRAVITATIONAL WAVES

I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be.

WILLIAM THOMSON, LORD KELVIN [(1889), p. 73]

§37.1. COORDINATE SYSTEMS AND IMPINGING WAVES

Linearized description of gravitational waves propagating past Earth

The detector is even easier to analyze than the generator or the transmission when one deals with gravitational waves within the framework of general relativity. Man's potential detectors all lie in the solar system, where gravity is so weak and spacetime so nearly flat that a plane gravitational wave coming in remains for all practical purposes a plane gravitational wave. (Angle of deflection of wave front passing limb of sun is only 1.″75.) Moreover, the nearest source of significant waves is so far away that, for all practical purposes, one can consider the waves as plane-fronted when they reach the Earth. Consequently, as they propagate in the z-direction past a detector, they can be described to high accuracy by the following transverse-traceless linearized expressions

$$\text{Metric perturbation: } h_{xx}^{TT} = -h_{yy}^{TT} = A_+(t-z),\ h_{xy}^{TT} = h_{yx}^{TT} = A_\times(t-z), \tag{37.1a}$$

$$\begin{aligned}\text{Riemann tensor: } R_{x0x0} &= -R_{y0y0} = -\frac{1}{2}\ddot{A}_+(t-z),\\ R_{x0y0} &= R_{y0x0} = -\frac{1}{2}\ddot{A}_\times(t-z).\end{aligned} \tag{37.1b}$$

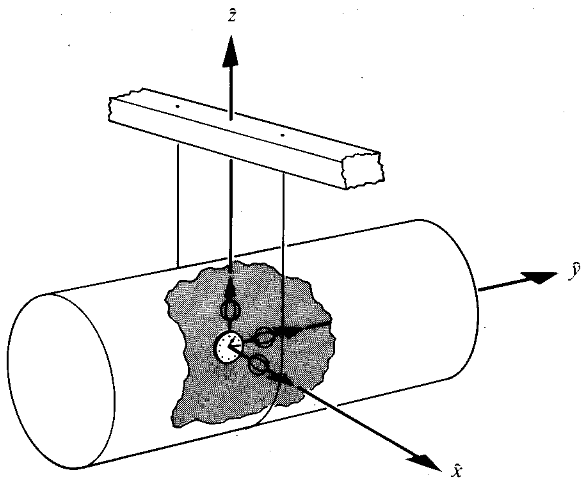


Figure 37.1.
The proper reference frame of a vibrating-bar detector. The bar hangs by a wire from a cross beam, which is supported by vertical posts (not shown) that are embedded in the Earth. Consequently, the bar experiences a 4-acceleration given, at the moment when this diagram is drawn, by $\boldsymbol{a} = g(\partial/\partial\hat{z})$, where g is the "local acceleration of gravity" ($g \sim 980$ cm/sec^2). Later, the spatial axes will have rotated relative to the bar ("Foucoult-pendulum effect" produced by Earth's rotation), so the components of $\boldsymbol{a}$ but not its magnitude will have changed.

The proper reference frame relies on an imaginary clock and three imaginary gyroscopes located at the bar's center of mass (and shown above in a cut-away view). Coordinate time is equal to proper time as measured by the clock, and the directions of the spatial axes $\partial/\partial x^{\hat{j}}$ are attached to the gyroscopes. The forces that prevent the gyroscopes from falling in the Earth's field must be applied at the centers of mass of the individual gyroscopes (no torque!).

$$\text{Stress-energy: } T^{(\mathrm{GW})}_{00} = T^{(\mathrm{GW})}_{zz} = -T^{(\mathrm{GW})}_{0z} = \frac{1}{16\pi}\langle \dot{A}_+^2 + \dot{A}_\times^2 \rangle_{\text{time avg.}} \qquad (37.1\text{c})$$

(See exercise 37.1.)

To analyze most easily the response of the detector to these impinging waves, use not the *TT* coordinate system $\{x^\alpha\}$ (which is specially "tuned" to the waves), but rather use coordinates $\{x^{\hat{\alpha}}\}$ specially "tuned" to the experimenter and his detector. The detector might be a vibrating bar, or the vibrating Earth, or a loop of tubing filled with fluid (see Figures 37.1 and 37.2). But whatever it is, it will have a center of mass. Attach the spatial origin, $x^{\hat{j}} = 0$, to this center of mass; and attach *orthonormal* spatial axes, $\partial/\partial x^{\hat{j}}$, to (possibly imaginary) gyroscopes located at this spatial origin (Figure 37.1). If the detector is accelerating (i.e., not falling freely), make the gyroscopes accelerate with it by applying the necessary forces at their centers of mass (no torque!). Use, as time coordinate, the proper time $x^{\hat{0}} = \tau$ measured by a clock at the spatial origin. Finally, extend these locally defined coordinates $x^{\hat{\alpha}}$ throughout all spacetime in the "straightest" manner possible. (See

Proper reference frame of a detector

Track 2's §13.6 for full details.) The metric in this "*proper reference frame of the detector*" will have the following form

$$ds^2 = -(1 + 2a_{\hat{j}}x^{\hat{j}})(dx^{\hat{0}})^2 + \delta_{\hat{j}\hat{k}}\, dx^{\hat{j}}\, dx^{\hat{k}} + O(|x^{\hat{j}}|^2)\, dx^{\hat{\alpha}}\, dx^{\hat{\beta}}. \tag{37.2}$$

[equation (13.71) with $\omega^{\hat{l}} = 0$]. Here $a_{\hat{j}}$ are the spatial components of the detector's 4-acceleration. (Since $\boldsymbol{a}$ must be orthogonal to the detector's 4-velocity, $a_{\hat{0}}$ vanishes.) Notice that, except for the acceleration term in $g_{\hat{0}\hat{0}}$ ("gravitational redshift term"; see §38.5 and exercise 6.6), this reference frame is locally Lorentz.

EXERCISES

Exercise 37.1. GENERAL PLANE WAVE IN TT GAUGE

Show that the most general linearized plane wave can be described in the transverse-traceless gauge of linearized theory by expressions (37.1). [*Hint:* Express the plane wave as a superposition (Fourier integral) of monochromatic plane waves, and describe each monochromatic plane wave by expressions (35.16). Use equations (35.10) and (35.23) to calculate $R_{\alpha\beta\gamma\delta}$ and $T^{(\mathrm{GW})}_{\mu\nu}$.]

Exercise 37.2. TEST-PARTICLE MOTION IN PROPER REFERENCE FRAME

Show that a slowly moving test particle, falling freely through the proper reference frame of equation (37.2), obeys the equation of motion (geodesic equation)

$$d^2x^{\hat{j}}/d\hat{t}^2 = -a_{\hat{j}} + O(|x^{\hat{k}}|).$$

Thus, one can interpret $-a_{\hat{j}}$ as the "local acceleration of gravity" (see caption of Figure 37.1).

---

§37.2. ACCELERATIONS IN MECHANICAL DETECTORS

Equations of motion for a mechanical detector

The proper reference frame of equation (37.2) is the closest thing that exists to the reference frame a Newtonian physicist would use in analyzing the detector. In fact, it is so nearly Newtonian that (according to the analysis of Box 37.1) *the equations of motion for a mechanical detector, when written in this proper reference frame, take their standard Newtonian form and can be viewed and dealt with in a fully Newtonian manner, with one exception: the gravitational waves produce a driving force of non-Newtonian origin, given by the familiar expression for geodesic deviation*

$$\begin{pmatrix}\text{force per unit mass (i.e., acceleration)}\\ \text{of a particle at } x^{\hat{j}} \text{ relative to detector's}\\ \text{center of mass at } x^{\hat{j}} = 0\end{pmatrix} = \left(\frac{d^2x^{\hat{j}}}{d\hat{t}^2}\right)_{\text{due to waves}} = -(R_{\hat{j}\hat{0}\hat{k}\hat{0}})_{\text{due to waves}}\, x^{\hat{k}}. \tag{37.3}$$

To use this equation, and to calculate detector cross sections later, one must know the components of the curvature tensor $R^{\hat{\alpha}}{}_{\hat{\beta}\hat{\gamma}\hat{\delta}}$, and of the waves' stress-energy tensor, $T^{(\mathrm{GW})}_{\hat{\mu}\hat{\nu}}$, in the detector's proper reference frame. One cannot calculate $R^{\hat{\alpha}}{}_{\hat{\beta}\hat{\gamma}\hat{\delta}}$ directly

Box 37.1 DERIVATION OF EQUATIONS OF MOTION FOR A MECHANICAL DETECTOR

Consider a "mass element" in a mechanical detector (e.g., a cube of aluminum one millimeter on each edge if the detector is the bar of Figure 37.1; or an element of fluid with volume 1 mm^3 if the detector is the tube filled with fluid shown in part h of Figure 37.2). This mass element gets pushed and pulled by adjacent matter and electromagnetic fields, as the medium of the detector vibrates or flows or does whatever it is supposed to do. Let

$$\boldsymbol{f} \equiv \begin{pmatrix} \text{4-force per unit mass exerted on mass-element} \\ \text{by adjacent matter and by electromagnetic fields} \end{pmatrix}. \tag{1}$$

This 4-force per unit mass gives the mass element a 4-acceleration $\boldsymbol{\nabla}_{\boldsymbol{u}} \boldsymbol{u} = \boldsymbol{f}$; or, in terms of components in the detector's proper reference frame, $f^{\hat{j}} = Du^{\hat{j}}/d\tau$. Assume that the mass element has a very small velocity ($v \ll 1$) in the detector's proper reference frame (i.e., relative to the detector's center of mass). Then, ignoring terms of $O(v^2)$, $O(|x^{\hat{j}}|^2)$, and $O(|x^{\hat{j}}|v)$, one has [see equation (37.2)]

$$d\hat{t}/d\tau = u^{\hat{0}} = 1 - a_{\hat{j}}x^{\hat{j}} \equiv 1 - \boldsymbol{a} \cdot \boldsymbol{x}, \tag{2}$$

and

$$f^{\hat{j}} = d^2x^{\hat{j}}/d\tau^2 + \Gamma^{\hat{j}}{}_{\hat{\alpha}\hat{\beta}}u^{\hat{\alpha}}u^{\hat{\beta}} = (d^2x^{\hat{j}}/d\hat{t}^2 + \Gamma^{\hat{j}}{}_{\hat{0}\hat{0}})(1 - 2\boldsymbol{a} \cdot \boldsymbol{x}). \tag{3}$$

Exercise 37.3 calculates $\Gamma^{\hat{j}}{}_{\hat{0}\hat{0}}$ to precision of $O(|x^{\hat{j}}|)$. Inserting its result and rearranging terms, one finds that

$$d^2x^{\hat{j}}/d\hat{t}^2 = (1 + 2\boldsymbol{a} \cdot \boldsymbol{x})f^{\hat{j}} - a^{\hat{j}}(1 + \boldsymbol{a} \cdot \boldsymbol{x}) - R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}}x^{\hat{k}} \tag{4}$$

("*equation of motion for mass element*").

Examine this equation, first from the viewpoint of an Einsteinian physicist, and then from the viewpoint of a Newtonian physicist.

The Einsteinian physicist recognizes $d^2x^{\hat{j}}/d\hat{t}^2$ as the "coordinate acceleration" of the mass element—but he keeps in mind that, to precision of $O(|x^{\hat{j}}|^2)$, coordinate lengths and proper lengths are the same [see equation (37.2)]. The coordinate acceleration $d^2x^{\hat{j}}/d\hat{t}^2$ has three causes: (1) *the externally applied force,*

$$\begin{aligned}(1 + 2\boldsymbol{a} \cdot \boldsymbol{x})f^{\hat{j}} &= (d^2x^{\hat{j}}/d\hat{t}^2)_{\text{external force}} \\ &= (1 + 2\boldsymbol{a} \cdot \boldsymbol{x})(d^2x^{\hat{j}}/d\tau^2)_{\text{external force}}\end{aligned} \tag{5a}$$

(the origin of the $\boldsymbol{a} \cdot \boldsymbol{x}$ correction is simply the conversion between coordinate time

Box 37.1 *(continued)*

and proper time); (2) the "inertial force" due to the acceleration of the reference frame,

$$-a^{\hat{j}}(1 + \boldsymbol{a} \cdot \boldsymbol{x}) = (d^2x^{\hat{j}}/d\hat{t}^2)_{\text{inertial force}} \tag{5b}$$

(see exercise 37.4 for explanation of the $\boldsymbol{a} \cdot \boldsymbol{x}$ correction); and (3) a "*Riemann curvature force,*" which will include Riemann curvature due to local, Newtonian gravitational fields (fields of Earth, sun, moon, etc.), plus Riemann curvature due to the impinging gravitational waves,

$$-(R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}})_{\text{waves}}x^{\hat{k}} - (R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}})_{\text{Newton fields}}x^{\hat{k}} = (d^2x^{\hat{j}}/d\hat{t}^2)_{\text{curvature}} \tag{5c}$$

(linear superposition because all gravitational fields in the solar system are so weak). This "Riemann curvature force" is not, of course, "felt" by the mass element; it does not produce any 4-acceleration. Rather, like the inertial force, it originates in the choice of reference frame: The spatial coordinates $x^{\hat{j}}$ measure proper distance and direction away from the detector's center of mass; and Riemann curvature tries to change this proper distance and direction ("relative acceleration;" "geodesic deviation").

A Newtonian physicist views the equation of motion (4) in a rather different manner. Having been told that the spatial coordinates $x^{\hat{j}}$ measure proper distance and direction away from the detector's center of mass, he thinks of them as the standard Euclidean spatial coordinates of Newtonian theory. He then rewrites equation (4) in the form

$$d^2x^{\hat{j}}/d\hat{t}^2 = F^{\hat{j}} - (R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}})_{\text{waves}}x^{\hat{k}}, \tag{6}$$

where

$$\begin{aligned} F^{\hat{j}} &\equiv \begin{pmatrix}\text{Newtonian force per unit mass}\\ \text{acting on mass element}\end{pmatrix} \qquad (7)\\ &= (1 + 2\boldsymbol{a} \cdot \boldsymbol{x})f^{\hat{j}} - a^{\hat{j}}(1 + \boldsymbol{a} \cdot \boldsymbol{x}) - (R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}})_{\text{Newton fields}}x^{\hat{k}}. \end{aligned}$$

The Newtonian physicist is free to express $F^{\hat{j}}$ in a form more familiar than this. He can ignore the subtleties of the $\boldsymbol{a} \cdot \boldsymbol{x}$ "redshift effects" because (1) they are small

$$|a^{\hat{j}}(\boldsymbol{a} \cdot \boldsymbol{x})| \sim |f^{\hat{j}}(\boldsymbol{a} \cdot \boldsymbol{x})| \lesssim |(R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}})_{\text{waves}}x^{\hat{k}}|; \tag{8}$$

and (2) they are steady in time, and therefore—by contrast with the equally small wave-induced forces—they cannot excite resonant motions of the detector. Also, he can separate the "inertial acceleration," $-a^{\hat{j}}$, into a contribution from the local acceleration of gravity at the detector's center of mass, $-(\partial\Phi/\partial x^{\hat{j}})_{x^{\hat{j}}=0}$, plus a contribution $-a^{\hat{j}}{}_{\text{absolute}}$ due to acceleration of the detector relative to the "absolute space" of Newtonian theory. Finally, he can rewrite the Riemann curvature due to Newtonian gravity in the familiar form $R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}} = \partial^2\Phi/\partial x^{\hat{j}}\,\partial x^{\hat{k}}$. The net result is

$$
\begin{aligned}
F^{\hat{j}} = \quad & & &\begin{bmatrix}\text{total Newtonian force per unit}\\ \text{mass acting on mass element}\end{bmatrix}\\
& +f^{\hat{j}} & &\begin{bmatrix}\text{Newtonian force per unit mass exerted by}\\ \text{adjacent matter and by electromagnetic fields}\end{bmatrix}\\
& -a^{\hat{j}}{}_{\text{absolute}} & &\begin{bmatrix}\text{inertial force per unit mass due to acceleration}\\ \text{of detector relative to Newtonian absolute space}\end{bmatrix}\\
& -\left(\frac{\partial\Phi}{\partial x^{\hat{j}}}\right)_{\text{at mass element}} & &\begin{bmatrix}= -(\partial\Phi/\partial x^{\hat{j}})_{x^{\hat{j}}=0} - (\partial^2\Phi/\partial x^{\hat{j}}\,\partial x^{\hat{k}})x^{\hat{k}}\\ = \text{Newtonian gravitational acceleration}\end{bmatrix}.
\end{aligned} \tag{9}
$$

Conclusion: The equation of motion for a mass element of a mechanical detector, when written in the detector's proper reference frame, has the standard Newtonian form (6), with standard Newtonian driving forces (9), plus a driving force due to the gravitational waves given by

$$(d^2x^{\hat{j}}/d\hat{t}^2)_{\text{due to waves}} = -(R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}})_{\text{waves}}x^{\hat{k}}. \tag{10}$$

from the metric coefficients $g_{\hat{\alpha}\hat{\beta}}$ of expression (37.2); to do so one would need the unknown corrections of $O(|x^{\hat{j}}|^2)$. However, one can easily obtain $R^{\hat{\alpha}}{}_{\hat{\beta}\hat{\gamma}\hat{\delta}}$ and $T^{(\text{GW})}_{\hat{\mu}\hat{\nu}}$ from the corresponding components in the *TT* coordinate frame [equations (37.1)] by applying the transformation matrix $||\partial x^{\alpha}/\partial x^{\hat{\mu}}||$. To make the transformation trivial, orient the *TT* coordinate frame so that, to a precision of $O(|h_{\mu\nu}|) \ll 1$, it coincides with the detector's proper reference frame near the detector's center of mass at the moment of interest, $t = \hat{t} = 0$. Then the transformation matrix will be

$$\partial x^{\alpha}/\partial x^{\hat{\mu}} = \delta^{\alpha}_{\mu} + \underbrace{O(h_{\mu\nu})}_{\begin{bmatrix}\text{corrections due to}\\ \text{ripples in spacetime}\\ \text{caused by waves}\end{bmatrix}} + \underbrace{O(a_{\hat{j}}\,x^{\hat{j}})}_{\begin{bmatrix}\text{redshift}\\ \text{corrections}\end{bmatrix}} + \underbrace{O(|\boldsymbol{a}|\,\hat{t})}_{\begin{bmatrix}\text{corrections due to relative}\\ \text{velocity of frames resulting}\\ \text{from detector's acceleration}\end{bmatrix}}. \tag{37.4}$$

The acceleration the detector experiences is typically

$$|\boldsymbol{a}| = \text{one "Earth gravity"} = 980\ \text{cm/sec}^2 \sim 1/(\text{light-year}).$$

Description of waves in frame of detector

Therefore to enormous precision $||\partial x^{\alpha}/\partial x^{\hat{\mu}}|| = ||\delta^{\alpha}_{\mu}||$, and components of tensors are the same in the two reference frames:

$$R_{\hat{x}\hat{0}\hat{x}\hat{0}} = -R_{\hat{y}\hat{0}\hat{y}\hat{0}} = -\frac{1}{2}\ddot{A}_{+}, \qquad R_{\hat{x}\hat{0}\hat{y}\hat{0}} = R_{\hat{y}\hat{0}\hat{x}\hat{0}} = -\frac{1}{2}\ddot{A}_{\times},$$

$$T^{(\mathrm{GW})}_{\hat{0}\hat{0}} = T^{(\mathrm{GW})}_{\hat{z}\hat{z}} = -T^{(\mathrm{GW})}_{\hat{0}\hat{z}} = \frac{1}{16\pi}\langle \dot{A}_{+}^{2} + \dot{A}_{\times}^{2}\rangle_{\text{time avg.}} \tag{37.5}$$

[see equation (37.1)].

Combining equations (37.3) and (37.5), one obtains for the wave-induced accelerations relative to the center of mass of the detector

Explicit form of accelerations due to waves

$$\left(\frac{d^2\hat{x}}{d\hat{t}^2}\right)_{\text{due to waves}} = -R_{\hat{x}\hat{0}\hat{x}\hat{0}}\hat{x} - R_{\hat{x}\hat{0}\hat{y}\hat{0}}\hat{y} = \frac{1}{2}(\ddot{A}_{+}\hat{x} + \ddot{A}_{\times}\hat{y}),$$

$$\left(\frac{d^2\hat{y}}{d\hat{t}^2}\right)_{\text{due to waves}} = -R_{\hat{y}\hat{0}\hat{y}\hat{0}}\hat{y} - R_{\hat{y}\hat{0}\hat{x}\hat{0}}\hat{x} = \frac{1}{2}(-\ddot{A}_{+}\hat{y} + \ddot{A}_{\times}\hat{x}), \tag{37.6}$$

$$\left(\frac{d^2\hat{z}}{d\hat{t}^2}\right)_{\text{due to waves}} = 0.$$

This analysis is valid only for "small" detectors ($L \ll \lambda\!\!\!^{-}$)

These expressions, like the equation of geodesic deviation, are valid only over regions small compared to one wavelength. Second derivatives of the metric (i.e., the components of the Riemann tensor) give a poor measure of geodesic deviation and of wave-induced forces over regions of size $L \gtrsim \lambda\!\!\!^{-}$. Thus, to analyze large detectors ($L \gtrsim \lambda\!\!\!^{-}$), one must abandon the "local mathematics" of the curvature tensor and replace it by "global mathematics"—e.g., an analysis in the TT coordinate frame using the metric components $h_{\mu\nu}$. For an example, see exercise 37.6.

All detectors of high sensitivity that have been designed up until now (1973) are small compared to a wavelength, and therefore can be analyzed using the techniques of Newtonian physics and the driving forces of equations (37.6).

It is useful to develop physical intuition for the driving forces, $-R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}}x^{\hat{k}}$, produced by waves of various polarizations. Figure 35.2 is one aid to such intuition; Box 37.2 is another. [The reader may find it interesting to examine, compare, and reconcile them!]

EXERCISES

Exercise 37.3. CONNECTION COEFFICIENTS IN PROPER REFERENCE FRAME

(a) Calculate $\Gamma^{\hat{\alpha}}{}_{\hat{\beta}\hat{\gamma}}$ for the metric (37.2), ignoring corrections of $O(|x^{\hat{j}}|)$. [Answer: Equations (13.69) with $\omega^{\hat{i}} = 0$.]

(b) Calculate $R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}}$ using the standard formula (8.44), and leaving spatial derivatives of the connection coefficients unevaluated because of the unknown corrections of $O(|x^{\hat{j}}|)$ in $\Gamma^{\hat{\alpha}}{}_{\hat{\beta}\hat{\gamma}}$. [Answer: $R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}} = \Gamma^{\hat{j}}{}_{\hat{0}\hat{0},\hat{k}} - a^{\hat{j}}a^{\hat{k}}$.]

(c) Use the answer to part (b) to evaluate the $O(|x^{\hat{j}}|)$ corrections to $\Gamma^{\hat{j}}{}_{\hat{0}\hat{0}}$. [Answer:

$$\Gamma^{\hat{j}}{}_{\hat{0}\hat{0}} = a^{\hat{j}}(1 + a_{\hat{k}}x^{\hat{k}}) + R^{\hat{j}}{}_{\hat{0}\hat{k}\hat{0}}x^{\hat{k}} + O(|x^{\hat{k}}|^2).] \tag{37.7}$$

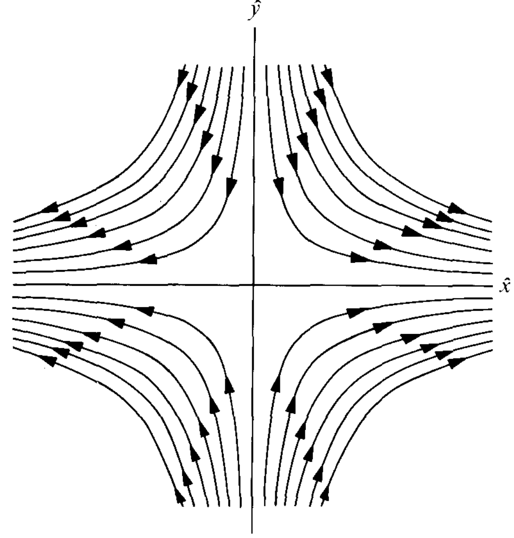
Box 37.2 LINES OF FORCE FOR GRAVITATIONAL-WAVE ACCELERATIONS

A. Basic Idea

Consider a plane wave propagating in the $\hat{z}$ direction. Discuss it entirely in the proper reference frame of a detector. The relative accelerations due to the wave are entirely transverse. Relative to the center of mass of the detector (origin of spatial coordinates) they are

$$\begin{aligned} d^2\hat{x}/d\hat{t}^2 &= \frac{1}{2}(\ddot{A}_+\hat{x} + \ddot{A}_\times\hat{y}), \\ d^2\hat{y}/d\hat{t}^2 &= \frac{1}{2}(-\ddot{A}_+\hat{y} + \ddot{A}_\times\hat{x}), \qquad (1) \\ d^2\hat{z}/d\hat{t}^2 &= 0. \end{aligned}$$

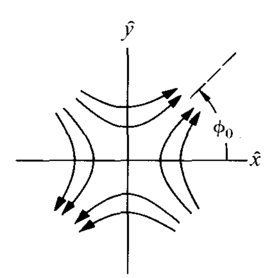
Notice that these accelerations are divergence-free. Consequently they can be represented by "lines of force," analogous to those of a vacuum electric field. At a value of $\hat{t} - \hat{z}$ where $\ddot{A}_\times = 0$ (so polarization is entirely $\boldsymbol{e}_+$), the lines of force are the hyperbolas shown here [sketch (a)]. The direction of the acceleration at any point is the direction of the arrow there; the magnitude of the acceleration is the density of force lines. Since acceleration is proportional to distance from center of mass, the force lines get twice as close together when one moves twice as far away from the origin in a given direction. When $\ddot{A}_+$ is positive, the arrows on the force lines are as shown in (a); when it is negative, they are reversed. As $|\ddot{A}_+|$ increases, the force lines move in toward the origin so their density goes up; as $|\ddot{A}_+|$ decreases, they move out toward infinity so their density goes down.



(a) Force lines for $\ddot{A}_\times = 0$, $\ddot{A}_+ > 0$

For polarization $\boldsymbol{e}_\times$ the force lines are rotated by 45° from the above diagram. For intermediate polarization (values of $\hat{t} - \hat{z}$ where $\ddot{A}_+$ and $\ddot{A}_\times$ are both nonzero), the diagram is rotated by an intermediate angle [sketch (b)]

$$\phi_0 = \frac{1}{2}\,\text{arc tan}\,(\ddot{A}_\times/\ddot{A}_+). \qquad (2)$$



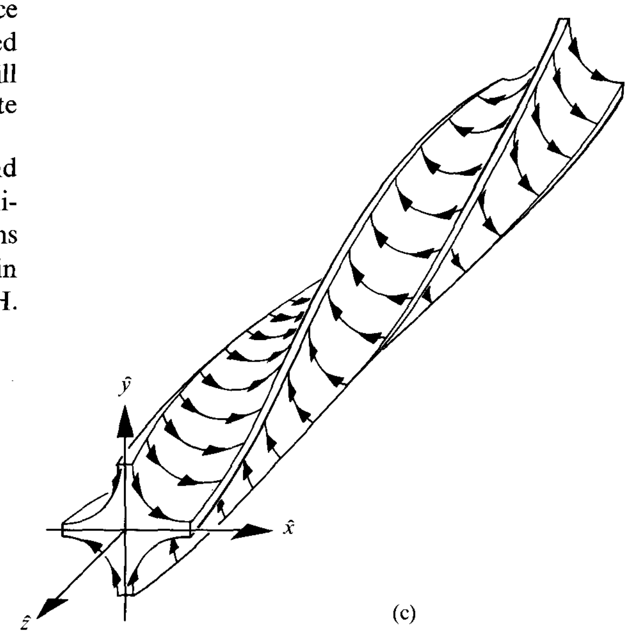
(b)

Box 37.2 ***(continued)***

B. Three-Dimensional Diagram

At each value of $\hat{t} - \hat{z}$, the wave-produced accelerations have a specific polarization [orientation angle ϕ_0 of sketch (b)] and a specific amplitude (density of lines of force). Draw the lines of force in a three-dimensional $(\hat{x}, \hat{y}, \hat{z})$ diagram for fixed $\hat{t}$. Then as time passes the over-all diagram will remain unchanged in form, but will propagate with the speed of light in the $\hat{z}$ direction.

Sketch (c) shows such a diagram for righthand circularly polarized waves of unchanging amplitude. *Note:* The authors are not aware of diagrams such as these [(a), (b), (c) above] and their use in analyzing detector response prior to William H. Press (1970).



(c)

Exercise 37.4. WHY THE $a \cdot x$?

Explain the origin of the $\boldsymbol{a} \cdot \boldsymbol{x}$ correction in equation (5b) of Box 37.1. [*Hint:* Take the viewpoint of an observer at rest at the spatial origin who watches two freely falling particles respond to the inertial force. At time $\hat{t} = 0$, put one particle at the origin and the other at $x^{\hat{j}}$. As time passes, the separation of the particles in their common Lorentz frame remains fixed; so there develops a Lorentz contraction from the viewpoint of the observer at $x^{\hat{j}} = 0$.]

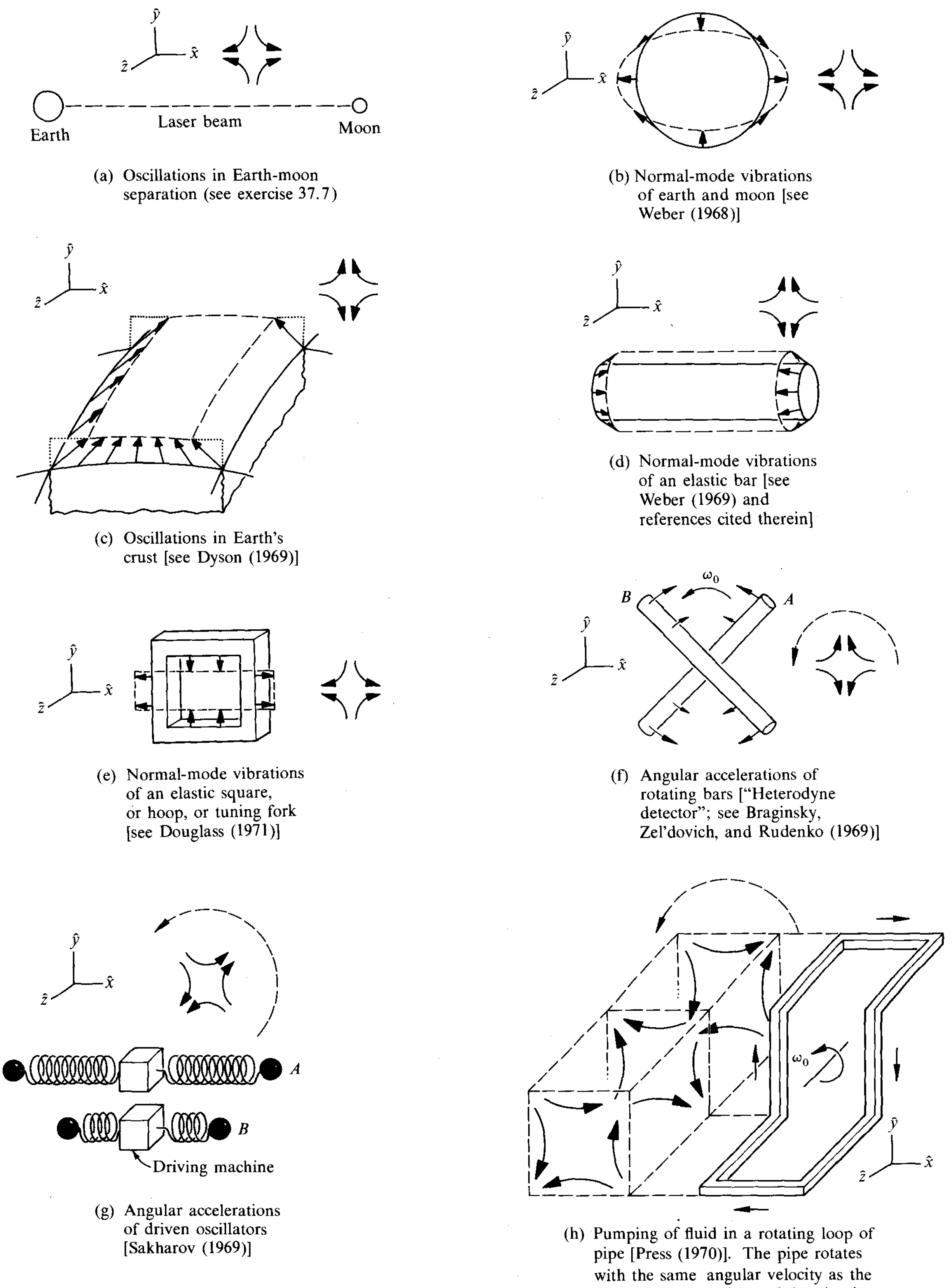
Exercise 37.5. ORIENTATION OF POLARIZATION DIAGRAM

Derive equation (2) of Box 37.2.

§37.3. TYPES OF MECHANICAL DETECTORS

Eight types of mechanical detectors:

Figure 37.2 shows eight different types of mechanical detectors for gravitational waves. (By "mechanical detector" is meant a detector that relies on the relative



(a) Oscillations in Earth-moon separation (see exercise 37.7)

(b) Normal-mode vibrations of earth and moon [see Weber (1968)]

(c) Oscillations in Earth's crust [see Dyson (1969)]

(d) Normal-mode vibrations of an elastic bar [see Weber (1969) and references cited therein]

(e) Normal-mode vibrations of an elastic square, or hoop, or tuning fork [see Douglass (1971)]

(f) Angular accelerations of rotating bars ["Heterodyne detector"; see Braginsky, Zel'dovich, and Rudenko (1969)]

(g) Angular accelerations of driven oscillators [Sakharov (1969)]

(h) Pumping of fluid in a rotating loop of pipe [Press (1970)]. The pipe rotates with the same angular velocity as the waves; so the position of the pipe in the righthand polarized lines of force remains forever fixed

Figure 37.2.
Various types of gravitational-wave detectors.

motions of matter. Nonmechanical detectors are described in §37.9, near end of this chapter.) These eight detectors, and others, can be analyzed easily using the force-line diagrams of Box 37.2. A *qualitative* discussion of each of the eight detectors is given below. (A full quantitative analysis for each one would entail experimental technicalities for which general relativity is irrelevant, and which are beyond the scope of this book. However, some quantitative details are spelled out in §§37.5–37.8.)

1. The Relative Motions of Two Freely Falling Bodies

(1) freely falling bodies

As a gravitational wave passes two freely falling bodies, their proper separation oscillates (Figure 37.3). This produces corresponding oscillations in the redshift and round-trip travel times for electromagnetic signals propagating back and forth between the two bodies. Either effect, oscillating redshift or oscillating travel time, could be used in principle to detect the passage of the waves. Examples of such detectors are the Earth-Moon separation, as monitored by laser ranging [Fig. 37.2(a)]; Earth-spacecraft separations as monitored by radio ranging; and the separation between two test masses in an Earth-orbiting laboratory, as monitored by redshift measurements or by laser interferometry. Several features of such detectors are explored in exercises 37.6 and 37.7. As shown in exercise 37.7, such detectors have so low a sensitivity that they are of little experimental interest.

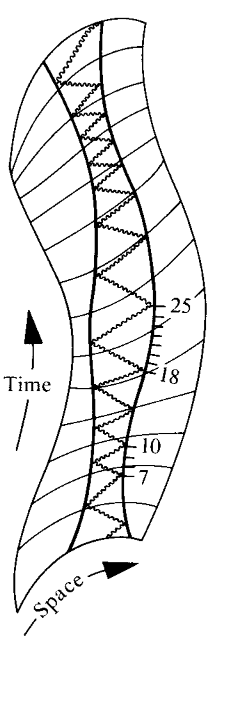


Figure 37.3.
Time of round-trip travel between two geodesics responds to oscillations in the curvature of spacetime (diagram is schematic only; symbolic of a laser pulse sent from the Earth to a corner reflector on the Moon and back at a time when a very powerful, long-wavelength gravitational wave passes by; the wave would have to be powerful because a direct measure of distance to better than 10 cm is difficult, and such precision produces a much less sensitive indicator of waves than the vibrations in length [10^{-14} cm or less] of a Weber bar; see exercise 37.7). The geodesics are curved toward each other in regions where the relevant component of the Riemann curvature tensor, call it $R_{\hat{x}\hat{0}\hat{x}\hat{0}}$, has one sign, and curved away from each other in regions where it has the opposite sign. The diagram allows one to see at a glance the answer to an often expressed puzzlement: Is not any change in round-trip travel time mere trumpery flummery? The metric perturbation, $\delta h_{\mu\nu}$, of the wave changes the scale of distances slightly but also correspondingly changes the scale of time. Therefore does not every possibility of any really meaningful and measurable effect cancel out? Answer: (1) The widened separation between the geodesics is not a local effect but a cumulative one. It does not arise from the local value of $\delta h_{\mu\nu}$ directly or even from the local value of the curvature. It arises from an accumulation of the bending process over an entire half-period of the gravitational wave. (2) The change in separation of the geodesics is a true change in proper distance, and shows up in a true change in proper time (see "ticks" on the world line of one of the particles). See exercise 37.6. Note: When one investigates the separation between the geodesics, not over a single period, as here, but over a large number of periods, he finds a cumulative, systematic, net slow bending of the rapidly wiggling geodesics toward each other. This small, attractive acceleration is evidence in gravitation physics of the effective mass-energy carried by the gravitational waves (see Chapter 35).

2. Normal-Mode Vibrations of the Earth and Moon

(2) Earth and Moon

A gravitational wave sweeping over the Earth will excite its quadrupole modes of vibration, since the driving forces in the wave have quadrupole spatial distributions [see Fig. 37.2(b)]. The fundamental quadrupole mode of the Earth has a period of 54 minutes, while that of the moon has a period of 15 minutes. Thus, the Earth and Moon should selectively pick out the 54-minute and 15-minute components of any passing wave train. Section 37.7 will analyze quantitatively the interaction between the wave and solid-body vibrations. By comparing that analysis with seismometer studies of the Earth's vibrations, Weber (1967) put the first observational limit ever on the cosmic flux of gravitational waves:

$$I_\nu \equiv \frac{d\text{ flux}}{d\text{ frequency}} < 3 \times 10^7 \text{ erg cm}^{-2}\text{ sec}^{-1}\text{ Hz}^{-1} \text{ at } \nu = 3.1 \times 10^{-4}\text{ Hz}. \quad (37.8)$$

3. Oscillations in the Earth's Crust

(3) Earth's crust

If the neutron star in a pulsar is slightly deformed from axial symmetry, its rotation will produce gravitational waves. The period of the waves is half the period of the pulsar (rotation of star through 180° produces one period of waves)—i.e., it should range from 0.017 sec for NP0532 (Crab Pulsar) to 1.87 sec. for NP0527. Such a wave train cannot excite the 54-minute quadrupole vibration or any of the other normal, low-frequency modes of vibration of the Earth. The kind of vibrations it *can* excite allow themselves in principle to be described in the language of normal modes. However, they are clearly and more conveniently envisaged as vibrations of localized regions of the Earth; or, more particularly, vibrations of the Earth's crust.

Dyson (1969) has analyzed the response of an elastic solid, such as the Earth, to an incident, off-resonance gravitational wave. He shows that the response depends on irregularities in the elastic modulus for shear waves, and that it is strongest at a free surface [Figure 37.2(c)]. For the fraction of gravitational-wave energy crossing a flat surface that is converted into energy of elastic motion of the solid, he finds the expression

$$(\text{fraction}) = (8\pi G\rho/\omega^2)(s/c)^3 \times \sin^2\theta|\cos\theta|^{-1}[1 + \cos^2\theta + (s/v)\sin^2\theta]. \quad (37.9)$$

Here s and v are the velocities of shear waves and compressive waves, respectively, and θ is the angle between the direction of propagation of the waves and the normal to the surface. Considering a flux of 2×10^{-5} erg/cm² sec (an optimistic but conceivable value for waves from a pulsar) incident horizontally ($\theta = \pi/2$; "divergent" factor $|\cos\theta|^{-1}$ cancels out in calculation!), and taking s to be 4.5×10^5 cm/sec and ω to be 6 rad/sec, he calculates that the 1-Hz horizontal displacement produced in the surface has an amplitude of $\xi_0 \sim 2 \times 10^{-17}$ cm, too small by a factor of the order of 10^5 to be detected against background seismic noise. He points to the possibilities of improvements, especially via resonance (elastic waves reflected back and forth between two surfaces; Antarctic ice sheet).

4. Normal-Mode Vibrations of an Elastic Bar

(4) elastic bar

As of 1972, the most often-discussed type of detector is the aluminum bar invented by Joseph Weber (1960, 1961) [see Figures 37.1 and 37.2(d)]. Weber's bars are cylindrical in shape, with length 153 cm, diameter 66 cm, and weight 1.4×10^6 g. Each bar is suspended by a wire in vacuum and is mechanically decoupled from its surroundings. Around its middle are attached piezoelectric strain transducers, which couple into electronic circuits that are sensitive to the bar's fundamental end-to-end mode of oscillation (frequency $\nu = 1{,}660$ Hz). When a gravitational wave hits the bar broadside, as shown in Figure 37.2(d), the relative accelerations carried by the wave will excite the fundamental mode of the bar. As of 1972, Weber observes sudden, simultaneous excitations in two such bars, one at the University of Maryland, near Washington, D.C.; the other at Argonne National Laboratory, near Chicago [see Weber (1969, 1970a,b)]. No one has yet come forward with a workable explanation for Weber's coincidences other than gravitational waves from outer space. However, the history of physics is rich with instances where supposedly new effects had to be attributed in the end to long familiar phenomena. Therefore it would seem difficult to rate the observed events as "battle-tested." To achieve that confidence rating would seem to require confirmation with different equipment, or under different circumstances, or both; experiments to provide that confirmation are now (1972) underway. If one makes this tentative assessment, one can be excused for expressing at the same time the greatest admiration for the experimental ingenuity, energy, and magnificent persistence that Joseph Weber has shown in his more than decade-long search for the most elusive radiation on the books of physics.

Mechanical detectors of the above four types represent systems on which measurements have been made; so practical difficulties and realizable noise levels can be estimated properly. In the continuing search for improved methods, more elaborate detectors are being studied, and in 1972 one can list a number of interesting proposals, as below. For these it is hard to know how much development would be required in order to achieve the desired performance.

5. Normal-Mode Vibrations of Elastic Bodies of Other Shapes

(5) elastic bodies of other shapes

The "bar" of a Weber detector need not be cylindrical in shape. For a discussion of a detector with the shape of a hollow square, a hoop, or a tuning fork, see Douglass (1971); such a detector might allow its fundamental frequency to be adjusted for the most favorable response, with given mass, or given maximum dimension, or both. Sections 37.4 and 37.7 and exercises 37.9 to 37.12 analyze in detail the operation of a "vibrating-bar" detector of arbitrary shape. See also Douglass and Tyson (1971).

6. Angular Accelerations of Rotating Bars

(6) rotating bars ("heterodyne detector")

All the potential detectors described thus far respond in the most obvious of manners to the tidal accelerations of a gravitational wave: relative distances oscillate in and out. But the tidal accelerations contain, in addition to a length-changing component, also a tangential, rotation-producing component. In picture (a) of Box 37.2, the length-changing component dominates near the $\hat{x}$ and $\hat{y}$ axes, whereas the rotation-producing component dominates half-way between the axes. Vladimir B. Braginsky was the first to propose a detector that responds to the rotation-producing accelerations [see Braginsky, Zel'dovich, and Rudenko (1969); Braginsky and Nazarenko (1971)]. It consists of two metal rods, oriented perpendicular to each other, and rotating freely with angular velocity ω_0 in their common plane [see Fig. 37.2(f)]. (The rotation is relative to the gyroscopes of the proper reference frame of the detector; equivalently, it is relative to the Lorentz frame local to the detector.) Let monochromatic gravitational waves of angular frequency $\omega = 2\omega_0$ (change of phase per unit of time equals twice the angular velocity at which the pattern of lines of force turns) impinge broadside on the rotating rods. The righthand circularly polarized component of the waves will then rotate with the rods; so their orientation in its lines-of-force diagram will remain forever fixed. With the orientation of Fig. 37.2(f), rod A will undergo angular acceleration, while rod B will decelerate. The experimenter can search for the constant relative angular acceleration of the two rods (constant so long as the angle between them does not depart significantly from 90°). Better yet, the experimenter can (all too easily) adjust the rotation rate ω_0 so it does not quite match the waves' frequency ω. Then for $\frac{1}{2}\omega_0/|\omega - 2\omega_0|$ rotations, rod A will accelerate and B will decelerate; then will follow $\frac{1}{2}\omega_0/|\omega - 2\omega_0|$ rotations in which A decelerates and B accelerates, and so on (frequency beating). The experimenter can search for oscillations in the relative orientation of the rods. One need not worry about the lefthand polarized waves marring the experiment. Since they do not rotate with the rods, their angular accelerations average out over one cycle.

Such a device is called a "heterodyne detector" by Braginsky. He envisages that such detectors might be placed in free-fall orbits about the Earth late in the 1970's. Heterodyne detectors would work most efficiently for long monochromatic wave trains such as those from pulsars; but even for short bursts of waves they may be more sensitive than vibrating bars [see Braginsky and Nazarenko (1971)].

7. Angular Accelerations of Driven Oscillators

(7) rotation of driven oscillators

Andrei D. Sakharov (1969) has proposed a different type of detector for the angular accelerations of a gravitational wave. Instead of two rotating bars, it consists of two identical, driven oscillators, initially parallel and nonrotating, but oscillating out of phase with each other. Each oscillator experiences angular accelerations in one direction at one phase of a passing wave, and in the opposite direction at the next phase, but the torques do not cancel out. When the oscillator is maximally distended, it experiences a greater torque (acceleration $\propto$ length; torque $\propto$ length^2) than when it is maximally contracted. Consequently, if the driven oscillations have the same angular frequency as a passing, monochromatic wave, and if the phases are as shown in Figure 37.2(g), then oscillator A will receive an angular acceleration in the righthand direction, while B receives an angular acceleration in the lefthand direction.

8. Pumping of Fluid in a Rotating Loop of Pipe

(8) fluid in pipe

A third type of detector that responds to angular accelerations has been described by William Press (1970). This detector would presumably be far less sensitive than others, and therefore not worth constructing; but it is intriguing in its novel design; and it illustrates features of gravitational waves ignored by other detectors. Press's detector consists of a loop of rotating pipe, containing a superfluid. The shape of the pipe and its constant rotation rate are chosen so that the gravitational waves will pump the fluid around inside the pipe. One conceivable pipe design (a bad one to build in practice, but an easy one to analyze) is shown in Fig. 37.2(h). Note that use is made of the variation in tidal acceleration along the direction of propagation of the wave as well as perpendicular to that direction. To analyze the response of the fluid to a righthand circularly polarized wave, one can mentally place the rotating pipe in the three-dimensional line-of-force diagram of Box 37.2(c).

EXERCISES

Exercise 37.6. RELATIVE MOTION OF FREELY FALLING BODIES AS A DETECTOR OF GRAVITATIONAL WAVES [see Figures 37.2(a) and 37.3.]

Consider two test bodies initially at rest with respect to each other in flat, empty spacetime. (The case where other, gravitating bodies are nearby can be treated without too much more difficulty; but this exercise concerns only the simplest example!) A plane, nearly monochromatic gravitational wave, with angular frequency ω and polarization $\boldsymbol{e}_+$, impinges on the bodies, coming from the $-z$ direction. As shown in exercise 35.5, the bodies remain forever at rest in those TT coordinates that constituted the bodies' global inertial frame before the wave arrived. Calculate, for arbitrary separations (Δx, Δy, Δz) of the test bodies, the redshift and the round-trip travel time of photons going back and forth between them. Compare the answer, for large Δx, Δy, Δz, with the answer one would have obtained by using (without justification!) the equation of geodesic deviation. Physically, why does the correct answer *oscillate* with increasing separation? Discuss the feasibility and the potential sensitivity of such a detector using current technology.

Exercise 37.7. EARTH-MOON SEPARATION AS A GRAVITATIONAL-WAVE DETECTOR

In the early 1970's one can monitor the Earth-moon separation using laser ranging to a precision of 10 cm, with successive observations separated by at least one round-trip travel time. Suppose that no oscillations in round-trip travel time are observed except those (of rather long periods) to be expected from the Earth-moon-sun-planets gravitational interaction. What limits can one then place on the energy flux of gravitational waves that pass the Earth? The mathematical formula for the answer should yield numerically

$$\text{Flux} \lesssim 10^{18}\ \text{erg/cm}^2\ \text{sec for } 0.3\ \text{cycle/sec} \lesssim \nu \lesssim 1\ \text{cycle/day}, \tag{37.10a}$$

corresponding to a limit on the mass density in gravitational waves of

$$\text{Density} \lesssim 10^{-13}\ \text{g/cm}^3. \tag{37.10b}$$

Why is this an uninteresting limit?

2

§37.4. VIBRATING, MECHANICAL DETECTORS: INTRODUCTORY REMARKS

1

2

The rest of this chapter is Track 2. No earlier track-2 material is needed as preparation for it, nor is it needed for any later chapter.

The remainder of this chapter (except for §37.9) gives a detailed analysis of vibrating, mechanical detectors (Earth; Weber bar; "bars" with complex shapes; and so on).

The details of the analysis and its applications depend in a crucial way on the values of two dimensionless numbers: (1) the ratio $\tau_{\rm GW}/\tau_0$, where

$$\tau_{\rm GW} \equiv \begin{pmatrix}\text{characteristic time scale for changes in}\\ \text{gravitational-wave amplitude and spectrum}\end{pmatrix}, \tag{37.11a}$$

$$\tau_0 \equiv \begin{pmatrix}e\text{-folding time for detector vibrations (in}\\ \text{normal mode of interest) to die out as}\\ \text{a result of internal damping}\end{pmatrix}; \tag{37.11b}$$

and (2) the ratio $\bar{E}_{\rm vibration}/kT$, where

$$\bar{E}_{\rm vibration} \equiv \begin{pmatrix}\text{mean value of detector's vibration energy (in}\\ \text{normal mode of interest) while waves are}\\ \text{passing and driving detector}\end{pmatrix}, \tag{37.12a}$$

$$\begin{aligned} kT &\equiv \begin{pmatrix}\text{Boltzman's}\\ \text{constant}\end{pmatrix} \times \begin{pmatrix}\text{detector's}\\ \text{temperature}\end{pmatrix} \\ &= \begin{pmatrix}\text{Mean energy in normal mode}\\ \text{of interest when gravitational}\\ \text{waves are not exciting it}\end{pmatrix}. \end{aligned} \tag{37.12b}$$

Definitions: "steady flux," "hammer-blow waves," "wave-dominated detector," "noisy detector"

When $\tau_{\rm GW} \gg \tau_0$, the detector views the radiation as having a "*steady flux,*" and it responds with steady-state vibrations; when $\tau_{\rm GW} \ll \tau_0$ (short burst of waves), the waves deal a "*hammer blow*" to the detector. When $\bar{E}_{\rm vibration} \gg kT$, the driving force of the waves dominates over the detector's random, internal, Brownian-noise forces ("*wave-dominated detector*"); when $\bar{E}_{\rm vibration} \lesssim kT$, the driving force of the waves must compete with the detector's random, internal, Brownian-noise forces ("*noisy detector*").

Sections 37.5 to 37.7 deal with wave-dominated detectors ($\bar{E}_{\rm vibration} \gg kT$). The key results of those sections are summarized in Box 37.3, which appears here as a quick preview (though it may not be fully understandable in advance). Section 37.8 treats noisy detectors.

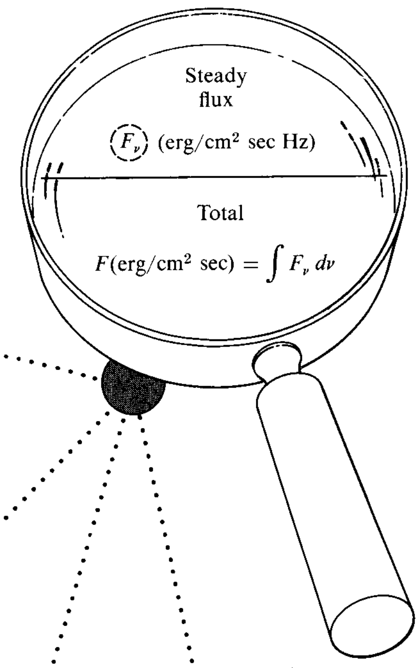
Design of detectors requires much more than the concept of cross section

Warning: Throughout the rest of this chapter prime attention focuses on the concept of cross section. This is fine for a first introduction to the theory of detectors. But cross section is not the entire story, especially when one wishes to study the detailed wave-form of the radiation. And sometimes (e.g., for the detector of Figure 37.2a), it is *none* of the story. A first-rate experimenter designing a new detector will not deal primarily in cross sections any more than a radio engineer will in designing a new radio telescope. Attention will also focus heavily on the bandwidth

(*continued on page 1022*)

Box 37.3 WAYS TO USE CROSS SECTION FOR WAVE-DOMINATED DETECTORS

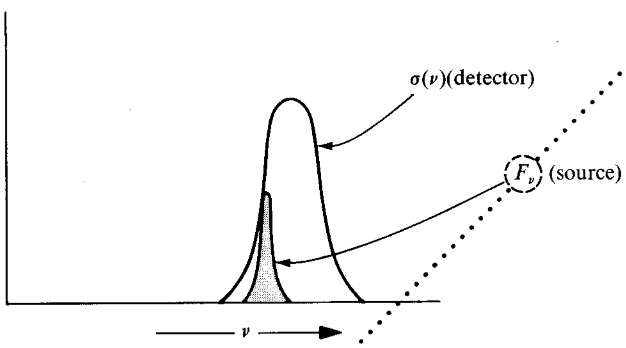
A. To Calculate *Rate* at which Detector Extracts Energy from a *Steady* Flux of Radiation
$(\tau_{GW} \gg \tau_0)$



1. Frequency distribution of radiation arbitrary:

$$\begin{pmatrix}\text{steady rate at which detector extracts}\\ \text{energy from gravitational waves}\end{pmatrix} = \int \underbrace{F_\nu(\nu)}_{\text{erg/cm}^2\text{ sec Hz}} \cdot \underbrace{\sigma(\nu)}_{\text{cm}^2} \underbrace{d\nu}_{\text{Hz}}$$

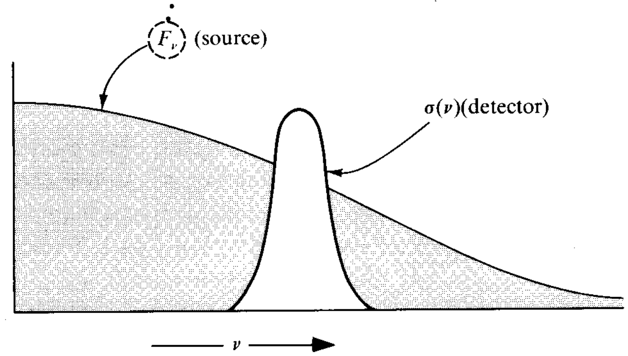
2. Frequency spread of radiation small compared to line width of detector:



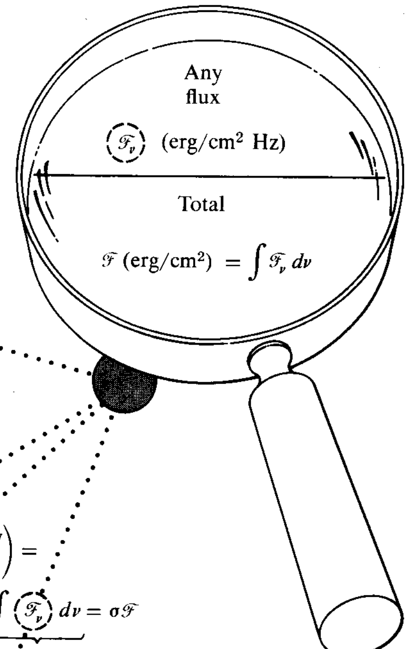
$$\begin{pmatrix}\text{Steady rate at which}\\ \text{detector extracts energy}\\ \text{from gravitational waves}\end{pmatrix} = \sigma(\nu_{\text{source}}) \int F_\nu\, d\nu = \sigma F$$

3. Frequency spread of radiation large compared to line width of detector:

$$\begin{pmatrix}\text{steady rate at which detector extracts}\\ \text{energy from gravitational waves}\end{pmatrix} = \underbrace{F_\nu(\nu_{\text{detector}})}_{\text{erg/cm}^2\text{ sec Hz}} \underbrace{\int \sigma(\nu)\, d\nu}_{\text{"resonance integral", cm}^2\text{ Hz}}$$



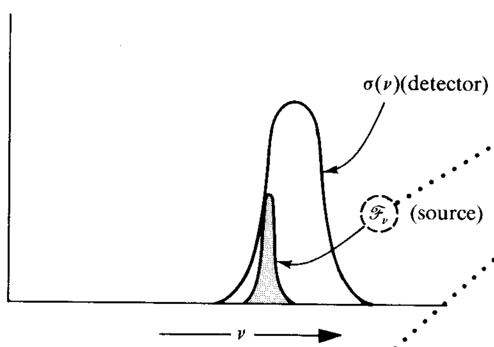
B. To Calculate *Total* Energy Deposited in Detector by *any* Passing Wave train



1. If frequency distribution of radiation is arbitrary:

$$\begin{pmatrix}\text{total energy}\\ \text{deposited}\end{pmatrix} = \int \underbrace{\mathscr{F}_\nu(\nu)}_{\text{erg/cm}^2\,\text{Hz}} \cdot \underbrace{\sigma(\nu)}_{\text{cm}^2} \underbrace{d\nu}_{\text{Hz}}$$

2. If frequency spread of radiation is small compared to line width of detector ("monochromatic waves"):

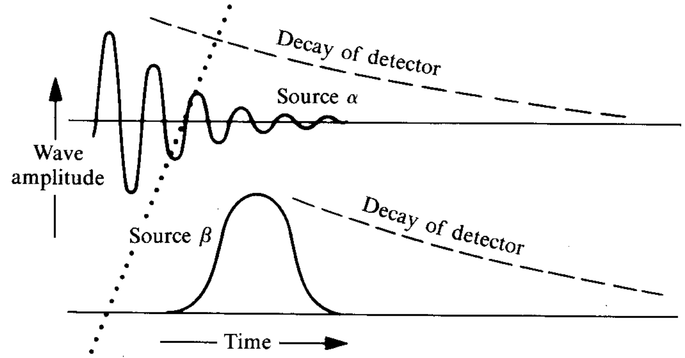


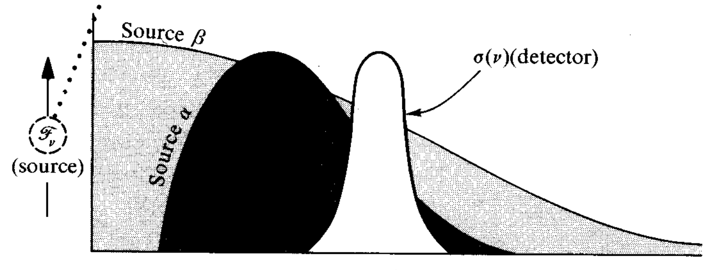
$$\begin{pmatrix}\text{total energy}\\ \text{deposited}\end{pmatrix} = \underbrace{\sigma(\nu_{\text{source}})}_{\text{cm}^2} \underbrace{\int \mathscr{F}_\nu \, d\nu}_{\text{erg/cm}^2} = \sigma\mathscr{F}$$

3. If frequency spread of radiation is large compared to line width of detector (as it *must* be for hammer-blow radiation, where

$\Delta\nu_{\text{source}} \gtrsim 1/4\pi\tau_{\text{GW}} \gg 1/4\pi\tau_0 = \Delta\nu_{\text{detector}}$):

$$(\text{total energy deposited}) = \underbrace{\mathscr{F}_\nu(\nu_{\text{detector}})}_{\text{erg/cm}^2\,\text{Hz}} \underbrace{\int \sigma(\nu)\, d\nu}_{\substack{\text{cm}^2\,\text{Hz},\\ \text{"resonance}\\ \text{integral"}}}$$





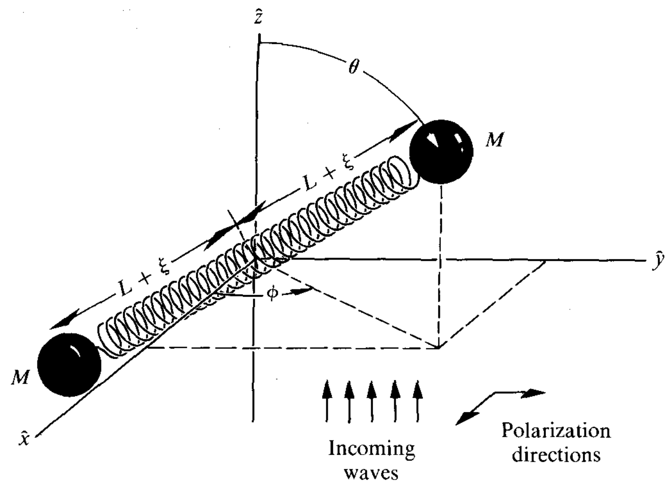


Figure 37.4.
An idealized detector (vibrator) responding to linearly polarized gravitational waves.

of the antenna, and on other, more detailed characteristics of its response, on coupling of the antenna to the displacement sensor, on response characteristics of the sensor, on antenna noise, on sensor noise, and so on. For an overview of these issues, and for discussions of detectors for which the concept of cross section is useless, see, e.g., Press and Thorne (1972).

§37.5. IDEALIZED WAVE-DOMINATED DETECTOR, EXCITED BY STEADY FLUX OF MONOCHROMATIC WAVES

Idealized detector: oscillator driven by a steady flux of monochromatic waves:

(1) derivation of equation of motion

Begin with the case of a *wave-dominated detector* ($\bar{E}_{\text{vibration}} \gg kT$) being driven by a steady flux of radiation ($\tau_{\text{GW}} \gg \tau_0$). Deal at first, not with a solid bar of arbitrary shape, but rather with the idealized detector of Figure 37.4: an oscillator made of two masses M on the ends of a spring of equilibrium length $2L$. Let the detector have a natural frequency of vibration ω_0 and a damping time $\tau_0 \gg 1/\omega_0$, so that its equation of motion (in the detector's proper reference frame) is

$$\ddot{\xi} + \dot{\xi}/\tau_0 + \omega_0^2\xi = \text{driving acceleration.} \tag{37.13}$$

Let gravitational waves of polarization $\boldsymbol{e}_+$ and angular frequency ω impinge on the detector from the $-\hat{z}$ direction; and let the polar angles of the detector relative to the wave-determined $\hat{x}, \hat{y}, \hat{z}$-axes be θ and ϕ.

The incoming waves are described by equations (37.1) with the amplitude

$$A_\times = 0, \qquad A_+ = \mathcal{A}_+ e^{-i\omega(t-z)}. \tag{37.14}$$

(Here and throughout one must take the real part of all complex expressions.)

Assume that the detector is much smaller than a wavelength, so that one can set $z \approx \hat{z} = 0$ throughout it. Then the tidal accelerations produced by the wave

$$\left(\frac{d^2\hat{x}}{d\hat{t}^2}\right)_{\text{due to wave}} = -R_{\hat{x}\hat{0}\hat{j}\hat{0}}x^{\hat{j}} = -\frac{1}{2}\omega^2 \mathscr{A}_+ e^{-i\omega\hat{t}}\hat{x},$$

$$\left(\frac{d^2\hat{y}}{d\hat{t}^2}\right)_{\text{due to wave}} = -R_{\hat{y}\hat{0}\hat{j}\hat{0}}x^{\hat{j}} = +\frac{1}{2}\omega^2 \mathscr{A}_+ e^{-i\omega\hat{t}}\hat{y},$$

have as their component along the oscillator

$$\frac{d^2\xi}{dt^2} = \frac{\hat{x}}{L}\frac{d^2\hat{x}}{dt^2} + \frac{\hat{y}}{L}\frac{d^2\hat{y}}{dt^2} + \frac{\hat{z}}{L}\frac{d^2\hat{z}}{dt^2} = -\frac{1}{2}\omega^2 \mathscr{A}_+ L e^{-i\omega\hat{t}}\frac{\hat{x}^2 - \hat{y}^2}{L^2}$$
$$= -\frac{1}{2}\omega^2 \mathscr{A}_+ L e^{-i\omega\hat{t}} \sin^2\theta \cos 2\phi.$$

Consequently, the equation of motion for the oscillator is

$$\ddot{\xi} + \dot{\xi}/\tau_0 + \omega_0{}^2\xi = -\frac{1}{2}\omega^2 \mathscr{A}_+ L e^{-i\omega\hat{t}} \sin^2\theta \cos 2\phi. \tag{37.15}$$

The driving force varies as $\cos 2\phi$ because of the "spin-2" nature of gravitational waves: a rotation through 180° in the transverse plane leaves the waves unchanged; a rotation through 90° reverses the phase. The $\sin^2\theta$ term results from the transverse nature of the waves [one factor of $\sin\theta$ to account for projection onto the detector's direction], plus their tidal-force nature [another factor of $\sin\theta$ to account for (relative force) $\propto$ (distance *in transverse plane*)].

The straightforward steady-state solution of the equation of motion (37.15) is

(2) oscillator amplitude as function of frequency and orientation

$$\xi = \frac{\frac{1}{2}\omega^2 \mathscr{A}_+ L \sin^2\theta \cos 2\phi}{\omega^2 - \omega_0{}^2 + i\omega/\tau_0} e^{-i\omega\hat{t}}. \tag{37.16}$$

When the incoming waves are near resonance with the detector, $|\omega \pm \omega_0| \lesssim 1/\tau_0$, the oscillator is excited to large amplitude. Otherwise the excitation is small. Focus attention henceforth on near-resonance excitations; then equation (37.16) can be simplified (*note:* ω_0 is positive, but ω may be negative or positive):

$$\xi = \frac{\frac{1}{4}\omega_0 \mathscr{A}_+ L \sin^2\theta \cos 2\phi}{|\omega| - \omega_0 + \frac{1}{2}\text{sgn}(\omega) i/\tau_0} e^{-i\omega\hat{t}}. \tag{37.16'}$$

One measure of the detector's usefulness is its cross section for absorbing gravitational-wave energy. The steady-state vibrational energy in a detector with the above amplitude and with 2 masses M is

$$E_{\text{vibration}} = 2 \cdot \frac{1}{2} \cdot M \cdot (\dot{\xi}^2)_{\max} = \frac{\frac{1}{16} ML^2\omega_0{}^4 \mathscr{A}_+{}^2 \sin^4\theta \cos^2 2\phi}{(|\omega| - \omega_0)^2 + (1/2\tau_0)^2}. \tag{37.17}$$

This energy is being dissipated internally at a rate $E_{\text{vibration}}/\tau_0$. If one ignores reradiation of energy as gravitational waves (a negligible process!), one can equate the dissipation rate to the rate at which the detector absorbs energy from the incoming waves—which in turn equals the "cross section" σ times the incoming flux:

$$E_{\text{vibration}}/\tau_0 = -dE_{\text{waves}}/dt = \sigma T^{0z(\text{GW})} = \frac{1}{32\pi}\sigma\omega^2 \mathcal{A}_+{}^2.$$

(3) cross sections for polarized radiation

Consequently, *near resonance* ($|\omega \pm \omega_0| \ll \omega_0$), *the cross section* for interception of gravitational-wave energy is

$$\sigma = \frac{2\pi ML^2(\omega_0{}^2/\tau_0)\sin^4\theta\cos^2 2\phi}{(|\omega| - \omega_0)^2 + (1/2\tau_0)^2}, \quad \text{for polarized radiation.} \tag{37.18}$$

This expression applies to monochromatic radiation. However, experience with many other kinds of waves has taught that one often has to deal with a broad continuum of frequencies, with the "bandwidth" of the incident radiation far greater than the width of the detector resonance (see Box 37.3). Under these conditions, the relevant quantity is not the cross section itself, but the "resonance integral" of the cross section,

$$\int_{\text{resonance}} \sigma\, d\nu = \int \sigma(d\omega/2\pi) = 2\pi ML^2\omega_0{}^2\sin^4\theta\cos^2 2\phi, \quad \text{for polarized radiation.} \tag{37.19}$$

Before examining the magnitude of this cross section, scrutinize its directionality (the "antenna-beam pattern"). The factor of $\sin^4\theta\cos^2 2\phi$ refers to linearly polarized, $\boldsymbol{e}_+$ radiation (see Figure 37.4). For the orthogonal mode of polarization, $\boldsymbol{e}_\times$, $\cos^2 2\phi$ is to be replaced by $\sin^2 2\phi$; and for unpolarized (incoherent mixture) radiation or circularly polarized radiation, the cross section is the average of these two expressions; thus

(4) cross sections for unpolarized radiation

$$\sigma = \frac{\pi ML^2(\omega_0{}^2/\tau_0)\sin^4\theta}{(|\omega| - \omega_0)^2 + (1/2\tau_0)^2} \quad \text{for unpolarized radiation.} \tag{37.20}$$

Notice that this unpolarized cross section is peaked, with half-width 33°, about the equatorial plane of the detector. Averaged over all possible directions of incoming waves, the cross section is

$$\begin{aligned}\langle\sigma\rangle_{\text{all directions}} &= \frac{1}{2}\int_0^\pi \sigma\sin\theta\, d\theta = \frac{8}{15}\sigma_{\max} \\ &= \frac{(8\pi/15)ML^2(\omega_0{}^2/\tau_0)}{(|\omega| - \omega_0)^2 + (1/2\tau_0)^2} \text{ for unpolarized radiation.}\end{aligned} \tag{37.21}$$

One can rewrite the above cross sections in several suggestive forms. For example, on resonance, the cross section (37.21) reads

$$\langle\sigma\rangle_{\text{all directions}} = \frac{4\pi^2}{15}\frac{4M}{2\pi/\omega_0}(\omega_0\tau_0)(2L)^2.$$

Recall that $\omega_0\tau_0$ defines the "Q" of a detector, $1/Q \equiv$ (fraction of *energy* dissipated per radian of oscillation). Note that $2\pi/\omega_0$ is the wavelength λ_0 of resonant radiation. Finally, denote by $r_g = 4M$ the gravitational radius of the detector. In terms of these three familiar quantities, find for the cross section the formula

Magnitude of cross sections for any resonant detector

$$\frac{\langle\sigma\rangle_{\text{all directions}}}{(2L)^2} = \frac{\text{(cross section for absorbing waves on resonance)}}{\text{("geometric" cross section of detector)}}$$
$$= (4\pi^2/15)(r_g/\lambda_0)Q \quad \begin{array}{l}\text{for unpolarized radiation}\\ \text{on resonance.}\end{array} \tag{37.22}$$

This relation holds in order of magnitude for any resonant detector. It shows starkly that gravitational-wave astronomy must be a difficult enterprise. How large could *you* make the factor r_g/λ_0, given a reasonable budget? Weber's 1970 detectors have $2L_{\text{effective}} \approx 1$ meter, $r_g \approx (0.74 \times 10^{-28}\ \text{cm/g}) \times (10^6\ \text{g}) \approx 10^{-22}$ cm, $\nu_0 = \omega_0/2\pi =$ 1,660 Hz, $\lambda_0 \approx 200$ km, $r_g/\lambda_0 \approx \frac{1}{2} \times 10^{-29}$, $\tau_0 \approx 20$ sec, $Q \approx 2 \times 10^5$; so that

$$\sigma_{\text{Weber}} \approx 3 \times 10^{-20}\ \text{cm}^2 \text{ on resonance.} \tag{37.23}$$

Flux required to excite a Weber-type detector

What flux of gravitational-wave energy would have to be incident to excite a cold detector ($\sim 0°$ K) into roughly steady-state vibrations with a vibration energy of (Boltzmann's constant) $\times$ (room temperature) $\sim 4 \times 10^{-14}$ erg? The vibrator, if cooled enough to be wave-dominated, dissipates its energy at the rate $E_{\text{vibration}}/\tau_0$ $\sim 2 \times 10^{-15}$ erg/sec. The incident flux has to make up this loss, at the rate

$$T_{00}^{(\mathrm{GW})}\sigma \sim 2 \times 10^{-15}\ \text{erg/sec}, \tag{37.24a}$$

implying an incident flux of the order of $2 \times 10^{-15}/3 \times 10^{-20} \sim 10^5$ erg/cm^2 sec. Moreover, this flux has to be concentrated in the narrow range of resonance

$$\nu \approx \nu_0 \pm 1/4\pi\tau_0 = (1660 \pm 0.004)\ \text{Hz}. \tag{37.24b}$$

By anybody's standards, this is a very high flux of gravitational radiation for such a small bandwidth ($\sim 10^7$ erg/cm^2 sec Hz, as compared to the flux of blackbody gravitational radiation, $8\pi\nu^2 kT/c^2 = 3 \times 10^{-27}$ erg/cm^2 sec Hz, that would correspond to Planck equilibrium at the same temperature; the large factor of difference is a direct reflection of the difference in rate of damping of the oscillator by friction and by gravitational radiation).

A large Q is not necessarily optimal

Equation (37.22) makes it seem that an optimal detector must have a large Q. This is not necessarily so. Recall that the bandwidth, $\Delta\omega \approx \omega_0/Q$, over which the cross section is large, decreases with increasing Q. When an incoming steady flux of waves of bandwidth $\Delta\omega \gg \omega_0/Q \equiv 1/\tau_0$ and of specific flux

$$F_\nu(\text{erg/cm}^2\ \text{sec Hz})$$

drives the detector, it deposits energy at the rate

2

$$\begin{pmatrix}\text{rate of deposit}\\ \text{of energy}\end{pmatrix} = \frac{dE}{dt} = \int_{\text{resonance}} F_\nu \sigma \, d\nu = F_\nu(\nu_0) \int_{\text{resonance}} \sigma \, d\nu.$$

$$\begin{bmatrix}\text{for radiation with}\\ \text{bandwidth } \Delta\nu \gg 1/\tau_0\end{bmatrix}$$

Consequently, the relevant measure of detector effectiveness will be the integral of the cross section over the resonance, $\int \sigma \, d\nu$ (37.19). (See next section.) This frequency-integrated cross section is independent of the detector's Q, so one must use more sophisticated reasoning (e.g., signal-to-noise theory) in deciding whether a large Q is desirable. (See §37.8).

§37.6. IDEALIZED, WAVE-DOMINATED DETECTOR, EXCITED BY ARBITRARY FLUX OF RADIATION

Response of idealized detector to an arbitrary, non-monochromatic flux:

(1) derivation

Let plane-polarized waves of polarization $\boldsymbol{e}_+$ but *arbitrary* spectrum [equation (37.1) with $A_\times = 0$] impinge on the idealized detector of Figure 37.4. Then the equation of motion for the detector is the same as for monochromatic waves [equation (37.15)], but with $-\omega^2 \mathcal{A}_+ e^{-i\omega t}$ replaced by $\ddot{A}_+$:

$$\ddot{\xi} + \dot{\xi}/\tau_0 + \omega_0{}^2 \xi = \frac{1}{2} \ddot{A}_+ L \sin^2\theta \cos 2\phi. \tag{37.26}$$

[By now one is fully accustomed to the fact that all analyses of detectors (when the detector is much smaller than the wavelength of the waves) are performed in the proper reference frame, with coordinates $\hat{t}, \hat{x}, \hat{y}, \hat{z}$. *Henceforth, for ease of eyesight, abandon the "hats" on these "proper coordinates," and denote them as merely t, x, y, z.*]

Fourier-analyze the waves and the detector displacement,

$$A_+(t) = (2\pi)^{-1/2} \int_{-\infty}^{+\infty} \tilde{A}_+(\omega) e^{-i\omega t}, \tag{37.27a}$$

$$\xi(t) = (2\pi)^{-1/2} \int_{-\infty}^{+\infty} \tilde{\xi}(\omega) e^{-i\omega t}; \tag{37.27b}$$

and conclude from equation (37.26) that

$$\tilde{\xi} = \frac{\frac{1}{2}\omega^2 \tilde{A}_+ L \sin^2\theta \cos 2\phi}{\omega^2 - \omega_0{}^2 + i\omega/\tau_0}.$$

This Fourier amplitude is negligible unless $|\omega \pm \omega_0| \ll \omega_0$; consequently, without loss of accuracy, one can rewrite it as

$$\tilde{\xi} = \frac{\frac{1}{4}\omega_0 \tilde{A}_+ L \sin^2\theta \cos 2\phi}{|\omega| - \omega_0 + \frac{1}{2}\operatorname{sgn}(\omega) i/\tau_0}. \tag{37.28}$$

[Compare with the steady-state amplitude (37.16′).]

Ask how much total energy is deposited in the detector by the gravitational waves. Do *not* seek an answer by examining the amplitude of the vibrations, $\xi(t)$, directly; since that amplitude is governed by *both* internal damping and the driving force of the waves, it does not reflect directly the energy deposited. To get the total energy deposited, integrate over time the force acting on each mass multiplied by its velocity:

$$\begin{pmatrix}\text{total energy}\\ \text{deposited}\end{pmatrix} = \int_{-\infty}^{+\infty} \underset{[2\text{ masses}]}{2} \underbrace{\left(\frac{1}{2} M \ddot{A}_+ L \sin^2\theta \cos 2\phi\right)}_{[\text{force on each mass}]} \underset{[\text{velocity of each mass}]}{\dot{\xi}}\, dt.$$

Use Parseval's theorem (one of the most powerful tools of mathematical physics!) to replace the time integral by a frequency integral

$$\begin{pmatrix}\text{total energy}\\ \text{deposited}\end{pmatrix} = \Re \int_{-\infty}^{+\infty} (ML \sin^2\theta \cos 2\phi)(-\omega^2 \tilde{A}_+^*)(-i\omega\tilde{\xi})\, d\omega.$$

Then use equation (37.28) to rewrite this entirely in terms of the wave amplitude

$$\begin{pmatrix}\text{total energy}\\ \text{deposited}\end{pmatrix} = \int_{-\infty}^{+\infty} \left[\frac{2\pi(\omega_0^2/\tau_0) ML^2 \sin^4\theta \cos^2 2\phi}{(|\omega| - \omega_0)^2 + (1/2\tau_0)^2}\right]\left[\frac{\omega^2 |\tilde{A}_+|^2}{16\pi}\right] d\omega. \tag{37.29}$$

The first term in this expression is precisely the cross section for monochromatic waves, derived in the last section (37.18). The second term has an equally simple interpretation: the total energy that the gravitational waves carry past a unit surface area of detector is

$$\begin{aligned}\mathcal{F}(\text{ergs/cm}^2) &= \int T_{00}^{(\mathrm{GW})}\, dt = \int \frac{1}{16\pi} \dot{A}_+^{\,2}\, dt \\ &= \int \frac{\omega^2 |\tilde{A}_+|^2}{16\pi}\, d\omega = \int \frac{\omega^2 |\tilde{A}_+|^2}{8}\, d\nu\end{aligned} \tag{37.30}$$

(Parseval's theorem again!). Consequently, the energy per unit frequency interval, per unit area carried by the waves is

$$\mathcal{F}_\nu(\text{ergs/cm}^2\ \text{Hz}) = \frac{1}{8}\omega^2 |\tilde{A}_+|^2 \tag{37.31}$$

[for $-\infty < \nu < +\infty$; double this for $0 < \nu < +\infty$, a convention we use for the rest of this chapter]. This is 2π times the second term in (37.29).

Combining equations (37.18), (37.29), and (37.31), then, one finds

$$\begin{pmatrix}\text{total energy}\\ \text{deposited}\end{pmatrix} = \int \sigma(\nu) \mathcal{F}_\nu(\nu)\, d\nu. \tag{37.32}$$

(2) answer—
$\begin{pmatrix}\text{energy}\\ \text{deposited}\end{pmatrix} = \int \sigma \mathcal{F}_\nu\, d\nu$

This is the total energy deposited, regardless of the spectrum of the waves, and regardless of whether they come in a steady flux for a long time, or in a short burst, or in any other form. It is perfectly general—so long as the detector is wave-dominated ($E_{\text{vibration}} \gg kT$) while the waves are driving it.

How one can measure energy deposited

How can an experimenter measure the total energy deposited? He cannot measure it directly, in general, but he can measure a quantity equal to it: the total energy that goes into internal damping, i.e., into "friction." Energy is removed by "friction" at a rate $E_{\text{vibration}}/\tau_0$, when the vibration energy is much greater than kT (during period of wave-dominance). Therefore, the experimenter can measure

$$\begin{pmatrix}\text{total energy}\\ \text{deposited}\end{pmatrix} = \frac{1}{\tau_0}\int E_{\text{vibration}}\, dt, \qquad \text{in general.} \tag{37.33}$$

[integrate over the period that $E_{\text{vibration}} \gg kT$]

In the special case of "hammer-blow waves" (τ_{GW} = duration of waves $\ll \tau_0$), the vibration energy is driven "instantaneously" from $\sim kT$ to a peak value, $E^{\text{peak}}_{\text{vibration}} \gg kT$, and then decays exponentially back to $\sim kT$; thus

$$\begin{pmatrix}\text{total energy}\\ \text{deposited}\end{pmatrix} = \frac{1}{\tau_0}\int_0^\infty E^{\text{peak}}_{\text{vibration}} e^{-t/\tau_0}\, dt = E^{\text{peak}}_{\text{vibration}} \tag{37.34}$$

for hammer-blow waves.

When the waves are steady for a long period of time ($\tau_{\text{GW}} \gg \tau_0$), with specific flux

$$F_\nu = \mathcal{F}_\nu/\tau_{\text{GW}} \qquad (\text{ergs/cm}^2\ \text{sec Hz}),$$

then the energy will be deposited at a constant rate

$$(dE/dt) = (\text{total energy deposited})/\tau_{\text{GW}};$$

and equation (37.32) can be rewritten

$$\begin{pmatrix}\text{rate of deposit}\\ \text{of energy}\end{pmatrix} = \int \sigma(\nu) F_\nu\, d\nu, \qquad \text{for steady waves } (\tau_{\text{GW}} \gg \tau_0). \tag{37.35}$$

Equations (37.32) and (37.35) are the key equations for application of the concept of cross section to realistic situations. They are applicable not only to polarized radiation, but also to unpolarized radiation and to radiation coming in from all directions, if one merely makes sure to use the appropriate cross section [equation (37.20) or (37.21) instead of (37.18)]. For examples of their application, see Box 37.3.

§37.7. GENERAL WAVE-DOMINATED DETECTOR, EXCITED BY ARBITRARY FLUX OF RADIATION

The cross sections of the idealized spring-plus-mass detector can be put into a form more elegant than equations (37.18) to (37.21)—a form that makes contact with many branches of physics, and is valid for *any* vibrating resonant detector whatsoever.

Cross sections reexpressed in terms of "Einstein A-coefficients"

Introduce the "Einstein A-coefficients," which describe the rate at which a unit amount of detector energy is lost to internal damping and to reradiation of gravitational waves:

$$A_{\rm diss} \equiv \left(\frac{\text{rate at which energy is dissipated internally}}{\text{energy in oscillations of detector}}\right) = \frac{1}{\tau_0}, \tag{37.36a}$$

$$A_{\rm GW} \equiv \left(\frac{\text{rate at which energy is reradiated}}{\text{energy in oscillations}}\right). \tag{37.36b}$$

For the idealized detector of Figure 37.4, the standard formula (36.1) for the emission of gravitational waves yields

$$(\text{power reradiated}) = \frac{32}{15}\omega^6 M^2 L^2 \langle \xi^2 \rangle_{\text{time avg.}} \tag{37.37}$$

(see exercise 37.8). Consequently

$$A_{\rm GW} = \frac{16}{15} M L^2 \omega^4. \tag{37.38}$$

One can use these relations to rewrite the detector cross sections in terms of $A_{\rm diss}$, $A_{\rm GW}$, and the reduced wavelength

$$\lambdabar \equiv 1/\omega \tag{37.39}$$

of the radiation. For example, the cross section (37.21)—now with $\omega \geq 0$—is

$$\langle \sigma \rangle_{\text{all directions}} = \frac{1}{2}\pi \lambdabar^2 \frac{A_{\rm GW} A_{\rm diss}}{(\omega - \omega_0)^2 + (A_{\rm diss}/2)^2} \quad \text{for unpolarized radiation} \tag{37.40}$$

(*recall the assumption* $|\omega - \omega_0| \ll \omega_0$ *in all cross-section formulas*) and the corresponding integral over the resonance is

$$\int \langle \sigma \rangle_{\text{all directions}}\, d\nu = \frac{1}{2}\pi \lambdabar_0{}^2 A_{\rm GW} \quad \text{for polarized radiation.} \tag{37.41}$$

Generality of the A-coefficient formalism

These expressions for the cross section are comprehensive in their application. They apply to any vibrating, resonant, gravitational-wave detector whatsoever, as one sees from the "detailed balance" calculation of exercise 37.9, and from the dynamic calculations of exercise 37.10. They also apply, with obvious changes in statistical factors and notation, to compound-nucleus reactions in nuclear physics ("Breit-Wigner formula"; see Blatt and Weisskopf, pp. 392–94, 408–10, 555–59), to the absorption of photons by atoms and molecules, to reception of electromagnetic waves by a television antenna, etc. Equation (37.41) says in effect, "Calculate the rate at which the oscillator is damped by emission of gravitational radiation; multiply that rate by the geometric factor familiar in all work with antennas, $\frac{1}{2}\pi\lambdabar_0{}^2$, and immediately obtain the resonance integral of the cross section. The result is expressed in geometric units (cm). To get the resonance integral in conventional units, multiply by the conversion factor $c = 3 \times 10^{10}$ cm Hz.

The 'dynamic analysis' of the idealized masses-on-spring detector, as developed in the last section, is readily extended to a vibrating detector of arbitrary shape (Earth; Weber's bar; an automobile fender; and so on). The extension is carried out in exercise 37.10 and its main results are summarized in Box 37.4.

Scattering of radiation by detector

Part of the energy that goes into a detector is reradiated as scattered gravitational radiation. For any detector of laboratory dimensions with laboratory damping coefficients, this fraction is fantastically small. However, in principle one can envisage a larger system and conditions where the reradiation is not at all negligible. In such an instance one is dealing with scattering. No attempt is made here to analyze such scattering processes. For a simple order-of-magnitude treatment, one can use the same type of Breit-Wigner scattering formula that one employs to calculate the scattering of neutrons at a nuclear resonance or photons at an optical resonance. A still more detailed account will analyze the correlation between the polarization of the scattered radiation and the polarization of the incident radiation. The kind of formalism useful here for gravitational radiation with its tensor character will be very much like that now used to treat polarization of radiation with a spin-1 character. Here notice especially the "Madison Convention" [Barschall and Haeberli (1971)] developed by the collaborative efforts of many workers after experience during many years with a variety of conflicting notations. Considering the way in which the best notation that is available today for spin-1 radiation was evolved, one can only feel that it is too early to canonize any one notation for describing the scattering parameters for an object that is scattering gravitational radiation.

EXERCISES

Exercise 37.8. POWER RERADIATED

The idealized gravitational wave detector of Figure 37.4 vibrates with angular frequency ω. Show that the power it radiates as gravitational waves is given by equation (37.37).

Exercise 37.9. CROSS SECTIONS CALCULATED BY DETAILED BALANCE

Use the principle of detailed balance to derive the cross sections (37.41) for a vibrating, resonant detector of any size, shape, or mass (e.g., for the vibrating Earth, or Weber's vibrating cylinder, or the idealized detector of Figure 37.4). [*Hints:* Let the detector be in thermal equilibrium with a bath of blackbody gravitational waves. Then it must be losing energy by reradiation as rapidly as it is absorbing it from the waves. (Internal damping can be ignored because, in true thermal equilibrium, energy loss by internal damping will match energy gain from random internal Brownian forces.) In detail, the balance of energy in and out reads [with I_ν = "specific intensity," equation (22.48)]

$$[4\pi I_\nu(\nu = \nu_0)]_{\text{blackbody}} \times \int \langle\sigma\rangle_{\text{all directions}}\, d\nu$$
$$= A_{\text{GW}} \times (\text{Energy in normal mode of detector}).$$

Solve for $\int\langle\sigma\rangle\, d\nu$, using the familiar form of the Planck spectrum and the fact that gravitational waves have two independent states of polarization.] *Note:* Because detailed balance

Box 37.4 VIBRATING, RESONANT DETECTOR OF ARBITRARY SHAPE

A. Physical Characteristics of Detector

1. Detector is a solid object (Earth, Weber bar, automobile fender, . . .) with density distribution $\rho(x)$ and total mass $M = \int \rho \, d^3x$.
2. Detector has normal modes of vibration. The nth normal mode is characterized by:

$$\omega_n = \text{angular frequency};$$

$$\tau_n = \begin{pmatrix}\text{e-folding time for vibration energy} \\ \text{to decay as result of internal damping}\end{pmatrix} \gg 1/\omega_n; \tag{1}$$

$$\boldsymbol{u}_n(\boldsymbol{x}) = \text{eigenfunction (defined here to be dimensionless and real).}$$

The eigenfunctions $\boldsymbol{u}_n$ are orthonormalized, so that

$$\int \rho \boldsymbol{u}_n \cdot \boldsymbol{u}_m \, d^3x = M\delta_{nm}. \tag{2}$$

3. During a normal-mode vibration with $E_{\text{vibration}} \gg kT$, a mass element originally at $\hat{x}$ receives the displacement

$$\delta \boldsymbol{x} = \boldsymbol{\xi} = \boldsymbol{u}_n(\boldsymbol{x})\underbrace{\mathscr{B}_n}_{\text{[constant amplitude]}} e^{-i\omega_n t - t/\tau_n}, \tag{3a}$$

the density at fixed $\boldsymbol{x}$ changes by

$$\delta\rho = -\boldsymbol{\nabla} \cdot (\rho \boldsymbol{u}_n)\mathscr{B}_n e^{-i\omega_n t - t/\tau_n}, \tag{3b}$$

and the moment of inertia tensor oscillates

$$\delta I_{jk} = I_{(n)jk}\mathscr{B}_n e^{-i\omega_n t - t/\tau_n}. \tag{3c}$$

Here $I_{(n)jk}$ is the "*moment of inertia factor for the nth normal mode*":

$$\begin{aligned} I_{(n)jk} &\equiv \int -(\rho u_n{}^l)_{,l} x^j x^k \, d^3x \\ &= \int \rho(u_n{}^j x^k + u_n{}^k x^j) \, d^3x \end{aligned} \tag{4}$$

[dimensions: mass $\times$ length, multiply by $\mathscr{B}_n$ (length) to get I_{jk}].

The corresponding "*reduced quadrupole factor for the nth normal mode*" is

$$\text{\it\char'40}\!\!I_{(n)jk} \equiv I_{(n)jk} - \frac{1}{3} I_{(n)ll}\,\delta_{jk}. \tag{5}$$

Box 37.4 (*continued*)

B. Cross Sections for Detector (exercise 37.10)

1. For *polarized radiation* with propagation direction $\boldsymbol{n}$ and polarization tensor $\boldsymbol{e}$:

$$h_{jk} = A(t - \boldsymbol{n} \cdot \boldsymbol{x})e_{jk},$$

$$e_{jk}n_k = 0, \qquad e_{jj} = 0, \qquad e_{jk}e_{jk} = 2; \tag{6}$$

$$\sigma_n(\nu) = \sigma_n(\omega/2\pi) = \frac{\pi}{4}\frac{|\not{I}_{(n)jk}e_{jk}|^2}{M}\frac{\omega_n{}^2/\tau_n}{(|\omega| - \omega_n)^2 + (1/2\tau_n)^2}, \tag{7a}$$

$$\int_{\text{resonance}} \sigma_n \, d\nu = \frac{\pi}{4}\frac{|\not{I}_{(n)jk}e_{jk}|^2}{M}\omega_n{}^2. \tag{7b}$$

2. For *unpolarized radiation* (random mixture of polarizations) with propagation direction $\boldsymbol{n}$, cross sections are

$$\sigma_n(\nu) = \sigma_n(\omega/2\pi) = \frac{\pi}{4}\frac{(\not{I}^{\text{TT}}_{(n)jk})^2}{M}\frac{\omega_n{}^2/\tau_n}{(|\omega| - \omega_n)^2 + (1/2\tau_n)^2}, \tag{8a}$$

$$\int_{\text{resonance}} \sigma_n \, d\nu = \frac{\pi}{4}\frac{(\not{I}^{\text{TT}}_{(n)jk})^2}{M}\omega_n{}^2. \tag{8b}$$

Here $\not{I}^{\text{TT}}_{(n)jk}$ is the transverse-traceless part of $\not{I}_{(n)jk}$ (transverse and traceless relative to the propagation direction $\boldsymbol{n}$):

$$\not{I}^{\text{TT}}_{(n)} = P\not{I}_{(n)}P - \frac{1}{2}P \text{ trace } (P\not{I}_{(n)}), \qquad P_{jk} \equiv \delta_{jk} - n_j n_k. \tag{9}$$

(See Box 35.1)

3. Cross sections for *unpolarized radiation, averaged over all directions,* are

$$\langle\sigma_n(\nu)\rangle_{\text{all directions}} = \frac{1}{2}\pi\bar{\lambda}^2\frac{A_{\text{GW}}A_{\text{diss}}}{(|\omega| - \omega_n)^2 + (A_{\text{diss}}/2)^2}, \tag{10a}$$

$$\int_{\text{resonance}} \langle\sigma_n\rangle_{\text{all directions}} \, d\nu = \frac{1}{2}\pi\bar{\lambda}^2 A_{\text{GW}}, \tag{10b}$$

where the Einstein A coefficients are

$$A_{\text{diss}} = 1/\tau_n, \tag{11}$$

$$A_{\text{GW}} = \frac{1}{5}\frac{(\not{I}_{(n)jk})^2}{M}\omega_n{}^4. \tag{12}$$

C. Spectrum Radiated by an Aperiodic Source (exercise 37.11)

It is instructive to compare these formulas with expressions for the radiation emitted by an aperiodic source.

1. Fourier-analyze the reduced quadrupole factor of the source

$$\bar{I}_{jk}(t) = (2\pi)^{-1/2} \int_{-\infty}^{+\infty} \tilde{\bar{I}}_{jk}(\omega) e^{-i\omega t}\, d\omega.$$

2. Then the total energy per unit frequency ($\nu \geq 0$) radiated over all time, into a unit solid angle about the direction $\boldsymbol{n}$, and with polarization tensor $\boldsymbol{e}$, is

$$\frac{dE}{d\nu\, d\Omega} = \frac{1}{8} \sum_{\omega = \pm 2\pi\nu} |\tilde{\bar{I}}_{jk} e_{jk}|^2 \omega^6 \tag{13a}$$

[compare with equations (7)]. Summed over polarizations, this is

$$\frac{dE}{d\nu\, d\Omega} = \frac{1}{2} \sum_{j,k} |\tilde{\bar{I}}_{jk}{}^{\mathrm{TT}}|^2 \omega^6 \tag{13b}$$

[compare with equations (8)]. Here $\nu \geq 0$.

3. The total energy radiated per unit frequency, integrated over all directions, still with $\nu \geq 0$, is

$$dE/d\nu = \frac{4\pi}{5} \sum_{j,k} |\tilde{\bar{I}}_{jk}|^2 \omega^6 \tag{14}$$

[compare with equations (10)–(12)].

can be applied to any kind of resonant system in interaction with any kind of thermal bath of radiation or particles, equations (37.40) and (37.41), with appropriate changes of statistical factors, have wide generality.

Exercise 37.10. NORMAL-MODE ANALYSIS OF VIBRATING, RESONANT DETECTORS

Derive all the results for vibrating, resonant detectors quoted in Box 37.4. Pattern the derivation after the treatment of the idealized detector in §37.6. [*Guidelines:* (a) Let the detector be driven by the polarized waves of equation (6), Box 37.4; and let it be wave-dominated ($E_{\mathrm{vibration}} \gg kT$). Show that the displacements $\delta x = \xi(x, t)$ of its mass elements are described by

$$\xi = \sum_n B_n(t) \boldsymbol{u}_n(x), \tag{37.42a}$$

where the time-dependent amplitude for the nth mode satisfies the driven-oscillator equation

$$\ddot{B}_n + \dot{B}_n/\tau_n + \omega_n{}^2 B_n = R_n(t), \tag{37.42b}$$

and where the curvature-induced driving term is

$$\begin{aligned} R_n(t) &= -R_{\hat{j}\hat{0}\hat{k}\hat{0}} \int (\rho/M) u_n{}^j x^k \, d^3x \\ &= \frac{1}{4} \ddot{A} (I\!\!\!-_{(n)jk} e_{jk})/M. \end{aligned} \tag{37.42c}$$

(See Box 37.4 for notation.)

(b) Fourier-analyze the amplitudes of the detector and waves,

$$B_n(t) = (2\pi)^{-1/2} \int_{-\infty}^{+\infty} \tilde{B}_n(\omega) e^{-i\omega t} \, d\omega, \tag{37.42d}$$

$$A(t) = (2\pi)^{-1/2} \int_{-\infty}^{+\infty} \tilde{A}(\omega) e^{-i\omega t} \, d\omega, \tag{37.42e}$$

and solve the equation of motion (37.42b,c) to obtain, in the neighborhood of resonance,

$$\tilde{B}_n = \frac{\frac{1}{8} \omega_n \tilde{A} (I\!\!\!-_{(n)jk} e_{jk})/M}{|\omega| - \omega_n + \frac{1}{2} i/\tau_n} \qquad \text{for } |\omega \pm \omega_n| \ll \omega_n. \tag{37.42f}$$

(c) Calculate the total energy deposited in the detector by integrating

$$\begin{pmatrix}\text{energy} \\ \text{deposited}\end{pmatrix} = \int (\textbf{Force} \text{ per unit volume}) \cdot (\textbf{Velocity}) \, d^3x \, dt.$$

Thereby obtain

$$\begin{pmatrix}\text{energy deposited in} \\ n\text{th normal mode}\end{pmatrix} = \frac{1}{4} (I\!\!\!-_{(n)jk} e_{jk}) \int \ddot{A} \dot{B}_n \, dt.$$

(d) Apply Parseval's theorem and combine with expression (37.42f) to obtain

$$\begin{pmatrix}\text{energy deposited in} \\ n\text{th normal mode}\end{pmatrix} = \int \sigma_n(\nu) \mathcal{F}_\nu(\nu) \, d\nu, \tag{37.43}$$

where σ_n is given by equation (7a) of Box 37.4, and (for $-\infty < \omega < +\infty$)

$$\mathcal{F}_\nu(\nu) = \mathcal{F}_\nu(\omega/2\pi) = \frac{1}{8} \omega^2 |\tilde{A}|^2. \tag{37.44}$$

(e) Show that $\mathcal{F}_\nu(\nu)$ is the total energy per unit area per unit frequency carried by the waves past the detector.

(f) Obtain all the remaining cross sections quoted in Box 37.4 by appropriate manipulations of this cross section. Use the mathematical tools for projecting out and integrating "transverse-traceless parts," which were developed in Box 35.1 and exercise 36.9.

Exercise 37.11. SPECTRUM OF ENERGY RADIATED BY A SOURCE

Derive the results quoted in the last section of Box 37.4.

Exercise 37.12. PATTERNS OF EMISSION AND ABSORPTION

The elementary dumbbell oscillator of Figure 37.4, initially unexcited, has a cross section for absorption of unpolarized gravitational radiation proportional to $\sin^4\theta$, and when excited radiates with an intensity also proportional to $\sin^4\theta$ (Chapter 36). The patterns of emission and absorption are identical. Any other dumbbell oscillator gives the same pattern, apart from a possible difference of orientation. Consider a nonrotating oscillator of general shape undergoing free vibrations in a single nondegenerate (and therefore nonrotatory) mode, or excited from outside by *unpolarized* radiation.

(a) Show that its pattern of emission is identical with its pattern of absorption. [*Hint:* Make the comparisons suggested in the last few parts of Box 37.4.]

(b) Show that this emission pattern ($\equiv$ absorption pattern), apart from three Euler angles that describe the orientation of this pattern in space, and apart from a fourth parameter that determines total intensity, is uniquely fixed by a single ("fifth") parameter.

(c) Construct diagrams for the pattern of intensity for the two extreme values of this parameter and for a natural choice of parameter intermediate between these two extremes.

(d) Define the parameter in question in terms of a certain dimensionless combination of the principal moments of the reduced quadrupole tensor.

Exercise 37.13. MULTIMODE DETECTOR

Consider a cylindrical bar of length very long compared to its diameter. Designate the fundamental mode of end-to-end vibration of the bar as "$n = 1$," and call the mode with $n - 1$ nodes in its eigenfunction the "nth mode." Show that the cross section for the interception of unpolarized gravitational waves at the nth resonance, integrated over that resonance, and averaged over direction, is given by the formula [Ruffini and Wheeler (1971b)]

$$\int_{\substack{n\text{th}\\ \text{resonance;}\\ \text{random}}} \sigma(\nu)d\nu = \frac{32}{15\pi}\frac{v^2}{c^2}\frac{M}{n^2} \text{ for } n \text{ odd (zero for even } n), \tag{37.45}$$

where v is the speed of sound in the bar expressed in the same units as the speed of light, c; and M is the mass of the bar (geometric units; multiply the righthand side by the factor $G/c = 2.22 \times 10^{-18}\ \text{cm}^2\ \text{Hz/g}$ when employing conventional units). Show that this expression gives $\int\sigma\, d\nu = 1.0 \times 10^{-21}\ \text{cm}^2$ Hz for the lowest mode of Weber's bar. Multimode detectors are (1973) under construction by William Fairbank and William Hamilton, and by David Douglass and John A. Tyson.

Exercise 37.14. CROSS SECTION OF IDEALIZED MODEL OF EARTH FOR ABSORPTION OF GRAVITATIONAL RADIATION

The observed period of quadrupole vibration of the earth is 54 minutes [see, e.g., Bolt (1964) or Press (1965) for survey and bibliography]. To analyze that mode of vibration, with all due allowance for elasticity and the variation of density in the earth, is a major enterprise. Therefore, for a first estimate of the cross section of the earth for the absorption of quadrupole radiation, treat it as a globe of fluid of uniform density held in the shape of a sphere by gravitational forces alone (zero rigidity). Let the surface be displaced from $r = a$ to

$$r = a + a\alpha P_2(\cos\theta), \tag{37.46a}$$

where θ is polar angle measured from the North Pole and α is the fractional elongation of the principal axis. The motion of lowest energy compatible with this change of shape is described by the velocity field

$$\xi^x = -\frac{1}{2}\alpha x, \qquad \xi^y = -\frac{1}{2}\alpha y, \qquad \xi^z = \alpha z \tag{37.46b}$$

(zero divergence, zero curl).

(a) Show that the sum of the kinetic energy and the gravitational potential energy is

$$E = -(3/5)(M^2/a)(1 - \alpha^2/5) + (3/20)Ma^2\dot{\alpha}^2. \tag{37.46c}$$

(b) Show that the angular frequency of the free quadrupole vibration is

$$\omega = (16\pi/15)^{1/2}\rho^{1/2}. \tag{37.46d}$$

(c) Show that the reduced quadrupole moments are

$$\mathcal{I}_{xx} = \mathcal{I}_{yy} = -Ma^2\alpha/5, \qquad \mathcal{I}_{zz} = 2Ma^2\alpha/5. \tag{37.46e}$$

(d) Show that the rate of emission of vibrational energy, averaged over a period, is

$$-\langle dE/dt \rangle = (3/125)M^2a^4\omega^6{\alpha_{\text{peak}}}^2. \tag{37.46f}$$

(e) Show that the exponential rate of decay of energy by reason of gravitational wave damping, or "gravitational radiation line broadening," is

$$A_{\text{GW}} = (4/25)Ma^2\omega^4. \tag{37.46g}$$

(f) Show that the resonance integral of the absorption cross section for radiation incident from random directions with random polarization is

$$\int_{\text{resonance}} \langle \sigma(\nu) \rangle \, d\nu = (\pi/2)\bar{\lambda}^2 A_{\text{GW}} = (2\pi/25)Ma^2/\bar{\lambda}^2. \tag{37.46h}$$

(g) Evaluate this resonance integral. Note: This model of a globe of fluid of uniform density would imply for the earth, with average density 5.517 g/cm^3, a quadrupole vibration period of 94 min, as compared to the observed 54 min; and a moment of inertia $(2/5)Ma^2$ as compared to the observed $0.33Ma^2$. Ruffini and Wheeler (1971b) have estimated correction factors for both effects and give for the final resonance integral $\sim$5 cm^2 Hz.

---

§37.8. NOISY DETECTORS

When the bandwidth of the incoming waves is large compared to the resonance width of the detector, the waves deposit a total energy in the detector given by

$$\underbrace{(\text{total energy deposited})}_{[\text{ergs}]} = \int \sigma \mathcal{F}_\nu \, d\nu = \underbrace{\mathcal{F}_\nu(\nu = \nu_0)}_{[\text{erg cm}^{-2}\text{Hz}^{-1}]} \underbrace{\int \sigma \, d\nu}_{[\text{cm}^2\text{Hz}]}.$$

At least, this is so if the detector is wave-dominated (i.e., if $E_{\text{vibration}} \gg kT$ while waves act; i.e., if initial amplitude of oscillation, produced by Brownian forces, is too small to interfere constructively or destructively with the amplitude due to waves).

Extraction of small signal from large noise—general remarks

Unfortunately, all experiments today (1973) are faced with noisy detectors. Nobody has yet found waves so strong, or constructed a detector so sensitive, that the detector is wave-dominated. Consequently, a key experimental task today is to pick a small signal out of large noise. Many techniques for doing this have been developed and used in a variety of fields of physics, as well as in astronomy, psychology and engineering [see, e.g., Davenport and Root (1958), Blackman and Tukey (1959), and

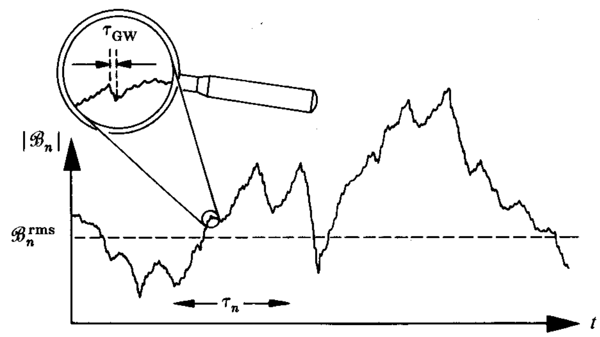


Figure 37.5.
Detection of hammer-blow gravitational waves with a noisy detector. Detection of even a weak pulse is possible if the time of the pulse is short enough. The amplitude $\mathscr{B}_n$ of the detector's vibrations changes by an amount $\sim\mathscr{B}_n^{rms}\,(\Delta t/\tau_n)^{1/2}$ during a time interval Δt, due to thermal fluctuations (random-walk, Brownian-noise forces). Depicted in the inset is a change in amplitude produced by a burst of waves of duration τ_{GW} arriving out of phase with the detector's thermal motions (energy extracted by waves!). The waves are detectable because

$$\Delta|\mathscr{B}_n|_{\text{due to waves}} \gg \mathscr{B}_n^{rms}\,(\tau_{GW}/\tau_n)^{1/2},$$

even though $\Delta|\mathscr{B}_n| \ll \mathscr{B}_n^{rms}$.

references given there]. The key point is always to find some feature of the signal that is statistically more prominent than the same feature of the noise, plus a correlation to show that it arises from the expected signal source and not from elsewhere ("protection from systematic error"). Thus to detect steady gravitational waves from a pulsar, one might seek to define very precisely two numbers $\langle N^2\rangle$ and $\langle (N+S)^2\rangle = \langle N^2\rangle + \langle S^2\rangle$, where N and S are the noise and signal amplitudes respectively. A long series of observations (with the pulsar out of the antenna beam) gives one value of $\langle N^2\rangle$. Another equally long series of observations, interspersed with the first series, will be expected in zeroth approximation to give the same value of $\langle N^2\rangle$. In the next approximation one recognizes and calculates the influence of normal statistical fluctuations. In an illustrative example, theory, confirmed by statistical tests of other parameters drawn from the same data, guarantees that the fluctuations are less than $10^{-5}\langle N^2\rangle$ with 95 per cent confidence (only 5 per cent chance of exceeding $10^{-5}\langle N^2\rangle$; this limit is set by time and money, not by absolute limitations of physics). Let the second series of observations be carried out only at times when the pulsar is in the antenna beam. Let it give

$$\{\langle N^2\rangle + \langle S^2\rangle\}_{\text{2d series}} = (1 + 7.3\times 10^{-5})\{\langle N^2\rangle\}_{\text{1st series}}.$$

Then in first approximation one can say that $\langle S^2\rangle$ lies with 95 per cent confidence in the limits $(7.3 \pm 1.0)\times 10^{-5}\langle N^2\rangle$.

Many conceivable sources of gravitational radiation produce bursts rather than a steady signal strength (Figure 37.5). Thus one is led to ask in what features "hammer-blow radiation" ($\tau_{GW} \ll \tau_0$) differs from noise. The "Brownian motion" noise in the detector may be thought of as arising from large numbers of small

Rate-of-change of detector amplitude as a tool for extracting burst signals from thermal noise (molecular) energy exchanges with a heat bath. The calculations below estimate the typical rate of change of amplitude that a series of such molecular "knocks" can produce in a detector, and compare it with the rapid amplitude change produced by a "hammer-blow" pulse of radiation. The calculations show that sudden thermally induced changes, even of very small amplitude, are rare. Thus sudden changes are a suitable feature for the observations to focus on. The actual detection of pulses requires a more extended analysis, however, which goes beyond the estimates made below. Such an analysis would calculate the probabilities that rare events (sudden changes in amplitude) occur by chance (i.e., due to thermal fluctuations) in specified periods of time, the still smaller probabilities that they occur in coincidence between two or more detectors, and the correlations with postulated sources.

Consider a realistic detector of the type described in Box 37.4. But examine it at a time when it is *not* radiation-dominated. Then its motions are being driven by internal Brownian forces (thermal fluctuations), and perhaps also by an occasional burst of gravitational waves. Focus attention on a particular normal mode (mode "n"), and describe that mode's contribution to the vibration of the detector by the vector field

$$\delta \boldsymbol{x} = \boldsymbol{\xi} = \mathscr{B}_n(t)e^{-i\omega_n t}\boldsymbol{u}_n(x). \tag{37.47}$$

Description of thermal noise in resonant detector

Since $\boldsymbol{u}_n$ is dimensionless with mean value unity ($\int \rho \boldsymbol{u}_n{}^2\, d^3x = M$), the complex number $\mathscr{B}_n(t)$ is the mass-weighted average of the amplitudes of motion of the detector's mass elements. This amplitude changes slowly with time (rate $\ll \omega_n$) as a result of driving by Brownian forces; but averaged over time it has a magnitude corresponding to a vibration energy of kT:

$$\langle E_{\text{vibration}}\rangle = 2\left\langle\frac{1}{2}\int \rho\dot{\boldsymbol{\xi}}^2\, d^3x\right\rangle = \frac{1}{2}M\omega_n{}^2\langle|\mathscr{B}_n|^2\rangle = kT; \tag{37.48}$$

i.e.,

$$\mathscr{B}_n^{\text{rms}} \equiv \langle|\mathscr{B}_n|^2\rangle^{1/2} = (2kT/M\omega_n{}^2)^{1/2}. \tag{37.49}$$

Example: for Weber's detector ($M \sim 10^3$ kg, $\omega_0 \sim 10^4$/sec), the fundamental mode at room temperature has

$$\mathscr{B}_0^{\text{rms}} = \left(\frac{2\times 1.38\times 10^{-16}\times 300 \text{ erg}}{10^6 \text{ g}\times 10^8 \text{ sec}^{-2}}\right)^{1/2} = 3\times 10^{-14}\text{ cm}. \tag{37.50}$$

One's hope for detecting weak hammer-blow radiation lies not in an examination of the detector's vibration amplitude (or energy), but in an examination of its rate of change (Figure 37.5). The time-scale for large Brownian fluctuations in amplitude ($|\Delta\mathscr{B}_n| \sim \mathscr{B}_n^{\text{rms}}$), when the detector is noisy, is the same as the time scale τ_n for internal forces to damp the detector, when it is driven to $E_{\text{vibration}} \gg kT$. Thus, *the amplitude* $\mathscr{B}_n$ *does a "random walk" under the influence of Brownian forces, with the mean time for "large walks"* ($|\Delta\mathscr{B}_n| \sim \mathscr{B}_n^{\text{rms}}$) *being* $\Delta t \approx \tau_n$. The change in $\mathscr{B}_n$ over shorter times Δt is smaller by the "$1/\sqrt{N}$ factor," which always enters into random-walk processes:

$$\sqrt{N} = \left(\frac{\text{number of vibration cycles in time } \tau_n}{\text{number of vibration cycles in time } \Delta t}\right)^{1/2} = \left(\frac{\tau_n}{\Delta t}\right)^{1/2}; \tag{37.51}$$

$$\langle|\Delta\mathscr{B}_n^{(\text{thermal})}|\rangle \approx \mathscr{B}_0^{\text{rms}}\left(\frac{\Delta t}{\tau_n}\right)^{1/2} = \left(\frac{2kT}{M\omega_n{}^2}\right)^{1/2}\left(\frac{\Delta t}{\tau_n}\right)^{1/2} \begin{array}{l}\text{during}\\ \text{time } \Delta t\end{array}. \tag{37.52*}$$

Effect of a burst of waves on a noisy, resonant detector

Now suppose that "hammer-blow" radiation (burst of duration $\Delta t = \tau_{\text{GW}} \ll \tau_n$) strikes the detector, producing a change $\Delta\mathscr{B}_n^{(\text{GW})}$ in the detector's amplitude. This change in amplitude, because it comes so quickly, (1) superposes linearly on any change in amplitude produced in the same time interval by the action of Brownian-motion forces; and (2) is therefore independent in value of the presence or absence of Brownian-motion forces, i.e., independent of all thermal agitation. Therefore $\Delta\mathscr{B}_n^{(\text{GW})}$ (a quantity with *both* magnitude and phase!) is identical to what it would have been if the detector were at zero temperature:

$$\underbrace{\frac{1}{2}M\omega_n^2|\Delta B_n^{(\text{GW})}|^2}_{\substack{\text{energy that would}\\ \text{be deposited if}\\ \text{detector were at}\\ \text{zero temperature}}} = \int \sigma_n(\nu)\mathscr{F}_\nu(\nu)\,d\nu \underset{\substack{\text{For hammer-blow radiation, bandwidth of}\\ \text{radiation is always} \gg \text{bandwidth of detector;}\\ \text{see Box 37.4}}}{=} \mathscr{F}_\nu(\omega_n/2\pi)\int \sigma_n(\nu)\,d\nu;$$

i.e.,

$$|\Delta\mathscr{B}_n^{(\text{GW})}| = \left(\frac{2\mathscr{F}_\nu(\omega_n/2\pi)\int\sigma_n\,d\nu}{M\omega_n{}^2}\right)^{1/2}. \tag{37.53}$$

This wave-induced change in amplitude will be distinguishable from thermal changes only if it is significantly bigger than the thermal changes (37.52) expected during the same length of time τ_{GW}:

Criteria for detectability of burst

$$\left.\begin{array}{l} |\Delta\mathscr{B}_n^{(\text{GW})}| \gg \langle|\Delta\mathscr{B}_n^{(\text{thermal})}|\rangle \text{ during time } \tau_{\text{GW}} \\[1ex] \text{equivalently: } F_\nu(\omega_n/2\pi) \gg \left(\dfrac{kT}{\int\sigma_n\,d\nu}\right)\left(\dfrac{\tau_{\text{GW}}}{\tau_n}\right) \end{array}\right\} \begin{array}{l}\text{criteria for}\\ \text{detectability}\end{array} \tag{37.54}$$

Of course, if one is equipped only to measure the magnitude of the detector's amplitude or energy, and not its phase, these criteria for detectability are not quite sufficient. The wave-induced change in squared amplitude (proportional to change in energy) will depend on the relative phases of the initial amplitude and amplitude change

*For a fuller derivation and discussion of this formula, see, e.g., Braginsky (1970). Two key points covered there are: (1) a statistical version of the formula, which describes the probability that in time Δt the amplitude will change by a given amount, from a given initial value; and (2) quantum-mechanical corrections, which come into play in the limit as $\tau_n \longrightarrow \infty$, but which are unimportant for detectors of the early 1970's.

$$
\begin{aligned}
\Delta|\mathcal{B}_n|^2 &= |\mathcal{B}_n^{(\text{initial})} + \Delta\mathcal{B}_n^{(\text{GW})}|^2 - |\mathcal{B}_n^{(\text{initial})}|^2 && \\
&\approx 2|\mathcal{B}_n^{(\text{initial})}||\Delta\mathcal{B}_n^{(\text{GW})}| && \text{if in phase} \\
&\approx 0 && \text{if phase difference is } \pm\pi/2 \\
&\approx -2|\mathcal{B}_n^{(\text{initial})}||\Delta\mathcal{B}_n^{(\text{GW})}| && \text{if phase difference is } \pi.
\end{aligned}
\tag{37.55}
$$

Thus, only a burst that arrives in phase with the initial motion of the detector or with reversed phase will be measurable. But for such a burst, the criteria (37.54) are sufficient.

Ways to improve sensitivity of detector

Equations (37.54) make it clear that *there are three ways to improve the sensitivity of vibratory detectors to hammer-blow radiation: (1) increase the detector's integrated cross-section* [*which can be done only by increasing the rate* A_{GW} *at which it reradiates gravitational waves; see equations* (*10b*) *and* (*11b*) *of* Box 37.4]; (*2*) *cool the detector;* (*3*) *increase the detector's damping time.*

Box 37.5 applies the above detectability criteria to some detectors that seem feasible in the 1970's, and to some bursts of waves predicted by theory. The conclusions of that comparison give one hope!

To be complete, the above discussion should have analyzed not only noise in the detector, but also the noise in the sensor which one uses to measure the amplitude of the detector's displacements. However, the theory of displacement sensors is beyond the scope of this book. For a brief discussion and for references, see Press and Thorne (1972).

§37.9. NON-MECHANICAL DETECTORS

Non-mechanical detectors

When gravitational waves flow through matter, they excite it into motion. Such excitations are the basis for all detectors described thus far. But gravitational waves interact not only with matter; they also interact with electromagnetic fields; and those interactions can also be exploited in detectors. One of the most promising detectors that may be built in the future, one designed by Braginsky and Menskii (1971), relies on a resonant interaction between gravitational waves and electromagnetic waves. It is described in Box 37.6.

§37.10. LOOKING TOWARD THE FUTURE

The future of gravitational-wave astronomy

As this book is being written, it is not at all clear whether the experimental results of Joseph Weber constitute a genuine detection of gravitational waves. (See §37.4, part 4.) But whether they do or not, gravitational-wave astronomy has begun, and seems to have a bright future. The technology of 1973 appears sufficient for the construction of detectors that will register waves from a star that collapses to form a black hole anywhere in our galaxy (Box 37.5); and detectors of the late 1970's and early 1980's may well register waves from pulsars and from supernovae in other galaxies. The technical difficulties to be surmounted in constructing such detectors are enormous. But physicists are ingenious; and with the impetus provided by Joseph Weber's pioneering work, and with the support of a broad lay public sincerely interested in pioneering in science, all obstacles will surely be overcome.

Box 37.5 DETECTABILITY OF HAMMER-BLOW WAVES FROM ASTROPHYSICAL SOURCES: TWO EXAMPLES
(The following calculations are accurate only to within an order of magnitude or so)

A. Waves from a Star of Ten Solar Masses Collapsing to Form a Black Hole; 1972 Detector with 1975 (?) Sensor

1. Predicted characteristics of radiation:

$$(\text{intensity at Earth}) = \mathscr{F}_\nu \sim \frac{M_\odot}{4\pi(\text{distance})^2\nu}$$
$$\sim(2\times 10^5\ \text{ergs/cm}^2\ \text{Hz})\ [(\text{distance to center of galaxy})/(\text{distance})]^2,$$
$$(\text{frequency of waves}) = \nu \sim 10^3\ \text{Hz},$$
$$(\text{bandwidth of waves}) = \Delta\nu \sim 10^3\ \text{Hz},$$
$$(\text{duration of burst}) = \tau_{\text{GW}} \sim 10^{-3}\ \text{sec to } 10^{-1}\ \text{sec}.$$

2. Detector properties: A Weber bar, vibrating in its fundamental mode, with

$$M = 10^6\ \text{g}, \qquad \int \sigma\, d\nu = 10^{-21}\ \text{cm}^2\ \text{Hz (exercise 37.13)},$$
$$\nu_0 = \omega_0/2\pi = 1{,}660\ \text{Hz}, \quad T = 3\ \text{K (liquid Helium temperature)},$$
$$\tau_0 = 20\ \text{seconds},$$
$$\mathscr{B}_0^{\text{rms}} = \left(\frac{2\times 1.37\times 10^{-16}\times 3\ \text{erg}}{10^6\ \text{g}\times 10^8\ \text{sec}^{-2}}\right)^{1/2} = 3\times 10^{-15}\ \text{cm},$$
$$|\Delta\mathscr{B}_0^{(\text{thermal})}| = (3\times 10^{-15}\ \text{cm})(10^{-3}/20)^{1/2} = 2\times 10^{-17}\ \text{cm}, \quad \text{during } \Delta t = 10^{-3}\ \text{sec},$$
$$|\Delta\mathscr{B}_0^{(\text{thermal})}| = 2\times 10^{-16}\ \text{cm, during } \Delta t = 0.1\ \text{sec}.$$

3. Effect of waves [equation (37.53)]:

$$|\Delta\mathscr{B}_0^{(\text{GW})}| = \left(\frac{2\times 2\times 10^5\times 10^{-21}\ \text{ergs}}{10^6\times 10^8\ \text{sec}^{-2}}\right)^{1/2}\left(\frac{\text{center of Galaxy}}{\text{distance}}\right)$$
$$= 2\times 10^{-15}\ \text{cm}\left(\frac{\text{distance to center of Galaxy}}{\text{distance}}\right).$$

4. Conclusion: Gravitational waves from a massive star collapsing to form a black hole anywhere in our galaxy are readily detectable, *if* one can construct a "*sensor*" to measure changes in vibration amplitudes of magnitude $\lesssim 10^{-15}$ cm on time scales <0.1 seconds. This does appear to be feasible with 1972 technology; see Press and Thorne (1972).

2

Box 37.5 (*continued*)

B. Waves from a Supernova Explosion in the Virgo Cluster of Galaxies; a Detector that might be constructable by late 1970's or early 1980's

1. Predicted characteristics of radiation:

$$(\text{intensity at Earth}) = F_\nu \sim \frac{0.03 M_\odot}{4\pi(11 \text{ megaparsecs})^2 \nu}$$
$$\sim 4 \times 10^{-3} \text{ ergs/cm}^2 \text{ Hz},$$
$$(\text{frequency of waves}) = \nu \sim 10^3 \text{ Hz},$$
$$(\text{bandwidth of waves}) \sim \nu \sim 10^3 \text{ Hz},$$
$$(\text{duration of burst}) = \tau_{GW} \sim 0.3 \text{ sec, or } \tau_{GW} \sim 2 \times 10^{-3} \text{ sec.*}$$

2. Detector: A Weber-type bar made not of metal, but of a 1,000-kg monocrystal of quartz, cooled to a temperature of 3×10^{-3} K. (For such a monocrystal, it is thought that the damping time would increase in inverse proportion to temperature, $\tau_0 \propto 1/T$.) Estimated properties of such a detector:

$$M \sim 10^6 \text{ g}, \qquad \int \sigma\, d\nu = 10^{-21} \text{ cm}^2 \text{ Hz (same as for Weber bar)},$$
$$\nu_0 = \omega_0/2\pi \sim 1{,}500 \text{ Hz}, \qquad T = 3 \times 10^{-3} \text{ K},$$
$$\tau_0 \sim 10^6 \text{ sec},$$
$$\mathscr{B}_0^{\text{rms}} = \left(\frac{2 \times 1.37 \times 10^{-16} \times 3 \times 10^{-3} \text{ erg}}{10^6 \text{ g} \times 10^8 \text{ sec}^{-2}}\right)^{1/2} = 1 \times 10^{-16} \text{ cm},$$
$$|\Delta\mathscr{B}_0^{(\text{thermal})}| = (1 \times 10^{-16} \text{ cm})\left(\frac{0.3 \text{ or } 2 \times 10^{-3}}{10^6}\right)^{1/2} = \begin{cases} 6 \times 10^{-20} \text{ cm}, \\ \text{or} \\ 5 \times 10^{-21} \text{ cm}. \end{cases}$$

3. Effect of waves [equation (37.53)]:

$$|\Delta\mathscr{B}_0^{(\text{GW})}| = \left(\frac{2 \times 4 \times 10^{-3} \times 10^{-21} \text{ ergs}}{10^6 \times 10^8 \text{ sec}^{-2}}\right)^{1/2} = 3 \times 10^{-19} \text{ cm}.$$

4. Conclusion: Gravitational waves are detectable from a supernova in the Virgo cluster, if one can construct a sensor to measure changes in vibration amplitudes of magnitude $\lesssim 10^{-19}$ cm on time scales of $\lesssim 0.1$ seconds; and if one can construct a detector with the above characteristics.

---

*For the duration of waves from a supernova explosion, two time scales appear to be relevant: (1) the time required for the final stages of the collapse of the white-dwarf core to a neutron star or a neutron-star pancake, $\tau \sim$ (dimensions of neutron star)/(speed of sound in nuclear matter) $\sim 2 \times 10^{-3}$ sec ("pulse of gravitational radiation"); and (2) the time required for a vibrating neutron star to lose its energy of vibration by gravitational radiation ("damped train of waves"), $\tau \sim 0.3$ sec.

Box 37.6 A NONMECHANICAL DETECTOR OF GRAVITATIONAL WAVES [Braginsky and Menskii (1971)]

The Idea in Brief

(see diagram at right)

A toroidal waveguide contains a monochromatic train of electromagnetic waves, traveling around and around it. Gravitational waves propagate perpendicular to the plane of the torus. If the circuit time for the EM waves is twice the period of the gravitational waves, then one circularly polarized component of the gravitational waves will stay always in phase with the traveling EM waves. Result: a resonance develops. In one region of the EM wave train, gravitational tidal forces always "push" the waves forward (*blue shift!*) in another region the tidal forces "push" backward (*red shift!*). An EM frequency difference builds up linearly with time; a phase difference builds up quadratically.

Outline of Quantitative Analysis

1. Let waveguide fall freely in an Earth orbit. Orient axes of waveguide's proper reference frame ($\equiv$ local Lorentz frame) so (1) waveguide lies in $\hat{x}, \hat{y}$-plane, and (2) gravitational waves propagate in $\hat{z}$ direction.

2. Let gravitational waves have amplitudes

$$A_+ - iA_\times = \mathcal{A} e^{-i\omega(t-z)} \tag{1}$$

[Recall: $\hat{t} \approx t$, $\hat{z} \approx z$; i.e., proper frame and TT coordinates almost agree.] Then in plane of waveguide ($z \doteq 0$),

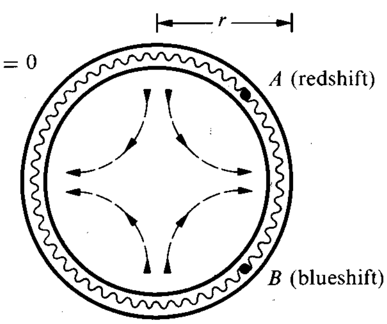
$$\begin{aligned} R_{\hat{x}\hat{0}\hat{x}\hat{0}} &= -R_{\hat{y}\hat{0}\hat{y}\hat{0}} = \frac{1}{2}\omega^2 \mathcal{A} \cos(\omega t) \\ R_{\hat{x}\hat{0}\hat{y}\hat{0}} &= R_{\hat{y}\hat{0}\hat{x}\hat{0}} = \frac{1}{2}\omega^2 \mathcal{A} \sin(\omega t) \end{aligned} \tag{2}$$

3. Consider two neighboring parts of the EM wave, one at $\phi = \alpha + \frac{1}{2}\omega t$; the other at $\phi = \alpha + \delta\alpha + \frac{1}{2}\omega t$. Treat them as photons. Each moves along a null geodesic, except for

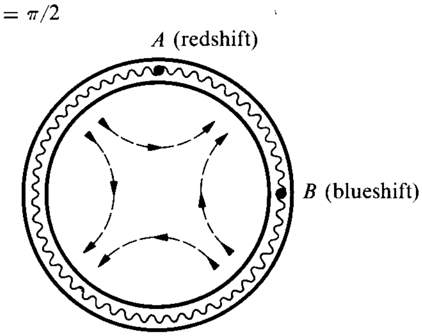
ω = (angular frequency of gravitational waves) = (rate of change of phase of waves with time) = (two times angular velocity with which pattern of "lines of force" rotates)

r = (radius of torus), is adjusted so the speed of propagation of EM waves in waveguide is $v = \frac{1}{2}\omega r$.

• $\omega t = 0$

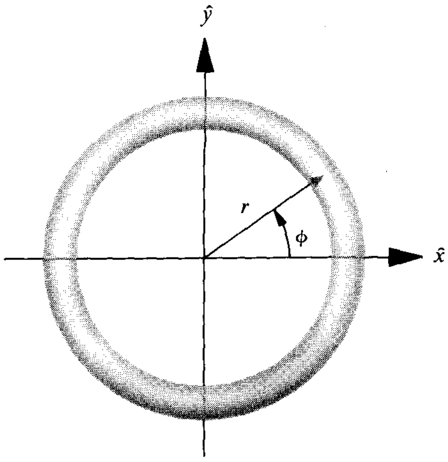


• $\omega t = \pi/2$



[EM waves propagate counterclockwise; gravitational line-of-force diagram rotates counterclockwise; they stay in phase.]

Box 37.6 (*continued*)



the deflective guidance of the wave guide. Thus, their wave vectors $\boldsymbol{k}$ satisfy

$$\nabla_{\boldsymbol{k}}\boldsymbol{k} = \begin{pmatrix}\text{deflective "acceleration"}\\ \text{of waveguide}\end{pmatrix}; \qquad (3)$$

and the difference $\delta\boldsymbol{k} = \nabla_{\boldsymbol{n}}\boldsymbol{k}$ between the wave vectors of the two parts of the wave (difference measured via parallel transport) satisfies the equation

$$\begin{aligned}\nabla_{\boldsymbol{k}}\delta\boldsymbol{k} &= \nabla_{\boldsymbol{k}}\nabla_{\boldsymbol{n}}\boldsymbol{k} = [\nabla_{\boldsymbol{k}}, \nabla_{\boldsymbol{n}}]\boldsymbol{k} + \nabla_{\boldsymbol{n}}\nabla_{\boldsymbol{k}}\boldsymbol{k} \qquad (4)\\ &= \boldsymbol{Riemann}\,(\ldots, \boldsymbol{k}, \boldsymbol{k}, \boldsymbol{n}) + \nabla_{\boldsymbol{n}}\underbrace{\nabla_{\boldsymbol{k}}\boldsymbol{k}}\end{aligned}$$

[deflective acceleration of wave guide]

The waveguide influences the direction of propagation of the waves, but not their frequency. Thus only ***Riemann*** enters into the 0 component of the above equation:

$$k^{\hat{\alpha}}\delta k^{\hat{0}}{}_{,\hat{\alpha}} = R^{\hat{0}}{}_{\hat{\alpha}\hat{\beta}\hat{\gamma}}k^{\hat{\alpha}}k^{\hat{\beta}}n^{\hat{\gamma}}. \qquad (5)$$

4. Let $k^{\hat{0}} = \omega_e$ be the angular frequency of the electromagnetic wave. The direction of the space component $\boldsymbol{k}$ of the propagation 4-vector is along the purely spatial vector $\boldsymbol{n}$; so

$$k^{\hat{0}} = \omega_e, \qquad \boldsymbol{k} = (v\omega_e/r\delta\alpha)\boldsymbol{n}, \qquad n^{\hat{0}} = 0. \qquad (6)$$

Use these relations to rewrite equation (5) as

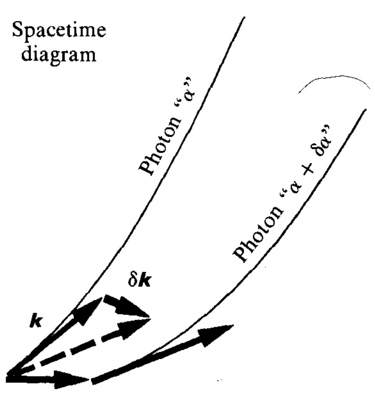
$$\begin{aligned}&(d\,\delta\omega_e/d\hat{t})_{\text{moving with photons}}\\ &\qquad = (v\omega_e/r\delta\alpha)R^{\hat{0}}{}_{\hat{i}\hat{0}\hat{j}}n^{\hat{i}}n^{\hat{j}}. \qquad (7)\end{aligned}$$

5. Combine the expression for $\boldsymbol{n}$ in the spacetime diagram with equations (2) and (7), and with the world line $\phi = \alpha + \frac{1}{2}\omega t$ for the photons, to obtain

$$\begin{aligned}&(d\,\delta\omega_e/d\hat{t})_{\text{moving with photons}}\\ &\qquad = -\frac{1}{2}v\omega_e\omega^2\mathscr{A}r\,(\cos 2\alpha)\,\delta\alpha. \qquad (8)\end{aligned}$$

6. Integrate over time and over α to obtain

$$\omega_e = \omega_{eo}\left[1 - \frac{1}{4}\mathscr{A}v\,(\sin 2\alpha)(\omega r)(\omega\hat{t})\right]. \qquad (9)$$



$\boldsymbol{n} = \begin{pmatrix}\text{vector connecting}\\ \text{photons}\end{pmatrix}$

$= r\,\delta\alpha\,(-\sin\phi\,\boldsymbol{e}_x + \cos\phi\,\boldsymbol{e}_y)$

PART IX

EXPERIMENTAL TESTS OF GENERAL RELATIVITY

Wherein the reader is tempted by a harem of charming gravitation theories (and some not so charming), is saved from his foolish passions by an army of experiments, cleaves unto his faithful spouse, Geometrodynamics, vows to lead an honest life hereafter, and becomes a True Believer.

CHAPTER 38

TESTING THE FOUNDATIONS OF RELATIVITY

Provando e riprovando
(Verify the one and disprove the other)
GALILEO

§38.1. TESTING IS EASIER IN THE SOLAR SYSTEM THAN IN REMOTE SPACE

For the first half-century of its life, general relativity was a theorist's paradise, but an experimentalist's hell. No theory was thought more beautiful, and none was more difficult to test.

The situation has changed. In the last few years general relativity has become one of the most exciting and fruitful branches of experimental physics. A half-century late, the march of technology has finally caught up with Einstein's genius—not only on the astronomical front, but also in laboratory experiments.

Clean tests of general relativity are currently confined to solar system

On the astronomical front, observers search for phenomena in which relativity is important, and study them: cosmology, pulsars, quasars, gravitational waves, black holes. Unfortunately, in pulsars and quasars, and in the sources of cosmological radiation and gravity waves, gravitational effects are tightly interwoven with the local hydrodynamics and local plasma physics. There is little hope of separating the several effects sufficiently sharply to get *clean* tests of the nature of gravity. Instead, astrophysicists must put the laws of gravity into their calculations along with all the other laws of physics and the observational data; and they must then seek, as output, information about the doings of matter and fields "way out there."

Thus, for clean tests of general relativity one turns to the laboratory—but to a laboratory that is much larger today than formerly: a laboratory that includes the entire solar system.

Capabilities of technology in 1970's

In the solar system all relativistic effects are tiny. Nonetheless, some of them are measurable with a precision, in the 1970's, of one part in 1,000 of their whole magnitude or better (see Box 38.1).

§38.2. THEORETICAL FRAMEWORKS FOR ANALYZING TESTS OF GENERAL RELATIVITY

There are now possible many experiments for testing general relativity. But most of them are expensive; very expensive. They involve atomic clocks flown on space-

Box 38.1 TECHNOLOGY OF THE 1970's CONFRONTED WITH RELATIVISTIC PHENOMENA

Quantity to be measured	*Magnitude of relativistic effects*	*Precision of a one-day measurement in the early 1970's*
Angular separation of two sources on the sky	Solar deflection of starlight (1) if light ray grazes edge of Sun, $1''.75$ (2) if light ray comes in perpendicular to Earth-sun line, $0''.004$	(a) With optical telescope, $\sim 1''$ (b) Angular separation of two quasars with radio telescope (differential measurement from day to day, not absolute measurement) in 1970, $\sim 0''.1$ in mid 1970's, $\sim 0''.001$
Distance between two bodies in solar system	(a) Perihelion shift per Earth year (1) for Mercury, 120 km (2) for Mars, 15 km (b) Relativistic time delay for radio waves from Earth, past limb of sun, to Venus (one way), 1×10^{-4} sec $= 30$ km	(a) Separation of another planet (Mercury, Venus, Mars) from Earth, by bouncing radar signals off it, ~ 0.3 km (b) Separation of a radio transponder (on another planet or in a space craft) from Earth, by measuring round-trip radio travel time, $\sim 3 \times 10^{-8}$ sec $= 10$ m $= 0.01$ km
	(c) Periodic relativistic effects in Earth-moon separation (1) in general relativity, 100 cm (2) in Jordan-Brans-Dicke theory, 100 cm; $(840\text{ cm})/(2 + \omega)$	(c) Earth-moon separation by laser ranging, ~ 10 cm
Difference in lapse of proper time between two world lines in solar system	(a) Clock on Earth vs. clock in synchronous Earth orbit, $\Delta t/t \sim 6 \times 10^{-10}$ (b) Clock on Earth vs. clock in orbit about sun, $\Delta t/t \sim 10^{-8}$	Stability of a hydrogen maser clock, $\Delta t/t \sim 10^{-13}$ for t up to one year

craft; radar signals bounced off planets; radio beacons and transponders landed on planets or orbited about them; etc. Because of the expense, it is crucial to have as good a theoretical framework as possible for comparing the relative values of various experiments—and for proposing new ones, which might have been overlooked.

Theoretical frameworks for analyzing gravitation experiments:

Such a framework must lie outside general relativity. It must scrutinize the foundations of Einstein's theory. It must compare Einstein's theory with other viable theories of gravity to see which experiments can distinguish between them. It must be a "theory of theories."

(1) Dicke framework

At present, in 1973, there are two different frameworks in broad use. One, devised largely by Dicke (1964b),* assumes almost nothing about the nature of gravity. It is used to design and discuss experiments for testing, at a very fundamental level, the nature of spacetime and gravity. Within it, one asks such questions as: Do all bodies respond to gravity with the same acceleration? Is space locally isotropic in its intrinsic properties? What are the theoretical implications of local isotropy? What types of fields, if any, are associated with gravity: scalar fields, vector fields, tensor fields, affine fields? Although some of the experiments that tackle these questions will be discussed below, this book will not attempt a detailed exposition of the Dicke framework.

(2) PPN framework

The second framework in broad use is the "parametrized post-Newtonian (PPN) formalism." It has been developed to higher and higher levels of sophistication by Eddington (1922), Robertson (1962), Schiff (1962, 1967), Nordtvedt (1968b, 1969), Will (1971c), and Will and Nordtvedt (1972).

The PPN formalism is an approximation to general relativity, and also to a variety of other contemporary theories of gravity, called "metric theories." It is a good approximation whenever, as in the solar system, the sources of the field gravitate weakly ($|\Phi|/c^2 \ll 1$) and move slowly ($v^2/c^2 \ll 1$). The PPN formalism contains a set of ten parameters whose values differ from one theory to another. Solar-system experiments (measurements of perihelion shift, light deflection, etc.) can be regarded as attempts to measure some of these PPN parameters, and thereby to determine which metric theory of gravity is correct—general relativity, Brans-Dicke (1961)-Jordan (1959) theory, one of Bergmann's (1968) scalar-tensor theories, one of Nordstrøm's theories, Whitehead's (1922) theory, or something else. [For reviews of Nordstrøm and Whitehead, see Whitrow and Morduch (1965), Will (1971b), and Ni (1972). For a significant nonmetric theory, see Cartan (1920) and Trautman (1972).]

Chapter 39 will discuss the concept of a metric theory of gravity and will construct the PPN formalism; and then Chapter 40 will use the PPN formalism to analyze the systematics of the solar system, and to discuss a variety of past and future experiments that distinguish between various metric theories of gravity. But first, as a prelude to those topics, this chapter will examine experiments that test the foundations of general relativity—foundations on which most other metric theories also rest. For a more detailed discussion of most of these experiments, see Dicke (1964b).

*See Thorne and Will (1971), or Will (1972), for expositions of both frameworks and a comparison of them.

§38.3. TESTS OF THE PRINCIPLE OF THE UNIQUENESS OF FREE FALL: EÖTVÖS-DICKE EXPERIMENT

The rest of this chapter is Track 2.
No earlier Track-2 material is needed as preparation for it, but Chapter 7 (incompatibility of gravity and special relativity) will be helpful.
This chapter is not needed as preparation for any later chapter, but it will be helpful in Chapters 39 and 40 (other theories; PPN formalism; experimental tests).

One fundamental building block common to Einstein's theory of gravity and to almost all other modern theories is the principle of *"uniqueness of free fall"*:* *"The world line of a freely falling test body is independent of its composition or structure."* By *"test body"* is meant an electrically neutral body, small enough that (1) its self-gravitational energy, as calculated using standard Newtonian theory, can be neglected compared to its rest mass ($M/R \ll 1$), and (2) the coupling of its multipole moments to inhomogeneities of the gravitational field can be neglected.†

The uniqueness of free fall permits one to regard spacetime as filled with a set of curves, the test-body trajectories, which are unique aside from parametrization. Through each event, along each timelike or null direction in spacetime, there passes one and only one test-body trajectory. Describe these trajectories mathematically: that is a central imperative of any theory of gravity.

Eötvös-Dicke experiment to test uniqueness of free fall

When translated into Newtonian language, the uniqueness of free fall states that any two test bodies must fall with the same acceleration in a given external gravitational field. Experimental tests of this principle search for differences in acceleration from one body to another. The most precise experiments to date are of a type devised by Baron Lorand von Eötvös (Box 38.2), redesigned and pushed to much higher precision by the Princeton group of Robert H. Dicke (Box 38.3), and extended with modifications by the Moscow group of Vladimir B. Braginsky. (See Figure 1.6 and Box 1.2 for experimental details.)

These Eötvös-Dicke experiments are "null experiments." They balance the acceleration of one body against the acceleration of another, and look for tiny departures from equilibrium. The reason is simple. Null experiments typically have much higher precision than experiments measuring the value of a nonzero quantity.

Eötvös, Pekar, and Fekete (1922) checked to an accuracy of 5 parts in 10^9 that the Earth imparts the same acceleration to wood, platinum, copper, asbestos, water, magnalium (90% Al, 10% Mg), copper sulphate, and tallow. Renner (1935) checked, to 7 parts in 10^{10}, the Earth's acceleration of platinum, copper, bizmuth, brass, glass, ammonium fluoride, and an alloy of 30% Mg, 70% Cu. Dicke, and later Braginsky, chose to use the sun's gravitational acceleration rather than the Earth's, since the alternation in the direction of the sun's pull every 12 hours lends itself to amplification by resonance. (See Figure 1.6.) Roll, Krotkov, and Dicke (1964) reported an

---

*R. H. Dicke calls this principle "The weak equivalence principle." We prefer to avoid confusion with the equivalence principle (Chapter 16).

†In general relativity, one often uses an alternative definition of test body, which places no constraint on the self-gravitational energy [abandon condition (1) while retaining (2)]. Such a definition is preferable, in principle, because the theory of matter has not been developed sufficiently to decide whether (and no objective test has ever been proposed to decide whether), gravitational energy at the subnuclear scale is a small fraction, a large fraction, or the entirety of the rest mass. But for present purposes a definition constraining test bodies to have $M/R \ll 1$ is preferable for two reasons. First, most theories of gravity that currently "compete" with Einstein's (a) agree with the principle of uniqueness of free fall when the macroscopic, Newtonian, self-gravitational energy is neglected ($M/R \ll 1$), but (b) violate that principle when macroscopic, Newtonian self-gravitational energy is taken into account. See §40.9 for details. Second, the test bodies used in the Eötvös-Dicke experiment have M/R so small that their macroscopic, Newtonian, self-gravitational energies are, in fact, negligible ($M/R \sim E_{\text{grav}}/M \sim 10^{-27}$).

agreement of 1 part in 10^{11} between the sun's acceleration of aluminum and gold, while Braginsky and Panov (1971) reported agreement to 1 part in 10^{12} for aluminum and platinum.

Theoretical implications of Eötvös-Dicke experiment

From this agreement, one can infer the response of neutrons, protons, electrons, virtual electron-positron pairs, nuclear binding energy, and electrostatic energy to the sun's gravity. Gold is 60% neutrons, while aluminum is only 50% neutrons. Therefore even from the 1964 results one could conclude that neutrons and protons must have the same acceleration to within $[0.6 - 0.5 = 0.1]^{-1}$ parts in $10^{11} = 1$ part in 10^{10}. Similarly, electrons must accelerate the same as nucleons to 2 parts in 10^7; virtual pairs (being more abundant in gold than in aluminum) must accelerate the same to 1 part in 10^4; nuclear binding energy, to 1 part in 10^7; and electrostatic energy to 3 parts in 10^9.

This accuracy of testing gives one confidence in the principle of the *uniqueness of free fall.*

(*continued on page 1054*)

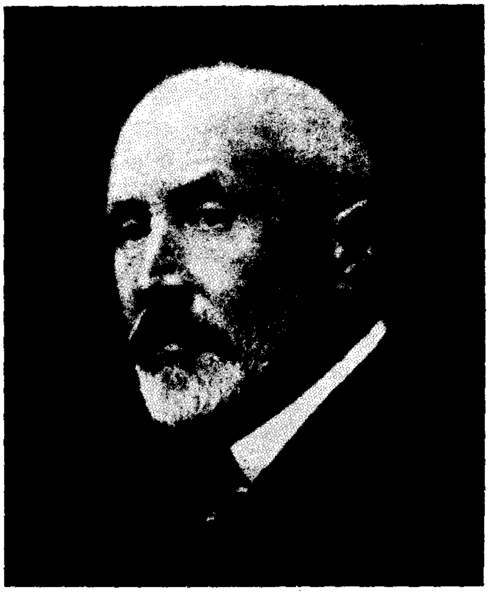
Box 38.2 BARON LORAND VON EÖTVÖS
Budapest, July 27, 1848—Budapest, April 8, 1919

Eötvös (pronounced ut'vûsh) studied at Heidelberg with Kirchhoff, Helmholtz, and Bunsen and at Königsberg with Neumann and Richelot. His 1870 Heidelberg Ph.D. thesis dealt with an issue of relativity: can the motion of a light source relative to an "ether" be detected by comparing the light intensities in the direction of the motion and in the opposite direction?

Studies of his at the same time resulted in the Eötvös law of capillarity, (surface tension) $\approx 2.12\,(T_{\rm crit} - T)/(\text{specific volume})^{2/3}$. Eötvös, made professor of physics at Budapest in 1872, concentrated on gravity from 1886 onward. He developed and extended the original Michell-Cavendish torsion balance, which measured not only $\Phi_{,xx}$ and $\Phi_{,xy}$ (where Φ is the gravitational potential) but also $\Phi_{,xz}$ and $\Phi_{,yz}$, all to a precision destined to be unexcelled for decades. He showed that the so-called "ratio between gravitational mass and inertial mass" cannot vary from material to material by more than 5 parts in 10^9. He investigated the paleomagnetism of bricks and other ceramic objects, and studied the shape of the earth. He served (June 1894—January 1895) as minister of public instruction and religious affairs (a cabinet position held in earlier years by his father). He founded a school which trained high-school teachers, to whose leavening influence one can give some of the credit for such outstanding scientists as von Karman, von Neuman, Teller, and Wigner. He served one year as rector of the University of Budapest.

> *"I can never forget the moment when my train rushed into the railroad station of Heidelberg along the banks of the Neckar. . . I cannot forget my happiness that now I could breathe the same air as those men of science whose fame attracted me there."*
>
> [EÖTVÖS IN 1887, AS QUOTED IN FEJÉR AND MIKOLA (1918), P. 259.]

Box 38.2 (*continued*)



Photograph by A. Szekely 1913

"Insofar as it is permitted on the basis of a few experiments, we can therefore declare that μ, that is, the weakening of the Earth's attraction through the intervening compensator quadrants, is less than one part in 5×10^{10}. . . . the absorption (of gravity) by the entire earth along a diameter is less than about one part in 800.

"We have carried out a series of observations which surpassed all previous ones in precision, but in no case could we discover any detectable deviation from the law of proportionality of gravitation and inertia."

[EÖTVÖS, PEKÁR, AND FEKETE (1922).]

"Science shall never find that formula by which its necessary character could be proved. Actually science itself might cease if we were to find the clue to the secret."

[EÖTVÖS, PRESIDENTIAL ADDRESS TO THE HUNGARIAN ACADEMY OF SCIENCES, 1890, AS QUOTED IN FEJÉR AND MIKOLA (1918), P. 280.]

"We should consider it as one of the most astonishing errors of the present age that so many people listen to the words of pseudoprophets who, in place of the dogmas of religion, offer scientific dogmas with medieval impatience but without historical justification."

[EÖTVÖS, 1877, AS QUOTED IN FEJÉR AND MIKOLA (1918), P. 280.]

2

Box 38.3 ROBERT HENRY DICKE May 6, 1916, St. Louis, Missouri
Cyrus Fogg Brackett Professor of Physics at Princeton University

During 1941-1946, Dicke was a leader in replacing the outmoded concept of lumped circuit elements by a new microwave analysis based on symmetry considerations, conservation laws, reciprocity relations, and the scattering matrix—concepts that led, among others, to the lock-in amplifier and the microwave radiometer. Searching for a means to reduce the Doppler width of spectral lines for precision measurements, Dicke discovered recoilless radiation in atomic systems held in a box or in a buffer gas. This development led to (1) the discovery of the basic idea of the gas-cell atomic clock and (2) a much more precise measurement of the gyromagnetic ratio of electrons in the 1s and 2s levels of hydrogen and of the hyperfine structure of atomic hydrogen.

A fundamental paper by Dicke in 1954 set forth the theory of coherent radiation processes and of the superradiant state, and laid the foundation for the future development of the laser and the maser, to which he also contributed. His patent no. 2,851,652 (filed May 21, 1956) was the first disclosure of a device for the generation of infrared radiation by a coherent process, and supplied the first suggestion for combining the use of an etalon resonator with an amplifying gas.

Beginning in the 1960's, Dicke brought his talent for precision measurement to the service of experimental cosmology, and with his collaborators: (1) checked the equivalence principle with the up-to-then unprecedented accuracy of 1 part in 10^{11}; (2) determined the solar oblateness; and (3) suggested that the primordial cosmic-fireball radiation, a tool for seeing deeper into the past history of the universe than has ever before been possible, should be observable, and therefore should be hunted down and found.

"For want of a better term, a gas which is radiating strongly because of coherence will be called 'superradiant.' . . . As the system radiates it passes to states of lower m *with* r *unchanged—to the 'superradiant' region* $m \sim 0$*"*

(1954)

"Possibilities are examined for the excitation of optical 'superradiant' states of gas"

(1957)

Box 38.3 (*continued*)

"A 'gravitational oblateness' of [the sun of] 5×10^{-5} would require the abandonment of Einstein's purely geometrical theory of gravitation. . . . Such a flattening [of the sun] could be understood as the effect of a rather rapidly rotating interior. . . . The answer appears to be that in the past, and to this day, the solar corona with its magnetic field has acted as a brake on the surface of the sun"

(1964a)

"New measurements of the solar oblateness have given a value for the fractional difference of equatorial and polar radii of $(5.0 \pm 0.7) \times 10^{-5}$"

[DICKE AND GOLDENBERG (1967)]

"[The universe must] have aged sufficiently for there to exist elements other than hydrogen. It is well-known that carbon is required to make physicists"

(1961)

"The question of the constancy of such dimensionless numbers is to be settled not by definition but by measurements"

[BRANS AND DICKE (1961)]

"The geophysical data lead to an upper limit of 3 parts in 10^{13} per year on the rate of change of the fine-structure constant"

[DICKE AND PEEBLES (1962)]

§38.4. TESTS FOR THE EXISTENCE OF A METRIC GOVERNING LENGTH AND TIME MEASUREMENTS, AND PARTICLE KINEMATICS

Special relativity, general relativity, and all other metric theories of gravity assume the existence of a metric field and predict that this field determines the rates of ticking of atomic clocks and the lengths of laboratory rods by the familiar relation $-d\tau^2 = ds^2 = g_{\alpha\beta}\, dx^\alpha\, dx^\beta$.

Experimental evidence for existence of a metric

The experimental evidence for a metric comes largely from elementary particle physics. It is of two types: *first,* experiments that measure time intervals directly, e.g., measurements of the time dilation of the decay times of unstable particles;* *second,* experiments that reveal the fundamental role played by the Lorentz group in particle kinematics and elsewhere in particle physics.† To cast out the metric tensor entirely would leave one with no theoretical framework adequate for interpreting such experiments.

*For a 2 per cent test of time dilation with muons of $(1 - v^2)^{-1/2} \sim 12$ in a storage ring, see Farley, Bailey, Brown, Giesch, Jöstlein, van der Meer, Picasso, and Tannenbaum (1966). For earlier time-dilation experiments see Frisch and Smith (1963); Durbin, Loar, and Havens (1952); and Rossi and Hall (1941).

†See p. 18 of Lichtenberg (1965) for a discussion of Lorentz invariance, spin and statistics, the TCP theorem, and relevant experiments.

Notice what particle-physics experiments *do* and *do not* tell one about the metric tensor, $\boldsymbol{g}$. *First,* they *do not* guarantee that there exist global Lorentz frames, i.e., coordinate systems extending throughout all of spacetime, in which $g_{\alpha\beta} = \eta_{\alpha\beta}$. However, they *do* suggest that at each event $\mathscr{P}$ there exist orthonormal frames with $\boldsymbol{e}_{\hat{\alpha}}(\mathscr{P}) \cdot \boldsymbol{e}_{\hat{\beta}}(\mathscr{P}) = \eta_{\alpha\beta}$, which are related to each other by Lorentz transformations. These orthonormal frames provide one with a definition of the inner product between any two vectors at a given event—and, thereby, they define the metric field.

Particle experiments do *not* guarantee existence of global Lorentz frames, or geodesic motion for test particles

Second, particle experiments *do not* guarantee that freely falling particles move along geodesics of the metric field, i.e., along straight lines in local Lorentz frames. (Here, in §§38.4 and 38.5, the phrase "local Lorentz frame" means a "normal" coordinate system at an event $\mathscr{P}$, in which $g_{\alpha\beta}(\mathscr{P}) = \eta_{\alpha\beta}$ *and* $g_{\alpha\beta,\gamma}(\mathscr{P}) = 0$. The term "inertial frame" is avoided because no assertions are made, yet, about test-body motion.) In particular, one does not know from elementary-particle experiments whether the local Lorentz frames in the laboratory are freely falling (so they fly up from the center of the earth and then fall back with Newtonian acceleration $g = 980\ \text{cm/sec}^2$), whether they are forever at rest relative to the laboratory walls, or whether they undergo some other type of motion. All one is led to believe is that a metric determines the nature of the spacetime intervals ($d\tau^2 = -g_{\mu\nu}\, dx^\mu\, dx^\nu$) measured by atomic clocks, that the various local Lorentz frames in the laboratory therefore move with uniform velocity relative to each other (they are connected by Lorentz transformations), and that electric and magnetic fields and the energies and momenta of particles undergo Lorentz transformations in the passage from one local Lorentz frame to another.

Third, elementary particle experiments *do* suggest that the times measured by atomic clocks depend only on velocity, not on acceleration. The measured squared interval is $ds^2 = g_{\alpha\beta}\, dx^\alpha\, dx^\beta$, independently of acceleration (until the acceleration becomes so great it disturbs the structure of the clock; see §16.4 and Box 16.3). Equivalently, but more physically, the time interval measured by a clock moving with velocity v^j relative to a local Lorentz frame is

Particle experiments *do* suggest proper time is independent of acceleration

$$d\tau = (-\eta_{\alpha\beta}\, dx^\alpha\, dx^\beta)^{1/2} = [1 - (v^x)^2 - (v^y)^2 - (v^z)^2]^{1/2}\, dt, \tag{38.1}$$

independently of the clock's acceleration d^2x^j/dt^2. If this were not so, then particles moving in circular orbits in strong magnetic fields would exhibit different decay rates than freely moving particles—which they do not [Farley *et al.* (1966)].*

§38.5. TESTS OF GEODESIC MOTION: GRAVITATIONAL REDSHIFT EXPERIMENTS

The uniqueness of free fall, as tested by the Dicke-Eötvös experiments, implies that spacetime is filled with a family of preferred curves, the test-body trajectories. There

*The experiment of Farley *et al.* is a 2 percent check of acceleration-independence of the muon decay rate for energies $E/m = (1 - v^2)^{-1/2} \sim 12$ and for accelerations, as measured in the muon rest frame, of $a = 5 \times 10^{20}\ \text{cm/sec}^2 = 0.6\ \text{cm}^{-1}$.

is also another family of preferred curves, the *geodesics* of the metric $\boldsymbol{g}$. It is tempting to identify these geodesics with the test-body trajectories. Einstein's geometric theory of gravity makes this identification ("equivalence principle"). One might conceive of theories that reject this identification. What is the experimental evidence on this point?

In order to see what kinds of experiments are relevant, it is helpful to elucidate the physical significance of the geodesics.

Physical meaning of a comparison between test-body trajectories and geodesics of metric

A geodesic of $\boldsymbol{g}$ is most readily identified locally by the fact that it is a straight line in the local Lorentz frames. Put differently, a body's motion is unaccelerated as measured in a local Lorentz frame if and only if the body moves along a geodesic of $\boldsymbol{g}$. Hence, to determine whether test-body trajectories are geodesics, one must compare experimentally the motion of the spatial origin of a local Lorentz frame (as defined by atomic-clock readings) with the motion of a test body (material particle).

It is easy to study experimentally the motions of test bodies; relative to an earth-bound laboratory, they accelerate downward with $g = 980\ \text{cm/sec}^2$; and this acceleration can be measured at a given location on the Earth to a precision of 1 part in 10^6.

Unfortunately, it is much more difficult to measure the motion of a local Lorentz frame, once again as defined by atomic-clock readings. The only direct experimental handle one has on this today, with sufficient precision to be interesting, is gravitational redshift experiments. (See §§7.2–7.5 and §25.4 for theoretical discussions of the gravitational redshift in the framework of general relativity.)

Pound-Rebka-Snider redshift experiment as a test of geodesic motion

The redshift experiment of highest precision is that of Pound and Rebka (1960), as improved by Pound and Snider (1965); see Figure 38.1. It used the Mossbauer effect to measure the redshift of 14.4 keV gamma rays from Fe^{57}. The emitter and absorber of the gamma rays were placed at rest at the bottom and top of a tower at Harvard University, separated by a height $h = 74$ feet $= 22.5$ meters. The measured redshift agreed, to 1 percent precision, with the general relativistic prediction of

$$\Delta\lambda/\lambda = gh = 2.5 \times 10^{-15}. \tag{38.2}$$

This result tells one that the local Lorentz frames are not at rest relative to the Earth's surface; rather, they are accelerating downward with the same acceleration, g, as acts on a free particle (to within 1 percent precision). To arrive at this conclusion, one analyzes the experiment in the laboratory reference frame, where everything (the experimental apparatus, the Earth, the Earth's gravitational field) is static. Relative to the laboratory a local Lorentz frame, momentarily at rest, accelerates downward (horizontal accelerations being ruled out by symmetry) with some unknown acceleration a. Equivalently, the laboratory accelerates upward (in $+z$ direction) with acceleration a relative to the local Lorentz frame. Consequently, the spacetime metric in the laboratory frame has the standard form

$$ds^2 = -(1 + 2az)\, dt^2 + dx^2 + dy^2 + dz^2 + O(|x^j|^2)\, dx^\alpha\, dx^\beta, \tag{38.3}$$

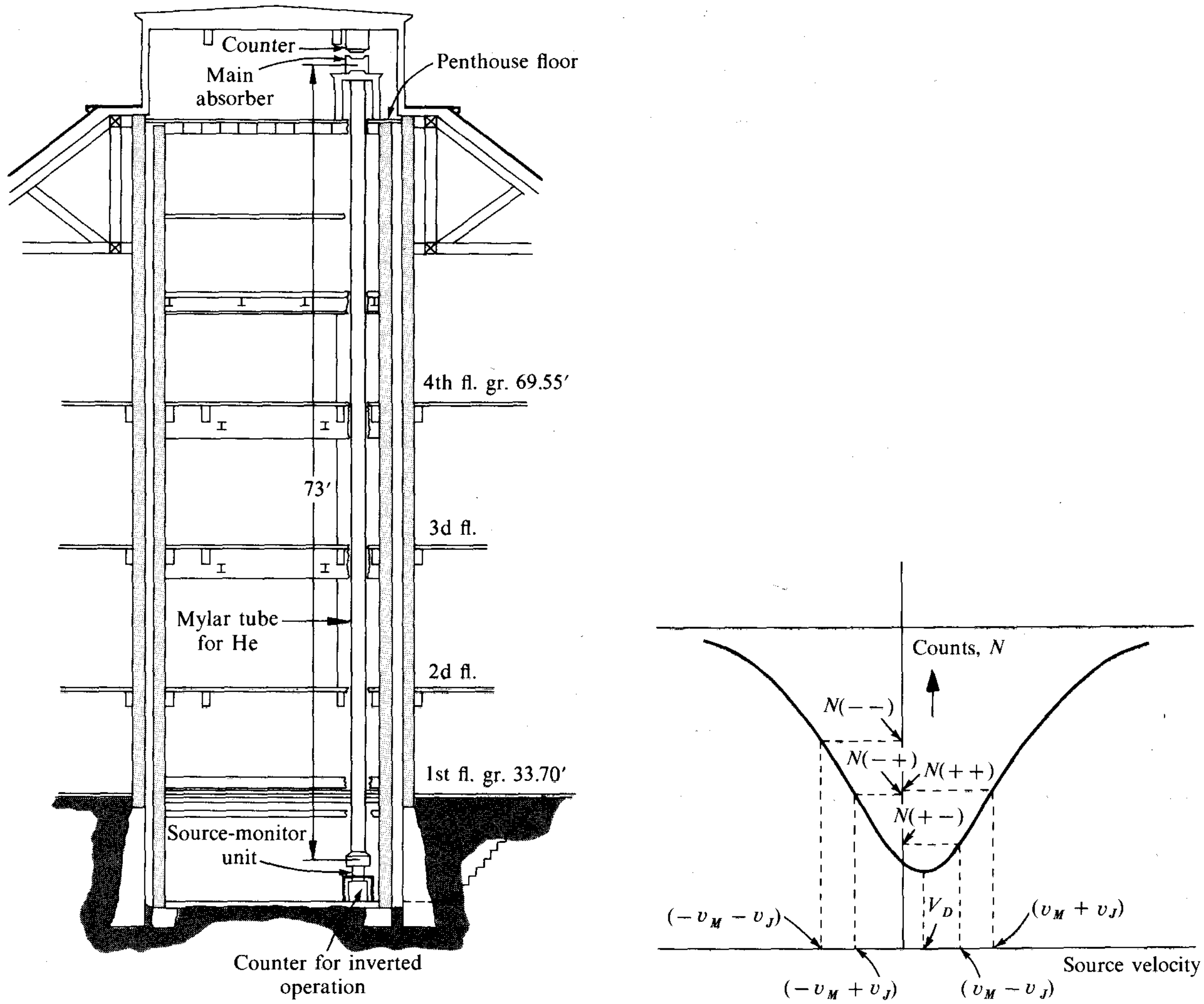


Figure 38.1.
The experiment of Pound and Rebka (1959) and Pound and Snider (1965) on the gravitational redshift of photons rising 22.5 meters against gravity through a helium-filled tube in a shaft in the Jefferson Physical Laboratory of Harvard University. The source of Co^{57} had an initial strength greater than a curie. The 14.4 keV gamma rays had to pass in through an absorber enriched in Fe^{57} to reach the large-window proportional counters. Both source and absorber were placed in temperature-regulated ovens. The velocity of the source consisted of two parts: one steady (v_M), to put the center of the emission line on the part of the transmission curve that is nearly straight; and the other alternating between $+v_J$ and $-v_J$, to sweep the transmission curve in this straight region; similarly when the steady velocity was $-v_M$. The departure from symmetry between the two cases $+v_M$ and $-v_M$ allows one to determine the offset v_D (effect of gravitational redshift) from the zero-gravity case of stationary emitter and stationary absorber. The final result for the redshift was (0.9990 ± 0.0076) times the value 4.905×10^{-15} of $2gh/c^2$ predicted from the principle of equivalence (difference between "up" experiment and "down" experiment). Diagrams adapted from Pound and Snider (1965).

which Track-2 readers have met in §§6.6 and 13.6; and Track-1 readers have met and used in Box 16.2. Moreover, in the laboratory frame the metric is static, gravity is static, and the experimental apparatus is static. Therefore the crest of each electromagnetic wave that climbs upward must follow a world line $t(z)$ identical in form to the world lines of the crests before and after it; thus,

$$\begin{aligned}
&\text{wave crest } \#0\text{: } t = t_0(z),\\
&\text{wave crest } \#1\text{: } t = t_0(z) + \Delta t,\\
&\qquad\vdots\\
&\text{wave crest } \#n\text{: } t = t_0(z) + n\,\Delta t.
\end{aligned}$$

[Here, as in Schild's argument (§7.3) that redshift implies spacetime curvature, no assumption is made about the form of the wave-crest world lines $t_0(z)$; see Figure 7.1.] Hence, expressed in *coordinate* time, the interval between reception of successive wave crests is the same as the interval between emission. Both are Δt. But the atomic clocks of the experiment (Fe^{57} nuclei) are assumed to measure proper time $\Delta\tau \equiv (-g_{\alpha\beta}\,\Delta x^\alpha\,\Delta x^\beta)^{1/2}$, not coordinate time. Thus

$$\frac{\lambda_{\text{received}}}{\lambda_{\text{emitted}}} = \frac{\Delta\tau_{\text{received}}}{\Delta\tau_{\text{emitted}}} = \frac{(1 + az_{\text{received}})\,\Delta t}{(1 + az_{\text{emitted}})\,\Delta t} = 1 + a(z_{\text{received}} - z_{\text{emitted}});$$

i.e.,

$$\frac{\Delta\lambda}{\lambda} = ah \quad \left[\begin{array}{l}\text{theoretical prediction based on assumptions}\\ \;(i)\ \text{that atomic clocks measure } \Delta\tau = (-g_{\alpha\beta}\Delta x^\alpha\,\Delta x^\beta)^{1/2};\\ (ii)\ \text{that electromagnetic radiation has the form of a}\\ \qquad\text{wave train;}\\ (iii)\ \text{that local Lorentz frames accelerate downward}\\ \qquad\text{with acceleration } a \text{ relative to the laboratory.}\end{array}\right] \tag{38.4}$$

Direct comparison with the experimental result (38.2) reveals that *local Lorentz frames in an Earthbound laboratory accelerate downward with the same acceleration g as acts on a test particle* (*to within 1 per cent precision*).

[The above discussion is basically a reworked version of Schild's proof (§7.2) that the redshift experiment implies spacetime is curved. After all, how could spacetime possibly be flat if Lorentz frames in Washington, Moscow, and Peking all accelerate toward the Earth's center with $g = 980\ \text{cm/sec}^2$?]

Other redshift experiments

Of all redshift experiments, the Pound-Rebka-Snider experiment is the easiest to interpret theoretically, because it was performed in a uniform gravitational field. Complementary to it is the experiment by Brault (1962), which measured the redshift of the sodium D_1 line emitted on the surface of the sun and received at Earth (Figure 38.2). To a precision of 5 per cent, he found a redshift of $GM_\odot/R_\odot c^2$, where $M_\odot$ and $R_\odot$ are the mass and radius of the sun. This is just the redshift to be expected if

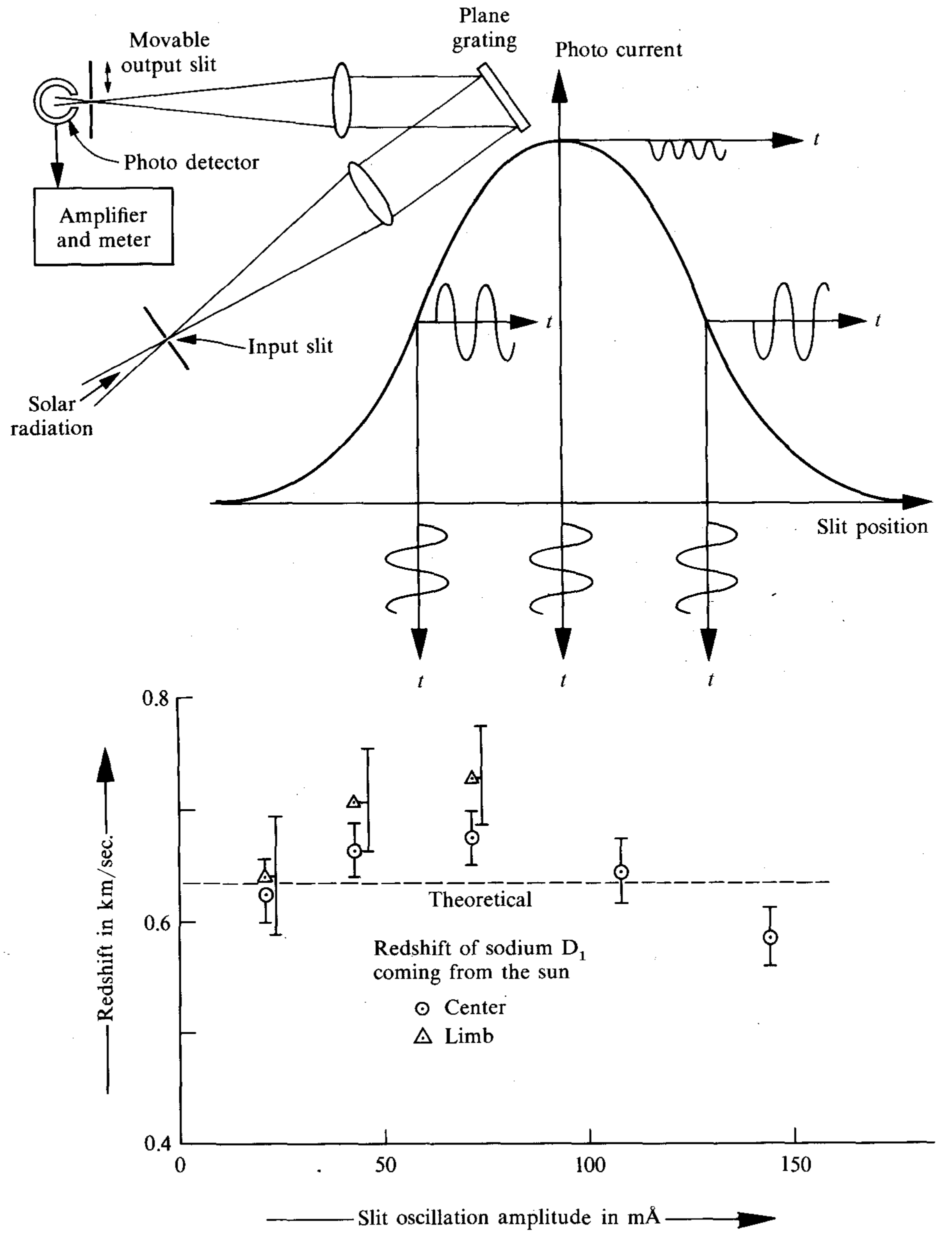


Figure 38.2.
The measurement by Brault (1962) of the redshift of the D_1 line of sodium gives 1.05 ± 0.05 of the gravitational redshift predicted by general relativity. This strong line, in contrast to the weak lines used by earlier investigators (1) is emitted high in the sun's atmosphere, above the regions strongly disturbed by the pressure and convective shifts, and yet lower than the chromosphere, and (2) comes closer to standing up cleanly above the background than any other line in the visible spectrum. Brault built a new photoelectric spectrometer (upper diagram), with its slit vibrated mechanically back and forth across a narrow region of the spectrum, to define the position of the line peak (1) electronically, (2) independently of subjective judgment, and (3) with a precision greater by a factor of the order of ten than that afforded by conventional visual methods. The slit is considered set on a line when its mean position is such that the photomultiplier current contains no signal at the frequency of the modulation. The redshift measured in this way is corrected for orbital motion and for rotation of the sun and the Earth to give the points in circles and triangles in the lower diagram. Extrapolation to zero vibration of the slit gives the cited number for the redshift. Figure adapted from thesis of Brault (1962).

the local Lorentz frames, at each point along the photon trajectory, fall in step with freely falling test bodies.*

In summary, redshift experiments reveal that, to a precision of several percent, the local Lorentz frames at the Earth's surface and near the sun are unaccelerated relative to freely falling test bodies. Equivalently, test bodies move along straight lines in the local Lorentz frames. Equivalently, *the test-body trajectories are geodesics of the metric* $\boldsymbol{g}$.

§38.6. TESTS OF THE EQUIVALENCE PRINCIPLE

Tests of the equivalence principle:

Of all the principles at work in gravitation, none is more central than the equivalence principle. As enunciated in §16.2, it states: *"In any and every local Lorentz frame, anywhere and anytime in the universe, all the* (*nongravitational*) *laws of physics must take on their familiar special-relativistic forms."*

(1) geodesic motion

That test bodies move along straight lines in local Lorentz frames (geodesic motion) is one aspect of the equivalence principle. Other aspects are the universality of Maxwell's equations

$$F^{\alpha\beta}{}_{,\beta} = 4\pi J^{\alpha} \text{ and } F_{\alpha\beta,\gamma} + F_{\beta\gamma,\alpha} + F_{\gamma\alpha,\beta} = 0 \tag{38.5}$$

in all local Lorentz frames; the universality of the law of local energy-momentum conservation

$$T^{\alpha\beta}{}_{,\beta} = 0; \tag{38.6}$$

and the universality of the values of the dimensionless constants that enter into the local laws of physics:

$$\begin{gathered}\alpha_e \equiv \frac{e^2}{\hbar c} = \frac{1}{137.0360\ldots} = \begin{pmatrix}\text{electromagnetic fine-}\\ \text{structure constant}\end{pmatrix};\\ \frac{m_{\text{neutron}}}{m_{\text{proton}}} = 1.00138\ldots, \qquad \frac{m_{\text{electron}}}{m_{\text{proton}}} = \frac{1}{1836.12\ldots}, \qquad \text{etc.}\end{gathered} \tag{38.7}$$

(Attention here is confined to dimensionless constants, since only they are independent of one's arbitrary choice of units of measure.)

(2) physical laws are locally Lorentz-invariant

If one focuses attention on a given event and asks about invariance of the form of the physical laws [equations (38.5), (38.6), etc.] from one Lorentz frame to another, one is then in the province of special relativity. Here a multitude of experiments verify the equivalence principle (see §38.4).

(3) laws do not vary from event to event

If one asks about variations in the form of the laws from one event to another, one opens up a Pandora's box of possibilities that one hardly dares to contemplate. However, no experimental evidence has ever given the slightest warrant to consider any such "departure from democracy" in the action of the laws of physics. Moreover, astronomical observations provide strong evidence that the laws of physics are the

*For a review of other, less-precise redshift experiments, see Bertotti, Brill, and Krotkov (1962).

same in distant stellar systems as in the solar system, and the same in distant galaxies as in our own Galaxy. (See, in Box 29.5, Edwin Hubble's expressions of joy upon discovering this.)

(4) fundamental constants do not vary from event to event

Constancy of the dimensionless "constants" from event to event can be tested to high precision, if one assumes constancy of the physical laws. Dirac (1937, 1938), Teller (1948), Jordan (1955, 1959), Gamow (1967), and others have proposed that the fine-structure "constant" α_e might be a slowly varying scalar field, perhaps governed by a cosmological equation. However, rather stringent limits on such variations follow from data on the fine-structure splitting of the spectral lines of quasars and radio galaxies. For the quasar 3C 191 with redshift $z = 1.95$, Bahcall, Sargent, and Schmidt (1967) find $\alpha_e(\text{3C 191})/\alpha_e(\text{Earth}) = 0.97 \pm 0.5$. With a cosmological interpretation of the quasar redshift, this corresponds to a limit $(1/\alpha_e)(d\alpha_e/dt) \lesssim 1/10^{11}$ years. An even tighter limit has been obtained from radio-galaxy data, where there is no question about the interpretation of the redshift. Bahcall and Schmidt (1967) measured fine-structure splitting in five radio galaxies with $z \approx 0.20$, corresponding to an emission of light 2×10^9 years ago. They obtained $\alpha_e(z = 0.20)/\alpha_e(\text{Earth}) = 1.001 \pm 0.002$, which yields the limit $|(1/\alpha_e)(d\alpha_e/dt)| \lesssim 1/10^{12}$ years.

Dyson (1972) points out that comparison of the rate of beta decay of Re^{187} in times past (via osmium-rhenium abundance ratios in old ores) with the rate of beta-decay today provides a means to check on any possible variation of α_e with time more sensitive than redshift data and more sensitive than any changes in rates of alpha decay and fission between early times and now. He summarizes the available data on Re^{187} and arrives at the limit

$$|(1/\alpha_e)(d\alpha_e/dt)| \lesssim 5/10^{15} \text{ years.}$$

For further evidence of the constancy of the fundamental constants see Minkowski and Wilson (1956), Dicke (1959a,b), Dicke and Peebles (1962b).

Eötvös-type experiments as tests for spatial variation of fundamental constants

Spatial variations of α_e, $m_{\text{neutron}}/m_{\text{proton}}$, and other "constants" in the solar system can be sought by means of Eötvös-type experiments. The reasoning [by Dicke (1969)] leading from such experiments to limits on any spatial variation of the constants is indirect. It recalls the reasoning used in standard treatises on polar molecules to deduce the acceleration of a polarizable molecule pulled on by an inhomogeneous electric field. It proceeds as follows.

Suppose one of the dimensionless "constants," "α," depends on position. This will lead to a position-dependence of the total mass-energy of a laboratory test body. For example, if α_e depends on position, then the coulomb energy of an atomic nucleus will also ($E_{\text{coul}} \propto e^4 \propto \alpha_e^{\,2}$; $\delta M/E_{\text{coul}} = 2\,\delta\alpha_e/\alpha_e$). One can calculate the change in a test body's mass-energy when it is moved from x^μ to $x^\mu + \delta x^\mu$ by assuming no change at all in the body's structure during its displacement:

$$\delta M = (\partial M/\partial\alpha)_{\text{fixed structure}}(\partial\alpha/\partial x^\mu)\,\delta x^\mu. \tag{38.8}$$

After the displacement, a weakening of internal forces (due, e.g., to a decrease of α)

may cause a change in structure, but that change will be accompanied by a conversion of internal potential energy into internal kinetic energy, which conserves M.

Now consider the following thought experiment [an elaboration of the argument by which Einstein first derived the gravitational redshift (§7.2)]: Take n particles, each with mass-energy μ. Make the particles with a structure such that a negligible fraction of μ is associated with the "constant" of interest, α:

$$(1/\mu)(\partial\mu/\partial\alpha) = 0. \tag{38.9}$$

Place these particles at a height h in a (locally) uniform Newtonian field. Combine them together there, releasing binding energy $E_B(h)$, to form a composite body of mass

$$M = n\mu - E_B(h) \tag{38.10}$$

which depends in a significant manner on the "constant" α,

$$(1/M)(\partial M/\partial\alpha) \neq 0. \tag{38.11}$$

Lower this body, *and* the released binding energy tied up in a little bag, a distance δh. The total force acting is (in Newtonian language)

$$F = Ma + E_B(h)g. \tag{38.12}$$

Here g is acceleration experienced by the type of mass-energy that is independent of α when it is in free fall. In contrast, "free" fall of the assembled body M is not really free fall, because of the supplementary "polarization force" pulling on this object. Hence the assembled body in "free" fall experiences an acceleration, a, a little different from g. However, the mass that is accelerated is precisely M, and therefore the force required to produce this acceleration is given by the product Ma. The energy gained in lowering the body and the bag is

$$E(\text{down}) = F\,\delta h = Ma\,\delta h + E_B(h)g\,\delta h.$$

Put this energy in the bag.

At $h - \delta h$ use some of the energy from the bag to pull the body apart into its component particles. The energy required is

$$\begin{aligned} E_B(h - \delta h) &= n\mu - M(h - \delta h) = n\mu - M(h) + \frac{\partial M}{\partial\alpha}\frac{d\alpha}{dh}\,\delta h \\ &= E_B(h) + \frac{\partial M}{\partial\alpha}\frac{\partial\alpha}{\partial h}\,\delta h; \end{aligned}$$

so an energy

$$\begin{aligned} E_{\text{bag}} &= E_B(h) + E(\text{down}) - E_B(h - \delta h) \\ &= \left[Ma + E_B(h)g - \frac{\partial M}{\partial\alpha}\frac{d\alpha}{dh}\right]\delta h \end{aligned} \tag{38.13}$$

is left in the bag. Use this energy to raise the n particles and the bag back up to height h. *Assume total energy conservation,* so that there will be no extra energy and no deficit when the n particles and bag have returned to the original state back at height h. This means that E_{bag} must be precisely the right amount of energy to do the raising:

$$E_{\text{bag}} = n\mu g\,\delta h = [M + E_B(h)]g\,\delta h. \tag{38.14}$$

Combining expressions (38.13) and (38.14) for E_{bag}, discover that

$$a - g = \frac{1}{M}\frac{\partial M}{\partial \alpha}\frac{d\alpha}{dh}. \tag{38.15}$$

Thus, under the assumption of total energy conservation (no perpetual-motion machines!), a spatial dependence of a physical "constant" α will lead to the anomaly (38.15) in the acceleration of a body whose mass depends on α.

Coulomb energy, which is proportional to $\alpha_e{}^2$, amounts in a gold nucleus to 0.4 per cent of the mass, and to 0.1 per cent in an aluminum nucleus. Hence, a spatial variation in α_e should lead to a fractional difference in the gravitational accelerations of these two nuclei equal to

$$\left|\frac{a_{Au} - a_{Al}}{g}\right| \approx \frac{1}{g}\,2\,\frac{0.003}{\alpha_e}\frac{d\alpha_e}{dh} \lesssim 1 \times 10^{-11};$$

i.e.,

$$\frac{1}{\alpha_e}\left|\frac{d\alpha_e}{dh}\right| \lesssim 1 \times 10^{-9}\,g \approx 1 \times 10^{-9}\ \text{cm/sec}^2 = 1 \times 10^{-30}/\text{cm} \quad \text{at the Earth due to the sun.} \tag{38.16}$$

Here use is made of the limit (1×10^{-11}) from Dicke's experiment (§38.3), and the acceleration $g = 0.6\ \text{cm/sec}^2$ due to the sun at Earth.

Notice that this says the gradient of $\ln \alpha_e$ is less than 1×10^{-9} the gradient of the Newtonian potential!

§38.7. TESTS FOR THE EXISTENCE OF UNKNOWN LONG-RANGE FIELDS

Whether or not one accepts the assumption that test bodies move on geodesics of the metric, it remains conceivable that previously unknown long-range fields (fields with "$1/r$" fall-off at large distances) are somehow associated with gravity.

Possible existence of new long-range fields associated with gravity

If "new" long-range fields (not metric, not electromagnetic) do exist, waiting to be discovered, then there are two ways by which they could influence matter. First, they could *couple directly to matter,* producing, for example, slight deviations from geodesic motion (deviations smaller than the limits of §38.5), or slight dependences of masses of particles on position (dependences smaller than the limits of §38.6). Second (and harder to detect), they could *couple indirectly to matter* by being mere

Direct vs. indirect coupling

2

participants in field equations that determine the geometry of spacetime. This section will describe tests for direct-coupling effects. Theories with fields that couple indirectly will be described in Box 39.1, and tests for such fields will be discussed in Chapter 40.

Experimental limits on direct-coupling fields:

Dicke (1964b), using his framework for analyzing tests of gravitation theories (§38.2), has shown that several null experiments place stringent limits on unknown, direct-coupling, long-range fields.

(1) Hughes-Drever experiment

One of these experiments is the "Hughes-Drever Experiment" [Hughes, Robinson, and Beltran-Lopez (1960); Drever (1961)]. It can be thought of as a search for a symmetric second-rank tensor field $h_{\alpha\beta}$ that produces slight deviations of test-body trajectories from geodesics of the metric $g_{\alpha\beta}$. Unless one's experiments happen to be made in a region of spacetime where $h_{\alpha\beta}$ is a constant multiple of $g_{\alpha\beta}$ ("mere rescaling of all lengths and times by a constant factor"), this tensor field must produce anisotropies in the properties of spacetime—which, in turn, will cause anisotropies in the inertial mass of a nucleon, and in turn will cause in an atomic nucleus relative shifts of degenerate energy levels with different magnetic quantum numbers. The Hughes-Drever experiment places stringent limits on such shifts, and thereby on a possible tensor field $h_{\alpha\beta}$. To quote Dicke (1964, p. 186), "If two [tensor] fields are present with the one strongly anisotropic in a coordinate system chosen to make the other isotropic, the strength of [direct] coupling to one must be only of the order of 10^{-22} that of the other. . . . [Moreover], on the moving Earth with ever-changing velocity, anisotropy would be expected at some season." From the experiments of Hughes and Drever, then, one concludes that there is not the slightest evidence for the presence of a second tensor field. For further details see Dicke and Peebles (1962a).

(2) ether-drift experiments

Another series of experiments, called *"ether-drift experiments,"* places stringent limits on any unknown, long-range vector field that couples directly to mass-energy. One can imagine such a field of cosmological origin. Being cosmological, the 4-vector would most naturally be expected to point in the same direction as the 4-vector $\boldsymbol{u}$ of the "cosmological fluid" (identical with the time direction $\boldsymbol{e}_0$ of a frame in which the cosmic microwave radiation is isotropic). The 4-vector of the new field would then have spatial components in any other frame. In principle an observer could use them to discern his direction of motion and speed relative to the mean rest frame of the universe. The ether-drift experiments search for effects of such a field.

For example, the experiment of Turner and Hill (1964) searches for a dependence of clock rates on such a vector field, by examining the transverse Doppler shift as a function of direction for an emitter on the rim of a centrifuge and a receiver at its center (Figure 38.3). If there is any effect, it would most naturally be expected to have the form

$$\frac{\begin{pmatrix}\text{rate of clock moving relative}\\ \text{to universe with speed } \boldsymbol{\beta}\end{pmatrix}}{\begin{pmatrix}\text{rate of clock at rest}\\ \text{relative to universe}\end{pmatrix}} = 1 + \gamma\beta^2, \qquad \gamma \text{ a small constant.} \tag{38.17}$$

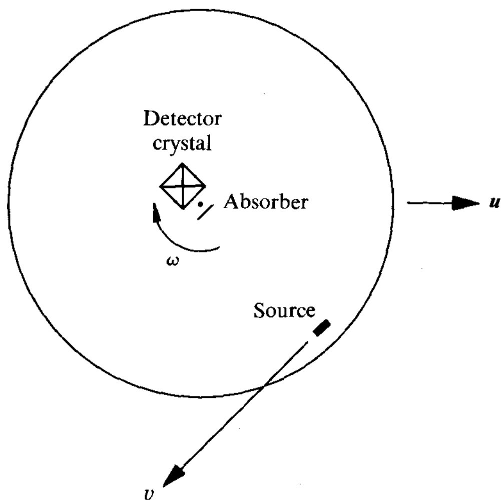


Figure 38.3.
The experiment of Turner and Hill (1964) looks for a dependence of proper clock rate (the clock being a Co^{57} source placed near the rim of the centrifuge) on velocity relative to the distant matter of the universe; or, in operational terms, relative to a "new local field" described by a 4-vector. The 14.4 keV gamma rays from the Co^{57} already experience a second-order Doppler shift of 1.3 parts in 10^{13}. One searches for an additional shift $\gamma\beta^2$ where $\boldsymbol{\beta} = \boldsymbol{u} + v(\boldsymbol{e}_x \cos \omega t + \boldsymbol{e}_y \sin \omega t)$ is the velocity relative to the frame in which the scalar field is purely timelike. The transmission of the gamma rays through the Fe^{57} absorber will drop linearly with any such additional shift, and will be noted as a drop in the counting rate of the NaI crystal. The source was 10 cm from the axis of rotation and the centrifuge turned at 15,000 rpm. The value of γ deduced from the experiment was $(1 \pm 4) \times 10^{-5}$.

A clock at the center of the centrifuge has $\boldsymbol{\beta} = \boldsymbol{u} = u\boldsymbol{e}_x$, whereas one on the rim has $\boldsymbol{\beta} = \boldsymbol{u} + v(\boldsymbol{e}_x \cos \omega t + \boldsymbol{e}_y \sin \omega t)$. Thus, the shift between rim and disk should vary with position

$$\Delta\lambda/\lambda = -\Delta\nu/\nu = -2\gamma uv \cos \omega t + \text{usual transverse shift.}$$

The data of Turner and Hill, using the Mössbauer effect, show that

$$|\gamma| < 4 \times 10^{-5}. \tag{38.18}$$

Hence, a cosmological vector field, if present, has only a weak direct coupling to matter.

For further discussion of these experiments and references on others like them, see Dicke (1964b).

1

CHAPTER 39

OTHER THEORIES OF GRAVITY AND THE POST-NEWTONIAN APPROXIMATION

§39.1. OTHER THEORIES

Among all bodies of physical law none has ever been found that is simpler or more beautiful than Einstein's geometric theory of gravity (Chapters 16 and 17); nor has any theory of gravity ever been discovered that is more compelling.

As experiment after experiment has been performed, and one theory of gravity after another has fallen by the wayside a victim of the observations, Einstein's theory has stood firm. No purported inconsistency between experiment and Einstein's laws of gravity has ever surmounted the test of time.

Role of alternative gravitation theories as foils for experimental tests

Query: Why then bother to examine alternative theories of gravity? *Reply:* To have "foils" against which to test Einstein's theory.

To say that Einstein's geometrodynamics is "battle-tested" is to say it has won every time it has been tried against a theory that makes a different prediction. How then does one select new antagonists for decisive new trials by combat?

Not all theories of gravity are created equal. Very few, among the multitude in the literature, are sufficiently viable to be worth comparison with general relativity or with future experiments. The "worthy" theories are those which satisfy *three criteria for viability: self-consistency, completeness, and agreement with past experiment.*

Criteria for viability of a theory:

(1) self-consistency

Self-consistency is best illustrated by describing several theories that fail this test. The classic example of an internally inconsistent theory is the spin-two field theory of gravity [Fierz and Pauli (1939); Box 7.1 here], which is equivalent to linearized general relativity (Chapter 18). The field equations of the spin-two theory imply that all gravitating bodies move along straight lines in global Lorentz reference frames, whereas the equations of motion of the theory insist that gravity deflects bodies away from straight-line motion. (When one tries to remedy this inconsistency, one finds oneself being "bootstrapped" up to general relativity; see route 5 of Box 17.2.) Another self-inconsistent theory is that of Kustaanheimo (1966). It predicts zero gravitational redshift when the wave version of light (Maxwell theory) is used, and nonzero redshift when the particle version (photon) is used.

(2) completeness

Completeness: To be complete a theory of gravity must be capable of analyzing from "first principles" the outcome of every experiment of interest. It must therefore mesh with and incorporate a consistent set of laws for electromagnetism, quantum mechanics, and all other physics. No theory is complete if it *postulates* that atomic clocks measure the "interval" $d\tau = (-g_{\alpha\beta}\, dx^{\alpha}\, dx^{\beta})^{1/2}$ constructed from a particular metric. Atomic clocks are complex systems whose behavior must be calculated from the fundamental laws of quantum theory and electromagnetism. No theory is complete if it *postulates* that planets move on geodesics. Planets are complex systems whose motion must be calculated from fundamental laws for the response of stressed matter to gravity. For further discussion see §§16.4, 20.6, and 40.9.

(3) agreement with past experiment

Agreement with past experiment: The necessity that a theory agree, to within several standard deviations, with the "four standard tests" (gravitational redshift, perihelion shift, electromagnetic-wave deflection, and radar time-delay) is obvious. Equally obvious but often forgotten is the need to agree with the expansion of the universe (historically the ace among all aces of general relativity) and with observations at the more everyday, Newtonian level. Example: Birkhoff's (1943) theory predicts the same redshift, perihelion shift, deflection, and time-delay as general relativity. But it requires that the pressure inside gravitating bodies equal the total density of mass-energy, $p = \rho$; and, as a consequence, it demands that sound waves travel with the speed of light. Of course, this prediction disagrees violently with experiment. Therefore, Birkhoff's theory is not viable. Another example: Whitehead's (1922) theory of gravity was long considered a viable alternative to Einstein's theory, because it makes exactly the same prediction as Einstein for the "four standard tests." Not until the work of Will (1971b) was it realized that Whitehead's theory predicts a time-dependence for the ebb and flow of ocean tides that is completely contradicted by everyday experience (see §40.8).

§39.2. METRIC THEORIES OF GRAVITY

Why attention focuses on metric theories of gravity

Two lines of argument narrow attention to a restricted class of gravitation theories, called *metric theories.*

The first line of argument constitutes the theme of the preceding chapter. It examined experiment after experiment, and reached two conclusions: (1) *spacetime possesses a metric; and* (2) *that metric satisfies the equivalence principle* (the standard special relativistic laws of physics are valid in each local Lorentz frame). *Theories of gravity that incorporate these two principles are called metric theories.** In brief, Chapter 38 says, "For any adequate description of gravity, look to a metric theory."

*For a slightly narrower definition of metric theories, see Thorne and Will (1971).

Exception: Cartan's (1922b, 1923) theory ["general relativity plus torsion"; see Trautman (1972)] is nonmetric, but agrees with experiment and is experimentally indistinguishable from general relativity with the technology of the 1970's.

The second line of argument pointing to metric theories begins with the issue of completeness (preceding section). To be complete, a theory must incorporate a self-consistent version of all the nongravitational laws of physics. No one has found a way to incorporate the rest of physics with ease except to introduce a metric, and then invoke the principle of equivalence. Other approaches lead to dismaying complexity, and usually to failure of the theory on one of the three counts of self-consistency, completeness, and agreement with past experiment. *All the theories known to be viable in 1973 are metric,* except Cartan's. [See Ni (1972b); Will (1972).]

How metric theories differ

In only one significant way do metric theories of gravity differ from each other: their laws for the generation of the metric. In general relativity theory, the metric is generated directly by the stress-energy of matter and of nongravitational fields. In Dicke-Brans-Jordan theory (Box 39.1, p. 1070), matter and nongravitational fields generate a scalar field ϕ; then ϕ acts together with the matter and other fields to generate the metric. Expressed in the language of §38.7, ϕ is a "new long-range field" that couples indirectly to matter. As another example, a theory devised by Ni (1970, 1972) (Box 39.1) possesses a flat-space metric $\boldsymbol{\eta}$ and a universal time coordinate t ("prior geometry"; see §17.6); $\boldsymbol{\eta}$ acts together with matter and nongravitational fields to generate a scalar field ϕ; and then $\boldsymbol{\eta}$, t, and ϕ combine to create the physical metric $\boldsymbol{g}$ that enters into the equivalence principle.

All three of the above theories—Einstein, Dicke-Brans-Jordan, Ni—were viable in the summer of 1971, when this section was written. But in autumn 1971 Ni's theory, and many other theories that had been regarded as viable, were proved by Nordtvedt and Will (1972) to disagree with experiment. This is an example of the rapidity of current progress in experimental tests of gravitation theory!

Henceforth, in this chapter and the next, attention will be confined to metric theories of gravity and their comparison with experiment.

§39.3. POST-NEWTONIAN LIMIT AND PPN FORMALISM

Weak-field, slow-motion expansion of a metric theory

The solar system, where experiments to distinguish between metric theories are performed, has weak gravity,

$$|\Phi| = |\text{Newtonian potential}| \lesssim 10^{-6}; \tag{39.1a}$$

moreover, the matter that generates solar-system gravity moves slowly

$$v^2 = (\text{velocity relative to solar-system center of mass})^2 \lesssim 10^{-7} \tag{39.1b}$$

and has small stress and internal energies

$$|T_{jk}|/\rho_o = (\text{stress divided by baryon "mass" density}) \lesssim 10^{-6}, \tag{39.1c}$$

$$\Pi = (\rho - \rho_o)/\rho_o = \begin{pmatrix}\text{internal energy density per}\\ \text{unit baryon "mass" density}\end{pmatrix} \lesssim 10^{-6}. \tag{39.1d}$$

[Here the *baryon "mass" density* ρ_o, despite its name, and despite the fact it is sometimes even more misleadingly called "density of rest mass-energy," is actually a measure of the number density of baryons n, and nothing more. It is defined as the product of n with some standard figure for the mass per baryon, μ_0, in some well-defined standard state; thus,

$$\rho_o \equiv n\mu_0.] \tag{39.1e}$$

Consequently, the analysis of solar-system experiments using any metric theory of gravity can be simplified, without significant loss of accuracy, by a simultaneous expansion in the small parameters $|\Phi|$, v^2, $|T_{jk}|/\rho_o$, and Π. Such a "weak-field, slow-motion expansion" gives: (1) flat, empty spacetime in "zero order"; (2) the Newtonian treatment of the solar system in "first order"; and (3) post-Newtonian corrections to the Newtonian treatment in "second order".

Post-Newtonian approximation

The formalism of Newtonian theory plus post-Newtonian corrections is called the "*post-Newtonian approximation.*" Each metric theory has its own post-Newtonian approximation. Despite the great differences between metric theories themselves, their post-Newtonian approximations are very similar. They are so similar, in fact, that one can construct a single post-Newtonian theory of gravity, devoid of any reference to indirectly coupling fields (ϕ in Dicke-Brans-Jordan; $\boldsymbol{\eta}$, t, and ϕ in Ni; see Box 39.1), that contains the post-Newtonian approximation of every conceivable metric theory as a special case. This all-inclusive post-Newtonian theory is called the "*Parametrized Post-Newtonian (PPN) Formalism.*" It contains a set of parameters (called "*PPN parameters*") that can be specified arbitrarily. One set of values for these parameters makes the PPN formalism identical to the post-Newtonian limit of general relativity; another set of values makes it the post-Newtonian limit of Dicke-Brans-Jordan theory, etc.

PPN formalism

Subsequent sections of this chapter present a version of the PPN formalism devised by Clifford M. Will and Kenneth Nordtvedt, Jr. (1972). [See also Will (1972).] This version, containing ten PPN parameters, encompasses as special cases nearly every metric theory of gravity known to the authors. The few exceptions [Whitehead (1922) and theories reviewed by Will (1973)] all disagree with experiment. One can include them in the PPN formalism by adding additional terms and parameters.

The ten parameters are described heuristically in Box 39.2, for the convenience of readers who would skip the full details of the formalism (§§39.4–39.12).

Accuracy of PPN formalism in solar system

How accurate is the PPN formalism? Or, stated more precisely, how accurately does the post-Newtonian approximation agree with the metric theory from which it comes? In the solar system, where $|\Phi|$, v^2, $|T_{jk}|/\rho_o$, and Π are all $\lesssim 10^{-6}$, the post-Newtonian approximation makes fractional errors of $\lesssim 10^{-6}$ in quantities of post-Newtonian order, and fractional errors of $\lesssim 10^{-12}$ in quantities of Newtonian order. For example, it misrepresents the deflection of light by $\lesssim 10^{-6} \times$ (post-Newtonian deflection) $\sim 10^{-6}$ seconds of arc. And it ignores relativistic deformations of the Earth's orbit of magnitude $< 10^{-12} \times$ (one astronomical unit) ~ 10 centimeters. Clearly, there is no need in the 1970's to use higher-order corrections to the post-Newtonian approximation; and hence no need to construct a "parametrized post-post-Newtonian framework." However, in the words of Shapiro (1971b): "If one projects from the achievements in the last decade, it is not unreasonable to predict

(*continued on page 1072*)

Box 39.1 THE THEORIES OF DICKE-BRANS-JORDAN AND OF NI

A. Dicke-Brans-Jordan

References: Brans and Dicke (1961); Jordan (1959). [*Notes:* This is the special case $\eta = -1$ of Jordan's theory. An alternative mathematical representation of the theory is given by Dicke (1962).]

Fields associated with gravity:

ϕ, a long-range scalar field;

$\boldsymbol{g}$, the metric of spacetime (from which are constructed the covariant derivative $\boldsymbol{\nabla}$ and the curvature tensors, in the usual manner).

Equations by which these fields are determined:

The trace of the stress-energy tensor generates ϕ via the curved-spacetime wave equation

$$\Box\phi = \phi^{,\alpha}{}_{;\alpha} = \frac{8\pi}{3+2\omega}T,$$

where ω is the dimensionless "Dicke coupling constant."

The stress-energy tensor and ϕ together generate the metric (i.e., the spacetime curvature) via the field equations

$$G_{\alpha\beta} = \frac{8\pi}{\phi}T_{\alpha\beta} + \frac{\omega}{\phi^2}\left(\phi_{,\alpha}\phi_{,\beta} - \frac{1}{2}g_{\alpha\beta}\phi_{,\mu}\phi^{,\mu}\right) + \frac{1}{\phi}(\phi_{;\alpha\beta} - g_{\alpha\beta}\Box\phi),$$

where $G_{\alpha\beta}$ is the Einstein tensor.

Variational principle for these equations:

$$\delta\int[\phi R - \omega(\phi_{,\alpha}\phi^{,\alpha}/\phi) + 16\pi L](-g)^{1/2}\,d^4x = 0,$$

where R is the scalar curvature and L is the matter Lagrangian.

Equivalence principle is satisfied:

The special-relativistic laws of physics are valid, without change, in the local Lorentz frames of the metric $\boldsymbol{g}$.

Consequence: the scalar field does not exert any direct influence on matter; its only role is that of participant in the field equations that determine the geometry of spacetime. It is an "indirectly coupling field" in the sense of §38.7.

This theory is self-consistent, complete, and for $\omega > 5$ in "reasonable" accord (two standard deviations or better) with all pre-1973 experiments.

B. Ni

References: Ni (1970, 1972)

Fields associated with gravity:

$\boldsymbol{\eta}$, a flat "background metric" ("prior geometry" in sense of §17.6). There exist, by assumption, coordinate systems ("background Lorentz frames") in which everywhere at once $\eta_{00} = -1$, $\eta_{0j} = 0$, and $\eta_{jk} = \delta_{jk}$.

t, a scalar field called the "universal time coordinate" ("prior geometry" in sense of §17.6), which is so "tuned" to the background metric that

$$t_{|\alpha\beta} = 0, \qquad t_{,\alpha}t_{,\beta}\eta^{\alpha\beta} = -1,$$

where "$_|$" denotes covariant derivative with respect to $\boldsymbol{\eta}$.

This means there exists a background Lorentz frame (the "rest frame of the universe") in which $x^0 = t$.

ϕ, a scalar field called the "scalar gravitational field".

$\boldsymbol{g}$, the metric of spacetime (from which are constructed the covariant derivative $\boldsymbol{\nabla}$ and the curvature tensors, in the usual manner).

Equations by which these fields are determined:

The stress-energy of spacetime generates the scalar gravitational field ϕ via the wave equation

$$\begin{aligned}\Box\phi \equiv \phi^{,\alpha}{}_{;\alpha} &= -2\pi T^{\alpha\beta}\,\partial g_{\alpha\beta}/\partial\phi \\ &= 4\pi T^{\alpha\beta}[\eta_{\alpha\beta}e^{-2\phi} + (e^{2\phi} + e^{-2\phi})t_{,\alpha}t_{,\beta}].\end{aligned}$$

ϕ, $\boldsymbol{\eta}$, and t together determine the metric of spacetime through the algebraic relation

$$\boldsymbol{g} = e^{-2\phi}\boldsymbol{\eta} + (e^{-2\phi} - e^{2\phi})\,\boldsymbol{d}t \otimes \boldsymbol{d}t.$$

Note: In the "rest frame of the universe" that is presupposed in this theory, this metric reduces to

$$ds^2 = g_{\alpha\beta}\,dx^\alpha\,dx^\beta = -e^{2\phi}\,dt^2 + e^{-2\phi}(dx^2 + dy^2 + dz^2).$$

Variational principle for the field equation for ϕ:

$$\delta\int(-2\phi^{,\alpha}\phi_{,\alpha} + 16\pi L)(-g)^{1/2}\,d^4x = 0,$$

where L is the matter Lagrangian.

Equivalence principle is satisfied:

The special-relativistic laws of physics are valid, without change, in the local Lorentz frames of the spacetime metric $\boldsymbol{g}$.

Consequence: ϕ, $\boldsymbol{\eta}$, and t do not exert any direct influence on matter; they are "indirectly coupling fields" in the sense of §38.7.

This theory is self-consistent and complete. If the solar system were at rest in the "rest frame of the universe", the theory would agree with all experiments to date—except, possibly, the expansion of the universe. But the motion of the solar system through the universe leads to serious disagreement with experiment (Will and Nordtvedt 1972; §40.8).

Box 39.2 HEURISTIC DESCRIPTION OF THE TEN PPN PARAMETERS

Parameter	*What it measures, relative to general relativity*[a]	*Value in General Relativity*	*Value in Dicke-Brans-Jordan Theory*[b]	*Value in Ni's Theory*[b]
γ	How much space curvature (g_{jk}) is produced by unit rest mass?	1	$\frac{1+\omega}{2+\omega}$	1
β	How much nonlinearity is there in the superposition law for gravity (g_{00})?	1	1	1
β_1	How much gravity (g_{00}) is produced by unit kinetic energy ($\frac{1}{2}\rho_o v^2$)?	1	$\frac{3+2\omega}{4+2\omega}$	1
β_2	How much gravity (g_{00}) is produced by unit gravitational potential energy ($\rho_o U$)?	1	$\frac{1+2\omega}{4+2\omega}$	1
β_3	How much gravity (g_{00}) is produced by unit internal energy ($\rho_o \Pi$)?	1	1	1
β_4	How much gravity (g_{00}) is produced by unit pressure (p)?	1	$\frac{1+\omega}{2+\omega}$	1
ζ	How much *more* gravity (g_{00}) is produced by radial kinetic energy $[\frac{1}{2}\rho_o(\boldsymbol{v}\cdot\hat{\boldsymbol{r}})^2]$—i.e., kinetic energy of motion toward observer—than by transverse kinetic energy?	0	0	0
η	How much *more* gravity (g_{00}) is produced by radial stress $[\hat{\boldsymbol{r}}\cdot\boldsymbol{t}\cdot\hat{\boldsymbol{r}}]$ than by transverse stress?	0	0	0
Δ_1	How much dragging of inertial frames (g_{0j}) is produced by unit momentum ($\rho_o v$)?	1	$\frac{10+7\omega}{14+7\omega}$	$-\frac{1}{7}$
Δ_2	How much easier is it for momentum ($\rho_o v$) to drag inertial frames radially (toward the observer) than in a transverse direction?	1	1	1

[a]These heuristic descriptions are based on equations (39.23).
[b]For expositions of these theories see Box 39.1. For derivation of their PPN values and of PPN values for other theories, see Ni (1972).

that in the 1980's techniques will be available to detect second-order effects of general relativity. At that point the ratio of theoretical to experimental relativists may take a sharp turn downwards."

Actually, there are a few exceptions to the claim that the post-Newtonian approximation suffices for the 1970's. These exceptions occur where the external universe impinges on and influences the solar system. For example, gravitational waves propagating into the solar system from distant sources (Chapters 35–37) are ignored by every post-Newtonian approximation and by the PPN framework. They must be treated using a full metric theory or a weak-field, "fast-motion" approximation to such a theory. Similarly, time-dependence of the "gravitational constant" (§40.8), induced in some theories by expansion of the universe, is beyond the scope of the PPN formalism, as is the expansion itself.

Applications of PPN formalism to astrophysics

The PPN formalism is used not only in interpreting experimental tests of gravitation theories, but also as a powerful tool in theoretical astrophysics. By specializing all the PPN parameters to unity, except $\zeta = \eta = 0$, one obtains the post-Newtonian approximation to Einstein's theory of gravity. This post-Newtonian approximation can then be used (and has been used extensively) to calculate general relativistic corrections to such phenomena as the structure and stability of stars.*

Historical and Notational Notes

History and notation of PPN formalism

The earliest parametrizations of the post-Newtonian approximation were performed, and used in interpreting solar system experiments, by Eddington (1922), Robertson (1962), and Schiff (1962, 1967). However, they dealt solely with the vacuum gravitational field outside an isolated, spherical body (the sun). Nordtvedt (1968b, 1969) devised the first full PPN formalism, capable of treating all aspects of the solar system; he treated the sun, planets, and moon as made from "gases" of point-particles (atoms) that interact gravitationally and electromagnetically. Will (1971c) later used techniques devised by Chandrasekhar (1965a) to modify Nordtvedt's formalism, so that it employs a stressed, continuous-matter description of celestial bodies. The version of the formalism presented here, devised by Will and Nordtvedt (1972), generalizes all previous versions to acquire "post-Galilean invariance" [see Chandrasekhar and Contopolous (1967)]. The most detailed and up-to-date review article on the PPN formalism is Will (1972).

In the literature of post-Newtonian physics and the PPN formalism, the Newtonian potential is described traditionally not by Φ, but by

$$U \equiv -\Phi \equiv +\int \frac{\rho_o(x')\, d^3x'}{|x - x'|}. \tag{39.2}$$

To avoid confusion, this chapter and the next will use U, although the rest of the book uses Φ.

Turn now to a detailed, Track-2 exposition of the PPN formalism.

§39.4. PPN COORDINATE SYSTEM

EXPOSITION OF PPN FORMALISM:

Coordinate system

The PPN formalism covers the solar system (or whatever system is being analyzed) with coordinates $(t, x_j) \equiv (t, x^j)$ that are as nearly globally Lorentz as possible:

$$g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta}, \qquad |h_{\alpha\beta}| \lesssim M_\odot/R_\odot \sim 10^{-6}. \tag{39.3}$$

*See, e.g., a long series of papers by Chandrasekhar and his associates in the Astrophysical Journal, beginning with Chandrasekhar (1965a,b,c).

2

The rest of this chapter is Track 2. No earlier Track-2 material is needed as preparation for it, but the following will be helpful:
(1) Chapter 7 (incompatibility of gravity and special relativity)
(2) §17.6 (no prior geometry);
(3) §§36.9–36.11 (generation of gravitational waves); and
(4) Chapter 38 (tests of foundations).
This chapter is not needed as preparation for any later chapter, but it will be helpful in Chapter 40 (solar-system tests)

(In this sense the PPN formalism is like linearized theory; see Chapter 18.) The velocity of the coordinate system (i.e., 4-velocity of its spatial origin) is so chosen that the solar system is approximately at rest in these coordinates. (Whether the center of mass of the solar system is precisely at rest, or is moving with some low velocity $v \lesssim (M_\odot/R_\odot)^{1/2} \sim 10^{-3} \sim 300$ km/sec, is a matter for the user of the formalism to decide. For more on the options, see §§39.9 and 39.12.)

The PPN coordinates provide one with a natural "3 + 1" split of spacetime into space plus time. That split is conveniently treated using the notation of three-dimensional, flat-space vector analysis—even though spacetime and the three-dimensional hypersurfaces $x^0 =$ constant are both curved. The resultant three-dimensional formalism will look more like Newtonian theory than like general relativity—as, indeed, one wishes it to; after all, one's goal is to study small relativistic corrections to Newtonian theory!

§39.5. DESCRIPTION OF THE MATTER IN THE SOLAR SYSTEM

Description of matter

Relative to the PPN coordinates, the matter of the solar system (idealized as a stressed medium) has a coordinate-velocity field

$$v_j \equiv dx_j/dx^0. \tag{39.4}$$

Choose an event $\mathscr{P}$, and in its neighborhood transform to an orthonormal frame that moves with the matter there. Orient the spatial axes $\boldsymbol{e}_{\hat{j}}$ of this comoving frame so that they coincide as accurately as possible with the PPN coordinate axes. (This requirement will be made more precise in §39.10.) In the orthonormal comoving frame, define the following quantities, which describe the state of the matter:

$$\text{(density of total mass-energy)} \equiv \rho; \tag{39.5a}$$

$$\begin{aligned}&\text{(baryon "mass" density)} \equiv \rho_o \\ &\equiv \begin{pmatrix}\text{number density} \\ \text{of baryons, } n\end{pmatrix} \times \begin{pmatrix}\text{standard rest mass per baryon, } \mu_0, \\ \text{for matter in some standard state}\end{pmatrix};\end{aligned} \tag{39.5b}$$

$$\text{(specific internal energy density)} \equiv \Pi \equiv (\rho - \rho_o)/\rho_o; \tag{39.5c}$$

$$\text{(components of stress tensor)} \equiv t_{\hat{i}\hat{j}} \equiv \boldsymbol{e}_{\hat{i}} \cdot \boldsymbol{T} \cdot \boldsymbol{e}_{\hat{j}}; \tag{39.5d}$$

$$\begin{aligned}\text{(pressure)} &\equiv p \equiv \frac{1}{3}(t_{\hat{x}\hat{x}} + t_{\hat{y}\hat{y}} + t_{\hat{z}\hat{z}}) \\ &\equiv \text{(average of stress over all directions).}\end{aligned} \tag{39.5e}$$

Anisotropies (i.e., shears) in the stress are important only in planets such as the Earth; and even there they are dominated by the isotropic pressure:

$$t_{\hat{i}\hat{j}} = p\,\delta_{ij} + p \times [\text{corrections} \ll 1]. \tag{39.6}$$

For many purposes, especially inside the sun, one can ignore the anisotropies, thereby approximating the solar-system matter as a perfect fluid.*

The isotropic part of the radiation field gives a significant contribution to the pressure, p, and the density of internal energy, $\rho_o \Pi$, inside the sun. However, the anisotropic radiation flux is ignored in the stress-energy tensor. This approximation is allowable because in the sun the outward energy flux carried by radiation is less than 10^{-15} of the internal energy density $\rho_o \Pi$; in planets it is even less.

§39.6. NATURE OF THE POST-NEWTONIAN EXPANSION

For any gravitationally bound configuration such as the solar system, the Newtonian approximation imposes limits on the sizes of various dimensionless physical quantities (see exercise 39.1):

Relative magnitudes of expansion parameters

$$\begin{aligned}\epsilon^2 &\equiv \text{maximum value of Newtonian potential } U \\ &\gtrsim \text{values anywhere of } U,\ v^2,\ p/\rho_o,\ |t_{\hat{i}\hat{j}}|/\rho_o,\ \Pi.\end{aligned} \tag{39.7}$$

(The Newtonian potential at the center of the sun is $\epsilon^2 \sim 10^{-5}$. The values of p/ρ_o, t_{ij}/ρ_o, and Π there are also $\sim 10^{-5}$, and they are much smaller elsewhere. The orbital velocities of the planets are all less than $100\ \text{km/sec} = 3 \times 10^{-4}$, so $v^2 < 10^{-7}$.) Moreover, changes with time of all quantities at fixed x_j are due primarily to the motion of the matter. As a result, time derivatives are small by $O(\epsilon)$ compared to space derivatives,

$$\left|\frac{\partial A/\partial t}{\partial A/\partial x_j}\right| \sim |v_j| \lesssim \epsilon \text{ for any quantity } A, \tag{39.8}$$

although *not* in the radiation zone, where outgoing gravitational waves flow (distance $\gtrsim$ one light year from Sun). Consequently, the radiation zone must be excluded from the analysis when one makes a post-Newtonian expansion. To treat it requires different techniques, e.g., those of Chapter 36.

Rules of the expansion

Conditions 39.7 and 39.8 suggest that one expand the metric coefficients in powers of the small parameter ϵ, treating U, v^2, p/ρ_o, t_{ij}/ρ_o, and Π as though they were all of $O(\epsilon^2)$ (often they are smaller!), and treating time derivatives as $O(\epsilon)$ smaller than space derivatives.

In this "post-Newtonian" expansion, terms odd in ϵ (i.e., terms such as

$$\int \frac{\rho_o(x', t) v_j(x', t)}{|x - x'|}\, d^3x' \sim \frac{M}{R} v \sim \epsilon^3 \tag{39.9}$$

whose total number of v's and $(\partial/\partial t)$'s is odd) change sign under time reversal,

*In the solar system, post-Newtonian corrections due to anisotropic stresses are so much smaller than other post-Newtonian corrections that there is no hope of measuring them in the 1970's. For this reason, elsewhere in the literature (but not in this book) the PPN formalism treats all stresses at the post-Newtonian level as isotropic pressures, thereby setting to zero the PPN parameter η of §§39.8–39.11.

whereas terms even in ϵ do not. Time reversal ($x^{\bar{0}} = -x^0$) also changes the sign of $g_{0j}(g_{\bar{0}j} = -g_{0j})$, but leaves g_{00} and g_{jk} unchanged. Therefore, *g_{0j} must contain only terms odd in ϵ; whereas g_{00} and g_{jk} must contain only even terms.* (Actually, this ceases to be the case when radiation damping enters the picture. In the real world one always insists on outgoing-wave boundary conditions. But time reversal converts outgoing waves to ingoing waves; so an extra sign change is required to convert back to out. Therefore, radiation damping terms in the near-zone metric are even in ϵ for g_{0j}, but odd for g_{00} and g_{jk}. However, radiation damping does not come into play until order ϵ^5 beyond Newtonian theory—see Chapter 36—so it will bc ignored here.)

Expanded form of metric

The form of the expansion is already known through Newtonian order (see §17.4, with Φ replaced by $-U$): Newtonian gravity is only obtained when one demands that

$$\begin{aligned} g_{00} &= -1 + 2U + [\text{terms} \lesssim \epsilon^4], \\ g_{0j} &= [\text{terms} \lesssim \epsilon^3], \\ g_{ij} &= \delta_{ij} + [\text{terms} \lesssim \epsilon^2]. \end{aligned} \tag{39.10}$$

The stated limits on the higher-order corrections are dictated by demanding that the space components of the geodesic equation agree with the Newtonian equation of motion:

$$\begin{aligned} \frac{d^2x_j}{dt^2} \approx \frac{d^2x_j}{d\tau^2} &= -\Gamma^j{}_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} \approx -\Gamma^j{}_{\alpha\beta}\frac{dx^\alpha}{dt}\frac{dx^\beta}{dt} \\ &= -\Gamma^j{}_{00} - 2\Gamma^j{}_{0k}v_k - \Gamma^j{}_{k\ell}v_k v_\ell \\ &= U_{,j} + \text{terms of order } \{\epsilon g_{0k,j};\ \epsilon^2 g_{k\ell,j}\}. \end{aligned} \tag{39.11}$$

One would get the wrong Newtonian limit if g_{0k} were $O(\epsilon)$ or greater, and if $g_{k\ell} - \delta_{k\ell}$ were $O(1)$ or greater.

The above pattern continues on to all orders in the expansion. Thus in the geodesic equation, and also in the law of local conservation of energy-momentum $T^{\alpha\beta}{}_{;\beta} = 0$, g_{00} always goes hand-in-hand with ϵg_{0k} and $\epsilon^2 g_{jk}$ (see exercise 39.2). Therefore, the post-Newtonian expansion has the form summarized in Box 39.3.

EXERCISES

Exercise 39.1. ORDERS OF MAGNITUDE IN GRAVITATIONALLY BOUND SYSTEMS

Use Newtonian theory to derive conditions (39.7) for any gravitationally bound system. [*Hint:* Such concepts as orbital velocities, the speeds of sound and shear waves, the virial theorem, and hydrostatic equilibrium are relevant.]

Exercise 39.2. PATTERN OF TERMS IN POST-NEWTONIAN EXPANSION

Verify the statements in the paragraph following equation (39.11). In particular, suppose that one wishes to evaluate the coordinate acceleration, d^2x_j/dt^2, to accuracy $\epsilon^{2N}U_{,j}$ for some integer N. Show that this undertaking requires a knowledge of g_{00} to accuracy ϵ^{2N+2}, of g_{0k} to ϵ^{2N+1}, and of g_{jk} to ϵ^{2N}. Also suppose that one knows T^{00} to accuracy $\rho_o\epsilon^{2N}$, T^{0j} to $\rho_o\epsilon^{2N+1}$, and T^{jk} to $\rho_o\epsilon^{2N+2}$ [see, e.g., equations (39.13) for $N = 0$ and (39.42) for $N = 2$]. Show that to calculate $T^{0\alpha}{}_{;\alpha}$ with accuracy $\epsilon^{2N+1}\rho_{o,j}$ and $T^{j\alpha}{}_{;\alpha}$ with accuracy $\epsilon^{2N+2}\rho_{o,j}$, one must know g_{00} to ϵ^{2N+2}, g_{0k} to ϵ^{2N+1}, and g_{jk} to ϵ^{2N}. This dictates the pattern of Box 39.3.

§39.7. NEWTONIAN APPROXIMATION

Newtonian approximation

At Newtonian order the metric has the form (39.10); and the 4-velocity and stress-energy tensor have components, relative to the PPN coordinate system,

$$u^0 = +1 + O(\epsilon^2), \qquad u^j = v_j + O(\epsilon^3); \tag{39.12}$$

$$\begin{aligned} T^{00} &= \rho_o + O(\rho_o\epsilon^2), \qquad T^{0j} = \rho_o v_j + O(\rho_o\epsilon^3), \\ T^{jk} &= t_{\hat{j}\hat{k}} + \rho_o v_j v_k + O(\rho_o\epsilon^4) \end{aligned} \tag{39.13}$$

(see exercise 39.3). Two sets of equations govern the structure and evolution of the solar system. (1) The Einstein field equations. As was shown in §18.4, and also in §17.4, in the Newtonian limit Einstein's equations reduce to Laplace's equation

$$U_{,jj} = -4\pi\rho_o, \tag{39.14a}$$

which has the "action-at-a-distance" solution

$$U(\boldsymbol{x}, t) = \int \frac{\rho_o(\boldsymbol{x}', t)}{|\boldsymbol{x} - \boldsymbol{x}'|}\, d^3x'. \tag{39.14b}$$

Box 39.3 POST-NEWTONIAN EXPANSION OF THE METRIC COEFFICIENTS

Level of approximation (and papers expanding general relativity to this level)	*Order or value of terms* g_{00}	g_{0j}	g_{jk}
flat, empty spacetime	-1	0	δ_{jk}
Newtonian approximation	$2U$	0	0
post-Newtonian approximation [(Fock (1959); Chandrasekhar (1965a)]	+ terms $\sim \epsilon^4$	+ terms $\sim \epsilon^3$	+ terms $\sim \epsilon^2$
post-post-Newtonian approximation [Chandrasekhar and Nutku (1969)]	+ terms $\sim \epsilon^6$	+ terms $\sim \epsilon^5$	+ terms $\sim \epsilon^4$
radiation damping [Chandrasekhar and Esposito (1970)]	+ terms $\sim \epsilon^7$	+ terms $\sim \epsilon^6$	+ terms $\sim \epsilon^5$

(2) The law of local energy-momentum conservation, $T^{\alpha\beta}{}_{;\beta} = 0$. The time component of this law reduces to the conservation of rest mass

$$\partial\rho_o/\partial t + \partial(\rho_o v_j)/\partial x_j = 0 + \text{fractional errors of } O(\epsilon^2); \tag{39.15a}$$

and the space components reduce to Newton's second law of motion, "$F = ma$":

$$\rho_o\, dv_j/dt = \rho_o(\partial U/\partial x_j) - \partial t_{\hat{j}\hat{k}}/\partial x_k + \text{fractional errors of } O(\epsilon^2), \tag{39.15b}$$

$$d/dt \equiv (\text{time derivative following the matter}) \equiv \partial/\partial t + v_k\, \partial/\partial x_k \tag{39.16}$$

(see exercise 39.3).

Equations (39.14)–(39.16), together with equations of state describing the planetary and solar matter, are the foundations for all Newtonian calculations of the structure and motion of the sun and planets. Notice that the internal energy density $\rho_o\Pi$ nowhere enters into these equations. It is of no importance to Newtonian hydrodynamics. It matters for the sun's thermal-energy balance; but that is irrelevant here.

EXERCISES

Exercise 39.3. NEWTONIAN APPROXIMATION

(a) Derive equations (39.13) for the components of the stress-energy tensor in the PPN coordinate frame. [*Hint:* In the rest frame of the matter ("comoving orthonormal frame") $T_{\hat{0}\hat{0}} = \rho = \rho_o + O(\epsilon^2)$, $T_{\hat{0}\hat{j}} = 0$, $T_{\hat{j}\hat{k}} = t_{\hat{j}\hat{k}}$; see equations (39.5). Lorentz-transform these components by a pure boost with ordinary velocity $-v_j$ to obtain $T_{\alpha\beta}$.]

(b) Show that, in the PPN coordinate frame, $T^{0\alpha}{}_{;\alpha} = 0$ reduces to equation (39.15a), and $T^{j\alpha}{}_{;\alpha} = 0$, when combined with (39.15a), reduces to equation (39.15b).]

Exercise 39.4. A USEFUL FORMULA

Derive from equations (39.15) the following useful formula, valid for any function $f(\boldsymbol{x}, t)$:

$$\frac{d}{dt}\int \rho_o(\boldsymbol{x}, t) f(\boldsymbol{x}, t)\, d^3x = \int \rho_o(\boldsymbol{x}, t)\frac{df(\boldsymbol{x}, t)}{dt}\, d^3x + \text{fractional errors of } O(\epsilon^2). \tag{39.17}$$

Here both integrals are extended over all of space; and df/dt is the derivative following the matter (39.16).

Exercise 39.5. STRESS TENSOR FOR NEWTONIAN GRAVITATIONAL FIELD

Define a "stress tensor for the Newtonian gravitational field U" as follows:

$$\mathrm{t}_{jk} \equiv \frac{1}{4\pi}\left(U_{,j}U_{,k} - \frac{1}{2}\delta_{jk}U_{,l}U_{,l}\right). \tag{39.18}$$

Show that the equations of motion for the matter (39.15b) can be rewritten in the forms

$$\rho_o\frac{dv_j}{dt} = -\frac{\partial}{\partial x^k}(\mathrm{t}_{jk} + t_{\hat{j}\hat{k}}) + \text{fractional errors of } O(\epsilon^2), \tag{39.19}$$

$$(\rho_o v_j)_{,t} + (\mathrm{t}_{jk} + t_{\hat{j}\hat{k}} + \rho_o v_j v_k)_{,k} = 0 + \text{fractional errors of } O(\epsilon^2). \tag{39.19'}$$

Exercise 39.6. NEWTONIAN VIRIAL THEOREMS

(a) From equation (39.19′) show that

$$d^2I_{jk}/dt^2 = 2\int (\mathrm{t}_{jk} + t_{\hat{j}\hat{k}} + \rho_o v_j v_k)\, d^3x + \text{fractional errors of } O(\epsilon^2), \qquad (39.20\text{a})$$

where I_{jk} is the second moment of the system's mass distribution,

$$I_{jk} = \int \rho_o x_j x_k \, d^3x.$$

This is called the "*time-dependent tensor virial theorem.*"

(b) From this infer that, if $\langle \quad \rangle_{\text{long time}}$ denotes an average over a long period of time, then

$$\left\langle \int (\mathrm{t}_{jk} + t_{\hat{j}\hat{k}} + \rho_o v_j v_k)\, d^3x \right\rangle_{\text{long time}} = O\left(\int \rho_o \epsilon^4 \, d^3x\right). \qquad (39.20\text{b})$$

This is called the "*tensor virial theorem.*"

(c) By contraction of indices and use of equations (39.18), (39.14a), and (39.5e), derive the (ordinary) *virial theorems:*

$$\frac{1}{2}\, d^2I/dt^2 = \int \rho_o v^2 \, d^3x - \int \frac{1}{2}\rho_o U \, d^3x + 3\int p\, d^3x + \text{fractional errors of } O(\epsilon^2), \qquad (39.21\text{a})$$

where I is the trace of the second moment of the mass distribution

$$I = I_{jj} = \int \rho_o r^2 \, d^3x;$$

and

$$\left\langle \underbrace{\int \rho_0 v^2 \, d^3x}_{2\times\begin{pmatrix}\text{kinetic}\\ \text{energy}\end{pmatrix}} - \underbrace{\frac{1}{2}\int \rho_0 U\, d^3x}_{+\begin{pmatrix}\text{gravitational}\\ \text{energy}\end{pmatrix}} + \underbrace{3\int p\, d^3x}_{+3\times\begin{pmatrix}\text{pressure}\\ \text{integral}\end{pmatrix}} \right\rangle_{\text{long time}} = O\left(\int \rho_0 \epsilon^4 \, d^3x\right). \qquad (39.21\text{b})$$

Exercise 39.7. PULSATION FREQUENCY FOR NEWTONIAN STAR

Use the ordinary, time-dependent virial theorem (39.21a) to derive the following equation for the fundamental angular frequency of pulsation of a nonrotating, Newtonian star:

$$\omega^2 = (3\bar{\Gamma}_1 - 4)\frac{|\text{star's self-gravitational energy}|}{(\text{trace of second moment of star's mass distribution})}; \qquad (39.22\text{a})$$

$$\bar{\Gamma}_1 = \begin{pmatrix}\text{pressure-weighted average}\\ \text{of adiabatic index}\end{pmatrix} \equiv \frac{\int \Gamma_1 p\, d^3x}{\int p\, d^3x}. \qquad (39.22\text{b})$$

In the derivation assume that the pulsations are "homologous"—i.e., that a fluid element with equilibrium position x^j (relative to center of mass $x^j = 0$) gets displaced to $x^j + \xi^j(x, t)$, where

$$\xi^j = (\text{small constant})x^j e^{-i\omega t}.$$

Assume nothing else. *Notes:* (1) The result (39.22) was derived differently in Box 26.2 and used in §24.4. (2) The assumption of homologous pulsation is fully justified if $|\Gamma_1 - 4/3| = \text{constant} \ll 1$; see Box 26.2. (3) The result (39.22) is readily generalized to slowly rotating Newtonian stars; see, e.g., Chandrasekhar and Lebovitz (1968). It can also be generalized to nonrotating post-Newtonian stars using general relativity (Box 26.2), or using the PPN formalism for any metric theory [Ni (1973)]. And it can be generalized to slowly rotating, post-Newtonian stars [see, e.g., Chandrasekhar and Lebovitz (1968)].

---

§39.8. PPN METRIC COEFFICIENTS

Post-Newtonian corrections to metric:

The post-Newtonian corrections $k_{\alpha\beta}$ to the metric coefficients $g_{\alpha\beta}$ are calculated, in any metric theory of gravity, by lengthy manipulations of the field equations. (See, e.g., exercise 39.14 near the end of this chapter for general relativity.) But without ever picking some one theory, and without ever writing down any set of field equations, one can infer the *forms* of the post-Newtonian corrections $k_{\alpha\beta}$. Their forms are fixed by the following constraints:

(1) rules governing forms

(1) They must be of post-Newtonian order ($k_{00} \sim \epsilon^4$, $k_{0j} \sim \epsilon^3$, $k_{ij} \sim \epsilon^2$). (2) They must be dimensionless. (3) k_{00} must be a scalar under rotations, k_{0j} must be components of a 3-vector, and k_{jk} must be components of a 3-tensor. (4) The corrections must die out at least as fast as $1/r$ far from the solar system, so that the coordinates become globally Lorentz and spacetime becomes flat at $r = \infty$. (5) For simplicity, one can assume that the metric components are generated only by ρ_o, $\rho_o \Pi$, $t_{\hat{i}\hat{j}}$, p, products of these with the velocity v_j, and time-derivatives of such quantities;* but not by their spatial gradients. [This assumption of simplicity is satisfied by all metric theories examined up to 1973, except Whitehead (1922) and theories reviewed by Will (1973)—which disagree with experiment.] Note the further justification for this assumption in exercise 39.8.

(2) construction of corrections

Begin with the corrections to the spatial components, $k_{ij} \sim \epsilon^2$. There are only two functionals of ρ_o, p, Π, $t_{\hat{j}\hat{k}}$, v_j, that die out at least as fast as $1/r$, are dimensionless, are $O(\epsilon^2)$, and are second-rank, symmetric 3-tensors; they are

$$\delta_{ij} U(\boldsymbol{x}, t); \quad U_{ij}(\boldsymbol{x}, t) = \int \frac{\rho_o(\boldsymbol{x}', t)(x_i - x_i')(x_j - x_j')}{|\boldsymbol{x} - \boldsymbol{x}'|^3}\, d^3x'. \tag{39.23a}$$

Thus, k_{ij} must be $k_{ij} = 2\gamma\, \delta_{ij} U + 2\Gamma U_{ij}$, for some constant "PPN parameters" γ and Γ. By an infinitesimal coordinate transformation [$x_{i\,\mathrm{NEW}} = x_{i\,\mathrm{OLD}} + \Gamma\, \partial\chi/\partial x_i$, with $\chi(\boldsymbol{x}, t) = -\int \rho_o(\boldsymbol{x}', t)|\boldsymbol{x} - \boldsymbol{x}'|\, d^3x'$] one can set $\Gamma = 0$, thereby obtaining

$$g_{ij} = \delta_{ij} + k_{ij} = \delta_{ij}(1 + 2\gamma U) + O(\epsilon^4). \tag{39.23b}$$

---

*One allows for time derivatives because retarded integrals contain such terms when expanded to post-Newtonian order; thus,

$$\int \frac{\rho_o(\boldsymbol{x}', t - |\boldsymbol{x} - \boldsymbol{x}'|)}{|\boldsymbol{x} - \boldsymbol{x}'|}\, d^3x' = \int \left[\frac{\rho_o(\boldsymbol{x}', t)}{|\boldsymbol{x} - \boldsymbol{x}'|} - \frac{\partial \rho_o(\boldsymbol{x}', t)}{\partial t} + \cdots\right] d^3x'.$$

However, it turns out that, with a suitable choice of coordinates ("gauge"), all time-odd retarded terms [e.g., $\int(\partial\rho_o/\partial t)\, d^3x$] vanish, except at "the post$^{5/2}$-Newtonian order" and at higher orders of approximation; there they lead to radiation damping (see Box 39.3). For example, $\int(\partial\rho_o/\partial t)\, d^3x = (d/dt)\int\rho_o\, d^3x$ vanishes by virtue of the conservation of baryon number.

Next consider $k_{0j} \sim \epsilon^3$. Trial and error yield only two vector functionals that die out as $1/r$ or faster, are dimensionless, and are $O(\epsilon^3)$. They are

$$V_j(\boldsymbol{x}, t) = \int \frac{\rho_o(\boldsymbol{x}', t) v_j(\boldsymbol{x}', t)}{|\boldsymbol{x} - \boldsymbol{x}'|} d^3x', \tag{39.23c}$$

$$W_j(\boldsymbol{x}, t) = \int \frac{\rho_o(\boldsymbol{x}', t)[(\boldsymbol{x} - \boldsymbol{x}') \cdot \boldsymbol{v}(\boldsymbol{x}', t)](x_j - x_j')\, d^3x'}{|\boldsymbol{x} - \boldsymbol{x}'|^3}. \tag{39.23d}$$

Thus, k_{0j} must be a linear combination of these, involving unknown constants (PPN parameters) Δ_1 and Δ_2:

$$g_{0j} = k_{0j} = -\frac{7}{2}\Delta_1 V_j - \frac{1}{2}\Delta_2 W_j + O(\epsilon^5). \tag{39.23e}$$

Finally consider $k_{00} \sim \epsilon^4$. Trial and error yields a variety of terms, which can all be combined together with the Newtonian part of g_{00} to give

$$g_{00} = -1 + 2U + k_{00} = -1 + 2U - 2\beta U^2 + 4\Psi - \zeta\mathscr{A} - \eta\mathscr{D}, \tag{39.23f}$$

where

$$\begin{aligned} \Psi(\boldsymbol{x}, t) &= \int \frac{\rho_o(\boldsymbol{x}', t)\psi(\boldsymbol{x}', t)}{|\boldsymbol{x} - \boldsymbol{x}'|} d^3x', \\ \psi &= \beta_1 v^2 + \beta_2 U + \frac{1}{2}\beta_3 \Pi + \frac{3}{2}\beta_4 p/\rho_o, \end{aligned} \tag{39.23g}^*$$

$$\mathscr{A}(\boldsymbol{x}, t) = \int \frac{\rho_o(\boldsymbol{x}', t)[(\boldsymbol{x} - \boldsymbol{x}') \cdot \boldsymbol{v}(\boldsymbol{x}', t)]^2}{|\boldsymbol{x} - \boldsymbol{x}'|^3} d^3x', \tag{39.23h}$$

$$\mathscr{D}(\boldsymbol{x}, t) = \int \frac{\left[t_{jk}(\boldsymbol{x}', t) - \frac{1}{3}\delta_{jk} t_{ll}(\boldsymbol{x}', t)\right](x_j - x_j')(x_k - x_k')}{|\boldsymbol{x} - \boldsymbol{x}'|^3} d^3x'. \tag{39.23i}$$

Also, β, β_1, β_2, β_3, β_4, ζ, η are unknown constants (PPN parameters). Elsewhere in the literature the term $-\eta\mathscr{D}$ in g_{00} is ignored (see footnote on p. 1075).

Yet another term is possible: one could have set

$$\begin{aligned} g_{00} = {} & [\text{value in equation (39.23f)}] \\ & - \Sigma \int\!\!\int \frac{\rho_o(\boldsymbol{x}', t)\rho_o(\boldsymbol{x}'', t)[(\boldsymbol{x} - \boldsymbol{x}') \cdot (\boldsymbol{x}' - \boldsymbol{x}'')]}{|\boldsymbol{x} - \boldsymbol{x}'|\,|\boldsymbol{x}' - \boldsymbol{x}''|^3} d^3x'\, d^3x'', \end{aligned} \tag{39.24}$$

where Σ is another PPN parameter. [It can be shown, using the Newtonian equations (39.14)–(39.16), that this expression dies out as $1/r$ far from the solar system.] If

*WARNING: *Throughout* the literature the notation Φ is used where we use Ψ for the functional (39.23g), and ϕ is used for our ψ. We are forced to violate the standard notation to avoid confusion with the Newtonian potential $\Phi = -U$. However, we urge that nobody else violate the standard notation!

2

such a Σ term had been included, then one could have removed it by making the infinitesimal coordinate transformation

$$x^0{}_{\text{new}} = x^0{}_{\text{old}} - \frac{1}{2}\Sigma\int\frac{\rho_o(x',t)[(x-x')\cdot v(x',t)]}{|x-x'|}d^3x' \qquad (39.25)$$

(see exercise 39.9). Thus, there is no necessity to include the Σ term.

Rigidity of coordinate system

The absence of the Σ term from g_{00} means that the time coordinate has been fixed rigidly up through post-Newtonian order:

$$x^0 \text{ has uncertainties only of } O(R_\odot\epsilon^5) \sim 10^{-14} \text{ seconds.} \qquad (39.26a)$$

The space coordinates are also fixed rigidly through post-Newtonian order:

$$x^j \text{ has uncertainties only of } O(R_\odot\epsilon^4) \sim 0.1 \text{ cm,} \qquad (39.26b)$$

because any transformation of the form

$$x^j{}_{\text{new}} = x^j{}_{\text{old}} + \text{position-dependent terms of } O(\epsilon^2 R_\odot)$$

would destroy the form (39.23b) of the space part of the metric.

Summary of PPN metric and parameters

In summary, for almost every metric theory of gravity yet invented, accurate through post-Newtonian order the metric coefficients have the form (39.23). One theory is distinguished from another by the values of its ten "post-Newtonian parameters" β, β_1, β_2, β_3, β_4, γ, ζ, η, Δ_1 and Δ_2. These are determined by comparing the field equations of the given theory with the form (39.23) of the post-Newtonian metric. The parameter values for general relativity and for several other theories are given in Box 39.2, along with a heuristic description of each parameter.

EXERCISES

Exercise 39.8. ABSENCE OF "METRIC-GENERATES-METRIC" TERMS IN POST-NEWTONIAN LIMIT

In writing down the post-Newtonian metric corrections, one might be tempted to include terms that are generated by the Newtonian potential acting alone, without any *direct* aid from the matter. After all, general relativity and other metric theories are nonlinear; so the two-step process (matter) $\longrightarrow$ U $\longrightarrow$ (post-Newtonian metric corrections) seems quite natural. *Show* that such terms are not needed, because the equations (39.14)–(39.16) of the Newtonian approximation enable one to reexpress them in terms of direct integrals over the matter distribution. In particular, show that

$$\int\frac{\partial^2 U(x',t)/\partial x_j'\,\partial t}{|x-x'|}d^3x' = 2\pi[V_j(x,t) - W_j(x,t)] \qquad (39.27)$$

where V_j and W_j are defined by equations (39.23c,d); also show that

$$\int\frac{[\partial U(x',t)/\partial x_j'][\partial U(x',t)/\partial x_j']}{|x-x'|}d^3x'$$
$$= -2\pi[U(x,t)]^2 + 4\pi\int\frac{\rho_o(x',t)U(x',t)}{|x-x'|}d^3x'. \qquad (39.28)$$

Note that the terms on the righthand sides of (39.27) and (39.28) are already included in the expressions (39.23e,f) for g_{0j} and g_{00}.

Exercise 39.9. REMOVAL OF Σ TERM FROM g_{00}

Show that the coordinate transformation (39.25) removes the Σ term from the metric coefficient g_{00} of equation (39.24), as claimed in the text.

Exercise 39.10. VERIFICATION OF FORMS OF POST-NEWTONIAN CORRECTIONS

Verify the claims in the text immediately preceding equations (39.23a,b,c,f).

---

§39.9. VELOCITY OF PPN COORDINATES RELATIVE TO "UNIVERSAL REST FRAME"

Thus far it has been assumed tacitly that the center of mass of the solar system is at rest in the PPN coordinate system. Is this really a permissible assumption? Put differently, can one always so adjust the PPN coordinate system that its origin moves with any desired velocity (e.g., that of the solar system); or is the PPN coordinate system rigidly and irrevocably attached to some "universal rest frame"?

In general relativity, the geometry of curved spacetime picks out no preferred coordinate frames (except in cases with special symmetry). Therefore, one expects the velocity of the PPN coordinate frame to be freely specifiable. Put differently, one expects the entire PPN formalism, for general relativity, to be invariant under Lorentz transformations of the PPN coordinates [combined, perhaps, with "infinitesimal coordinate transformations" to maintain the gauge conditions that the "Σ" and "U_{jk}" terms of (39.24) and (39.23a) be absent]. By contrast, in Ni's theory of gravity (Box 39.1) the geometry of spacetime *always* picks out a preferred coordinate frame: the "rest frame of the universe." One would not be surprised, in this case, to find the PPN coordinate frame rigidly attached to the universal rest frame.

Preferred-frame theories of gravity

The above intuition is correct, according to calculations by Will (1971d) and by Will and Nordtvedt (1972). When dealing with general relativity and other theories with little or no "prior geometry," one can freely specify the velocity of the PPN coordinate system (at some initial instant of time). But for theories like Ni's, with a preferred "universal rest frame" ("preferred-frame theories"), only in the preferred frame can the post-Newtonian metric assume the form derived in the last section [equations (39.23)]. This restriction on the PPN metric does not mean that one is confined, in preferred-frame theories, to perform all calculations in the universal rest frame. Rather, it means that for such theories the PPN metric requires generalization to take account of coordinate-frame motion relative to the universal rest frame.

Generalization of PPN metric to moving frames

The required generalization can be achieved by subjecting the PPN metric (39.23) to (1) a Lorentz boost from the preferred frame $\{x^{\alpha}_{\text{OLD}}\}$ to a new PPN frame $\{x^{\alpha}_{\text{NEW}}\}$, which moves with velocity $\boldsymbol{w}$, plus (2) a change of gauge designed to keep the metric coefficients as simple as possible. The boost-plus-gauge-change is [Will and Nordtvedt (1972)]

$$x_{\text{OLD}} = x_{\text{NEW}} + \frac{1}{2}(x_{\text{NEW}} \cdot w)w + \left(1 + \frac{1}{2}w^2\right)wt_{\text{NEW}} + O(\epsilon^5 t_{\text{NEW}} + \epsilon^4 x_{\text{NEW}}), \tag{39.29a}$$

$$t_{\text{OLD}} = t_{\text{NEW}}\left(1 + \frac{1}{2}w^2 + \frac{3}{8}w^4\right) + \left(1 + \frac{1}{2}w^2\right)x_{\text{NEW}} \cdot w + \underbrace{\left(\frac{1}{2}\Delta_2 + \zeta - 1\right)w_j\frac{\partial\chi}{\partial x^j_{\text{NEW}}}}_{\text{[gauge change]}} + O(\epsilon^6 t_{\text{NEW}} + \epsilon^5 x_{\text{NEW}}), \tag{39.29b}$$

$$\chi(t_{\text{NEW}}, x_{\text{NEW}}) \equiv -\int \rho_o(t_{\text{NEW}}, x'_{\text{NEW}})|x_{\text{NEW}} - x'_{\text{NEW}}|\, d^3x'_{\text{NEW}}. \tag{39.29c}$$

[*Note:* One insists, in the spirit of the post-Newtonian approximation, that the velocity w of the new PPN frame relative to the universal rest frame be no larger than the characteristic internal velocities of the system:

$$|w| \lesssim \epsilon.] \tag{39.30}$$

This change of coordinates produces corresponding changes in the velocity of the matter

$$v_{\text{OLD}} = \frac{dx_{\text{OLD}}}{dt_{\text{OLD}}} = v_{\text{NEW}}\left(1 - w \cdot v_{\text{NEW}} - \frac{1}{2}w^2\right) + w\left(1 - \frac{1}{2}w \cdot v_{\text{NEW}}\right) + O(\epsilon^5). \tag{39.31}$$

A long but straightforward calculation (exercise 39.11) yields the following components for the metric in the new PPN coordinates. [Note: The subscripts NEW are here and hereafter dropped from the notation.]

Final form of metric

$$g_{jk} = \delta_{jk}(1 + 2\gamma U) + O(\epsilon^4), \tag{39.32a}$$

$$g_{0j} = -\frac{7}{2}\Delta_1 V_j - \frac{1}{2}\Delta_2 W_j + \left(\alpha_2 - \frac{1}{2}\alpha_1\right)w_j U - \alpha_2 w_k U_{kj} + O(\epsilon^5), \tag{39.32b}$$

$$g_{00} = -1 + 2U - 2\beta U^2 + 4\Psi - \zeta\mathcal{A} - \eta\mathcal{D} + (\alpha_2 + \alpha_3 - \alpha_1)w^2 U + (2\alpha_3 - \alpha_1)w_j V_j - \alpha_2 w_j w_k U_{jk} + O(\epsilon^6). \tag{39.32c}$$

Here α_1, α_2, and α_3 are certain combinations of PPN parameters

$$\alpha_1 = 7\,\Delta_1 + \Delta_2 - 4\gamma - 4, \tag{39.33a}$$

$$\alpha_2 = \Delta_2 + \zeta - 1, \tag{39.33b}$$

$$\alpha_3 = 4\beta_1 - 2\gamma - 2 - \zeta. \tag{39.33c}$$

The "gravitational potentials" U, V_j, W_j, Ψ, $\mathcal{A}$, and $\mathcal{D}$ appearing here are to be calculated in the new, "moving" PPN coordinate system by the same prescriptions as one used in the universal rest frame. Thus, their functional forms are the same as previously, but their values at any given event are different (see exercise 39.11):

$$U(\boldsymbol{x}, t) = \int \frac{\rho_o(\boldsymbol{x}', t)}{|\boldsymbol{x} - \boldsymbol{x}'|}\, d^3x'; \tag{39.34a}$$

$$V_j(\boldsymbol{x}, t) = \int \frac{\rho_o(\boldsymbol{x}', t) v_j(\boldsymbol{x}', t)}{|\boldsymbol{x} - \boldsymbol{x}'|}\, d^3x'; \tag{39.34b}$$

$$W_j(\boldsymbol{x}, t) = \int \frac{\rho_o(\boldsymbol{x}', t)[(\boldsymbol{x} - \boldsymbol{x}') \cdot \boldsymbol{v}(\boldsymbol{x}', t)](x_j - x_j')\, d^3x'}{|\boldsymbol{x} - \boldsymbol{x}'|^3}; \tag{39.34c}$$

$$\Psi(\boldsymbol{x}, t) = \int \frac{\rho_o(\boldsymbol{x}', t)\psi(\boldsymbol{x}', t)}{|\boldsymbol{x} - \boldsymbol{x}'|}\, d^3x',$$

$$\psi = \beta_1 v^2 + \beta_2 U + \frac{1}{2}\beta_3 \Pi + \frac{3}{2}\beta_4 p/\rho_o; \tag{39.34d}$$

$$\mathscr{A}(\boldsymbol{x}, t) = \int \frac{\rho_o(\boldsymbol{x}', t)[(\boldsymbol{x} - \boldsymbol{x}') \cdot \boldsymbol{v}(\boldsymbol{x}', t)]^2}{|\boldsymbol{x} - \boldsymbol{x}'|^3}\, d^3x'; \tag{39.34e}$$

$$\mathscr{D}(\boldsymbol{x}, t) = \int \frac{\left[t_{\hat{j}\hat{k}}(\boldsymbol{x}', t) - \frac{1}{3}\delta_{jk} t_{\hat{l}\hat{l}}(\boldsymbol{x}', t)\right](x_j - x_j')(x_k - x_k')}{|\boldsymbol{x} - \boldsymbol{x}'|^3}\, d^3x'. \tag{39.34f}$$

The quantity U_{jk} is the gravitational potential defined in equation (39.23a):

$$U_{jk}(\boldsymbol{x}, t) = \int \frac{\rho_o(\boldsymbol{x}', t)(x_j - x_j')(x_k - x_k')}{|\boldsymbol{x} - \boldsymbol{x}'|^3}\, d^3x'. \tag{39.34g}$$

Notice that the velocity $\boldsymbol{w}$ of the PPN coordinate system relative to the universal rest frame appears explicitly in the PPN metric only if one or more of the coefficients $\alpha_1, \alpha_2, \alpha_3$, is nonzero. Thus, theories with $\alpha_1 = \alpha_2 = \alpha_3 = 0$ (e.g., general relativity) possess *no* preferred universal rest frame in the post-Newtonian limit; all their PPN frames are "created equal." By contrast, theories with at least one of $\alpha_1, \alpha_2, \alpha_3$, nonzero (e.g., Ni's theory) *do* possess a preferred frame.

Post-Galilean invariance

The generalized form (39.32) of the PPN metric, by virtue of the process used to construct it, is invariant under a Lorentz boost plus a gauge adjustment ["Post-Galilean transformation"; see Chandrasekhar and Contopolous (1967)]:

$$\begin{aligned} \boldsymbol{x}_{\rm OLD} &= \boldsymbol{x}_{\rm NEW} + \frac{1}{2}(\boldsymbol{x}_{\rm NEW} \cdot \boldsymbol{\beta})\boldsymbol{\beta} + \left(1 + \frac{1}{2}\beta^2\right)\boldsymbol{\beta} t_{\rm NEW} \\ &\quad + O(\epsilon^5 t_{\rm NEW} + \epsilon^4 \boldsymbol{x}_{\rm NEW}), \\ t_{\rm OLD} &= \left(1 + \frac{1}{2}\beta^2 + \frac{3}{8}\beta^4\right) t_{\rm NEW} + \left(1 + \frac{1}{2}\beta^2\right) \boldsymbol{x}_{\rm NEW} \cdot \boldsymbol{\beta} \\ &\quad + \left(\frac{1}{2}\Delta_2 + \zeta_1 - 1\right)\boldsymbol{\beta} \cdot \boldsymbol{\nabla}_{\rm NEW}\chi + O(\epsilon^6 t_{\rm NEW} + \epsilon^5 \boldsymbol{x}_{\rm NEW}). \end{aligned} \tag{39.35}$$

Of course, it is also invariant under spatial rotations.

EXERCISE

Exercise 39.11. TRANSFORMATION TO MOVING FRAME

Show that the change of coordinates (39.29) changes the PPN metric coefficients from the form (39.23) to the form (39.32). [*Hints:* (1) Keep firmly in mind the fact that the potentials U, V_j, W_j, $\mathscr{A}$, and Φ are not scalar fields. Each coordinate system possesses its own potentials. For example, by using equations (39.29) in the integral for U_{OLD}, one finds

$$U_{\text{OLD}}(x_{\text{OLD}}, t_{\text{OLD}}) = \int \frac{\rho_o(x'_{\text{OLD}}, t_{\text{OLD}})}{|x_{\text{OLD}} - x'_{\text{OLD}}|}\, d^3x'_{\text{OLD}} \tag{39.36}$$

$$= \left[U_{\text{NEW}} - w_j(V_{j\text{NEW}} - W_{j\text{NEW}}) + \frac{1}{2} w_j\, w_k X_{,jk} \right]_{x_{\text{NEW}}, t_{\text{NEW}}} + O(\epsilon^6).$$

(2) The law of baryon conservation (39.44) may be useful.]

§39.10. PPN STRESS-ENERGY TENSOR

The motion of the solar system is governed by the equations $T^{\alpha\beta}{}_{;\beta} = 0$. Before studying them, one must calculate the post-Newtonian corrections to the stress-energy tensor in the PPN coordinate frame. This requires a transformation from the comoving, orthonormal frame $\boldsymbol{\omega}^{\hat{\alpha}}$, where

$$T^{\hat{0}\hat{0}} = \rho_o(1 + \Pi), \qquad T^{\hat{0}\hat{j}} = 0, \qquad T^{\hat{j}\hat{k}} = t_{\hat{j}\hat{k}}, \tag{39.37}$$

to the coordinate frame. One can effect this transformation in two stages: stage 2 is a transformation

$$\boldsymbol{\omega}^{\tilde{0}} \equiv [1 - U + O(\epsilon^4)]\, \boldsymbol{d}t \tag{39.38a}$$

$$+ \left[\frac{7}{2}\Delta_1 V_j + \frac{1}{2}\Delta_2 W_j + \left(\frac{1}{2}\alpha_1 - \alpha_2\right) w_j U + \alpha_2 w_k U_{kj} + O(\epsilon^5) \right] \boldsymbol{d}x^j,$$

$$\boldsymbol{\omega}^{\tilde{j}} \equiv [(1 + \gamma U)\,\delta_{jk} + O(\epsilon^4)]\, \boldsymbol{d}x^k + O(\epsilon^5)\, \boldsymbol{d}t, \tag{39.38b}$$

between the coordinate frame and an orthonormal frame attached to it; stage 1 is a pure Lorentz transformation (boost) between the two orthonormal frames $\boldsymbol{\omega}^{\tilde{\alpha}}$ and $\boldsymbol{\omega}^{\hat{\alpha}}$. The 4-velocity of this boost is minus the 4-velocity of the matter, which has components

$$u^j = v_j u^0, \qquad u^0 = 1 + \frac{1}{2}v^2 + U + O(\epsilon^4) \qquad \text{in coord. frame;} \tag{39.39}$$

$$\left.\begin{aligned} u^{\tilde{j}} &= v_{\tilde{j}} u^{\tilde{0}}, \qquad u^{\tilde{0}} = 1 + \frac{1}{2}v^2 + O(\epsilon^4), \\ v_{\tilde{j}} &= v_j[1 + (1 + \gamma)U] \end{aligned}\right\} \text{ in } \boldsymbol{\omega}^{\tilde{\alpha}} \text{ frame.} \tag{39.40}$$

Combining the boost, which has ordinary velocity $\beta_{\tilde{j}} = -v_{\tilde{j}}$, with the transformation (39.38), and then inverting, one obtains the result (exercise 39.12)

Transformation from rest frame of matter to PPN coordinate frame

$$
\boldsymbol{d}x^\alpha = A^\alpha{}_{\hat\beta}\,\boldsymbol{\omega}^{\hat\beta}, \qquad \begin{cases} \boldsymbol{\omega}^{\hat\beta} = \text{orthonormal comoving basis,} \\ \boldsymbol{d}x^\alpha = \text{PPN coordinate basis;} \end{cases}
$$

$$
\begin{aligned}
A^0{}_{\hat 0} &= 1 + \frac{1}{2}v^2 + U + O(\epsilon^4), \\
A^0{}_{\hat j} &= v_j\left[1 + \frac{1}{2}v^2 + (2+\gamma)U\right] - \frac{7}{2}\Delta_1 V_j - \frac{1}{2}\Delta_2 W_j \\
&\quad + \left(\alpha_2 - \frac{1}{2}\alpha_1\right) w_j U - \alpha_2 w_k U_{kj} + O(\epsilon^5), \qquad (39.41) \\
A^j{}_{\hat 0} &= v_j\left[1 + \frac{1}{2}v^2 + U\right] + O(\epsilon^5), \\
A^j{}_{\hat k} &= (1 - \gamma U)\,\delta_{jk} + \frac{1}{2}v_j v_k + O(\epsilon^4).
\end{aligned}
$$

Stress-energy tensor in coordinate frame

This transformation, when applied to the stress-energy tensor (39.37) yields, in the PPN coordinate frame,

$$
T^{00} = \rho_o(1 + \Pi + v^2 + 2U) + O(\rho_o\epsilon^4), \qquad (39.42\text{a})
$$

$$
T^{0j} = \rho_o(1 + \Pi + v^2 + 2U)v_j + t_{\hat j\hat m}v_m + O(\rho_o\epsilon^5), \qquad (39.42\text{b})
$$

$$
\begin{aligned}
T^{jk} &= t_{\hat j\hat k}(1 - 2\gamma U) + \rho_o(1 + \Pi + v^2 + 2U)v_j v_k \\
&\quad + \frac{1}{2}(v_j t_{\hat k\hat m}v_m + v_k t_{\hat j\hat m}v_m) + O(\rho_o\epsilon^6). \qquad (39.42\text{c})
\end{aligned}
$$

EXERCISE

Exercise 39.12. THE TRANSFORMATION BETWEEN COMOVING FRAME AND PPN FRAME

Carry out the details of the derivation of the transformation matrix (39.41); and in the process calculate the correction of $O(\epsilon^4)$ to $A^0{}_{\hat 0}$.

---

§39.11. PPN EQUATIONS OF MOTION

The post-Newtonian corrections to the Newtonian equations of motion (39.15) and (39.16) are derived from the law of conservation of baryon number $(\rho_o u^\alpha)_{;\alpha} = 0$, and from the law of conservation of local energy-momentum, $T^{\alpha\beta}{}_{;\beta} = 0$. The simplest of the equations of motion is the conservation of baryon number. Its exact expression is $(\rho_o u^\alpha)_{;\alpha} = (1/\sqrt{-g})(\sqrt{-g}\rho_o u^\alpha)_{,\alpha} = 0$. Define a new quantity

$$
\begin{aligned}
\rho^* &\equiv \rho_o\left(1 + \frac{1}{2}v^2 + 3\gamma U\right) \qquad (39.43) \\
&= \rho_o u^0\sqrt{-g} + O(\rho_o\epsilon^4)
\end{aligned}
$$

[see (39.39) for u^0, and (39.32) for the metric]. Then rest-mass conservation takes on the same form as at the Newtonian order (39.15a), except now it is more accurate:

Law of baryon conservation

$$\rho^*{}_{,t} + (\rho^* v_j)_{,j} = 0 + \text{errors of } O(\rho_{o,j}\epsilon^5). \tag{39.44}$$

The next simplest equation of motion is $T^{0\alpha}{}_{;\alpha} = 0$. Straightforward evaluation, using the metric of equations (39.32) and the stress-energy tensor of equations (39.42), yields

$$\begin{aligned}[\rho_o(1 + \Pi + v^2 + 2U)]_{,t} &+ [\rho_o(1 + \Pi + v^2 + 2U)v_j + t_{\hat{j}\hat{m}}v_m]_{,j} \\ &+ (3\gamma - 2)\rho_o U_{,t} + (3\gamma - 3)\rho_o v_k U_{,k} = O(\rho_{o,j}\epsilon^5).\end{aligned} \tag{39.45}$$

By subtracting equation (39.44) from this, and using the Newtonian equations of motion (39.15) and (39.16) to simplify several terms where the Newtonian approximation is adequate, one obtains

Law of energy conservation

$$\rho_o\, d\Pi/dt + t_{\hat{j}\hat{k}}v_{j,k} = 0 + \text{errors of } O(\rho_{o,j}\epsilon^5). \tag{39.46}$$

Notice that this is nothing but the first law of thermodynamics (local energy conservation) with energy flow through the matter being neglected. (Neglecting energy flow was justified in §39.5.) This first law of thermodynamics is actually a post-Newtonian equation in the context of hydrodynamics, rather than a Newtonian equation, because Π does not affect the hydrodynamic motion at Newtonian order (see §39.7).

The last of the equations of motion, $T^{j\alpha}{}_{;\alpha} = 0$, reduces to the post-Newtonian Euler equation

Post-Newtonian Euler equation

$$\begin{aligned}\rho^*\frac{dv_j}{dt} &- \rho^* U_{,j} + [t_{\hat{j}\hat{k}}(1 + 3\gamma U)]_{,k} - t_{\hat{j}\hat{k},k}\left(\frac{1}{2}v^2 + \Pi\right) - \frac{t_{\hat{j}\hat{k}}t_{\hat{k}\hat{l},l}}{\rho^*} \\ &+ \rho^*\frac{d}{dt}\left[(2\gamma + 2)Uv_j - \frac{1}{2}(7\,\Delta_1 + \Delta_2)V_j - \frac{1}{2}\alpha_1 U w_j\right] - v_j\rho^* U_{,t} + v_k t_{\hat{k}\hat{j},t} \\ &+ \frac{1}{2}\Delta_2\rho^*(V_j - W_j)_{,t} + \frac{1}{2}\rho^*[(7\,\Delta_1 + \Delta_2)v_k + (\alpha_1 - 2\alpha_3)w_k]V_{k,j} \\ &- \rho^*\left[2\Psi - \frac{1}{2}\zeta\mathcal{A} - \frac{1}{2}\eta\mathcal{D} - \frac{1}{2}\alpha_2 w_i w_k U_{ik} + \alpha_2 w_i(V_i - W_i)\right]_{,j} \\ &- \rho^* U_{,j}\left[\gamma\boldsymbol{v}^2 - \frac{1}{2}\alpha_1\boldsymbol{w}\cdot\boldsymbol{v} + \frac{1}{2}(\alpha_2 + \alpha_3 - \alpha_1)\boldsymbol{w}^2 - (2\beta - 2)U + 3\gamma p/\rho^*\right] \\ &+ \frac{1}{2}(v_{j,k}t_{\hat{k}\hat{m}}v_m - t_{\hat{j}\hat{m}}v_{m,k}v_k) + \frac{1}{2}[v_m(t_{\hat{m}\hat{j}}v_k)_{,k} - v_j(t_{\hat{k}\hat{l}}v_k)_{,l}] = 0.\end{aligned} \tag{39.47}$$

Partial derivatives are denoted by commas; d/dt is the time-derivative following the matter [equation (39.16)].

Equations (39.44), (39.46), and (39.47) are a complete set of equations of motion at the post-Newtonian order.

EXERCISES

Exercise 39.13. EQUATIONS OF MOTION

Carry out the details of the derivation of the equations of motion (39.44), (39.46), and (39.47). As part of the derivation, calculate the following values of the Christoffel symbols in the PPN coordinate frame:

$$
\begin{aligned}
\Gamma^0{}_{00} &= -U_{,t} + O(U_{,j}\epsilon^3), \qquad \Gamma^0{}_{0j} = -U_{,j} + O(U_{,j}\epsilon^2),\\
\Gamma^0{}_{jk} &= \gamma U_{,t}\,\delta_{jk} + \frac{7}{2}\Delta_1 V_{(j,k)} + \frac{1}{2}\Delta_2 W_{(j,k)} + \left(\frac{1}{2}\alpha_1 - \alpha_2\right) w_{(j}U_{,k)}\\
&\quad + \alpha_2 w_i U_{i(j,k)} + O(U_{,j}\epsilon^3).\\
\Gamma^j{}_{00} &= -U_{,j} + \Big[(\beta+\gamma)U^2 - 2\Psi + \frac{1}{2}\zeta\mathcal{A} + \frac{1}{2}\eta\mathcal{D} + \frac{1}{2}(\alpha_1 - \alpha_2 - \alpha_3)w^2 U\\
&\quad + \frac{1}{2}(\alpha_1 - 2\alpha_3)w_i V_i + \frac{1}{2}\alpha_2 w_i w_k U_{ik}\Big]_{,j} - \frac{7}{2}\Delta_1 V_{j,t} - \frac{1}{2}\Delta_2 W_{j,t} \qquad (39.48)\\
&\quad + \left(\alpha_2 - \frac{1}{2}\alpha_1\right) w_j U_{,t} - \alpha_2 w_i U_{ij,t} + O(U_{,j}\epsilon^4),\\
\Gamma^j{}_{0k} &= \gamma U_{,t}\,\delta_{jk} - \left(\frac{7}{2}\Delta_1 + \frac{1}{2}\Delta_2\right) V_{[j,k]} - \frac{1}{2}\alpha_1 w_{[j}U_{,k]} + O(U_{,j}\epsilon^3),\\
\Gamma^j{}_{kl} &= -\gamma(U_{,j}\,\delta_{kl} - 2U_{,(k}\,\delta_{l)j}) + O(U_{,j}\epsilon^2).
\end{aligned}
$$

Here square brackets on tensor indices denote antisymmetrization, and round brackets denote symmetrization. As part of the derivation, it may be useful to prove and use the relations

$$\chi(t, \boldsymbol{x}) = -\int \rho_o(t, \boldsymbol{x}')|\boldsymbol{x} - \boldsymbol{x}'|\, d^3x', \qquad (39.49\text{a})$$

$$\chi_{,jk} = -\delta_{jk}U + U_{jk}, \qquad (39.49\text{b})$$

$$\chi_{,it} = V_i - W_i + O(\epsilon^5), \qquad (39.49\text{c})$$

$$W_{[k,j]} = V_{[k,j]}. \qquad (39.50)$$

Here χ is the function originally defined in equation (39.29c).

Exercise 39.14. POST-NEWTONIAN APPROXIMATION TO GENERAL RELATIVITY

Perform a post-Newtonian expansion of Einstein's field equations, thereby obtaining the values cited in Box 39.2 for the PPN parameters of general relativity. The calculations might best follow the approach of Chandrasekhar (1965a): Set $g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta}$, and assume

$$h_{00} = O(\epsilon^2) + O(\epsilon^4), \qquad h_{0j} = O(\epsilon^3), \qquad h_{jk} = O(\epsilon^2). \qquad (39.51)$$

Choose the space and time coordinates so that the four "gauge conditions"

$$\left.\begin{aligned} h_{jk,k} - \frac{1}{2}h_{,j} &= O(\epsilon^4/R_\odot)\\ h_{0k,k} - \frac{1}{2}h_{kk,0} &= O(\epsilon^5/R_\odot)\end{aligned}\right\}\text{ with } h = h_{\alpha\beta}\eta^{\alpha\beta} = -h_{00} + h_{ll} \qquad (39.52)$$

are satisfied.

(a) Show that the spatial gauge conditions are the post-Newtonian approximations to those (35.1a) used in the study of weak gravitational waves, but that the temporal gauge condition is not.

2

(b) Use these gauge conditions and the post-Newtonian limit in equations (8.24) and (8.47) to obtain for the Ricci tensor, accurate to linearized order,

$$R_{00} = -\frac{1}{2} h_{00,mm} + O(\epsilon^4/R_\odot{}^2), \qquad R_{jk} = -\frac{1}{2} h_{jk,mm} + O(\epsilon^4/R_\odot{}^2), \tag{39.53a}$$

$$R_{0j} = -\frac{1}{2} h_{0j,mm} - \frac{1}{4} h_{00,0j} + O(\epsilon^5/R_\odot{}^2). \tag{39.53b}$$

(c) Combine these with the Newtonian form (39.13) of the stress-energy tensor, and with equation (39.27), to obtain the following metric coefficients, accurate to linearized order:

$$h_{00} = 2U + k_{00} + O(\epsilon^6), \qquad h_{0j} = -\frac{7}{2} V_j - \frac{1}{2} W_j + O(\epsilon^5),$$

[unknown post-Newtonian correction] (39.54)

$$h_{jk} = 2U\,\delta_{jk} + O(\epsilon^4).$$

Here U, V_j, and W_j are to be regarded as defined by equations (39.34a,b,c). By comparing these metric coefficients with equations (39.32), discover that

$$\gamma = 1, \qquad \Delta_1 = 1, \qquad \Delta_2 = 1 \tag{39.55}$$

for general relativity.

(d) With this knowledge of the metric in linearized order, one can carry out the analysis of §39.10 (using $\gamma = \Delta_1 = \Delta_2 = 1$ throughout), to obtain the post-Newtonian corrections to the stress-energy tensor [equation (39.42) with $\gamma = 1$].

(e) Calculate, similarly, the post-Newtonian corrections to the Ricci tensor component R_{00}, using $g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta}$, using $h_{\alpha\beta}$ as given in equations (39.54), and using the gauge conditions (39.52). The answer should be

$$R_{00} = \left(-U - \frac{1}{2} k_{00} - U^2\right)_{,mm} + 4UU_{,mm} + O(\epsilon^6/R_\odot{}^2). \tag{39.56}$$

(f) Evaluate the Einstein equation $R_{00} = 8\pi(T_{00} - \frac{1}{2}g_{00}T)$, accurate to post-Newtonian order, and solve it to obtain the post-Newtonian metric correction

$$k_{00} = -2U^2 + 4\Psi, \tag{39.57}$$

where Ψ is given by equation (39.43d) with $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$. By comparing with equations (39.32c) and (39.34d), discover that

$$\beta = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 1, \qquad \zeta = \eta = 0 \tag{39.58}$$

for general relativity.

(g) Knowing the full post-Newtonian metric, and the full post-Newtonian stress-energy tensor, one can carry out the calculations of §39.11 (using $\gamma = \beta = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \Delta_1 = \Delta_2 = 1$, $\zeta = \eta = 0$) to obtain the post-Newtonian equations of motion for the matter [equations (39.44), (39.46), and (39.47)].

§39.12. RELATION OF PPN COORDINATES TO SURROUNDING UNIVERSE

One crucial issue remains to be clarified: What is the orientation of the PPN coordinate system relative to the surrounding universe? More particularly: Does the PPN coordinate system rotate relative to the "fixed stars on the sky;" or is it "rigidly attached" to them, in some sense? In order to answer this question, imagine using the PPN formalism to analyze the solar system. Make no assumptions about the solar system's velocity through the PPN coordinate frame. Then, as one moves outward from the Sun, past the Earth's orbit, past Pluto's orbit, and on out toward interstellar space, one sees the PPN coordinate frame become more and more Lorentz in its global properties [$g_{\alpha\beta} = \eta_{\alpha\beta} + O(M_\odot/r)$]. Thus, *far from the solar system the PPN coordinates become a "Lorentz frame moving through the galaxy."* This means, of course, that the spatial axes of the PPN coordinate frame behave as though they were attached to gyroscopes far outside the solar system. Equivalently: The PPN coordinate system Fermi-Walker-transports its spatial axes through the spacetime geometry of the galaxy and universe.

Solar system's PPN coordinate frame is attached to a local Lorentz frame of Galaxy

§39.13. SUMMARY OF PPN FORMALISM

The PPN formalism, as constructed in this chapter, is summarized in Box 39.4. Much of the recent literature uses a different set of PPN parameters than are used in this book; for a translation from one parameter set to the other, see Box 39.5.

EXERCISE

Exercise 39.15. MANY-BODY SYSTEM IN POST-NEWTONIAN LIMIT OF GENERAL RELATIVITY

Consider, in the post-Newtonian limit of general relativity, a system made up of many gravitationally interacting bodies with separations large compared to their sizes (*example:* the solar system). Idealize each body to be spherically symmetric, to be free of internal motions, and to have isotropic internal stresses, $t_{\hat{j}\hat{k}} = \delta_{jk}p$. Let the world line of the center of body A, in some chosen PPN coordinate frame, be $\boldsymbol{x}_A(t)$; and let the (coordinate) velocity of the center of body A be

$$\boldsymbol{v}_A(t) = d\boldsymbol{x}_A/dt. \tag{39.59a}$$

The total mass-energy of body A as measured in its neighborhood (rest mass-energy plus internal energy plus self-gravitational energy) is given by

$$M_A = \int_{\mathcal{V}_A} \left(1 + \Pi - \frac{1}{2} U_{\text{self}}\right) d(\text{rest mass}) + O(M_A \epsilon^4), \tag{39.59b}$$

where U_{self} is the body's own Newtonian potential (no contributions from other bodies), and $\mathcal{V}_A$ is the interior of the body.

(*continued on page 1094*)

Box 39.4 SUMMARY OF THE PPN FORMALISM

I. Variables

$\rho_o(\boldsymbol{x}, t)$: baryon "mass" density (§39.3), as measured in rest frame
$\Pi(\boldsymbol{x}, t)$: specific internal energy (dimensionless; §39.3), as measured in rest frame
$t_{\hat{j}\hat{k}}(\boldsymbol{x}, t)$: components of stress referred to orthonormal axes of rest frame
$v_j(\boldsymbol{x}, t)$: coordinate velocity of matter (i.e., rest frame) relative to PPN coordinates
$U(\boldsymbol{x}, t)$, $\Psi(\boldsymbol{x}, t)$, $\mathcal{A}(\boldsymbol{x}, t)$, $\mathcal{D}(\boldsymbol{x}, t)$, $V_j(\boldsymbol{x}, t)$, $W_j(\boldsymbol{x}, t)$, $U_{jk}(\boldsymbol{x}, t)$: gravitational potentials
γ, β, β_1, β_2, β_3, β_4, Δ_1, Δ_2, ζ, η: parameters whose values distinguish one theory from another (see Box 39.2)
$\boldsymbol{w}$: velocity of PPN coordinate frame relative to "universal rest frame" [relevant only for theories with nonzero α_1, α_2, or α_3; see eq. (39.33)].

II. Equations governing evolution of these variables

ρ_o: conservation of rest mass, equation (39.44)
Π: first law of thermodynamics, equation (39.46)
$t_{\hat{j}\hat{k}}$: determined in terms of ρ_o, Π, and other material variables (chemical composition, strains, etc.) by equations of state and the usual theory of a stressed medium—which is not discussed here
v_j: equations of motion ("$\boldsymbol{F} = m\boldsymbol{a}$"), equations (39.47)
U, Ψ, $\mathcal{A}$, $\mathcal{D}$, V_j, W_j, U_{jk}: source equations (39.34)

III. Quantities to be calculated from these variables

$g_{00}(\boldsymbol{x}, t)$, $g_{0j}(\boldsymbol{x}, t)$, $g_{jk}(\boldsymbol{x}, t)$: these components of metric in PPN coordinate frame are expressed in terms of gravitational potentials by equations (39.32)
$u^0(\boldsymbol{x}, t)$, $u^j(\boldsymbol{x}, t)$: these components of matter 4-velocity in PPN coordinate frame are given by equations (39.39)
$T^{00}(\boldsymbol{x}, t)$, $T^{0j}(\boldsymbol{x}, t)$, $T^{jk}(\boldsymbol{x}, t)$: these components of stress-energy tensor in PPN coordinate frame are given by equations (39.42)

IV. Relation between rest frame, PPN coordinates, and the universe

1. Orthonormal basis $\boldsymbol{\omega}^{\hat{\alpha}}$ of rest frame, where $t_{\hat{j}\hat{k}}$ are defined, is related to PPN coordinate basis $\boldsymbol{d}x^{\alpha}$ by equations (39.41)
2. Far from the sun, the PPN coordinates become asymptotically Lorentz; i.e., they form an inertial frame moving through the spacetime geometry of the galaxy and the universe.
3. Gives no account of expansion of universe or of cosmic gravitational waves impinging on solar system.

Box 39.5 PPN PARAMETERS USED IN LITERATURE: A TRANSLATOR'S GUIDE

The original "point-particle version" of the PPN formalism [Nordtvedt (1968b)], and the original "perfect-fluid version" [Will (1971c)] used different sets of PPN parameters. This book has adopted Will's set, and has added the parameter η characterizing effects of anisotropic stresses. More recently, Will and Nordtvedt have jointly adopted a revised set of parameters, described below.

A. Translation Table

Will-Nordtvedt revised parameters[a]	*Revised parameters in notation of this book*[b]	*Revised parameters in notation of Nordtvedt (1968b)*[c]
γ	γ	γ
β	β	β
α_1	$7\Delta_1 + \Delta_2 - 4\gamma - 4$	$8\Delta - 4\gamma - 4$
α_2	$\Delta_2 + \zeta - 1$	$\alpha''' - 1$
α_3	$4\beta_1 - 2\gamma - 2 - \zeta$	$4\alpha'' - \alpha''' - 2\gamma - 1$
ζ_1	ζ	$\alpha''' - \chi$
ζ_2	$2\beta + 2\beta_2 - 3\gamma - 1$	$2\beta - \alpha' - 1$
ζ_3	$\beta_3 - 1$	absent
ζ_4	$\beta_4 - \gamma$	absent

[a]Revised parameters are used by Will and Nordtvedt (1972), Nordtvedt and Will (1972), Will (1972), and Ni (1973).
[b]Notation of this book is used by Will (1971a,b,c,d), Ni (1972), and Thorne, Ni, and Will (1971).
[c]Nordtvedt's original "point-particle" parameters were used by Nordtvedt (1968b, 1970, 1971a,b).

B. Significance of Revised Parameters

$\alpha_1, \alpha_2, \alpha_3$ measure the extent of and nature of "preferred-frame effects"; see §39.9. Any theory of gravity with at least one α nonzero is called a *preferred-frame theory.*

$\zeta_1, \zeta_2, \zeta_3, \zeta_4, \alpha_3$ measure the extent of and nature of breakdowns in global conservation laws. A theory of gravity possesses, at the post-Newtonian level, all 10 global conservation laws (4 for energy-momentum, 6 for angular momentum; see Chapters 19 and 20) if and only if $\zeta_1 = \zeta_2 = \zeta_3 = \zeta_4 = \alpha_3 = 0$. See Will (1971d), Will and Nordtvedt (1972), Will (1972), for proofs and discussion. Any theory with $\zeta_1 = \zeta_2 = \zeta_3 = \zeta_4 = \alpha_3 = 0$ is called a *conservative theory.*

In general relativity and the Dicke-Brans-Jordan theory, all α's and ζ's vanish. Thus, general relativity and Dicke-Brans-Jordan are conservative theories with no preferred-frame effects.

(a) Show that, when written in the chosen PPN coordinate frame, this expression for M_A becomes

$$M_A = \int_{\mathcal{V}_A} \rho_o \left(1 + \Pi + \frac{1}{2} v_A{}^2 + 3U - \frac{1}{2} U_{\text{self}}\right) d^3x + O(M_A \epsilon^4). \tag{39.59c}$$

Use equations (39.43), (39.44), and (39.46) to show that M_A is conserved as the bodies move about, $dM_A/dt = 0$.

(b) Pick an event $(t, \boldsymbol{x})$ outside all the bodies, and at time t denote

$$\boldsymbol{r}_A \equiv \boldsymbol{x}_A - \boldsymbol{x}, \qquad \boldsymbol{r}_{AB} \equiv \boldsymbol{x}_A - \boldsymbol{x}_B, \qquad r_A \equiv |\boldsymbol{r}_A|, \qquad r_{AB} \equiv |\boldsymbol{r}_{AB}|. \tag{39.59d}$$

Show that the general-relativistic, post-Newtonian metric (39.32) at the chosen event has the form

$$g_{ik} = \delta_{jk}\left(1 + 2\sum_A \frac{M_A}{r_A}\right) + O(\epsilon^4), \tag{39.60a}$$

$$g_{0j} = -\sum_A \frac{M_A}{r_A}\left[\frac{7}{2} v_{Aj} + \frac{1}{2}\frac{(\boldsymbol{v}_A \cdot \boldsymbol{r}_A) r_{Aj}}{r_A{}^2}\right] + O(\epsilon^5), \tag{39.60b}$$

$$\begin{aligned} g_{00} = {} & -1 + 2\sum_A \frac{M_A}{r_A} - 2\left(\sum_A \frac{M_A}{r_A}\right)^2 + 3\sum_A \frac{M_A v_A{}^2}{r_A} \\ & -2\sum_A \sum_{B \neq A} \frac{M_A M_B}{r_A r_{AB}} + O(\epsilon^6). \end{aligned} \tag{39.60c}$$

[*Hint:* From the Newtonian virial theorem (39.21a), applied to body A by itself in its own rest frame, conclude that

$$\int_{\mathcal{V}_A} \left(3p - \frac{1}{2}\rho_o U_{\text{self}}\right) d^3x = O(M_A \epsilon^4), \tag{39.61}$$

where the integral is performed in the PPN frame.]

(c) Perform an infinitesimal coordinate transformation,

$$t_{\text{OLD}} = t_{\text{NEW}} - \frac{1}{2}\sum_A \frac{M_A(\boldsymbol{r}_A \cdot \boldsymbol{v}_A)}{r_A}, \qquad \boldsymbol{x}_{\text{OLD}} = \boldsymbol{x}_{\text{NEW}}, \tag{39.62}$$

to bring the metric (39.60) into the standard form originally devised by Einstein, Infeld, and Hoffman (1938), and by Eddington and Clark (1938):

$$g_{jk} = \delta_{jk}\left(1 + 2\sum_A \frac{M_A}{r_A}\right) + O(\epsilon^4), \tag{39.63a}$$

$$g_{0j} = -4\sum_A \frac{M_A}{r_A} v_{Aj} + O(\epsilon^5), \tag{39.63b}$$

$$\begin{aligned} g_{00} = {} & -1 + 2\sum_A \frac{M_A}{r_A} - 2\left(\sum_A \frac{M_A}{r_A}\right)^2 + 3\sum_A \frac{M_A v_A{}^2}{r_A} \\ & -2\sum_A \sum_{B \neq A} \frac{M_A M_B}{r_A r_{AB}} - \frac{\partial^2 \chi}{\partial t^2} + O(\epsilon^6), \end{aligned} \tag{39.63c}$$

where χ [equation (39.49a)] is given by

$$\chi = -\sum_A M_A r_A.$$

(d) The equations of motion for the bodies can be obtained in either of two ways: by performing a volume integral of the Euler equation (39.48) over the interior of each body; or by invoking the general arguments of §20.6. The latter way is the easier. Use it to conclude that any chosen body K moves along a geodesic of the metric obtained by omitting the terms $A = K$ from the sums in (39.63). Show that the geodesic equation for body K reduces to

$$\begin{aligned}\frac{d^2\boldsymbol{x}_K}{dt^2} \equiv \frac{d\boldsymbol{v}_K}{dt} = \sum_{A \neq K} \boldsymbol{r}_{AK} \frac{M_A}{{r_{AK}}^3}\Bigg[1 - 4\sum_{B \neq K} \frac{M_B}{r_{BK}} - \sum_{C \neq A} \frac{M_C}{r_{CA}}\left(1 - \frac{\boldsymbol{r}_{AK}\cdot\boldsymbol{r}_{CA}}{2{r_{CA}}^2}\right) \\ + {\boldsymbol{v}_K}^2 + 2{\boldsymbol{v}_A}^2 - 4\boldsymbol{v}_A\cdot\boldsymbol{v}_K - \frac{3}{2}\left(\frac{\boldsymbol{v}_A\cdot\boldsymbol{r}_{AK}}{r_{AK}}\right)^2\Bigg] \\ - \sum_{A \neq K} (\boldsymbol{v}_A - \boldsymbol{v}_K)\frac{M_A\boldsymbol{r}_{AK}\cdot(3\boldsymbol{v}_A - 4\boldsymbol{v}_K)}{{r_{AK}}^3} \\ + \frac{7}{2}\sum_{A \neq K}\sum_{C \neq A} \boldsymbol{r}_{CA}\frac{M_A M_C}{r_{AK}{r_{CA}}^3}.\end{aligned} \tag{39.64}$$

Equations (39.63) and (39.64) are called the Einstein-Infeld-Hoffman ("EIH") equations for the geometry and evolution of a many-body system. They are used widely in analyses of planetary orbits in the solar system. For example, the Caltech Jet Propulsion Laboratory uses them, in modified form, to calculate ephemerides for high-precision tracking of planets and spacecraft. The above method of deriving the EIH equations and metric was devised by Fock (1959). For a similar calculation in the Dicke-Brans-Jordan theory, see Estabrook (1969); and for a derivation of the analogous many-body equations in the full PPN formalism, see Will (1972).

CHAPTER 40

SOLAR-SYSTEM EXPERIMENTS

§40.1. MANY EXPERIMENTS OPEN TO DISTINGUISH GENERAL RELATIVITY FROM PROPOSED METRIC THEORIES OF GRAVITY

No audience will show up for a fight if in everyone's eyes the challenger has zero chance to win. No battle-hungry promoter desperately trying to finance the fight can afford to put into the ring against the champion any but the best contender that he can find. Against Einstein's metric theory of gravity, the judgment of the day (as §39.2 showed) leaves one no option except to put up another theory of gravity that is also metric (or metric plus torsion).

This chapter analyzes experiments using PPN formalism

To put on a contest, then, is to design and perform an experiment that distinguishes general relativity from some not completely implausible metric theory of gravity. This chapter describes such experiments—some already performed; some to be performed in the future—and analyses their significance using the PPN formalism of Chapter 39.

Complexity of solar system's spacetime geometry

In most of the experiments to be described, one investigates the motion of the moon, planets, spacecraft, light rays, or gyroscopes through the spacetime geometry of the solar system. That spacetime geometry is very complicated. It includes the spherical fields of the sun and all the planets, nonspherical fields due to their quadrupolar and higher-order deformations, and fields due to their momentum and angular momentum. Moreover, the spacetime geometry results—or at least in the post-Newtonian formalism it is viewed as resulting—from a *nonlinear* superposition of all these fields.*

*Of course, from the point of view of Einstein's full general relativity theory, all that legalistically counts is the one and only curved-spacetime geometry of the real physical world. All these "individual fields" are mere bookkeepers' discourse, and they are best abandoned (they cease to be useful) when one passes from the post-Newtonian limit to the full Einstein theory.

Fortunately for this discussion, several of the most important experiments are free of almost all these complications. The effects they measure are associated entirely with the spherical part of the sun's gravitational field. A description of these experiments will come first (§§40.2–40.5), and then attention will turn to experiments that are more complex in principle.

Idealization of geometry to that of isolated, static, spherical sun:

To discuss central-field experiments, one needs an expression for the external gravitational field of an idealized, isolated, static, spherical sun. In general relativity, such a gravitational field is described by the Schwarzschild line element,

(1) in Schwarzschild coordinates

$$ds^2 = -\left(1 - \frac{2M_\odot}{r}\right) dt^2 + \frac{dr^2}{1 - 2M_\odot/r} + r^2(d\theta^2 + \sin^2\phi\, d\phi^2).$$

But this line element is not what one wants, for two reasons: (1) it is "too accurate"; (2) it is written in the "wrong" coordinate system.

Why too accurate? Because it is simple only when unperturbed and unmodified; whereas some modified theories show up new effects that are so complex they are tractable only in the post-Newtonian approximation. Why wrong coordinate system? Because physicists, astronomers, and other celestial mechanics have adopted the fairly standard convention of using "isotropic coordinates" rather than "Schwarzschild coordinates" when analyzing the solar system. Example: post-Newtonian expansions, including the PPN formalism of Chapter 39, almost always use isotropic coordinates. Another example: the relativistic ephemeris for the solar system, prepared by the Caltech Jet Propulsion Laboratory [Ohandley *et al.* (1969); Anderson (1973)] and used extensively throughout the world, employs isotropic coordinates.

Modify the Schwarzschild line element, then. First transform to isotropic coordinates (Exercise 31.7); then expand the metric coefficients in powers of $M_\odot/r$, to post-Newtonian accuracy. Thereby obtain

(2) in isotropic coordinates

$$\begin{aligned} ds^2 &= -\left[1 - 2\frac{M_\odot}{r} + 2\left(\frac{M_\odot}{r}\right)^2\right] dt^2 + \left[1 + 2\frac{M_\odot}{r}\right][dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)] \\ &= -\left[1 - 2\frac{M_\odot}{r} + 2\left(\frac{M_\odot}{r}\right)^2\right] dt^2 + \left[1 + 2\frac{M_\odot}{r}\right][dx^2 + dy^2 + dz^2]. \end{aligned} \tag{40.1}$$

Here r, θ, ϕ are related to x, y, z in the usual manner:

$$r = (x^2 + y^2 + z^2)^{1/2}, \qquad \theta = \tan^{-1}[z/(x^2 + y^2)^{1/2}], \qquad \phi = \tan^{-1}(y/x); \tag{40.2}$$

and r is the new, "isotropic" radial coordinate, not to be confused with the Schwarzschild r. (The reader who has not studied §39.6 will discover in the next section why one keeps terms of order $M_\odot{}^2/r^2$ in g_{00} but not in g_{jk}.) *Note:* this post-Newtonian expression for the metric is a special case of the result derived in exercise 19.3.

If one calculates the gravitational field of the same source (the sun) in the same post-Newtonian approximation in other metric theories of gravity, one obtains a very similar result:

(3) in PPN formalism

$$ds^2 = -\left[1 - 2\frac{M_\odot}{r} + 2\beta\left(\frac{M_\odot}{r}\right)^2\right]dt^2 + \left[1 + 2\gamma\frac{M_\odot}{r}\right][dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)]$$
$$= -\left[1 - 2\frac{M_\odot}{r} + 2\beta\left(\frac{M_\odot}{r}\right)^2\right]dt^2 + \left[1 + 2\gamma\frac{M_\odot}{r}\right][dx^2 + dy^2 + dz^2] \tag{40.3}$$

(see exercise 40.1). Here γ and β are two of the ten PPN parameters described in Box 39.2. Recall from that box that γ measures "the amount of space curvature produced by unit rest mass," while β measures "the amount of nonlinearity in the superposition law for g_{00}." These heuristic descriptions find their mathematical counterparts in the above form for the idealized metric surrounding a spherically symmetric center of attraction.

By measuring the parameter γ to high precision, one can distinguish between general relativity ($\gamma = 1$) and the Dicke-Brans-Jordan theory [$\gamma = (1 + \omega)/(2 + \omega)$, where ω is the "Dicke coupling constant"]; see Box 39.2. But general relativity and Dicke-Brans-Jordan predict the same value for β ($\beta = 1$). This identity does not mean that β is unworthy of measurement. A value $\beta \neq 1$ is predicted by other theories [see Ni (1972)]; so measurements of β are useful in distinguishing such theories from general relativity.

Actually, the above form (40.3) for the sun's metric is not fully general. In any theory with a preferred "universal rest frame" (e.g., Ni's theory; Box 39.1), there are additional terms in the metric due to motion of the sun relative to that preferred frame (exercise 40.1):

(4) including preferred-frame effects

$$ds^2 = (\text{expression 40.3}) + (\alpha_2 + \alpha_3 - \alpha_1)\frac{M_\odot}{r} w^2\, dt^2 + 2\left(\alpha_2 - \frac{1}{2}\alpha_1\right)\frac{M_\odot}{r} w_j\, dx^j\, dt$$
$$-\alpha_2\left[\frac{M_\odot}{r^3}x^j x^k - \frac{I_\odot}{r^5}\left(x^j x^k - \frac{1}{3}r^2\,\delta_{jk}\right)\right] w_j\, dt(2\, dx^k + w_k\, dt). \tag{40.3$'$}$$

In these "preferred-frame terms" $I_\odot \equiv I_{jj} = \int \rho r^2\, d^3x$ is the trace of the second moment of the sun's mass distribution;

$$\alpha_1 = 7\Delta_1 + \Delta_2 - 4\gamma - 4,$$
$$\alpha_2 = \Delta_2 + \zeta - 1,$$
$$\alpha_3 = 4\beta_1 - 2\gamma - 2 - \zeta$$

are combinations of PPN parameters; and $\boldsymbol{w}$ is the sun's velocity ($\equiv$ velocity of coordinate system) relative to the preferred frame. (Theories such as general relativity and Dicke-Brans-Jordan, which possess no preferred frame, have $\alpha_1 = \alpha_2 = \alpha_3 = 0$, and therefore have no preferred-frame terms in the metric.) For ease of exposition, all equations and calculations in this chapter will ignore the preferred-frame terms; but the consequences of those terms will be discussed and references analyzing them will be cited.

EXERCISES

Exercise 40.1. PPN METRIC FOR IDEALIZED SUN [Track 2]

Show that for an isolated, static, spherical sun at rest at the origin of the PPN coordinate system, the PPN metric (39.32) reduces to expressions (40.3), (40.3′). As part of the reduction, show that the sun's total mass-energy is given by

$$M_\odot = \int_0^{R_\odot} \rho_0(1 + 2\beta_2 U + \beta_3 \Pi + 3\beta_4 p/\rho_0) 4\pi r^2\, dr. \tag{40.4}$$

[*Warning:* One must not look at this formula and immediately think: "The contribution of rest mass is $\int \rho_0 4\pi r^2\, dr$, the contribution of gravitational energy is $\int 2\beta_2 \rho_0 U 4\pi r^2\, dr$, etc." Rather, in making any such interpretation one must remember that (1) spacetime is curved, so $4\pi r^2\, dr$ is not *proper volume* as measured by physical meter sticks; also (2) virial theorems (exercise 39.6) and other integral theorems can be used to change the form of the integrand. For further discussion see exercises 40.9 and 40.10 below.]

§40.2. THE USE OF LIGHT RAYS AND RADIO WAVES TO TEST GRAVITY

Light rays and radio waves give "clean" tests of relativity

In the Newtonian limit, planetary and spacecraft orbits are strongly influenced by gravity; but light propagation and radio-wave propagation (at "infinite" velocity) are not influenced at all. For this reason, experimental studies of orbits are beset by the problem of separating the relativistic effects from much larger standard Newtonian effects. By contrast, experimental studies of light and radio-wave propagation do not contend with any such overpowering Newtonian background. Not surprisingly, they are to date (1973) the clearest and most definitive of the solar-system experiments.

Mathematically, the parameter that distinguishes a light ray from a planet is its high speed. In the geodesic equation, the magnitude of the velocity determines which metric coefficients can influence the motion. Consider, for example, a weak, static field $g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta}$, and a particle at $(x, y, z) = (r, 0, 0)$ moving with velocity $(v_x, v_y, v_z) = (0, v, 0)$; see Figure 40.1. Here the effect of gravity on the trajectory of the particle can be characterized by the quantity

$$\begin{aligned}
\begin{pmatrix}\text{curvature of trajectory in 3-dimensional,}\\ \text{nearly Euclidean, space}\end{pmatrix} &= \begin{pmatrix}\text{radius of curvature}\\ \text{of trajectory}\end{pmatrix}^{-1} \\
&= \frac{d^2x}{dy^2} = \frac{d\tau}{dy}\frac{d}{d\tau}\left(\frac{d\tau}{dy}\frac{dx}{d\tau}\right) = \frac{1}{u^y}\frac{d}{d\tau}\left(\frac{u^x}{u^y}\right) = \frac{1}{(u^y)^2}\frac{du^x}{d\tau} \\
&= -\frac{(1-v^2)}{v^2}\Gamma^x{}_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = -\frac{1}{v^2}\Gamma^x{}_{\alpha\beta}\frac{dx^\alpha}{dt}\frac{dx^\beta}{dt} \\
&= -\Gamma^x{}_{00}v^{-2} - 2\Gamma^x{}_{0y}v^{-1} - \Gamma^x{}_{yy} \\
&= \frac{1}{2}h_{00,x}v^{-2} + (h_{0y,x} - h_{0x,y})v^{-1} + \left(\frac{1}{2}h_{yy,x} - h_{xy,y}\right).
\end{aligned}$$

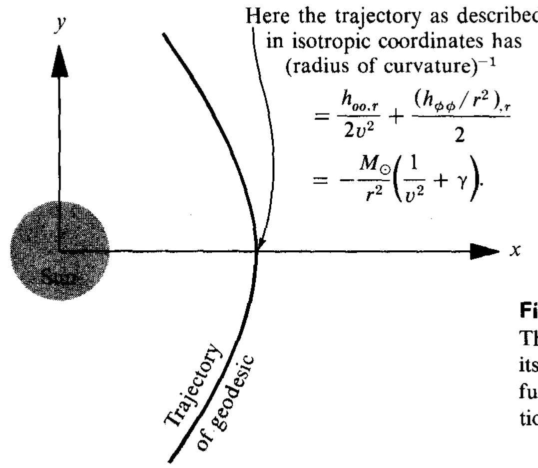


Figure 40.1.
The bending of the trajectory of a test body at its point of closest approach to the sun, as a function of its 3-velocity. (See text for computation and discussion.)

Reexpressed in spherical coordinates, in the terminology of the idealized solar line element (40.3), this formula says

$$\begin{pmatrix}\text{curvature of trajectory}\\ \text{in 3-space}\end{pmatrix} = \frac{1}{2}h_{00,r}v^{-2} + \frac{1}{2}(h_{\phi\phi}/r^2)_{,r} \approx -(M_\odot/r^2)(v^{-2}+\gamma) \tag{40.5}$$

for a particle at its point of closest approach to the sun. (Compare with exercise 25.21.) Note that here γ is a PPN parameter; it is *not* $(1 - v^2)^{-1/2}$.

Notice what happens as one boosts the velocity of the particle. For slow velocities [$v^2 \sim$ (post-Newtonian expansion parameter ϵ^2) $\approx M_\odot/R_\odot$], the Newtonian part of h_{00} dominates completely; and the tiny post-Newtonian corrections come equally from the ϵ^4 part of h_{00}, the ϵ^3 part of h_{0j}, and the ϵ^2 part of h_{jk}. [This was the justification for expanding h_{00} to $O(\epsilon^4)$, h_{0j} to $O(\epsilon^3)$, and h_{jk} to $O(\epsilon^2)$ in the post-Newtonian limit; see §39.6.] But as v increases, the ordering of the terms changes. In the high-v regime ($v \sim 1 \gg \epsilon^2$), the bending of the trajectory has become almost imperceptible because of the high forward momentum of the particle and the short time it receives transverse momentum from the sun. What bending is left is due to the ϵ^2 (Newtonian) part of h_{00}, and the ϵ^2 (post-Newtonian) part of h_{jk}. Nothing else can have a significant influence. Notice, moreover, that—even when one allows for "preferred-frame" effects—these dominant terms,

$$h_{00} = 2U = 2M_\odot/r \text{ and } h_{jk} = 2\gamma U\delta_{jk} = 2\gamma(M_\odot/r)\,\delta_{jk},$$

depend only on the Newtonian potential $U \equiv -\Phi$ and the PPN parameter γ.

Light rays are governed solely by Newtonian potential and PPN parameter γ

This is a special case of a more general result: *Aside from fractional corrections of $\epsilon^2 \lesssim 10^{-6}$, relativistic effects on light and radio-wave propagation are governed entirely by the Newtonian potential U and the PPN parameter* γ. These relativistic effects include the gravitational redshift (discussed in the last chapter; independent of γ), the gravitational deflection of light and radio waves (discussed below; dependent on γ), and the "relativistic time-delay" (discussed below; dependent on γ).

§40.3. "LIGHT" DEFLECTION

Light deflection:

Consider a light or radio ray coming into a telescope on Earth from a distant star or quasar. Do not assume, as in the usual discussion (exercises 18.6 and 25.24), that the ray passes near the sun. The deflection by the sun's gravitational field will probably be measurable, in the middle or late 1970's, even when the ray passes far from the sun! [The calculation that follows is due to Ward (1970), but Shapiro (1967) first derived the answer.]

(1) derivation

Orient the PPN spherical coordinates of equation (40.3) so that the ray lies in the "plane" $\theta = \pi/2$. By symmetry, if it starts out in this plane far from the Earth, it must lie in this plane always. Let the incoming ray enter the solar system along the line $\phi = 0$; and let the Earth be located at $r = r_E$, $\phi = \phi_E$ when the ray reaches it. (See Figure 40.2.) One wishes to calculate the angle α between the incoming light ray and the center of the sun, as measured in the orthonormal frame $(\boldsymbol{e}_{\hat{r}}, \boldsymbol{e}_{\hat{\phi}})$ of an observer on Earth. If the sun had zero mass (flat, Euclidean space), α would be $\pi - \phi_E$ (see Figure 40.2). However, the sun produces a deflection: $\alpha = \pi - \phi_E + \delta\alpha$. The deflection angle $\delta\alpha$ is the true objective of the calculation.

In the calculation, ignore the Earth's orbital and rotational motions. They lead to aberration, for which correction can be made by the usual formula of special relativity (Lorentz transformation in the neighborhood of the telescope.) Also ignore deflection of the light ray due to the Earth's gravitational field (deflection angle $\sim 2M_E/R_E \sim 0''.0003$), which might be detectable in the late 1970's.

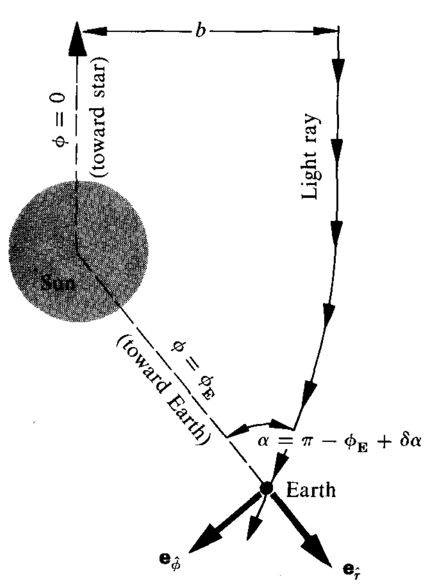


Figure 40.2.
Coordinates used in the text for calculating the deflection of light. Notice that in this diagram ϕ increases in the clockwise direction.

As the first step in calculating the deflection angle, determine the trajectory of the ray in the r, ϕ-plane. This can be calculated either using the geodesic equation, or using the eikonal method of geometric optics (Hamilton-Jacobi method; §22.5 and Box 25.4). The result of such a calculation (exercise 40.2) is an equation connecting r with ϕ; thus,

$$\frac{b}{r} = \sin\phi + \frac{(1+\gamma)M_\odot}{b}(1 - \cos\phi). \tag{40.6}$$

Notice that b has a simple geometric interpretation: far from the sun, the ray trajectory is $\phi = b/r + O(M_\odot b/r^2)$. Consequently, b is the impact parameter in the usual sense of classical scattering theory (see Figure 40.2). The ray makes its closest approach to the sun (assuming it is not intercepted by the Earth first) at the PPN coordinate radius

$$r_{\text{min}} = b\left[1 - \frac{(1+\gamma)M_\odot}{b}\right] \approx b. \tag{40.7}$$

Thus, b can also be thought of as the radius of the ray's "perihelion."

Notice that the ray returns to $r = \infty$, not at an angle $\phi = \pi$, but rather at

$$\phi(r = \infty) = \pi + 2(1+\gamma)M_\odot/b. \tag{40.8a}$$

Thus, the total deflection angle is

$$\begin{aligned}(\text{angle of total deflection}) &= 2(1+\gamma)M_\odot/b \\ &= \frac{1}{2}(1+\gamma)1''.75 \text{ for a ray that} \\ &\qquad\qquad\qquad \text{just grazes the sun.}\end{aligned} \tag{40.8b}$$

But this is not the quantity of primary interest. Rather, one seeks the position of the star as seen by an astronomer on Earth. The angle $\alpha = \pi - \phi_E + \delta\alpha$ between the sun and the star as measured by the astronomer is given by (see Figure 40.2)

$$\begin{aligned}\tan(\pi - \phi_E + \delta\alpha) &= -\tan\phi_E + \delta\alpha/\cos^2\phi_E \\ &= \frac{u^{\hat\phi}}{u^{\hat r}} = \left[\frac{(1+\gamma M_\odot/r)r\,d\phi/d\lambda}{(1+\gamma M_\odot/r)\,dr/d\lambda}\right]_E = \left[\frac{r\,d\phi}{dr}\right]_E \\ &= -\left[\frac{(b/r)\,d\phi}{d(b/r)}\right]_E,\end{aligned} \tag{40.9}$$

where $u^\beta = dx^\beta/d\lambda$ are the components of a tangent to the ray at the Earth. By inserting into this equation expression (40.6) for the trajectory of the ray, one obtains

$$\begin{aligned}\tan\phi_E - \frac{\delta\alpha}{\cos^2\phi_E} &= \frac{\sin\phi_E + [(1+\gamma)M_\odot/b](1-\cos\phi_E)}{\cos\phi_E + [(1+\gamma)M_\odot/b]\sin\phi_E} \\ &= \tan\phi_E - [(1+\gamma)M_\odot/b](1-\cos\phi_E)/\cos^2\phi_E.\end{aligned} \tag{40.10}$$

Thus, the deflection angle measured at the Earth is

$$\delta\alpha = \frac{(1+\gamma)M_\odot}{b}(1+\cos\alpha) = \frac{(1+\gamma)M_\odot}{r_E}\left(\frac{1+\cos\alpha}{1-\cos\alpha}\right)^{1/2}. \tag{40.11}$$

(2) formula for deflection angle

It ranges from zero when the ray comes in opposite to the sun's direction ($\alpha = \pi$), through the value

$$(1+\gamma)M_\odot/r_E = \frac{1}{2}(1+\gamma)0''.0041 \tag{40.12}$$

when the ray comes in perpendicular to the Earth-Sun line ($\alpha = \pi/2$), to the "classical value" of $\frac{1}{2}(1+\gamma) \times 1''.75$ when the ray comes in grazing the sun's limb.

Experimental measurements of deflection

All experiments to date (1972) have examined the case of grazing passage. The experimental results are stated and discussed in Box 40.1. They show that the PPN parameter γ has its general relativistic value of 1 to within an uncertainty of about 20 percent.

By the middle or late 1970's, measurements of the deflection of radio waves from quasars should determine γ to much better than 1 percent. Also, by that time radio astronomers may be making progress toward setting up high-precision coordinates on the sky using very long baseline interferometry. If so, they will have to use equation (40.11) to compensate for the "warping" of the coordinates caused by the sun's deflection of radio waves in all regions of the sky, not just near the solar limb.

EXERCISE

Exercise 40.2. TRAJECTORY OF LIGHT RAY IN SUN'S GRAVITATIONAL FIELD

Derive equation (40.6) for the path of a light ray in isotropic coordinates (40.3) in the sun's "equatorial plane." Use one or more of three alternative approaches: (1) direct integration of the geodesic equation (the hardest approach!); (2) computation based on the three integrals of the motion

$$\boldsymbol{k}\cdot\boldsymbol{k} = 0, \qquad \boldsymbol{k}\cdot(\partial/\partial t) = k_0, \qquad \boldsymbol{k}\cdot(\partial/\partial\phi) = k_\phi = -bk_0;$$
$$\boldsymbol{k} \equiv d/d\lambda = \text{tangent vector to geodesic}$$

(see §§25.2 and 25.3); (3) computation based on the Hamilton-Jacobi method (Box 25.4), which for photons (zero rest mass) reduces to the "eikonal method" of geometric optics (see §22.5).

---

§40.4. TIME-DELAY IN RADAR PROPAGATION

Radar time delay:

Another effect of spacetime curvature on electromagnetic waves is a relativistic delay in the round-trip travel time for radar signals. It was first pointed out by Shapiro (1964); see also Muhleman and Reichley (1964, 1965).

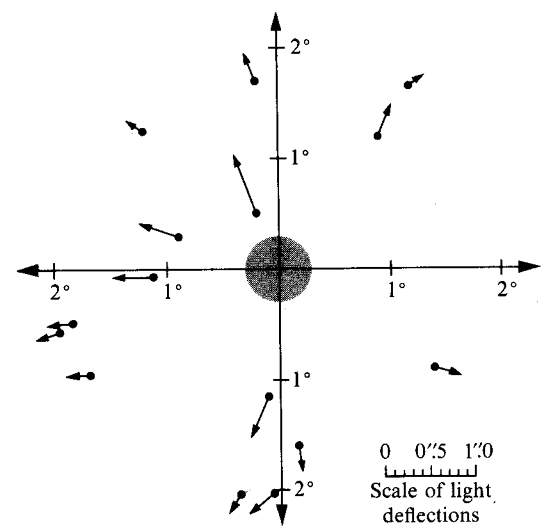
(continued on page 1106)

Box 40.1 DEFLECTION OF LIGHT AND RADIO WAVES BY SUN: EXPERIMENTAL RESULTS

Eclipse Measurements

Until 1968 every experiment measured the deflection of starlight during total eclipse of the sun. The measurements were beset by difficulties such as poor weather, optical distortions due to temperature changes, and the strange propensity of eclipses to attain maximum time of totality in jungles, in the middles of oceans, in deserts, and in arctic tundras. Lists of all the results and references are given by Bertotti, Brill and Krotkov (1962), and by Klüber (1960). Dicke (1964b) summarizes the results as follows:

"The analyses [of the experimental data] scatter from a deflection at the limb of the sun of 1.43 seconds of arc to 2.7 seconds [compared to a general relativistic value of 1.75 seconds]. The scatter would not be too bad if one could believe that the technique was free of systematic errors. It appears that one must consider this observation uncertain to at least 10 percent, and perhaps as much as 20 percent." This result corresponds to an uncertainty in γ of 20 to 40 percent.

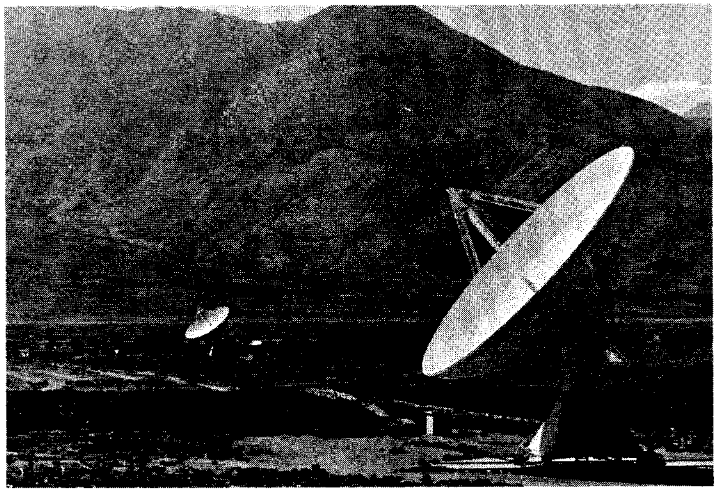


Observed light deflections (mean of two instruments) of the 15 best measured stars within 2°.5 of the sun's center in the total solar eclipse of September 21, 1922 at Wallal, Western Australia, as determined by Campbell and Trumpler (1928). The arrows represent in size and direction the observed light deflections relative to the reference stars (5° to 10° from the sun's center). (See Box 1.6 for Einstein's description of the deflection in terms of the curvature of geometry near the sun).

Measurements on the Deflection of Radio Waves

Each October 8 the sun, as seen from the Earth, passes in front of the quasar 3C279. By monitoring the angular separation between 3C279 and a nearby quasar 3C273, radio astronomers can measure the deflection by the sun of the 3C279 radio waves. The monitoring uses radio interferometers. [See references cited in table for discussion of the technique.] Technology of the early 1970's should permit measurements to a precision 0.001 seconds of arc or better, if the two ends of the interferometer are separated by several thousand kilometers ("transcontinental" or "transworld" baseline). But as of 1971 the only successful experiments were less ambitious: they used baselines of less than 10 kilometers. A summary of these pre-1971, short-baseline results is shown in the table.

they used baselines of less than 10 kilometers. A summary of these pre-1971, short-baseline results is shown in the table.



The 90-foot (background) and 130-foot (foreground) radio interferometer system at Caltech's Owens Valley Radio Observatory. These were used by Seielstadt, Sramek, and Weiler (1970) in their pioneering measurement of the deflection of quasar radio waves by the sun. During the experiment the two antennas were separated by 1.07 kilometers. (Photo by Alan Moffet.)

					Experimental results[a]		
Dates of observation	Observatory	Experimenters and reference	Number of telescopes and separations	Wave lengths	$\frac{1}{2}(1+\gamma) = \frac{\text{(Observed deflection)}}{\text{(Einstein prediction)}}$	Formal standard error	One-sigma error
Sept. 30–Oct. 15 1969	Owens Valley (Caltech)	Seielstadt, Sramek, Weiler (1970)	2, 1.07 km	3.1 cm	1.01	±0.12	±0.12
Oct. 2–Oct. 10 1969	Goldstone (Caltech-JPL)	Muhleman, Ekers, Fomalont (1970)	2, 21.56 km	12.5 cm	1.04	±0.05	+0.15 −0.10
Oct. 2–Oct. 12 1970	National Radio Astronomy Observatory (USA)	Sramek (1971)	3, 0.80 km, 1.90 km, 2.70 km	11.1 cm, 3.7 cm	0.90	±0.05	±0.05
Sept. 30–Oct. 15 1970	Mullard Radio Astronomy Observatory (Cambridge Univ.)	Hill (1971)	3, 0.66 km, 1.41 km	11.1 cm, 6.0 cm	1.07	±0.17	±0.17

[a]Here (observed deflection)/(Einstein prediction) is the number $\frac{1}{2}(1+\gamma)$ obtained by fitting the observational data to the PPN prediction (40.11). [For these experiments the ray passes near the solar limb; so (40.11) reduces to $\delta\alpha = \frac{1}{2}(1+\gamma)(M_\odot/b)$.] The "formal standard error" is obtained from the data by standard statistical techniques. However, it is not usually a good measure of the certainty of the result, because it fails to take account of systematic errors. The quoted "one-sigma error" is the experimenters' best estimate of the combined statistical and systematic uncertainties. The experimenters estimate a probability of 68 percent that the true result is within "1σ" of their measured value; a probability of 95 percent that it is within "2σ"; etc.

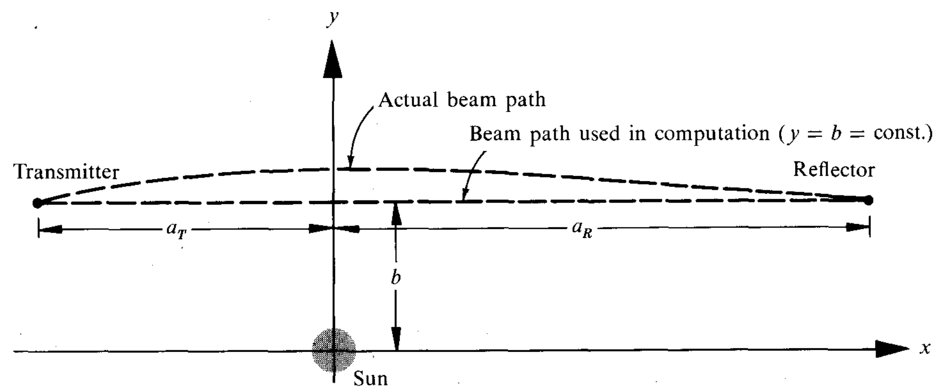


Figure 40.3.
Diagram, in the PPN coordinate system, for the calculation of the relativistic time delay.

(1) foundations for calculation; Fermat's principle

Let a radar transmitter on Earth send a radio wave out to a reflector elsewhere in the solar system, and let the reflector return the wave to Earth. Calculate the round-trip travel time, as measured by a clock on Earth. For simplicity of calculation, idealize both Earth and reflector as nonrotating and as at rest in the static, spherical gravitational field of the Sun. At the end of the calculation, the effects of rotation and motion will be discussed separately. Also ignore time dilation of the transmitter's clock due to the Earth's gravitational field; it is easily corrected for, and it is so small that it will not come into play in these radar experiments before the middle or late 1970's. The gravitational effects of the other planets on the radio waves are too small to be discernible in the foreseeable future, unless the beam grazes the limb of one of them. However, the effects of dispersion in the solar wind and corona are discernible and must be corrected for. These corrections will not be discussed here, since they are free of any general-relativistic influence.

The calculation of the round-trip travel time can be simplified by using a general-relativistic version of Fermat's principle: *In any static field* ($g_{0j} = 0$, $g_{\alpha\beta,0} = 0$) *consider all null curves between two points in space,* $x^j = a^j$ *and* $x^j = b^j$*. Each such null curve,* $x^j(t)$*, requires a particular coordinate time interval* Δt *to get from* a^j *to* b^j*. The curves of extremal* Δt *are the null geodesics of spacetime.* The proof of this theorem is outlined in exercise 40.3.

Because of Fermat's principle, the lapse of coordinate time between transmission of the radar beam and reflection at the reflector, t_{TR}, is the same for a straight path in the PPN coordinates as for the slightly curved path which the beam actually follows. (The two differ by a fractional amount $\Delta t_{\mathrm{TR}}/t_{\mathrm{TR}} \sim$ (angle of deflection)$^2 \lesssim 10^{-12}$, which is far from discernable.) Hence, in the computation one can ignore the gravitational bending of the beam.

(2) details of calculation

Adopt Cartesian PPN coordinates with the sun at the origin; the transmitter, sun, and reflector in the $z = 0$ "plane"; and the transmitter-reflector line along the x direction (see Figure 40.3). The transmitter is at $(x, y) = (-a_T, b)$ in the PPN coordinates, and the reflector is at $(x, y) = (a_R, b)$. Recall that for a null ray $ds^2 = 0 =$

$g_{00}\, dt^2 - g_{xx}\, dx^2$. It follows that the lapse of coordinate time between transmission and reflection is

$$t_{\mathrm{TR}} = \int_{-a_T}^{a_R} \left(\frac{g_{xx}}{-g_{00}}\right)^{1/2} dx = \int_{-a_T}^{a_R} \left[1 + \frac{(1+\gamma)M_\odot}{\sqrt{x^2+b^2}}\right] dx.$$

Integration yields

$$t_{\mathrm{TR}} = a_R + a_T + (1+\gamma)M_\odot \ln\left[\frac{(a_R + \sqrt{a_R^2+b^2})(a_T + \sqrt{a_T^2+b^2})}{b^2}\right]. \tag{40.13}$$

The lapse of coordinate time in round-trip travel has twice this magnitude. The lapse of proper time measured by an Earth-based clock is

(3) formula for delay

$$\begin{aligned} \Delta\tau &= |g_{00}|^{1/2}_{\mathrm{Earth}} 2 t_{\mathrm{TR}}, \\ \Delta\tau &= 2(a_R + a_T)\left(1 - \frac{M_\odot}{\sqrt{a_T^2+b^2}}\right) \\ &\quad + 2(1+\gamma)M_\odot \ln\left[\frac{(a_R + \sqrt{a_R^2+b^2})(a_T + \sqrt{a_T^2+b^2})}{b^2}\right] \end{aligned} \tag{40.14}$$

This is the lapse of time on an Earth-based clock, aside from corrections for the orbital and rotational motion of the Earth, for the orbital motion of the reflector, for dispersion of radiation traversing the solar wind and corona, and for time dilation in the Earth's gravitational field.

(4) comparison with experiment

Any reader is reasonable who objects to the form (40.14) in which the time-delay has been written. The quantities a_R, a_T, and b are coordinate positions in the PPN coordinate system, rather than numbers the astronomer can measure directly. They differ from coordinate positions in other, equally good coordinate systems by amounts of the order of $M_\odot \sim 1.5$ km. The objection is not mathematical in its origin. The quantities a_R, a_T, and b are perfectly well-defined [with post-post-Newtonian uncertainties of order $b(M_\odot/b)^2 \lesssim 10^{-6}$ km], because the PPN coordinate system is perfectly well-defined. But they are not quantities which the experimenter can measure directly, with precision anywhere near that required to test the relativistic terms in the time-delay formula (40.14).

In practice, fortunately, the experimenter does not need to measure a_R, a_T, or b with high precision. Instead, he checks the time-delay formula by measuring the changes in $\Delta\tau$ as the Earth and reflector move in their orbits about the Sun; i.e., he measures $\Delta\tau$ as a function of Earth-based time τ. Notice that when the beam is passing near the sun ($b \ll a_R$, $b \ll a_T$; but $db/d\tau \gg da_R/d\tau$ and $db/d\tau \gg da_T/d\tau$ because the Earth's and reflector's orbits are nearly circular), the change of b in the ln term of (40.14) dominates all other relativistic corrections to the Newtonian delay; consequently (using $db/d\tau \sim 10$ km/sec for typical experiments)

$$\frac{d\Delta\tau}{d\tau} - \begin{pmatrix}\text{Constant Newtonian} \\ \text{part}\end{pmatrix} \approx -4(1+\gamma)\frac{M_\odot}{b}\frac{db}{d\tau} \tag{40.15}$$

$$\sim 4(1+\gamma)\left(\frac{1.5\ \mathrm{km}}{10^6\ \mathrm{km}}\right)\left(\frac{10\ \mathrm{km}}{\mathrm{sec}}\right) \sim \frac{30\ \mu\mathrm{sec}}{\mathrm{day}}.$$

Such differential shifts in round-trip travel time—which rise as the Earth-reflector line moves toward the Sun and falls as it moves away—are readily observable.

In practice, in order to obtain precisions better than about 20 percent in the determination of the parameter γ by time-delay measurements, one must carefully collect and analyze data for a large fraction of a year—from a time when the beam is far from the sun ($b \sim a_T \sim 10^8$ km), to the time of superior conjunction ($b \sim R_\odot \sim 10^6$ km), and on around to a time of distant beam again. Such a long "arc" of data is needed to determine the reflector's orbit with high precision, and to take full advantage of the slow, logarithmic falloff of $\Delta\tau$ with b (40.14). When the beam is far from the sun ($b \gg R_\odot$), the simplifying assumptions behind equation (40.15) are not valid; and the relativistic time-delay gets intertwined with the orbital motions of the Earth and the reflector. The analysis then remains straightforward, but its details are so complex that one resorts to numerical integrations on a computer to carry it out. Because the orbital motions enter, the time-delay data then contain information about other metric parameters (β is the dominant one) in addition to γ.

The experimental results as of 1971 are described in Box 40.2. They yield a value for the PPN parameter γ that is more accurate than the value from light and radio-wave deflection experiments:

(5) experimental result for γ

$$\gamma = 1.02 \pm 0.08. \tag{40.16}$$

Future experiments using spacecraft may improve the precision of γ to ± 0.001 or better.

EXERCISE

Exercise 40.3. FERMAT'S PRINCIPLE

Prove Fermat's principle for a static gravitational field. [*Hint:* The proof might proceed as follows. Write down the geodesic equation in four-dimensional spacetime using an affine parameter λ. Convert from the parameter λ to coordinate time t, and use $ds^2 = 0$ to obtain

$$g_{jk}\frac{d^2x^k}{dt^2} + \Gamma_{jkl}\frac{dx^k}{dt}\frac{dx^l}{dt} - \Gamma_{j00}\frac{g_{kl}}{g_{00}}\frac{dx^k}{dt}\frac{dx^l}{dt} + \frac{d^2t/d\lambda^2}{(dt/d\lambda)^2}g_{jk}\frac{dx^k}{dt} = 0.$$

Combine with the time part of the geodesic equation

$$\frac{d^2t/d\lambda^2}{(dt/d\lambda)^2} = -2\Gamma_{0k0}\frac{dx^k/dt}{g_{00}}$$

and use the expression for the Christoffel symbols in terms of the metric to obtain

$$\gamma_{jk}\frac{d^2x^k}{dt^2} + \frac{1}{2}(\gamma_{jk,l} + \gamma_{jl,k} - \gamma_{kl,j})\frac{dx^k}{dt}\frac{dx^l}{dt} = 0, \qquad \gamma_{jk} \equiv -\frac{g_{jk}}{g_{00}}.$$

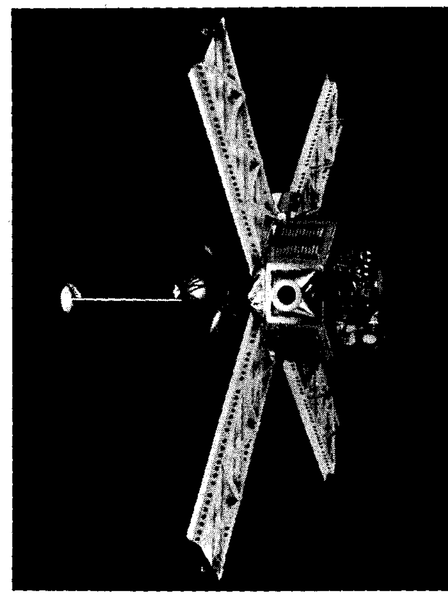
Then notice that this is a geodesic equation with affine parameter t in a three-dimensional manifold with metric γ_{jk}. The familiar extremum principle for this geodesic is

$$\delta\int_{a^j}^{b^j}(\gamma_{jk}\,dx^j\,dx^k)^{1/2} = \delta\int_{a^j}^{b^j}dt = 0,$$

which is precisely Fermat's principle!]

Box 40.2 RADAR TIME DELAY IN THE SOLAR SYSTEM: EXPERIMENTAL RESULTS

Two types of experiments have been performed to measure the relativistic effects [proportional to $\frac{1}{2}(1 + \gamma)$; equation (40.14)] in the round-trip radar travel time in the solar system. In one type (*"passive"* experiment) the reflector is the surface of the planet Venus or the planet Mercury. In the other type (*"active"* experiment) the "reflector" is electronic equipment on board a spacecraft that receives the signal and transmits it back to Earth ("transponder"). Passive experiments suffer from noise due to topography of the reflecting planet (earlier radar return from mountain tops than from valley floors), and they suffer from weakness of the returned signal. Active experiments suffer from buffeting of the spacecraft by solar wind, buffeting by fluctuations in solar radiation pressure, and buffeting by leakage from gas jets ("outgassing"). Experiments of the future will solve these problems by placing a transponder on the surface of a planet or on a "drag-free" (buffeting-free) spacecraft. But experiments of the present and future must both contend with fluctuating time delays due to dispersion in the fluctuating solar wind and corona. Fortunately, these are smaller than the relativistic effects, except when the beam passes within 2 or 3 solar radii of the sun.



The Mariner VI spacecraft (mock-up), which was the reflector in a 1970 measurement of $\frac{1}{2}(1 + \gamma)$ by radar time delay [photo courtesy the Caltech Jet Propulsion Laboratory].

The results of experiments performed before 1972 are listed in the table.

					Experimental result[a]		
Dates of observation	*Radar telescopes*	*Reflector*	*Experimenters and reference*	*Wave length*	$\frac{1}{2}(1+\gamma) = \dfrac{(\text{Observed delay})}{(\text{Einstein prediction})}$	*Formal standard error*	*One-sigma error*
November 1966 to August 1967	Haystack (MIT)	Venus and Mercury	Shapiro (1968)	3.8 cm	0.9		±0.2
1967 through 1970	Haystack (MIT), and Arecibo (Cornell)	Venus and Mercury	Shapiro, Ash, *et al.* (1971)	3.8 cm, and 70 cm.	1.015	±0.02	±0.05
October 1969 to January 1971	Deep Space Network (NASA)	Mariner VI and VII spacecraft	Anderson, *et al.* (1971)	14 cm.	1.00	±0.014	±0.04

[a] Here (observed delay)/(Einstein prediction) is the value of $\frac{1}{2}(1 + \gamma)$ obtained by fitting the observational data, $\Delta\tau(\tau)$, to a more sophisticated version of the PPN prediction (40.14). This more sophisticated version includes the gravitational influences of all the planets on the orbits of reflector and Earth: also the effect of the moon on the Earth's orbit and the effect of the Earth's rotation on the travel time; also, to as good an extent as possible, the delay due to dispersion in the solar corona and wind. "Formal standard error" and "one-sigma error" are defined in the table in Box 40.1.

§40.5. PERIHELION SHIFT AND PERIODIC PERTURBATIONS IN GEODESIC ORBITS

The light-deflection and time-delay experiments both measured γ. To measure other PPN parameters, one must examine the effects of gravity on slowly moving bodies; this was the message of §40.2.

Perihelion shift for geodesic orbits around spherical sun, ignoring preferred-frame effects

Begin with the simplest of cases: the geodesic orbit of a test body in the sun's spherical gravitational field, ignoring all gravitational effects of the planets, of solar oblateness, and of motion relative to any preferred frame. The PPN metric then has the form (40.3):

$$\begin{aligned} ds^2 = &-\left[1 - 2\frac{M_\odot}{r} + 2\beta\frac{M_\odot{}^2}{r^2}\right]dt^2 \\ &+ \left[1 + 2\gamma\frac{M_\odot}{r}\right][dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)]. \end{aligned} \tag{40.3}$$

Orient the coordinates so the test body moves in the equatorial "plane" $\theta = \pi/2$; and calculate the shape $r(\phi)$ of its nearly Keplerian, nearly elliptical geodesic orbit. The result, accurate to order $M_\odot/r$ beyond Newtonian theory, is

$$r = \frac{(1 - e^2)a}{1 + e\cos[(1 - \delta\phi_0/2\pi)\phi]}, \tag{40.17}$$

where a and e are constants of integration, and $\delta\phi_0$ is defined by

$$\begin{aligned} \delta\phi_0 &= \frac{(2 - \beta + 2\gamma)}{3}\frac{6\pi M_\odot}{a(1 - e^2)} \\ &= \frac{6\pi M_\odot}{a(1 - e^2)} \text{ in general relativity.} \end{aligned} \tag{40.18}$$

(For derivation, see exercise 40.4.)

Notice that, if $\delta\phi_0$ were zero—as it is in the Newtonian limit—then the orbit (40.18) would be an ellipse with semimajor axis a and eccentricity e (see Box 25.4). The constant $\delta\phi_0$ merely makes the ellipse precess: for r to go through a complete circuit, from perihelion to aphelion to perihelion again, $(1 - \delta\phi_0/2\pi)\phi$ must change by 2π; so ϕ must change by $2\pi + \delta\phi_0$. Thus, *the perihelion shifts forward by an angle* $\delta\phi_0$ *with each circuit around the ellipse.*

Relative to what does the perihelion shift? (1) Relative to the PPN coordinate system; (2) relative to inertial frames at the outskirts of the solar system (since the PPN coordinates are tied to those frames; see §39.12); (3) relative to a frame determined by the "fixed stars in the sky" (since the inertial frames at the outskirts of the solar system, inertial frames elsewhere in our galaxy, and inertial frames in our cluster of galaxies should not rotate significantly relative to each other); (4) relative to the perihelia of (other) planets, which themselves are shifting at calculable rates that decrease as one moves outward in the solar system from Mercury to Venus to Earth to. . . .

Periodic perturbations in geodesic orbits

The perihelion shift is not the only relativistic effect contained in the orbital motion for a test body. There are other effects, but they are all periodic rather than cumulative with time; so, with the limited technology of the pre-space-age era, it was impossible to detect them. But the technology of the 1970's is bringing them within reach. Moreover, many space-age experiments are necessarily of short duration ($\lesssim$ one orbit)—particularly those involving spacecraft and transponders landed on planets. For these, the periodic perturbations in an orbit are of almost as much experimental value as the cumulative perihelion shift. The periodic effects are not obvious in the PPN orbital equation (40.17); it looks like the simplest of precessing ellipses. But the quantities the observer measures directly are not a, e, and $\delta\phi_0$. Instead, he measures the time evolution of round-trip radar travel times, $\Delta\tau(\tau)$, and of angular positions on the sky $[\theta_0(\tau), \phi_0(\tau)]$. To compute these quantities is perfectly straightforward in principle, but in practice is a very complex task. The calculations predict relativistic effects periodic with the frequency of the orbit and all its harmonics. The amplitudes of these effects, for the lower harmonics, must obviously be of the order of $M_\odot \sim 1$ km $\sim 10\,\mu$sec $\sim 0''.01$ arc on the sky. (The distance $M_\odot = 1.48$ km is the characteristic length for all relativistic effects in the sun's spherical field!)

Comparison of theory with planetary orbits

The most favorable orbits for experimental tests of the perihelion shift and of periodic effects are those that go nearest the sun and have the highest eccentricity [see equation (40.18)]—the orbits of Mercury, Venus, Earth, Mars, and the asteroid Icarus. But how does one know that these orbits are geodesics? After all, planets are not "test bodies"; they themselves produce nonnegligible curvature in spacetime. It turns out (see §40.9 for full discussion) that there should exist tiny deviations from geodesic motion, but they are too small to compete with the perihelion shift or with the periodic effects discussed above, at least for these five bodies.

Extensive astronomical observations of planetary orbits, dating back to the mid-nineteenth century and aided by radar since 1966, have yielded accurate values for planetary perihelion shifts (accurate to ± 0.4 seconds of arc per century for Mercury). From the data, which are summarized and discussed in Box 40.3, one obtains the value

$$\frac{1}{3}(2 - \beta + 2\gamma) = 1.00\begin{cases} +0.01 \\ -0.10 \end{cases} \tag{40.19a}$$

for the ratio of observed relativistic shift to general relativistic prediction. Combining this result with the radar-delay value for γ (40.16), one obtains a value

Experimental result for β

$$\beta = 1.0\begin{cases} +0.4 \\ -0.2 \end{cases} \tag{40.19b}$$

for the PPN parameter β. (Recall: β measures the "amount of nonlinearity in the superposition law for g_{00}.")

The periodic effects in the planetary orbits have not yet (1973) been studied experimentally.

The above discussion and Box 40.3 have ignored the motion of the solar system relative to the preferred frame (if one exists); i.e., they have ignored the terms (40.3′)

Box 40.3 PERIHELION SHIFTS; EXPERIMENTAL RESULTS

Relativistic corrections to Newtonian theory are not the only cause of shift in the perihelion of a planetary orbit. Any departure of the Newtonian gravitational field from its idealized, spherical, inverse-square-law form also produces a shift. Such nonsphericities and resulting shifts are brought about by (1) the gravitational pulls of other planets, and (2) deformation of the sun ("solar oblateness"; "quadrupole moment"). In addition, when the primary data are optical positions of planets on the sky (right ascension and declination as functions of time), there is an apparent perihelion shift caused by the precession of the Earth's axis ("general precession"; observer not on a "stable platform"; see exercise 16.4).

The perihelion shifts due to a general precession and to the gravitational pulls of other planets can be calculated with high precision. But in 1973 there is no fully reliable way to determine the solar quadrupole moment. It is conventional to quantify the sun's quadrupole moment by a dimensionless parameter J_2, which appears in the following expression for the Newtonian potential,

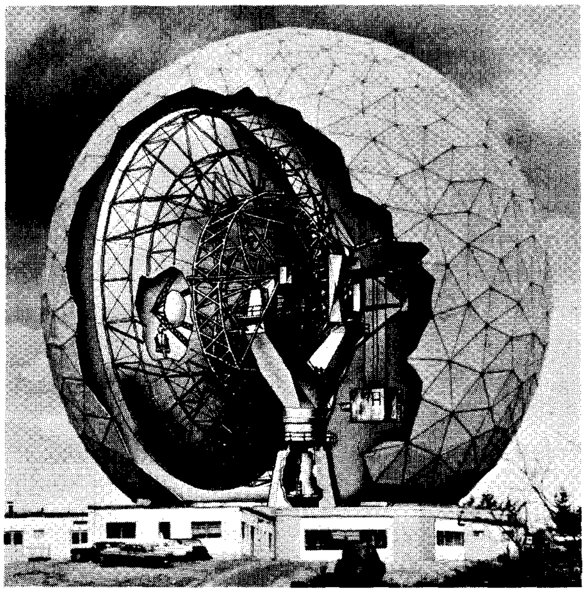
$$U = \frac{M_\odot}{r}\left[1 - J_2 \frac{R_\odot{}^2}{r^2}\left(\frac{3\cos^2\theta - 1}{2}\right)\right].$$

If the sun were rotating near breakup velocity, J_2 would be near 1. Very careful measurements of the optical shape of the sun [Dicke and Goldenberg (1967)] show a flattening, which suggests J_2 may be near 3×10^{-5}.

The total perihelion shift produced by relativity plus solar quadrupole moment is (see exercise 40.5)

$$\delta\phi = \frac{6\pi M_\odot}{a(1-e^2)}\lambda_p,$$

$$\lambda_p \equiv \frac{2 - \beta + 2\gamma}{3} + J_2 \frac{R_\odot{}^2/M_\odot}{2a(1-e^2)}.$$



The Haystack radar antenna, which Irwin Shapiro and his group have used to collect extensive data on the systematics of the inner part of the solar system. Those data are rapidly becoming the most important source of information about perihelion shifts. (Picture courtesy of Lincoln Laboratories, MIT.)

Note that relativistic and quadrupole shifts have different dependences on the semimajor axis a and ecentricity e of the orbit. This difference in dependence allows one to obtain values for both the quadrupole moment parameter J_2, and the PPN parameter $\frac{1}{3}(2 - \beta + 2\gamma)$ by combining measurements of $\delta\phi$ for more than one planet.

The experimental results, as of 1972, are as follows.

I. Data for Mercury from optical studies [Clemence (1943, 1947)]*
(general relativity with no solar oblateness predicts 43.03″/century)

Quantity	Value
(a) Total observed shift per century	5599″.74 ± 0″.41
(b) Contribution to shift caused by observer not being in an inertial frame far from the sun ("general precession" as evaluated in 1947)	5025″.645 ± 0″.50
(c) Shift per century produced by Newtonian gravitation of other planets	531″.54 ± 0″.68
(d) Residual shift per century to be attributed to general relativity plus solar oblateness	42″.56 ± 0″.94
(e) Residual shift if one uses the 1973 value for the "general precession"	41″.4 ± 0″.90
(f) Corresponding value of λ_p (see above)	$\lambda_p = 0.96 \pm 0.02$

II. 1970 Results of Shapiro (1970, 1971a,b), Shapiro *et al.* (1972)

(a) Values of λ_p obtained by reanalyzing all the world's collection of optical data, and combining it with radar data	$\begin{cases} (\lambda_p)_{\text{Mercury}} = 1.00 \pm 0.01 \\ (\lambda_p)_{\text{Mars}} = 1.07 \pm 0.10 \end{cases}$
(b) Value of J_2 obtained by comparing the observed shifts for Mercury and Mars	$J_2 \lesssim 3 \times 10^{-5}$

III. Theoretical implications of Shapiro's results

(a) Value of $(2 - \beta + 2\gamma)/3$	$1.00 \left\{ \begin{matrix} +0.01 \\ -0.10 \end{matrix} \right.$
(b) Value of β obtained by combining with γ from time delay experiments [equation (40.16)]	$1.0 \left\{ \begin{matrix} +0.4 \\ -0.2 \end{matrix} \right.$

*Clemence (1947) notes, "The observations cannot be made in a Newtonian frame of reference. They are referred to the moving equinox, that is, they are affected by the precession of the equinoxes, and the determination of the precessional motion is one of the most difficult problems of observational astronomy, if not the most difficult. In the light of all these hazards, it is not surprising that a difference of opinion could exist regarding the closeness of agreement between the observed and theoretical motions."

in the sun's metric. When one takes account of these terms, one finds an additional contribution to the perihelion shift, given for small eccentricities $e \ll 1$ by

Perihelion shift due to preferred-frame forces

$$\begin{aligned}\delta\phi_0 = &-\alpha_1 \frac{\pi}{2e}\left(\frac{M_\odot}{a}\right)^{1/2} \boldsymbol{w}\cdot\boldsymbol{Q} - \alpha_2 \frac{\pi}{4}[(\boldsymbol{w}\cdot\boldsymbol{P})^2 - (\boldsymbol{w}\cdot\boldsymbol{Q})^2] \\ &+ \alpha_3 \frac{\pi}{e}\left(\frac{|\Omega_\odot|}{M_\odot}\right)\left(\frac{\omega_\odot a^2}{M_\odot}\right)\boldsymbol{w}\cdot\boldsymbol{Q}\end{aligned} \tag{40.20}$$

[see Nordtvedt and Will (1972)]. Here $M_\odot$, $\Omega_\odot$, and $\omega_\odot$ are the sun's mass, self-gravitational energy, and rotational angular velocity; $\boldsymbol{w}$ is the sun's velocity relative to the preferred frame; a and e are the semimajor axis and eccentricity of the orbit; $\boldsymbol{P}$ is the unit vector pointing from the sun to the perihelion; and $\boldsymbol{Q}$ is a unit vector orthogonal to $\boldsymbol{P}$ and lying in the orbital plane. Comparison with observations for

Mercury—and combination with limits on α_1 and α_2 discussed below [equations (40.46b) and (40.48)]—yields the limit

Experimental result for α_3

$$\left|\alpha_3 \frac{\boldsymbol{w}\cdot\boldsymbol{Q}}{200\ \mathrm{km/sec}}\right| \lesssim 2\times 10^{-5}. \tag{40.21a}$$

Since the velocity of the sun around the Galaxy is $\sim$200 km/sec, and the peculiar motion of the Galaxy relative to other nearby galaxies is $\sim$200 km/sec, a value $w \sim 200$ km/sec is reasonable. Moreover, there is no reason to believe that $\boldsymbol{w}$ and $\boldsymbol{Q}$ are orthogonal, so one is fairly safe in concluding

$$|\alpha_3| = |4\beta_1 - 2\gamma - 2 - \zeta| \lesssim 2\times 10^{-5} \tag{40.21b}$$

This is a stringent limit on theories that possess universal rest frames. For example, with great certainty it rules out a theory devised by Coleman (1971), which has $\beta = \gamma = 1$, but $\alpha_3 = -4$; see Ni (1972).

The future of orbital experiments

Looking toward the future, one cannot expect data on orbits of spacecraft to give decisive tests of general relativity, despite the high precision ($\sim$10 meters in 1972) with which spacecraft can be tracked. Spacecraft are buffeted by the solar wind. They respond to fluctuations in this wind and in the pressure of solar radiation, and respond also to "outgassing" from leaky jets. Unless one can develop excellent "drag-free" or "conscience-guided" spacecraft, one must therefore continue to rely on planets as the source of data on geodesics. However, planetary data themselves can be greatly improved in the future by placing radar transponders on the surfaces of planets or in orbit about them, by improvements in radar technology, and by the continued accumulation of more and more observations.

EXERCISES

Exercise 40.4. DERIVATION OF PERIHELION SHIFT IN PPN FORMALISM

[See exercise 25.16 for a derivation in general relativity, accurate when gravity is strong ($2M/r$ as large as $\frac{1}{3}$) but the orbital eccentricity is small. The present exercise applies to any "metric theory" and to any eccentricity, but it assumes gravity is weak ($2M/r \ll 1$) and ignores motion relative to any universal rest frame.] Derive equation (40.17) for the shape of any bound orbit of a test particle moving in the equatorial plane of the PPN gravitational field (40.3). Keep only "first-order" corrections beyond Newtonian theory (first order in powers of $M_\odot/r$). [*Sketch of solution* using Hamilton-Jacobi theory (Box 25.4): (1) Hamilton-Jacobi equation, referred to a test body of unit mass, is

$$\begin{aligned}-1 &= g^{\alpha\beta}\tilde{S}_{,\alpha}\tilde{S}_{,\beta}\\ &= -\left[1 + 2\frac{M_\odot}{r} + (4-2\beta)\left(\frac{M_\odot}{r}\right)^2\right]\left(\frac{\partial\tilde{S}}{\partial t}\right)^2 + \left[1 - 2\gamma\frac{M_\odot}{r}\right]\left[\left(\frac{\partial\tilde{S}}{\partial r}\right)^2 + \frac{1}{r^2}\left(\frac{\partial\tilde{S}}{\partial\phi}\right)^2\right].\end{aligned}$$

(2) Solution to Hamilton-Jacobi equation is

$$\begin{aligned}\tilde{S} = -\tilde{E}t + \tilde{L}\phi \pm \int^r \Big\{&-(1-\tilde{E}^2) + \frac{2M_\odot}{r}[1-(1+\gamma)(1-\tilde{E}^2)]\\ &- \frac{\tilde{L}^2}{r^2}\left[1 - \frac{2M_\odot^{\,2}}{\tilde{L}^2}(2-\beta+2\gamma)\right]\Big\}^{1/2} dr,\end{aligned} \tag{40.22}$$

where post-post-Newtonian corrections have been discarded. In discarding post-post-Newtonian corrections, recall that $\tilde{E}$ is the conserved energy per unit rest mass and $\tilde{L}$ is the angular momentum per unit rest mass (see Box 25.4). Consequently one has the order-of-magnitude relations

$$1 - \tilde{E}^2 \sim (\text{velocity of test body})^2 \sim M_\odot/r$$

and

$$(M_\odot/\tilde{L})^2 \sim (M_\odot/rv)^2 \sim M_\odot/r.$$

(3) The shape of the orbit is determined by the "condition of constructive interference," $\partial\tilde{S}/\partial\tilde{L} = 0$:

$$\phi = \pm \int \left\{ -\frac{1 - \tilde{E}^2}{\tilde{L}^2} + \frac{2M_\odot}{\tilde{L}^2 r}[1 - (1 + \gamma)(1 - \tilde{E}^2)] - \frac{1}{r^2}\left[1 - \frac{2M_\odot{}^2}{\tilde{L}^2}(2 - \beta + 2\gamma)\right]\right\}^{-1/2} d(1/r).$$

(4) This integral is readily evaluated in terms of trigonometric functions. For a bound orbit ($\tilde{E} < 1$), the integral is

$$\phi = \left(1 + \frac{\delta\phi_0}{2\pi}\right)\cos^{-1}\left[\frac{(1 - e^2)a}{er} - \frac{1}{e}\right]$$

where

$$\begin{aligned} a &\equiv \frac{M_\odot}{1 - \tilde{E}^2}[1 - (1 + \gamma)(1 - \tilde{E}^2)], \\ 1 - e^2 &\equiv \left(\frac{\tilde{L}}{M_\odot}\right)^2 (1 - \tilde{E}^2)\left[1 + 2(1 + \gamma)(1 - \tilde{E}^2) - 2\left(\frac{M_\odot}{\tilde{L}}\right)^2 (2 - \beta + 2\gamma)\right], \\ \delta\phi_0 &\equiv \frac{1}{3}(2 - \beta + 2\gamma)6\pi(M_\odot/\tilde{L})^2. \end{aligned} \tag{40.23}$$

(5) Straightforward manipulations bring this result into the form of equations (40.17) and (40.18).]

Exercise 40.5. PERIHELION SHIFT FOR OBLATE SUN

(a) The Newtonian potential for an oblate sun has the form

$$U = \frac{M_\odot}{r}\left(1 - J_2 \frac{R_\odot{}^2}{r^2}\frac{3\cos^2\theta - 1}{2}\right), \tag{40.24}$$

where J_2 is the "quadrupole-moment parameter." One knows that $J_2 \lesssim 3 \times 10^{-5}$. Show that if an oblate sun is at rest at the origin of the PPN coordinate system, the metric of the surrounding spacetime [equations (39.32)] can be put into the form

$$\begin{aligned} ds^2 = &-\left[1 - 2\frac{M_\odot}{r} - 2J_2\left(\frac{M_\odot R_\odot{}^2}{r^3}\right)\left(\frac{3\cos^2\theta - 1}{2}\right) + 2\beta\left(\frac{M_\odot}{r}\right)^2\right]dt^2 \\ &+ \left[1 + 2\gamma\frac{M_\odot}{r}\right][dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)] \\ &+ \text{corrections of post-post-Newtonian magnitude.} \end{aligned} \tag{40.25}$$

(b) Let a test particle move in a bound orbit in the equatorial plane. Use Hamilton-Jacobi theory to show that its orbit is a precessing ellipse [equation (40.17)] with a precession per orbit given by

$$\delta\phi_0 = \frac{2 - \beta + 2\gamma}{3} \frac{6\pi M_\odot}{a(1 - e^2)} + J_2 \frac{3\pi R_\odot{}^2}{a^2(1 - e^2)^2}. \tag{40.26}$$

For the significance of this result, see Box 40.3.

2

The rest of this chapter is Track 2. No earlier Track-2 material is needed as preparation for it, but the following will be helpful:
(1) Chapter 6 (accelerated observers);
(2) §17.6 (no prior geometry); and
(3) Chapters 38 and 39 (tests of foundations; other theories; PPN formalism).
It is not needed as preparation for any later chapter.

§40.6. THREE-BODY EFFECTS IN THE LUNAR ORBIT

The relativistic effects discussed thus far all involve the spherical part of the sun's external gravitational field, and thus they can probe only the PPN parameters β and γ plus the "preferred-frame" parameters α_1, α_2, and α_3. Attempts to measure other PPN parameters can focus on three-body interactions (discussed here), the dragging of inertial frames by a rotating body (§40.7), anomalies in the locally measured gravitational constant (§40.8), and deviations of planetary and lunar orbits from geodesics (§40.9).

3-body effects in lunar orbit:

There is no better place to study three-body interactions than the Earth-moon orbit. The pulls of the Earth, the moon, and the sun all contribute. Perturbations in the motion of Earth and moon about their common center of gravity can be measured with high precision using laser ranging (earth-moon separation measured to $\sim$10 cm in early 1970's) and using a radio beacon on the moon's surface (angular position on sky potentially measurable to better than 0″.001 of arc).

(1) theory

Over and above any Newtonian three-body interactions, the Earth and the sun, acting together in a nonlinear manner, should produce relativistic perturbations in the lunar orbit that are barely within the range of this technology. These effects depend on the familiar parameters γ (measuring space curvature) and β [measuring amount of nonlinear superposition, $(U_{\text{Earth}} + U_{\text{sun}})^2$, in g_{00}]. In addition, they depend on β_2, which regulates the extent to which the sun's potential, U_{sun}, acting inside the Earth, affects the strength of the Earth's gravitational pull, causing it to vary as the Earth moves nearer and farther from the sun. These effects are expected to depend also on ζ, Δ_1, and Δ_2, which regulate the extent to which the Earth's orbital momentum and anisotropies in kinetic energy (caused by the sun) gravitate.

Bromberg (1958), Baierlein (1967), and Krogh and Baierlein (1968) have calculated the three dominant three-body effects in the Earth-moon orbit using general relativity and the Dicke-Brans-Jordan theory. These effects are noncumulative and have amplitudes of $\sim$100 cm, $\sim$20 cm, and $\sim$10 cm. The 100-cm effect [which was originally discovered by de Sitter (1916)] is known to depend only on γ. The precise dependence of the other effects on the PPN parameters is not known.

(2) prospects for measurement

The prospects for measuring these effects in the 1970's are dim; they are masked by peculiarities in the orbit of the moon that have nothing to do with relativity.

2

§40.7. THE DRAGGING OF INERTIAL FRAMES

Gyroscope precession:

The experiments discussed thus far study the motion of electromagnetic waves, spacecraft, planets, and asteroids through the solar system. An entirely different type of experiment measures changes in the orientation of a gyroscope moving in the gravitational field of the Earth. This experiment is particularly important because it can measure directly the "dragging of inertial frames" by the angular momentum of the Earth.

It is useful, before specializing to a rotating Earth, to derive a general expression for the precession of a gyroscope in the post-Newtonian limit. (Track-1 readers, and others who have not studied Chapters 6 and 39, may have difficulty following the derivation. No matter. It is the answer that counts!)

(1) general analysis

Let S^α be the spin of the gyroscope (i.e., its angular momentum vector), and let u^α be its 4-velocity. The spin is always orthogonal to the 4-velocity, $S^\alpha u_\alpha = 0$ (see Box 5.6). *Assume that any nongravitational forces acting on the gyroscope are applied at its center of mass,* so that there is no torque in its proper reference frame. Then the gyroscope will "Fermi-Walker transport" its spin along its world line (see §6.5):

$$\nabla_{\boldsymbol{u}} \boldsymbol{S} = \boldsymbol{u}(\boldsymbol{a} \cdot \boldsymbol{S}), \qquad \boldsymbol{a} \equiv \nabla_{\boldsymbol{u}} \boldsymbol{u} = \text{4-acceleration}. \tag{40.27}$$

The objective of the calculation is to write down and analyze this transport equation in the post-Newtonian limit.

The gyroscope moves relative to the PPN coordinate grid with a velocity $v_j \equiv dx^j/dt \equiv dx_j/dt$. Assume that $v_j \lesssim \epsilon$, where ϵ is the post-Newtonian expansion parameter ($\epsilon^2 \approx M_\odot/R_\odot$). As the gyroscope moves, it carries with itself an orthonormal frame $\boldsymbol{e}_{\hat{\alpha}}$, which is related to the PPN coordinate frame by a pure Lorentz boost, plus a renormalization of the lengths of the basis vectors [transformation (39.41)]. The spin is a purely spatial vector ($S^{\hat{0}} = 0$) in this comoving frame; its length $(S_{\hat{j}} S_{\hat{j}})^{1/2}$ remains fixed (conservation of angular momentum); and its direction is regulated by the Fermi-Walker transport law.

The basis vectors $\boldsymbol{e}_{\hat{\alpha}}$ of the comoving frame are *not* Fermi-Walker transported, by contrast with the spin. Rather, they are tied by a pure boost (no rotation!) to the PPN coordinate grid, which in turn is tied to an inertial frame far from the solar system, which in turn one expects to be fixed relative to the "distant stars." Thus, by calculating the precession of the spin relative to the comoving frame,

$$dS_{\hat{j}}/d\tau \equiv \epsilon_{\hat{j}\hat{k}\hat{l}} \Omega_{\hat{k}} S_{\hat{l}}, \tag{40.28}$$

one is in effect evaluating the spin's angular velocity of precession, $\Omega_{\hat{j}}$, relative to a frame fixed on the sky by the distant stars.

Calculate $dS_{\hat{j}}/d\tau$:

$$\frac{dS_{\hat{j}}}{d\tau} = \nabla_{\boldsymbol{u}}(\boldsymbol{S} \cdot \boldsymbol{e}_{\hat{j}}) = (\nabla_{\boldsymbol{u}} \boldsymbol{S}) \cdot \boldsymbol{e}_{\hat{j}} + \boldsymbol{S} \cdot (\nabla_{\boldsymbol{u}} \boldsymbol{e}_{\hat{j}}) = \boldsymbol{S} \cdot \nabla_{\boldsymbol{u}} \boldsymbol{e}_{\hat{j}}. \tag{40.29}$$

Here use is made of the fact that $\nabla_{\boldsymbol{u}} \boldsymbol{S}$ is in the $\boldsymbol{u}$ direction [equation (40.27)] and is thus orthogonal to $\boldsymbol{e}_{\hat{j}}$. The quantity $\boldsymbol{S}\cdot\boldsymbol{\nabla}_{\boldsymbol{u}}\boldsymbol{e}_{\hat{j}}$ is readily evaluated in the PPN coordinate frame. In the evaluation, one uses as metric coefficients [equations (39.32)] the expressions

$$g_{00} = -1 + 2U + O(\epsilon^4), \qquad g_{jk} = \delta_{jk}(1 + 2\gamma U) + O(\epsilon^4),$$
$$g_{0j} = -\frac{7}{2}\Delta_1 V_j - \frac{1}{2}\Delta_2 W_j + \begin{pmatrix}\text{“preferred-}\\ \text{frame terms”}\end{pmatrix} + O(\epsilon^5); \tag{40.30}$$

one takes as the components of $\boldsymbol{e}_{\hat{j}}$ and $\boldsymbol{S}$ [obtained via the transformation (39.41)] the expressions

$$e_{\hat{j}}{}^0 = v_j + O(\epsilon^3), \qquad e_{\hat{j}}{}^k = (1 - \gamma U)\,\delta_{jk} + \frac{1}{2}v_k v_j + O(\epsilon^4),$$
$$S^0 = v_j S_{\hat{j}} + O(\epsilon^3 S_j), \tag{40.31}$$
$$S^k = (1 - \gamma U)S_{\hat{k}} + \frac{1}{2}v_k v_j S_{\hat{j}} + O(\epsilon^4 S_j);$$

and one uses the relation

$$dv_j/d\tau = a_j + U_{,j} + O(\epsilon^2 U_{,j}) \tag{40.32}$$

where a_j (assumed $\lesssim U_{,j}$) are the components of the 4-acceleration. One finds (see exercise 40.6) for the precession of the spin the result

$$dS_{\hat{j}}/d\tau = \boldsymbol{S}\cdot\boldsymbol{\nabla}_{\boldsymbol{u}}\boldsymbol{e}_{\hat{j}} = S_{\hat{k}}[v_{[j}a_{k]} + g_{0[k,j]} - (2\gamma + 1)v_{[j}U_{,k]}].$$

Rewritten in *three-dimensional vector form* this result becomes

(2) general PPN formula for precession

$$d\boldsymbol{S}/d\tau = \boldsymbol{\Omega}\times\boldsymbol{S}, \tag{40.33a}$$

$$\boldsymbol{\Omega} \equiv -\frac{1}{2}\boldsymbol{v}\times\boldsymbol{a} - \frac{1}{2}\boldsymbol{\nabla}\times\boldsymbol{g} + \left(\gamma + \frac{1}{2}\right)\boldsymbol{v}\times\boldsymbol{\nabla}U, \tag{40.33b}$$

$$\boldsymbol{g} \equiv g_{0j}\boldsymbol{e}_{\hat{j}}. \tag{40.33c}$$

In this final answer it does not matter whether the 3-vectors entering into $\boldsymbol{\Omega}$ are evaluated in the coordinate frame or in the comoving orthonormal frame, since $\boldsymbol{e}_{\hat{j}}$ and $\partial/\partial x_j$ differ only by corrections of order ϵ^2.

Equations (40.33) describe in complete generality at the post-Newtonian level of approximation the precession of the gyroscope spin $\boldsymbol{S}$ *relative to the comoving orthonormal frame that is rotationally tied to the distant stars.*

(3) specialization: Thomas precession

For an electron with spin $\boldsymbol{S}$ in orbit around a proton, only the first term, $-\frac{1}{2}\boldsymbol{v}\times\boldsymbol{a}$, is present (no gravity). This term leads to the Thomas precession, which plays an important role in the fine structure of atomic spectra [see, e.g., Ruark and Urey (1930)]. For other ways of deriving the Thomas precession, see exercise 6.9 and §41.4.

The Thomas precession comes into play for a gyroscope on the surface of the Earth ($\boldsymbol{a}$ = Newtonian acceleration of gravity), but not for a gyroscope in a freely moving satellite.

(4) specialization: precessions due to acceleration and Earth's Newtonian potential

If one ignores the rotation of the Earth and preferred-frame effects, and puts the PPN coordinate frame at rest relative to the center of the Earth, then g_{0j} vanishes and $\boldsymbol{\Omega}$ is given by

$$\begin{aligned}\boldsymbol{\Omega} &= \boldsymbol{v} \times \left[-\frac{1}{2}\boldsymbol{a} + \left(\gamma + \frac{1}{2}\right)\boldsymbol{\nabla} U\right] \\ &= \gamma \boldsymbol{v} \times \boldsymbol{\nabla} U \text{ for gyroscope on Earth's surface} \\ &= \left(\gamma + \frac{1}{2}\right)\boldsymbol{v} \times \boldsymbol{\nabla} U \text{ for gyroscope in orbit.}\end{aligned} \tag{40.34}$$

The general-relativistic term $(\gamma + \frac{1}{2})\boldsymbol{v} \times \boldsymbol{\nabla} U$ is caused by the motion of the gyroscope through the Earth's curved, static spacetime geometry. Notice that it depends solely on the same parameter γ as is tested by electromagnetic-wave experiments. In order of magnitude, for a gyroscope in a near-Earth, polar orbit,

$$\Omega \approx \frac{3}{2}\left(\frac{M_E}{R_E}\right)^{1/2}\left(\frac{M_E}{R_E^2}\right) \approx 8 \text{ seconds of arc per year.} \tag{40.35}$$

The general-relativistic precession $\frac{3}{2}\boldsymbol{v} \times \boldsymbol{\nabla} U$ was derived by W. de Sitter (1916) for the "Earth-moon gyroscope" orbiting the sun. Eleven years later L. H. Thomas (1927) derived the special relativistic precession $-\frac{1}{2}\boldsymbol{v} \times \boldsymbol{a}$ for application to atomic physics.

(5) specialization: precession due to Earth's rotation

The Earth's rotation produces off-diagonal terms, g_{0j}, in the PPN metric (exercise 40.7):

$$\boldsymbol{g} = g_{0j}\boldsymbol{e}_{\hat{j}} = -\left(\frac{7}{4}\Delta_1 + \frac{1}{4}\Delta_2\right)\frac{\boldsymbol{J} \times \boldsymbol{r}}{r^3}. \tag{40.36}$$

Here $\boldsymbol{J}$ is the Earth's angular momentum. These off-diagonal terms contribute an amount

$$\boldsymbol{\Omega} = -\frac{1}{2}\boldsymbol{\nabla} \times \boldsymbol{g} = \left(\frac{7}{8}\Delta_1 + \frac{1}{8}\Delta_2\right)\frac{1}{r^3}\left[-\boldsymbol{J} + \frac{3(\boldsymbol{J}\cdot\boldsymbol{r})\boldsymbol{r}}{r^2}\right] \tag{40.37}$$

to the precession of the gyroscope. Notice that this contribution, unlike the others, is independent of the linear velocity of the gyroscope. One can think of it in the following way.

The gyroscope is rotationally at rest relative to the inertial frames in its neighborhood. It and the local inertial frames rotate relative to the distant galaxies with the angular velocity $\boldsymbol{\Omega}$ because the Earth's rotation "drags" the local inertial frames along with it. Notice that near the north and south poles the local inertial frames rotate in the same direction as the Earth does ($\boldsymbol{\Omega}$ parallel to $\boldsymbol{J}$), but near the equator they rotate in the opposite direction ($\boldsymbol{\Omega}$ antiparallel to $\boldsymbol{J}$; compare $\boldsymbol{\Omega}$ with the magnetic field of the Earth!). Although this might seem paradoxical at first, an analogy devised by Schiff makes it seem more reasonable.* Consider a rotating, solid sphere immersed in a viscous fluid. As it rotates, the sphere will drag the fluid along with it. At various points in the fluid, set down little rods, and watch how the fluid

*This analogy can be made mathematically rigorous; see footnote on p. 255 of Thorne (1971); see also, §21.12 on Mach's principle.

rotates them as it flows past. Near the poles the fluid will clearly rotate the rods in the same direction as the star rotates. But near the equator, because the fluid is dragged more rapidly at small radii than at large, the end of a rod closest to the sphere is dragged by the fluid more rapidly than the far end of the rod. Consequently, the rod rotates in the direction opposite to the rotation of the sphere.

In order of magnitude, the precessional angular velocity caused by the Earth's rotation is

$$\Omega \sim J_E/R_E^3 \sim 0.1 \text{ seconds of arc per year.} \tag{40.37'}$$

(6) prospects for measuring precession

Both this precession, and the larger one [equation (40.35)] due to motion through the Earth's static field, may be detectable in the 1970's. Equipment aimed at detecting them via a satellite experiment is now (1973) under construction at Stanford University; see Everitt, Fairbank, and Hamilton (1970); also O'Connell (1972).*

The gyroscope precession produced by motion of the Earth relative to the preferred frame (if any) is too small to be of much interest.

*The dragging of inertial frames by a rotating body plays important roles elsewhere in gravitation physics, e.g., in the definition of angular momentum for a gravitating body (§19.2), and in black-hole physics (Chapter 33). The effect was first discussed and calculated by Thirring and Lense (1918). More recent calculations by Brill and Cohen (1966) of idealized situations where the effect may be large give insight into the mechanism of the effect. See also the discussion of Mach's principle in §21.12.

EXERCISES

Exercise 40.6. PRECESSIONAL ANGULAR VELOCITY

Derive equations (40.33) for the precession of a gyroscope in the post-Newtonian limit. Base the derivation on equations (40.29)-(40.32).

Exercise 40.7. OFF-DIAGONAL TERMS IN METRIC ABOUT THE EARTH

Idealize the Earth as an isolated, rigidly rotating sphere with angular momentum $\boldsymbol{J}$. Use equations (39.34b,c) and (39.27) to show that (in three-dimensional vector notation)

$$\boldsymbol{V} \equiv V_j \boldsymbol{e}_{\hat{j}} = \boldsymbol{W} \equiv W_j \boldsymbol{e}_{\hat{j}} = \frac{1}{2} \boldsymbol{J} \times \boldsymbol{r}/r^3 \tag{40.38}$$

outside the Earth, in the Earth's PPN rest frame. From this, infer equation (40.36).

Exercise 40.8. SPIN-CURVATURE COUPLING

Consider a spinning body (e.g., the Earth or a gyroscope or an electron) moving through curved spacetime. Tidal gravitational forces produced by the curvature of spacetime act on the elementary pieces of the spinning body. These forces should depend not only on the positions of the pieces relative to the center of the object, but also on their relative velocities. Moreover, the spin of the body,

$$\boldsymbol{S} \equiv \int (\rho \boldsymbol{r} \times \boldsymbol{v})\, d(\text{volume}) \qquad \text{in comoving orthonormal frame,}$$

is a measure of the relative positions and velocities of its pieces. Therefore one expects the spin to couple to the tidal gravitational forces—i.e., to the curvature of spacetime—producing deviations from geodesic motion. Careful solution of the PPN equations of Chapter 39 for general relativity reveals [Papapetrou (1951), Pirani (1956)] that such coupling occurs and causes a deviation of the worldline from the course that it would otherwise take; thus,

$$m\frac{Du^\alpha}{d\tau} = -S_\mu u_\nu \frac{D^2 u_\beta}{d\tau^2}\epsilon^{\alpha\mu\nu\beta} + \frac{1}{2}(\epsilon^{\lambda\mu\rho\tau}R^{\alpha\nu}{}_{\lambda\mu})u_\nu S_\rho u_\tau. \tag{40.39}$$

Evaluate, in order of magnitude, the effects of the supplementary term on planetary orbits in the solar system.

[*Answer:* They are much too small to be detected. However, there are two other material places to look for the effect: (1) when a rapidly spinning neutron star, or a black hole endowed with substantial angular momentum enters the powerful tidal field of another neutron star or black hole; and (2) when an individual electron, or the totality of electrons in the "Dirac sea of negative energy states," enter a still more powerful tidal field (late phase of gravitational collapse). Such a tidal field, or curvature, pulls oppositely on electrons with the two opposite directions of spin [Pirani (1956); DeWitt (1962), p. 338; Schwinger (1963a,b)] just as an electric field pulls oppositely on electrons with the two opposite signs of charge ["vacuum polarization"; see especially Heisenberg and Euler (1936)]. In principle, the tidal field pulling on the spin of an electron need not be due to "background" spacetime curvature; it might be due to a nearby massive spinning object, such as a "live" black hole (chapter 33) ["gravitational spin-spin coupling"; O'Connell (1972)].

§40.8. IS THE GRAVITATIONAL CONSTANT CONSTANT?

The title and subject of this section are likely to arouse confusion. Throughout this book one has used geometrized units, in which $G = c = 1$. Therefore, one has locked oneself into a viewpoint that forbids asking whether the gravitational constant changes from event to event.

False! One can perfectly well ask the question in the context of $G = c = 1$, if one makes clear what is meant by the question.

In §§1.5 and 1.6, c was *defined* to be a certain conversion factor between centimeters and seconds; and G/c^2 was defined to be a certain conversion factor between grams and centimeters. These definitions by fiat do not guarantee, however, that a Cavendish experiment* to measure the attraction between two bodies will yield

$$\text{Force} = -Gm_1m_2/r^2 = -m_1m_2/r^2.$$

If general relativity correctly describes classical gravity, and if the values of the conversion factors G and c have been chosen precisely right, then any Cavendish experiment, anywhere in the universe, *will* yield "Force $= -m_1m_2/r^2$". But if the

*See any standard textbook for a description of Cavendish experiments. By his original version of the experiment, with two separated spheres suspended by fine wires, Henry Cavendish (1798) inferred the mass and hence the density of the Earth. He reported: "By a mean of the experiments made with the wire first used, the density of the Earth comes out 5.48 times greater than that of water; and by a mean of those made with the latter wire it comes out the same; and . . . the extreme results do not differ from the mean more than 0.38, or 1/14 of the whole." The most precise method of measuring G today [Rose *et al.* (1969)] gives $G_C = (6.674 \pm .004) \times 10^{-8}\ \text{cm}^3/\text{g sec}^2$ (one standard deviation).

Dicke-Brans-Jordan theory, or almost any other metric theory gives the correct description of gravity, the force in the Cavendish experiment will depend on where and when the experiment is performed, as well as on m_1, m_2, and r. To discuss Cavendish experiments as tests of gravitation theory, then, one must introduce a new proportionality factor

"Cavendish gravitational constant," G_C, defined

$$G_C \equiv G_{\text{Cavendish}} \equiv (\text{“Cavendish gravitational constant”}), \tag{40.40}$$

which enters into the Newtonian force law

$$\text{Force} = -G_C m_1 m_2 / r^2. \tag{40.41}$$

This Cavendish constant will be unity in general relativity, but in most other metric theories it will vary from event to event in spacetime.

Changes of G_C with time

In some theories, such as Dicke-Brans-Jordan, the Cavendish constant is determined by the distribution of matter in the universe. As a result, the expansion of the universe changes its value:

$$\frac{1}{G_C}\frac{dG_C}{dt} \sim -\left(\frac{0.1 \text{ to } 1}{\text{age of universe}}\right) \sim \frac{-1}{10^{10} \text{ or } 10^{11} \text{ years}}$$

[see, e.g. Brans and Dicke (1961)]. A variety of observations place limits on such time variations. Big time changes in G_C during the last 4.6 billion years would have produced marked effects on the Earth, the sun, and the entire solar system. The expected geophysical effects have been summarized and compared with observations by Dicke and Peebles (1965). It is hard to draw firm conclusions because of the complexity of the geophysics involved, but a fairly certain limit is

$$(1/G_C)(dG_C/dt) \lesssim 1/10^{10} \text{ years} \qquad \text{(geophysical)}. \tag{40.42a}$$

Eventually, high-precision measurements of the orbital motions of planets will yield a better limit. For the present, planetary observations show

$$(1/G_C)(dG_C/dt) \lesssim 4/10^{10} \text{ years} \qquad \text{(planetary orbits)} \tag{40.42b}$$

[Shapiro, Smith, *et al.* (1971)]. These limits are tight enough to begin to be interesting, but not yet tight enough to disprove any otherwise viable theories of gravity.

Spatial variations in G_C

If G_C is determined by the distribution of matter in the universe, then it should depend on where in the universe one is, as well as when. In particular, as one moves from point to point in the solar system, closer to the Sun and then farther away, one should see G_C change. Indeed this is the case in most metric theories of gravity, though not in general relativity. Analyses of Cavendish experiments using the PPN formalism reveal spatial variation in G_C given by

$$\Delta G_C = -2G_C(\beta + \gamma - \beta_2 - 1)U \tag{40.43}$$

[Nordtvedt (1970, 1971a); Will (1971b)].

The amplitude of these variations along the Earth's elliptical orbit is $\Delta G_C/G_C \sim 10^{-10}$, if $\beta + \gamma - \beta_2 - 1 \sim 1$. This is far too small to measure directly in the

1970's. Despite great ingenuity and effort, the most accurate experiments measuring the value of G_C have precisions in 1972 no better than 1 part in 10^4 [see Beams (1971)]. Experiments to search for yearly variations in G_C on Earth without measuring the actual value ("null-type experiments") can surely be performed with better precision than 1 in 10^4—but not with precisions approaching 1 in 10^{10}. On the other hand, *indirect* consequences of a spatial variation of G_C in the solar system are almost certainly measurable (see §40.9 below).

Dependence of G_C on velocity

In Ni's theory of gravity (Box 39.1), and other two-tensor or vector-tensor theories like it, where the prior geometry picks out a preferred "universal rest frame," the Cavendish constant G_C can depend on velocity relative to the preferred frame. For Cavendish experiments with two equal masses separated by distances large compared to their sizes, G_C varies as

$$\Delta G_C = G_C\left[\frac{1}{2}(\alpha_2 + \alpha_3 - \alpha_1)\boldsymbol{v}^2 - \frac{1}{2}\alpha_2(\boldsymbol{v}\cdot\boldsymbol{n})^2\right] \tag{40.44}$$

[Will (1971b)]. Here $\boldsymbol{v}$ is the velocity of the Cavendish apparatus relative to the preferred frame, and $\boldsymbol{n}$ is the unit vector between the two masses. For experiments where one body is a massive sphere (e.g., the Earth), and the other is a small object on the sphere's surface, G_C varies as

$$\begin{aligned}\Delta G_C/G_C = {}&\frac{1}{2}[(\alpha_3 - \alpha_1) + \alpha_2(1 - I/MR^2)]\boldsymbol{v}^2\\ &- \frac{1}{2}\alpha_2(1 - 3I/MR^2)(\boldsymbol{v}\cdot\boldsymbol{n})^2\end{aligned} \tag{40.44$'$}$$

[Nordtvedt and Will (1972)]. Here M and R are the mass and radius of the sphere, and

$$I = \int (\rho r^2) 4\pi r^2\, dr$$

is the trace of the second moment of its mass distribution. Consequences of these effects for planetary orbits have not yet been spelled out, but consequences for Earthbound experiments have.

Anomalies in Earth tides due to anisotropies in G_C:

Think of a Cavendish experiment in which one mass is the Earth, and the other is a gravimeter on the Earth's surface. The gravimeter gives a reading for the "local acceleration of gravity,"

$$g = G_C m_{\text{Earth}}/r_{\text{Earth}}{}^2. \tag{40.45}$$

As the Earth turns, so the unit vector $\boldsymbol{n}$ between its center and the gravimeter rotates, G_C and hence g will fluctuate with a period of 12 sidereal hours and an amplitude

$$(\Delta g/g)_{\text{amplitude}} = \frac{1}{4}\alpha_2 \boldsymbol{v}^2 \cos^2\theta_m.$$

Here θ_m is the minimum, as the Earth rotates, of the angle between $\boldsymbol{v}$ (constant vector) and $\boldsymbol{n}$ (rotating vector). (Note: we have used the value $I/MR^2 \simeq 0.5$ for the Earth.) These fluctuations will produce tides in the Earth of the same type as are produced by the moon and sun. As of 1972, gravimeter measurements near the Earth's equator show no sign of any anomalous 12-sidereal-hour effects down to an amplitude of $\sim 10^{-9}$ [Will (1971b)]. Consequently,

(1) experimental value of α_2

$$|\alpha_2|^{1/2} v \cos\theta_m = |\Delta_2 + \zeta - 1|^{1/2} v \cos\theta_m \lesssim 6 \times 10^{-5} \sim 20 \text{ km/sec}. \quad (40.46\text{a})$$

Using a rough estimate of $v \sim 200$ km/sec for the Earth's velocity relative to the universal rest frame, and $\theta_m \lesssim 60°$ for the angle between $\boldsymbol{v}$ and the Earth's equatorial plane, one obtains the rough limit

$$|\alpha_2| = |\Delta_2 + \zeta - 1| \lesssim 0.03. \quad (40.46\text{b})$$

[This limit does not affect the three theories in Box 39.1; of them, only Ni's theory has prior geometry and a universal rest frame; and it predicts isotropic effects in $\Delta G_C/G_C$ [equation (40.44)], but no anisotropic effects. However, other theories with universal rest frames—e.g. Papapetrou's (1954a,b,c) theory—are ruled out by this limit; see Ni (1972), Nordtvedt and Will (1972).]

(2) experimental disproof of Whitehead theory

Whitehead's theory of gravity (which is a two-tensor theory with a rather different type of prior geometry from Ni's) predicts that the galaxy should produce velocity-independent anisotropies in G_C. These, in turn, would produce Earth tides with periods of 12 sidereal hours and amplitudes of

$$\Delta g/g \sim 2 \times 10^{-7} \sim 100 \times \begin{pmatrix}\text{experimental limit on}\\ \text{such amplitudes}\end{pmatrix}$$

[Will (1971b)]. The absence of such tides proves Whitehead's theory to be incorrect—a feat of disproof beyond the power of all redshift, light-deflection, time-delay, and perihelion-shift measurements. (For all these "standard experiments," the predictions of Whitehead and Einstein are identical!)

Anomalies in Earth rotation rate due to dependence of G_C on velocity

Equation (40.44′) predicts a periodic annual variation of the Cavendish constant on Earth, as the Earth moves around the sun:

$$\boldsymbol{v} = \begin{pmatrix}\text{velocity of Earth}\\ \text{relative to sun}\end{pmatrix} + \begin{pmatrix}\text{velocity of sun relative}\\ \text{to preferred frame}\end{pmatrix} \equiv \boldsymbol{v}_E + \boldsymbol{w};$$

$$(\Delta G_C/G_C)_{\substack{\text{averaged over}\\ \text{Surface of Earth}}} = \frac{1}{2}\left(\frac{2}{3}\alpha_2 + \alpha_3 - \alpha_1\right)\left(\boldsymbol{w}^2 + \boldsymbol{v}_E{}^2 + \underbrace{2\boldsymbol{w}\cdot\boldsymbol{v}_E}_{\text{[varies sinusoidally with period of one year]}}\right). \quad (40.47)$$

This annual variation, assuming all PPN parameters are of order unity, is 1,000 times larger than the one produced by the Earth's motion in and out through the sun's gravitational potential [equation (40.43)]. In response to this changing Cavendish constant, the Earth's self-gravitational pull should change, and the Earth should "breathe" inward (greater pull) and outward (relaxed pull). The resulting annual variations in the Earth's moment of inertia should produce annual changes in its rotation rate ω (changes in "length of day" as measured by atomic clocks):

$$\delta\omega/\omega \sim 0.1\left(\frac{2}{3}\alpha_2 + \alpha_3 - \alpha_1\right)\boldsymbol{w}\cdot\boldsymbol{v}_E$$

[Nordtvedt and Will (1972)]. Comparison with the measured annual variations of rotation rate (all of which geophysicists attribute to seasonal changes in the Earth's atmosphere) yields the following limit

$$\left|\frac{2}{3}\alpha_2 + \alpha_3 - \alpha_1\right| = \lesssim 0.2. \tag{40.48}$$

Experimental value of $\frac{2}{3}\alpha_2 + \alpha_3 - \alpha_1$

[See Nordtvedt and Will (1972)]. This limit rules out several preferred-frame theories of gravity, including that of Ni (Boxes 39.1 and 39.2).

The experimental results (40.21), (40.46), and (40.48), when combined, place the following very rough limits on any theory that possesses a Universal rest frame:

$$\begin{aligned} |\alpha_1| &= |7\,\Delta_1 + \Delta_2 - 4\gamma - 4| \lesssim 0.2, \\ |\alpha_2| &= |\Delta_2 + \zeta - 1| \lesssim 0.03, \\ |\alpha_3| &= |4\beta_1 - 2\gamma - 2 - \zeta| \lesssim 2 \times 10^{-5}. \end{aligned} \tag{40.49}$$

These limits completely disprove all theories with preferred frames that have been examined to date except one devised by Will and Nordtvedt [see Ni (1972); Nordtvedt and Will (1972)].

Dependence of G_C on chemical composition

In some theories of gravity, the result of a Cavendish experiment depends on the chemical composition and internal structure of the test bodies (exercises 40.9 and 40.10). Kruezer (1968) has performed the most accurate search for such effects to date. He finds that G_C is the same for fluorine and bromine to a precision of

$$\left|\frac{G_C(\text{bromine}) - G_C(\text{fluorine})}{G_C}\right| \lesssim 5 \times 10^{-5}. \tag{40.50}$$

EXERCISES

Exercise 40.9. CAVENDISH CONSTANT FOR IDEALIZED SUN

Idealize the sun as a static sphere of perfect fluid at rest at the origin of the PPN coordinates. Then its external gravitational field has the form (40.3), with $M_\odot$ given by (40.4). Consequently, a test body of mass m, located far away at radius r, is accelerated by a gravitational force

$$\text{Force} = -mM_\odot/r^2. \tag{40.51a}$$

(a) Calculate the mass of the sun, M, in the sense of the amount of energy required to construct it by adding one spherical shell of matter on top of another, working from the inside outward. [*Answer:*

$$\begin{aligned} M &= \underbrace{\int_0^{R_\odot} \rho_0(1 + \Pi + 3\gamma U)4\pi r^2\,dr}_{\text{rest mass + internal energy}} - \underbrace{\frac{1}{2}\int_0^{R_\odot} \rho_0 U 4\pi r^2\,dr}_{\text{gravitational potential energy}} \\ &= \int_0^{R_\odot} \rho_0\left[1 + \Pi + \left(3\gamma - \frac{1}{2}\right)U\right]4\pi r^2\,dr.] \end{aligned} \tag{40.51b}$$

(b) Use the virial theorem [equation (39.21b)] to rewrite equation (40.4) in the form

$$M_\odot = \int_0^{R_\odot} \rho_0\left[1 + \beta_3\Pi + \left(2\beta_2 + \frac{1}{2}\beta_4\right)U\right]4\pi r^2\,dr. \tag{40.51c}$$

(c) Combine the above equations with the definition

$$\text{Force} = -G_C mM/r^2 \tag{40.51d}$$

of the Cavendish constant for r far outside the sun, to obtain

$$G_C = \frac{\begin{pmatrix}\text{mass of sun as defined by its effect in}\\ \text{bending world line of a faraway test particle}\end{pmatrix}}{\begin{pmatrix}\text{mass-energy as defined by applying law of}\\ \text{conservation of energy to the steps in the}\\ \text{construction of the sun}\end{pmatrix}} \tag{40.52}$$

$$= 1 + \int (\rho_0/M_0)[(\beta_3 - 1)\Pi + \frac{1}{2}(4\beta_2 + \beta_4 - 6\gamma + 1)U]4\pi r^2\, dr.$$

Unless $\beta_3 = 1$, and $4\beta_2 + \beta_4 - 6\gamma + 1 = 0$ (as they are, of course, in Einstein's theory), G_C will depend on the sun's internal structure! Specialize equation (40.52) to "conservative theories of gravity (Box 39.5), and explain why the result is what one would expect from equation (40.43).

Exercise 40.10. CAVENDISH CONSTANT FOR ANY BODY

Extend the analysis of exercise 40.9 to a source that is arbitrarily stressed and has arbitrarv shape and internal velocities (subject to the constraints $v^2 \ll 1$, $|t_{\hat{j}\hat{k}}|/\rho_0 \ll 1$, $U \ll 1$, $\Pi \ll 1$, of the post-Newtonian approximation). Assume that the body is at rest relative to the universal rest frame. Show that G_C depends on the internal structure of the source unless

$$2\beta_1 - \beta_4 = 1, \quad 4\beta_2 + \beta_4 - 6\gamma = -1, \quad \beta_3 = 1, \quad \zeta = 0, \quad \eta = 0. \tag{40.53}$$

Of course, these PPN constraints are all satisfied by Einstein's theory.

§40.9. DO PLANETS AND THE SUN MOVE ON GEODESICS?

Crucial to solar-system experiments is the question of whether the sun and the planets move on geodesics of spacetime. This question is complicated by the contributions to the spacetime curvature made by the moving body itself.

The sense in which general relativity predicts geodesic motion for planets and sun

To elucidate the question—and to obtain an answer in the framework of general relativity—consider an "Einstein elevator" type of argument. The astronomical object under consideration has an outer boundary, and each point on this boundary describes a world line. These world lines define a world tube. Some distance outside of this world tube construct a "buffer zone" as in §20.6. Tailor its inner and outer dimensions, according to the mass and moments of the object and the curvature of the enveloping space ("strength of the tide-producing force of the external gravitational field"), in such a way that the departure ϵ (cf. §20.6) of the metric from flatness in this buffer zone takes on values equal at most to twice the extremal achievable value ϵ_{extrem} (a minimum with respect to variations in r, a maximum with respect to variations in direction; in other words, a minimax). Then, apart from errors of order ϵ_{extrem}, the object can be regarded as moving in an asymptotically flat space. The law of conservation of total 4-momentum applies. It assures one that the object moves in a (locally) straight line with uniform velocity. Consider, next, a "background geometry" that agrees just outside the buffer zone with the actual geometry to accuracy ϵ_{extrem} or better, but that inside is a source-free solution of Einstein's field equation. Then, to an accuracy governed by the magnitude of ϵ_{extrem}, the locally straight line along which the astronomical object moves will be a geodesic of this background geometry.

Deviations from geodesic motion:

Insofar as one can give any well-defined meaning to the departure of the actual motion from this geodesic (a task complicated by the fact that the background geometry does not actually exist!), one can calculate this departure by making use of the PPN formalism or some other approximation scheme [see, e.g., Taub (1965)]. This deviation springs ordinarily in substantial measure, and sometimes almost wholly, from a coupling between the Riemann curvature tensor of the external field and the multipole moments of the astronomical object (angular momentum associated with rotation; quadrupole and higher moments associated with deformation; see, e.g., exercises 40.8 and 16.4). This coupling is important for the Earth-moon system, but one need not use relativity to calculate it; Newtonian theory does the job to far greater accuracy than needed—or would, if one understood the interiors of the Earth and the moon well enough! For the planets and sun, the effect is negligible. (Exercise: use Newtonian theory to prove so!).

(1) due to curvature coupling

Thus, in general relativity as applied to the solar system, one can approximate the orbit of the sun, the Earth-moon mass center, and each other planet, as a geodesic of that "background spacetime geometry" which would exist if its own curvature effects were absent. This is the approach used to analyze the perihelion shift for planets in §40.5 in the context of general relativity, and to derive in exercise 39.15 the post-Newtonian "many-body equations of motion."

In most other metric theories of gravity, including the Dicke-Brans-Jordan theory, there are substantial departures from geodesic motion. The "Einstein elevator" argument fails in these theories because spacetime is endowed not only with a metric, but also with a long-range field that couples indirectly (cf. §§38.7 and 39.2) to massive, gravitating bodies.

(2) due to spatial dependence of gravitational constant (Nordtvedt effect)

This phenomenon is best understood in terms of Dicke's argument about the influence of spatial variations of the fundamental constants on experiments of the Eötvös-Dicke type (see §38.6). In a theory where the Cavendish gravitational constant G_C depends on position (as it does not and cannot in general relativity), a body with significant self-gravitational energy E_{grav} must fall, in a perfectly uniform external Newtonian gravitational field, with an anomalous acceleration:

$$\begin{aligned}\begin{pmatrix}\text{acceleration of}\\ \text{massive body}\end{pmatrix} - \begin{pmatrix}\text{acceleration of}\\ \text{test body}\end{pmatrix} &= \frac{1}{M}\left(\frac{\partial E_{\text{grav}}}{\partial G_C}\right)\nabla G_C \\ &= \frac{E_{\text{grav}}}{MG_C}\nabla G_C\end{aligned} \tag{40.54}$$

[see equation (38.15)]. In Dicke-Brans-Jordan theory, G_C is essentially the reciprocal of the scalar field; and it contains a small part that is proportional to the Newtonian potential, U [equation (40.43) with the appropriate values of the parameters from Box 39.2]. As a result, the sun falls with an acceleration smaller by one part in 10^6 than the acceleration of a test body; Jupiter falls with an acceleration one part in 10^9 smaller; and the Earth, one part in 10^{10} smaller. Translated into relativistic language: the scalar field, by influencing the gravitational self-energy of a massive body, produces deviations from geodesic motion.

One can use the full PPN formalism of Chapter 29 to calculate the motion of massive bodies in any metric theory of gravity. Nordtvedt (1968b) and Will (1971a) have done this. They find that a massive body at rest in a uniform external field experiences a (Newtonian-type) PPN coordinate acceleration given by

$$\frac{d^2x_j}{dt^2} = E_{jk}\frac{\partial U}{\partial x_k},$$

where E_{jk} is a quantity depending on the body's structure:

$$E_{jk} = \delta_{jk}\left\{1 - (7\,\Delta_1 - 3\gamma - 4\beta)\frac{E_{\text{grav}}}{m}\right\} - (2\beta + 2\beta_2 - 3\gamma + \Delta_2 - 2)\frac{\Omega_{jk}}{m},$$

$$\Omega_{jk} = -\frac{1}{2}\int\frac{\rho_o\rho_o'(x_j - x_j')(x_k - x_k')}{|\boldsymbol{x} - \boldsymbol{x}'|^3}\,d^3x\,d^3x', \qquad E_{\text{grav}} = \sum \Omega_{jj}. \tag{40.55}$$

Here m is the body's total mass-energy, Ω_{jk} is the "Chandrasekhar potential-energy tensor," and E_{grav} is the body's self-gravitational energy. [Note: Dicke's method of calculating the anomalous acceleration (40.54) breaks down in theories that are not "conservative" (Box 39.5).]

In general relativity, the combinations of PPN coefficients appearing in E_{jk} vanish; so $E_{jk} = \delta_{jk}$, and the body falls with the usual acceleration—i.e., it moves along a geodesic. But in most other theories of gravity $E_{jk} \neq \delta_{jk}$; the body does *not* move on a geodesic; and its acceleration may even be in a different direction than the gradient of the Newtonian potential!

This predicted departure from geodesic motion is called the *"Nordtvedt effect."* The possibility of such an effect was first noticed in passing by Dicke (1964c), but was discovered independently and explored in great detail by Nordtvedt (1968a,b). The Nordtvedt effect in a theory other than general relativity produces a number of phenomena in the solar system that are potentially observable. [See Nordtvedt (1971b) for an enumeration and references.] The effect most suitable for a test is a "polarization" of the Earth-moon orbit due to the fact that the moon should fall toward the sun with a greater acceleration than does the Earth. This "polarization" results in an eccentricity in the orbit that points always toward the Sun and has the amplitude

Nordtvedt effect in Earth-moon orbit

$$\delta r = 840\left[3\gamma + 4\beta - 7\,\Delta_1 - \frac{1}{3}(2\beta + 2\beta_2 - 3\gamma + \Delta_2 - 2)\right]\text{cm} \tag{40.56}$$

$= 67$ meters in Ni's theory (Boxes 39.1 and 39.2)

$= \dfrac{8.4}{2+\omega}$ meters in Dicke-Brans-Jordan theory (Boxes 39.1 and 39.2)

$= 0$ in Einstein's theory.

Box 40.4 CATALOG OF EXPERIMENTS

Type of experiment	*Description of experiment*	*Where discussed*
I. Tests of foundations of general relativity	1. Tests of uniqueness of free fall (Eötvös-Dicke-Braginsky experiments	§38.3; Figure 1.6; Box 1.1
	2. Tests for existence of metric (time dilation of particle decays; role of Lorentz group in particle kinematics; etc.)	§38.4
	3. Searches for new, direct-coupling, long-range fields (Hughes-Drever experiment; ether-drift experiments)	§38.7; Figure 38.3
	4. Gravitational redshift experiments	§38.5; Figures 38.1 and 38.2; §§7.2, 7.3, and 7.4
	5. Constancy, in space and time, of the nongravitational physical constants	§38.6
II. Post-Newtonian ("solar-system") experiments	1. Deflection of light and radio waves by Sun	§40.3; Box 40.1
	2. Relativistic delay in round-trip travel time for radar beams passing near Sun	§40.4; Box 40.2
	3. Perihelion shifts and periodic perturbations in planetary orbits	§40.5; Box 40.3
	4. Three-body effects in the Lunar orbit	§40.6
	5. Precession of gyroscopes ("geodetic precession" and precession due to dragging of inertial frames by Earth's rotation)	§40.7
	6. Spatial variation of the Cavendish gravitational constant in the solar system	§§40.8 and 40.9
	7. Dependence of the Cavendish gravitational constant on the chemical composition of the gravitating body	§40.8
	8. Earth tides with sidereal periods	§40.8
	9. Annual variations in Earth rotation rate	§40.8
	10. Periodicities in Earth-Moon separation due to breakdown of geodesic motion	§40.9
III. Cosmological observations	1. Change of Cavendish gravitational constant with time in solar system	§40.8
	2. Large-scale features of universe (expansion, isotropy, homogeneity; existence and properties of cosmic microwave radiation; . . .)	Chapters 27–30; especially Chapter 29
	3. Agreement of various measures of age of universe (age from expansion; ages of oldest stars; age of solar system)	§29.7
IV. Gravitational-Wave experiments	Existence of waves; propagation speed; polarization properties; . . .	Chapters 35–37; especially Chapter 37

2

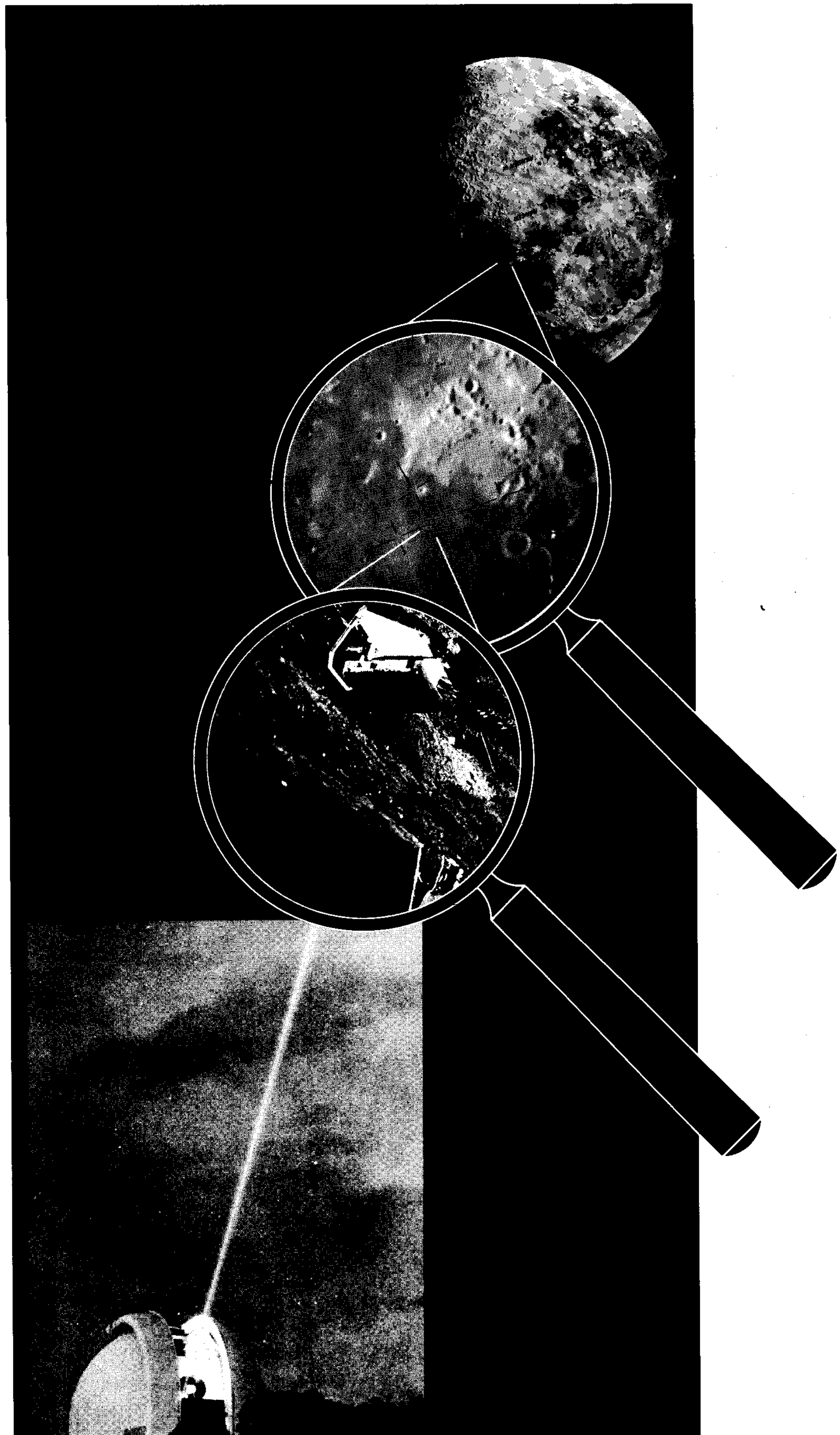


Figure 40.4. ***(facing page)***
Measuring the separation between earth and moon by determining the time-delay (about 2.5 sec) between the emission of light from a laser on the earth and the return of this light to the earth. A key element in the program is a corner reflector, the first of which was landed on the moon July 20, 1969, by the Apollo 11 flight crew. In November 1971, there were three such reflectors on the moon: two American, and one French-built and Soviet-landed. A pulsed ruby laser projects a beam out of the 107-inch reflecting telescope of the McDonald Observatory of the University of Texas, on Mount Locke, 119 miles east of El Paso. This beam makes a spot of light on the moon's surface about 3.2 km in diameter. Laser light is bounced straight back to the earth by the "laser ranging retroreflectors" (LR^3). Each consists of an aluminum panel of 46 cm by 46 cm with 100 fused silica corner cubes each 3.8 cm in diameter. The first reflector ever set up appears in the first inset, near the lunar landing module. It is tilted with respect to the landscape of the moon. The photograph was made shortly before astronauts Neil A. Armstrong and Edwin E. Aldrin, Jr., took off for the earth. The second inset is a photograph made by D. G. Currie of the field of view in the guiding eyepiece of the McDonald 107-inch telescope in an interval when the laser was not firing at the Apollo 11 site. One guides the telescope to Tranquility Base (small circle) by aligning fiducial marks on more visible moonscape features. In November 1971, the LR^3 experiment and continuing time-of-flight measurements were the responsibility of the National Aeronautics and Space Administration and a Lunar Retroreflecting Ranging Team of representatives from several centers of research. One of the members of this team, Carroll Alley, of the University of Maryland, is hereby thanked for his kindness in providing the photographs used in this montage. Thanks to this NASA work, the distance between the laser source on the earth and the reflectors on the moon is known with an accuracy now better than half a meter. The astronauts left behind on the moon not only LR^3 and a seismometer and other equipment, but also a plaque: "We came in peace for all mankind."

By the mid 1970's, lunar laser-ranging data will probably be able to determine the amplitudes of this polarization to a precision of one meter or better [see Bender *et al.* (1971); also Figure 40.4].

§40.10. SUMMARY OF EXPERIMENTAL TESTS OF GENERAL RELATIVITY

No longer is general relativity "a theorist's Paradise, but an experimentalist's Hell." It is now a Paradise for all—as one can see quickly by perusing the catalog of experiments given in Box 40.4 on page 1129. Moreover, general relativity has emerged from each of its tests unscathed—a remarkable 1973 tribute to the 1915 genius of Albert Einstein.

PART **X**

FRONTIERS

Wherein the reader—who, during a life of continued variety for forty chapters (besides the Preface), was eight chapters a mathematician, four times enticed (once by an old friend), four chapters a cosmologist, and four chapters a transported astrophysicist in the land of black holes, and who at last inherited a wealth of experiments, lived honest, and became a True Believer—now ventures forth in search of new frontiers to conquer.

CHAPTER 41

SPINORS

§41.1. REFLECTIONS, ROTATIONS, AND THE COMBINATION OF ROTATIONS

This chapter is entirely Track 2. No earlier Track-2 material is needed as preparation for it, nor is it needed as preparation for any later chapter.

Spinors and their applications in relativity grew out of the analysis of "rotations," first in space, then in spacetime. Take a cube (Figure 41.1). Rotate it about one axis through 90°. Then pick another axis at right angles to the first. About it rotate the cube again through 90°. In this way the cube is carried from the orientation marked "Initial" to that marked "Final." How can one make this net transformation in a single step, with a single rotation? In other words, what is the law for the combination of rotations?

The problem of combining rotations

Were rotations described by vectors, then one could apply the law of combination of vectors. The resultant of two vectors of the same magnitude (90°) separated by a right angle, is a single vector that (1) lies in the same plane and (2) has the magnitude $2^{1/2} \times 90° = 127.28°$. Both predictions are wrong. To turn the cube from initial to final orientation in a single turn, (1) take an axis running from the center through the vertex A and (2) rotate through 120°.

What computational algorithm can ever reproduce a law of combination of rotations apparently so strange? On the evening of October 16, 1843, William Rowan Hamilton was walking with his wife along the Royal Canal in Dublin when the answer leaped to his mind, the fruit of years of reflection. With his knife he then and there carved on a stone on Brougham Bridge the formulas*

$$\boldsymbol{i}^2 = \boldsymbol{j}^2 = \boldsymbol{k}^2 = \boldsymbol{ijk} = -1,$$

*In the same city on June 21, 1972 President Eamon de Valera told one of the authors that, while in jail one evening in 1916, scheduled to be shot the next morning, he wrote down the formula of which he was so fond, $\boldsymbol{i}^2 = \boldsymbol{j}^2 = \boldsymbol{k}^2 = \boldsymbol{ijk} = -1$.

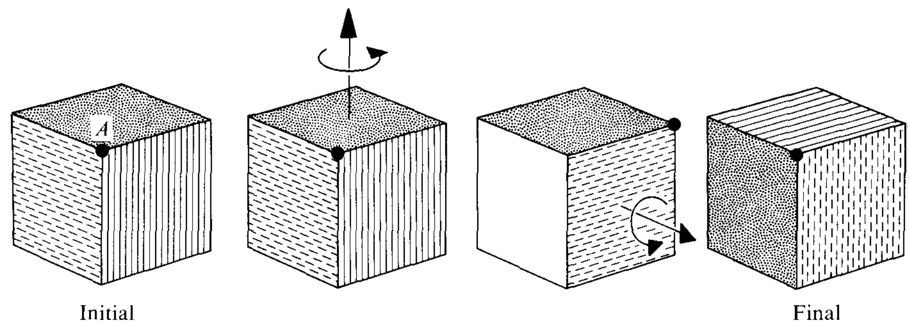


Figure 41.1.
Rotation about the vertical axis through 90°, followed by rotation about the horizontal axis through 90°, gives a net change in orientation that can be achieved by a single rotation through 120° about an axis emergent from the center through the corner A.

which in today's notation,

$$\sigma_x = \begin{Vmatrix} 0 & 1 \\ 1 & 0 \end{Vmatrix} = ii, \qquad \sigma_y = \begin{Vmatrix} 0 & -i \\ i & 0 \end{Vmatrix} = ij, \qquad \sigma_z = \begin{Vmatrix} 1 & 0 \\ 0 & -1 \end{Vmatrix} = ik, \tag{41.1}$$

take the form

$$\begin{gathered} \sigma_x{}^2 = \sigma_y{}^2 = \sigma_z{}^2 = 1, \\ \sigma_x\sigma_y = -\sigma_y\sigma_x = i\sigma_z \text{ (and cyclic permutations).} \end{gathered} \tag{41.2}$$

Rotation operators:
(1) defined

To any rotation is associated a quantity (Hamilton's "quaternion;" today's "spin matrix" or "spinor transformation" or "rotation operator")

$$R = \cos(\theta/2) - i\sin(\theta/2)(\sigma_x \cos\alpha + \sigma_y \cos\beta + \sigma_z \cos\gamma), \tag{41.3}$$

(2) as tools in combining rotations

where θ is the angle of rotation and α, β, γ are the angles between the axis of rotation and the coordinate axes. A rotation described by R_1 followed by a rotation described by R_2 gives a net change in orientation described by the single rotation

$$R_3 = R_2R_1. \tag{41.4}$$

This is Hamilton's formula for the combination of two rotations (steps toward it by Euler in 1776; obtained by Gauss in 1819 but never published by him).

In the example in Figure 41.1,

$$\begin{aligned} R_1(\text{rotation by } \theta = 90^\circ \text{ about } z\text{-axis}) &= (1 - i\sigma_z)/2^{1/2}, \\ R_2(\text{rotation by } \theta = 90^\circ \text{ about } x\text{-axis}) &= (1 - i\sigma_x)/2^{1/2}, \end{aligned}$$

and the product of the two is

$$\begin{aligned} R_2R_1 &= (1 - i\sigma_x + i\sigma_y - i\sigma_z)/2 \\ &= \cos 60^\circ - i\sin 60^\circ(\sigma_x/3^{1/2} - \sigma_y/3^{1/2} + \sigma_z/3^{1/2}). \end{aligned}$$

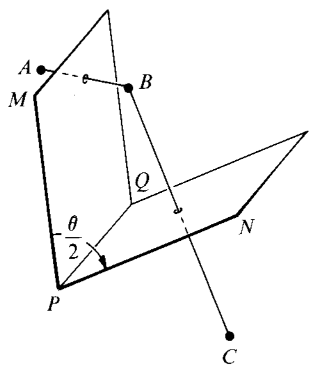


Figure 41.2.
Reflection in the plane MPQ carries A to B. Reflection in the plane NPQ carries B to C. The combination of the two reflections in the two planes separated by the angle $\theta/2$ produces the same end result (transformation from A to C) as rotation through the angle θ about the line PQ.

According to Hamilton's rule (41.3), this result implies a net rotation through 120° about a line that makes equal angles with the x-axis, the y-axis, and the z-axis, in conformity with what one already saw in Figure 41.1 (axis of rotation running from center of cube through the corner A).

What one has just done in the special example one can do in the general case: obtain the parameters θ_3, α_3, β_3, γ_3 of the net rotation (four unknowns!) by identifying the four coefficients of the four Hamilton units 1, $-i\sigma_x$, $-i\sigma_y$, $-i\sigma_z$ on both sides of the equation $R_3 = R_2R_1$. In this way one arrives at the four prequaternion formulas of Olinde Rodrigues (1840) for the combination of the two rotations.

Geometric reason that half angles appear in rotation operators

Why do half-angles put in an appearance? And what is behind the law of combination of rotations? The answer to both questions is the same: a rotation through the angle θ about a given axis may be visualized as the consequence of successive reflections in two planes that meet along that axis at the angle $\theta/2$ (Figure 41.2). Two rotations therefore imply four reflections. However, it can be arranged that reflections no. 2 and no. 3 take place in the same plane, the plane that includes the two axes of rotation. Then reflection no. 3 exactly undoes reflection no. 2. By now there remain only reflections no. 1 and no. 4, which together constitute one rotation: the net rotation that was desired (Figures 41.3 and 41.4).

Algebraic properties of rotation operators

The rotation

$$R = \cos(\theta/2) - i\sin(\theta/2)(\sigma_x\cos\alpha + \sigma_y\cos\beta + \sigma_z\cos\gamma) \tag{41.3}$$

is undone by the inverse rotation

$$R^{-1} = \cos(\theta/2) + i\sin(\theta/2)(\sigma_x\cos\alpha + \sigma_y\cos\beta + \sigma_z\cos\gamma). \tag{41.3'}$$

Thus the product of the two rotation operators

$$RR^{-1} = R^{-1}R = 1 \tag{41.5}$$

is an operator, the unit operator, that leaves unchanged everything that it acts on. The reciprocal R^{-1} of the combination $R = R_2R_1$ of two rotations is

$$R^{-1} = R_1^{-1}R_2^{-1} \tag{41.5'}$$

(reverse order of factors!), as one verifies by substitution into (41.5).

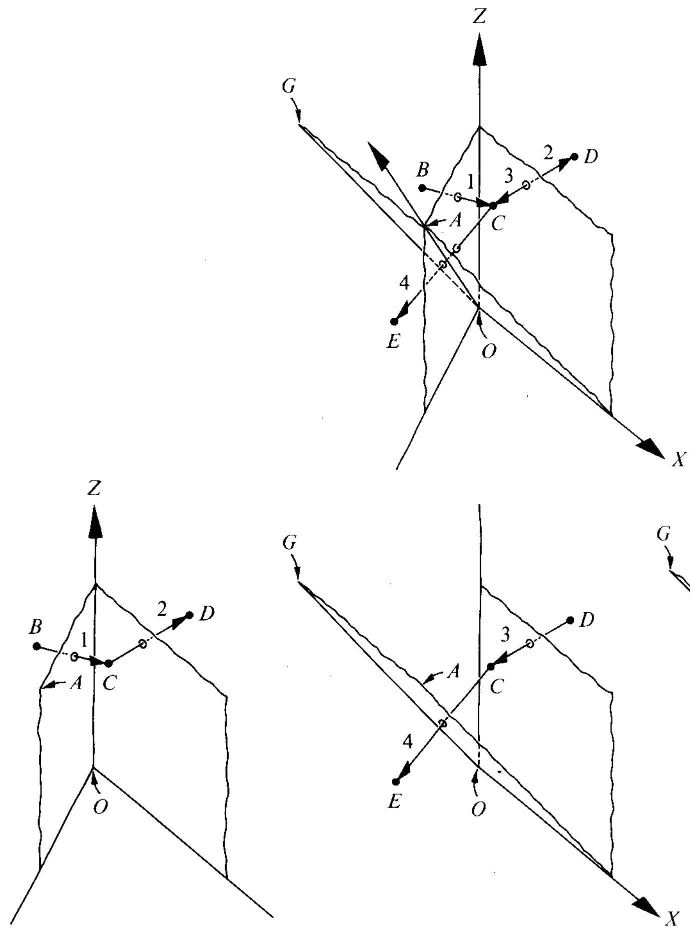


Figure 41.3.
Composition of two rotations seen in terms of reflections. The first rotation (for instance, 90° about OZ in the example of Figure 41.1.) is represented in terms of reflection 1 followed by reflection 2 (the planes of the two reflections being separated by $90°/2 = 45°$ in the example). The second reflection appears as the resultant of reflections 3 and 4. But the reflections 2 and 3 take place in the common plane ZOX. Therefore one reflection undoes the other. Thus the sequence of four operations 1234 collapses to the two reflections 1 and 4. Their place in turn is taken by a single rotation about the axis OA.

The conjugate transpose, M^*, of a matrix M is obtained by taking the conjugate complex of every element in the matrix and then interchanging rows and columns. By direct inspection of matrix expressions (41.1) one sees that $\sigma_x{}^* = \sigma_x$, $\sigma_y{}^* = \sigma_y$, $\sigma_z{}^* = \sigma_z$. Such matrices are said to be Hermitian. The conjugate transpose of the product $M = PQ$ of two matrices is the product $M^* = Q^*P^*$ of the individual conjugate transposed matrices taken in the reverse order. For the rotation matrix written down above, note that $R^* = R^{-1}$. Such a matrix is said to be unitary. The

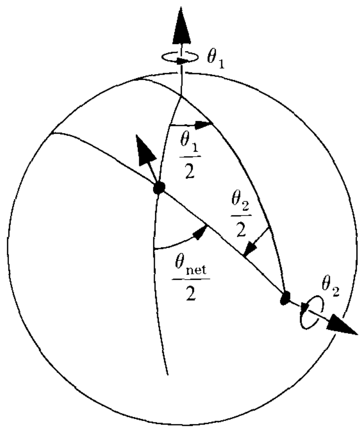


Figure 41.4.
Law of composition of rotations epitomized by a spherical triangle in which each of the three important angles represents *half* an angle of rotation.

determinant of a unitary matrix may be seen to have absolute value unity from the following line of argument:

$$\begin{aligned}1 &= \det(\text{unit matrix}) = \det(RR^{-1})\\ &= \det(RR^*) = \det R \det R^*\\ &= |\det R|^2. \end{aligned} \tag{41.6}$$

In actuality the determinant of the rotation spin matrix is necessarily unity ("unimodular matrix") as shown in the following exercises

---

EXERCISES

Exercise 41.1. ELEMENTARY FEATURES OF THE ROTATION MATRIX

Write equation (41.3) in the form

$$R(\theta) = \cos(\theta/2) - i \sin(\theta/2)(\boldsymbol{\sigma} \cdot \boldsymbol{n}),$$

and establish the following properties:

(a) $$(\boldsymbol{\sigma} \cdot \boldsymbol{n})^2 = 1 \equiv \text{unit matrix};$$

(b) $\operatorname{tr}(\boldsymbol{\sigma} \cdot \boldsymbol{n}) = 0$ (tr means "trace," i.e., sum of diagonal elements);

(c) $$\underbrace{[R, (\boldsymbol{\sigma} \cdot \boldsymbol{n})]}_{\text{[commutator]}} \equiv R(\boldsymbol{\sigma} \cdot \boldsymbol{n}) - (\boldsymbol{\sigma} \cdot \boldsymbol{n})R = 0;$$

(d) $$\frac{dR}{d\theta} = -\frac{i}{2}(\boldsymbol{\sigma} \cdot \boldsymbol{n})R. \tag{41.7}$$

[Note that if one thinks of θ as increasing with angular velocity ω, so $d\theta/dt = \omega =$ constant, then this last equation reads

$$\frac{dR}{dt} = -\frac{i}{2}(\boldsymbol{\sigma} \cdot \boldsymbol{\omega})R \tag{41.7'}$$

where $\boldsymbol{\omega} = \omega \boldsymbol{n}$.]

Exercise 41.2. ROTATION MATRIX HAS UNIT DETERMINANT

Recall from exercise 5.5 that for any matrix M one has

$$d[\ln(\det M)] = \mathrm{tr}\,(M^{-1}\,dM)$$

and use this to show that $\det R$ in (41.7) is constant, and therefore equal to $(\det R)_{\theta=0} = 1$.

§41.2. INFINITESIMAL ROTATIONS

Infinitesimal rotations

A given rotation can be obtained by performing in turn two rotations of half the magnitude, or four rotations of a fourth the magnitude, or eight of an eighth the magnitude, and so on. Thus one arrives in the limit at the concept of an infinitesimal rotation described by the spin matrix

$$R = 1 - (i/2)(\sigma_x\,d\theta_{yz} + \sigma_y\,d\theta_{zx} + \sigma_z\,d\theta_{xy})$$

or

$$R = 1 - (i\,d\theta/2)(\boldsymbol{\sigma}\cdot\boldsymbol{n}). \tag{41.8}$$

Here the quantities

$$\begin{aligned} d\theta_{yz} &= -d\theta_{zy} = n^x\,d\theta = \cos\alpha\,d\theta,\\ d\theta_{zx} &= -d\theta_{xz} = n^y\,d\theta = \cos\beta\,d\theta,\\ d\theta_{xy} &= -d\theta_{yx} = n^z\,d\theta = \cos\gamma\,d\theta, \end{aligned} \tag{41.9}$$

are the components of the infinitesimal rotation in the three indicated planes. An infinitesimal rotation in the (x, y)-plane through the angle $d\theta_{xy}$ transforms the vector $\boldsymbol{x} = (x, y, z)$ into a new vector with changed components x' and y' but with unchanged component $z' = z$. More generally, the infinitesimal rotation (41.8) considered in this same "active" sense* produces the transformation

$$\boldsymbol{x} \longrightarrow \boldsymbol{x}',$$

with

$$\begin{aligned} x' &= x - (d\theta_{xy})y - (d\theta_{xz})z,\\ y' &= -(d\theta_{yx})x + y - (d\theta_{yz})z,\\ z' &= -(d\theta_{zx})x - (d\theta_{zy})y + z. \end{aligned} \tag{41.10}$$

Representation of a 3-vector as a spin matrix

Spinor calculus provides an alternative (and shorthand!) means to calculate the foregoing effect of a rotation on a vector. Associate with the vector $\boldsymbol{x}$ the spin matrix

$$X = x\sigma_x + y\sigma_y + z\sigma_z = (\boldsymbol{x}\cdot\boldsymbol{\sigma}), \tag{41.11}$$

*An "*active*" transformation changes one vector into another, while leaving unchanged the underlying reference frame (if there is one). By contrast, a "*passive*" transformation leaves all vectors unchanged, but alters the reference frame. All transformations in previous chapters of this book were passive.

2

and with the vector $\boldsymbol{x}'$ a corresponding spin matrix or quaternion X'. Then the effect of the rotation is summarized in the formula

Rotation of a 3-vector described in spin-matrix language

$$X \longrightarrow X' = RXR^*. \tag{41.12}$$

Test this formula for the general infinitesimal rotation (41.10). It reads

$$(\boldsymbol{x}' \cdot \boldsymbol{\sigma}) = [1 - (i\,d\theta/2)(\boldsymbol{\sigma} \cdot \boldsymbol{n})](\boldsymbol{x} \cdot \boldsymbol{\sigma})[1 + (i\,d\theta/2)(\boldsymbol{\sigma} \cdot \boldsymbol{n})]$$

or, to the first order in the quantity $d\theta$,

$$(\boldsymbol{x}' \cdot \boldsymbol{\sigma}) = (\boldsymbol{x} \cdot \boldsymbol{\sigma}) + (i\,d\theta/2)[(\boldsymbol{x} \cdot \boldsymbol{\sigma})(\boldsymbol{\sigma} \cdot \boldsymbol{n}) - (\boldsymbol{\sigma} \cdot \boldsymbol{n})(\boldsymbol{x} \cdot \boldsymbol{\sigma})]. \tag{41.13}$$

The product of spin matrices $A = (\boldsymbol{a} \cdot \boldsymbol{\sigma})$ and $B = (\boldsymbol{b} \cdot \boldsymbol{\sigma})$ built from two distinct vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is

$$AB = (\boldsymbol{a} \cdot \boldsymbol{\sigma})(\boldsymbol{b} \cdot \boldsymbol{\sigma}) = a^x b^x \sigma_x{}^2 + a^x b^y \sigma_x \sigma_y + \ldots,$$

or, according to (41.2),

$$AB = (\boldsymbol{a} \cdot \boldsymbol{b}) + i(\boldsymbol{a} \times \boldsymbol{b}) \cdot \boldsymbol{\sigma}. \tag{41.14}$$

Employ this formula to evaluate the right-hand side of (41.13). In the square brackets, the terms in $(\boldsymbol{x} \cdot \boldsymbol{n})$ have opposite signs and cancel. In contrast, the terms in $(\boldsymbol{n} \times \boldsymbol{x})$ have the same sign. They combine to cancel the factor 2 in $(d\theta/2)$. End up with

$$(\boldsymbol{x}' \cdot \boldsymbol{\sigma}) = (\boldsymbol{x} \cdot \boldsymbol{\sigma}) + d\theta(\boldsymbol{n} \times \boldsymbol{x}) \cdot \boldsymbol{\sigma}$$

or

$$\boldsymbol{x}' = [1 + (d\theta)\boldsymbol{n} \times]\, \boldsymbol{x} \tag{41.15}$$

in agreement with (41.10), as was to be shown.

A finite rotation about a given axis can be considered as the composition of infinitesimal rotations about that axis. To see this composition in simplest form, rewrite the spin matrix (41.8) associated with the general infinitesimal rotation as

Composition of finite rotation from infinitesimal rotations

$$R(d\theta) = e^{-(i\,d\theta/2)(\boldsymbol{\sigma} \cdot \boldsymbol{n})} \tag{41.16}$$

(exponential function defined by its power-series expansion). Note that $(\boldsymbol{\sigma} \cdot \boldsymbol{n})$ commutes (a) with unity and (b) with itself, and in addition (c) has a unit square. Therefore the calculation of the exponential function proceeds no differently here, for spin matrices, than for everyday algebra. The composition of the spin matrices for infinitesimal rotations about an unchanging axis proceeds by adding exponents, to give

$$R(\theta) = e^{-i(\theta/2)(\boldsymbol{\sigma} \cdot \boldsymbol{n})}, \tag{41.17}$$

which can also be obtained immediately from equation (41.7). This expression can be put in another form by developing the power series; thus,

$$R(\theta) = \sum_{p=0}^{\infty} (1/p!)(-i\theta\boldsymbol{\sigma}\cdot\boldsymbol{n}/2)^p$$

$$= \sum_{\text{even}\,p} (1/p!)(-i\theta/2)^p + (\boldsymbol{\sigma}\cdot\boldsymbol{n}) \sum_{\text{odd}\,p} (1/p!)(-i\theta/2)^p \tag{41.18}$$

$$= \cos(\theta/2) - i\sin(\theta/2)(\boldsymbol{\sigma}\cdot\boldsymbol{n})$$

in agreement with the expression (41.3) originally given for a spinor transformation. The effect of one infinitesimal rotation after another after another . . . on a vector is given by

$$X' = R(d\theta)\ldots R(d\theta)XR^*(d\theta)\ldots R^*(d\theta),$$

with the consequence that even for a finite rotation $R = R(\theta)$ one is correct in employing the formula

$$X' = RXR^*. \tag{41.19}$$

EXERCISE

Exercise 41.3. MORE PROPERTIES OF THE ROTATION MATRIX

Show that for $X = \boldsymbol{x}\cdot\boldsymbol{\sigma}$ one has the commutation relation

$$[(\boldsymbol{\sigma}\cdot\boldsymbol{n}), X] = 2i(\boldsymbol{n}\times\boldsymbol{x})\cdot\boldsymbol{\sigma}.$$

Use this to obtain, from equation (41.19) in the form $X = RX_0R^*$ [where X_0 is constant, while $R(\theta)$ is given by equation (41.17)], the formula

$$\frac{d}{d\theta}(\boldsymbol{x}\cdot\boldsymbol{\sigma}) = (\boldsymbol{n}\times\boldsymbol{x})\cdot\boldsymbol{\sigma}.$$

Why is this equivalent to the standard definition

$$\frac{d\boldsymbol{x}}{dt} = \omega\times\boldsymbol{x}$$

for the angular velocity? Reverse the argument to show that equation (41.7′) correctly defines the rotation $R(t)$ resulting from a time-dependent angular velocity $\omega(t)$, even though the simple solution $R = \exp[-\frac{1}{2}it(\boldsymbol{\sigma}\cdot\omega)]$ of this equation can no longer be written when ω is not constant.

---

§41.3. LORENTZ TRANSFORMATION VIA SPINOR ALGEBRA

4-vectors and Lorentz transformations in spin-matrix language

Generate a rotation by two reflections in space? Then why not generate a Lorentz transformation by two reflections in spacetime? If for this purpose one has to turn from a real half-angle between the two planes of reflection to a complex half-angle, that development will come as no surprise; nor will it be a surprise that one can still represent the effect of the Lorentz transformation by a matrix multiplication of the form

$$X \longrightarrow X' = LXL^*. \tag{41.20}$$

Here the "Lorentz spin transformation matrix" L is a generalization of the rotation matrix, R. Also the "coordinate-generating spin matrix" X is now generalized from (41.11) to

$$X = t + (\boldsymbol{x} \cdot \boldsymbol{\sigma}) \tag{41.21}$$

or

$$X = \begin{Vmatrix} t+z & x-iy \\ x+iy & t-z \end{Vmatrix}. \tag{41.22}$$

It is demanded that this matrix be Hermitian

$$X = X^*. \tag{41.23}$$

Then and only then are the coordinates (t, x, y, z) real. The conjugate transpose of the transformed spin matrix must also be Hermitian—and is:

$$\begin{aligned}(X')^* &= (LXL^*)^* \\ &= (L^*)^*(X)^*(L)^* = LXL^* = X'.\end{aligned} \tag{41.24}$$

Therefore the new coordinates (t', x', y', z') are guaranteed to be real, as desired. This reality requirement is a rationale for the form of the spin-matrix transformation (41.20), with L appearing on one side of X and L^* on the other.

A Lorentz transformation is defined by the circumstance that it leaves the interval invariant:

$$t'^2 - x'^2 - y'^2 - z'^2 = t^2 - x^2 - y^2 - z^2. \tag{41.25}$$

Note that the determinant of the matrix X as written out above has the value

$$\det X = t^2 - x^2 - y^2 - z^2. \tag{41.26}$$

Consequently the requirement for the preservation of the interval may be put in the form

$$\det X' = \det X \tag{41.27}$$

or

$$(\det L)\,(\det X)\,(\det L^*) = \det X. \tag{41.28}$$

This requirement is fulfilled by demanding

$$\det L = 1 \tag{41.29}$$

[it is not a useful generalization to multiply every element of L here by a common phase factor $e^{i\delta}$, and therefore multiply $\det L$ by $e^{2i\delta}$, because the net effect of this phase factor is nil in the formula $X' = LXL^*$].

Infinitesimal Lorentz transformations

The spin matrix associated with a rotation, whether finite or infinitesimal, already satisfied the condition $\det L = 1$ [proved in exercise (41.2)]. This condition, being algebraic, will continue to hold when the real angles $d\theta_{yz}$, $d\theta_{zx}$, $d\theta_{xy}$, are replaced by complex angles, $d\theta_{yz} + i\,d\alpha_x$, $d\theta_{zx} + i\,d\alpha_y$, $d\theta_{xy} + i\,d\alpha_z$. The spin-transformation matrix acquires in this way a total of six parameters, as needed to describe the general infinitesimal Lorentz transformation. Thus the spin matrix for the general infinitesimal Lorentz transformation can be put in the form

$$\begin{aligned} L &= 1 - (i/2)(\sigma_x\,d\theta_{yz} + \sigma_y\,d\theta_{zx} + \sigma_z\,d\theta_{xy}) \\ &\quad + (1/2)(\sigma_x\,d\alpha_x + \sigma_y\,d\alpha_y + \sigma_z\,d\alpha_z) \\ &= 1 - (i\,d\theta/2)(\boldsymbol{\sigma}\cdot\boldsymbol{n}) + (\boldsymbol{\sigma}\cdot d\boldsymbol{\alpha}/2). \end{aligned} \tag{41.30}$$

The effect of this transformation upon the coordinates is to be read out from the formula

$$X \longrightarrow X' = LXL^*$$

or

$$\begin{aligned} t' + (\boldsymbol{\sigma}\cdot\boldsymbol{x}') = {} & [1 - (i\,d\theta/2)(\boldsymbol{\sigma}\cdot\boldsymbol{n}) + (\boldsymbol{\sigma}\cdot d\boldsymbol{\alpha}/2)] \\ & \times [t + (\boldsymbol{\sigma}\cdot\boldsymbol{x})][1 + (i\,d\theta/2)(\boldsymbol{\sigma}\cdot\boldsymbol{n}) + (\boldsymbol{\sigma}\cdot d\boldsymbol{\alpha}/2)] \end{aligned} \tag{41.31}$$

Employ equation (41.14) for $(\boldsymbol{\sigma}\cdot\boldsymbol{A})(\boldsymbol{\sigma}\cdot\boldsymbol{B})$ to reduce the right side to the form

$$t + (\boldsymbol{\sigma}\cdot\boldsymbol{x}) + (\boldsymbol{\sigma}\cdot d\boldsymbol{\alpha})t + d\theta(\boldsymbol{n}\times\boldsymbol{x})\cdot\boldsymbol{\sigma} + (\boldsymbol{x}\cdot d\boldsymbol{\alpha}).$$

Now compare coefficients of 1, $\boldsymbol{\sigma}_x$, $\boldsymbol{\sigma}_y$ and $\boldsymbol{\sigma}_z$, respectively, on both sides of the equation, and find

$$\begin{aligned} t' &= t + (\boldsymbol{x}\cdot d\boldsymbol{\alpha}) \\ \boldsymbol{x}' &= \boldsymbol{x} + t\,d\boldsymbol{\alpha} + d\theta(\boldsymbol{n}\times\boldsymbol{x}), \end{aligned} \tag{41.32}$$

in agreement with the conventional expression for an infinitesimal Lorentz transformation or "boost" of velocity $d\boldsymbol{\alpha}$, in active form, as was to be shown.

Composition of finite Lorentz transformations from infinitesimal transformations

The composition of such infinitesimal Lorentz transformations gives a finite Lorentz transformation. The result, however, can be calculated easily only when all infinitesimal transformations commute. Thus assume that $d\boldsymbol{\theta}$ and $d\boldsymbol{\alpha}$ are in a fixed ratio, so

$$\boldsymbol{\omega} \equiv \boldsymbol{n}\frac{d\theta}{d\tau} \text{ and } \boldsymbol{a} \equiv \frac{d\boldsymbol{\alpha}}{d\tau}$$

are constants, with τ a parameter. Then integration with respect to τ (composition of infinitesimal transformations) gives a finite transformation $L = \exp[-\frac{1}{2}i\tau\boldsymbol{\sigma}\cdot(\boldsymbol{\omega} + i\boldsymbol{a})]$. For $\tau = 1$, so $\theta\boldsymbol{n} = \boldsymbol{\omega}\tau$, $\boldsymbol{\alpha} = \boldsymbol{a}\tau$, this reads

$$L = \exp[(\boldsymbol{\alpha} - i\theta\boldsymbol{n})\cdot\boldsymbol{\sigma}/2]. \tag{41.33}$$

In the special case of a pure boost (no rotation; $\theta = 0$), the exponential function is evaluated along the lines indicated in (41.18), with the result

$$L = \cosh(\alpha/2) + (\boldsymbol{n}_\alpha\cdot\boldsymbol{\sigma})\sinh(\alpha/2). \tag{41.34}$$

Here $\boldsymbol{n}_\alpha = \boldsymbol{\alpha}/\alpha$ is a unit vector in the direction of the boost. The corresponding Lorentz transformation itself is evaluated from the formula

$$X' = LXL^*$$

or

$$\begin{aligned} t' + (\boldsymbol{x}' \cdot \boldsymbol{\sigma}) = {} & [\cosh \alpha/2 + (\boldsymbol{n}_\alpha \cdot \boldsymbol{\sigma}) \sinh \alpha/2][t + (\boldsymbol{x} \cdot \boldsymbol{\sigma})] \\ & \times [\cosh \alpha/2 + (\boldsymbol{n}_\alpha \cdot \boldsymbol{\sigma}) \sinh \alpha/2]. \end{aligned} \tag{41.35}$$

Simplify with the help of the relations

$$\cosh^2(\alpha/2) + \sinh^2(\alpha/2) = \cosh \alpha,$$
$$2 \sinh (\alpha/2) \cosh (\alpha/2) = \sinh \alpha,$$

and

$$(\boldsymbol{n}_\alpha \cdot \boldsymbol{\sigma})(\boldsymbol{x} \cdot \boldsymbol{\sigma})(\boldsymbol{n}_\alpha \cdot \boldsymbol{\sigma}) = 2(\boldsymbol{n}_\alpha \cdot \boldsymbol{x})(\boldsymbol{n}_\alpha \cdot \boldsymbol{\sigma}) - (\boldsymbol{n}_\alpha \cdot \boldsymbol{n}_\alpha)(\boldsymbol{x} \cdot \boldsymbol{\sigma}),$$

and on both sides of the equation compare coefficients of 1 and $\boldsymbol{\sigma}$, to find

$$\begin{aligned} t' &= (\cosh \alpha)t + (\sinh \alpha)(\boldsymbol{n}_\alpha \cdot \boldsymbol{x}), \\ \boldsymbol{x}' &= [(\sinh \alpha)\boldsymbol{n}_\alpha t + (\cosh \alpha)(\boldsymbol{n}_\alpha \cdot \boldsymbol{x})\boldsymbol{n}_\alpha] \text{ (“in-line part of transformation”)} \\ &\quad + [\boldsymbol{x} - (\boldsymbol{x} \cdot \boldsymbol{n}_\alpha)\boldsymbol{n}_\alpha] \text{ (“perpendicular part of } \boldsymbol{x} \text{ unchanged”)}. \end{aligned} \tag{41.36}$$

In this way one verifies that the quantity α is the usual “velocity parameter,” connected with the velocity itself by the relations

$$\begin{aligned} (1 - \beta^2)^{-1/2} &= \cosh \alpha, \\ \beta(1 - \beta^2)^{-1/2} &= \sinh \alpha, \\ \beta &= \tanh \alpha. \end{aligned} \tag{41.37}$$

That velocity parameters add for successive boosts in the same direction shows nowhere more clearly than in the representation (41.33) of the spin-transformation matrix:

$$\begin{aligned} L(\alpha_2)L(\alpha_1) &= \exp [\alpha_2(\boldsymbol{n}_\alpha \cdot \boldsymbol{\sigma})/2] \exp [\alpha_1(\boldsymbol{n}_\alpha \cdot \boldsymbol{\sigma})/2] = \exp [(\alpha_2 + \alpha_1)(\boldsymbol{n}_\alpha \cdot \sigma)/2] \\ &= L(\alpha_2 + \alpha_1). \end{aligned} \tag{41.38}$$

Turn from this special case, and ask how to get the resultant of two arbitrary Lorentz transformations, each of which is a mixture of a rotation and a boost. No simpler method offers itself to answer this question than to use formula (41.33) together with the equation

$$L(\text{resultant}) = L_2 L_1. \tag{41.39}$$

§41.4. THOMAS PRECESSION VIA SPINOR ALGEBRA

A spinning object, free of all torque, but undergoing acceleration, changes its direction as this direction is recorded in an inertial frame of reference. This is the

Thomas precession [see exercise 6.9 and first term in equation (40.33b)]. This precession accounts for a factor two in the effective energy of coupling of spin and orbital angular momentum of an atomic electron. In a nucleus it contributes a little to the coupling of the spin and orbit of a nucleon. The evaluation of the Thomas precession affords an illustration of spin-matrix methods in action.

Origin of Thomas precession: composition of two boosts is not a pure boost

The precession in question can be discussed quite without reference either to angular momentum or to mass in motion. It is enough to consider a sequence of inertial frames of reference $S(t)$ with these two features. (1) To whatever point the motion has taken the mass at time t, at that point is located the origin of the frame $S(t)$. (2) The inertial frame $S(t + dt)$ at the next succeeding moment has undergone no rotation with respect to the inertial frame $S(t)$, as rotation is conceived by an observer in that inertial frame. However, it has undergone a rotation ("Thomas precession") as rotation is conceived and defined in the laboratory frame of reference.

How is it possible for "no rotation" to appear as "rotation"? The answer is this: one pure boost, followed by another pure boost in another direction, does not have as net result a third pure boost; instead, the net result is a boost plus a rotation. This idea is not new in kind. Figure 41.1 illustrated how a rotation about the z-axis followed by a rotation about the x-axis had as resultant a rotation about an axis with not only an x-component and a z-component but also a y-component. What is true of rotations is true of boosts: they defy the law for the addition of vectors.

Derivation of Thomas precession using spin matrices

Let the frame S_0 coincide with the laboratory frame, and let the origin of this laboratory frame be where the moving frame is at time t. Let $S(t)$ be a Lorentz frame moving with this point at time t. Let one pure boost raise its velocity relative to the laboratory from $\boldsymbol{\beta}$ to $\boldsymbol{\beta} + d\boldsymbol{\beta}$. The resulting final configuration cannot be reached from S_0 by a pure boost. Instead, first turn S_0 relative to the laboratory frame ("rotation R associated with the Thomas precession") and then send it by a simple boost to the final configuration. Only one choice of this rotation will be right to produce match-up. Thus, distinguishing the spin matrices for pure boosts and pure rotations by the letters B and R, one has the relation

$$B(\boldsymbol{\beta} + d\boldsymbol{\beta})R(\boldsymbol{\omega}\, dt) = \text{“}B(d\boldsymbol{\beta})\text{”}B(\boldsymbol{\beta}) \tag{41.40}$$

out of which to find the angular velocity $\boldsymbol{\omega}$ of the Thomas precession. The quotation marks in "$B(d\boldsymbol{\beta})$" carry a double warning: (1) the velocity of transformation that boosts $S(t)$ to $S(t + dt)$ is not $(\boldsymbol{\beta} + d\boldsymbol{\beta}) - \boldsymbol{\beta} = d\boldsymbol{\beta}$ (law of vector addition—or subtraction—not applicable to velocity), and (2) "$B(d\boldsymbol{\beta})$" does not appear as a pure boost in the laboratory frame. It appears as a pure boost only in the comoving frame.

Take care of the second difficulty first. It is only a difficulty because the formalism for combination of transformations, $R_3 = R_2R_1$, as developed in §41.1 presupposes all operations $R_1, R_2, \ldots$, to be defined and carried out in the laboratory reference frame. In contrast, the quantity "$B(d\boldsymbol{\beta})$" is understood to imply a pure boost as defined and carried out in the comoving frame. Such an operation can be fitted into the formalism as follows. (1) Undo any velocity that the object already has. In other words apply the operator $B^{-1}(\boldsymbol{\beta})$. Then the object is at rest in the laboratory frame. Then apply the necessary small pure boost, $B(\boldsymbol{a}_{\text{comoving}}\, d\tau)$, where $\boldsymbol{a}_{\text{comoving}}$

is the acceleration as it will be sensed by the object and $d\tau$ is the lapse of proper time as it will be sensed by the object. At the commencement of this brief acceleration the object is at rest relative to the laboratory. What is a pure boost to it is a pure boost relative to the laboratory. It is also a pure boost in the spin-matrix formalism. Then transform back from laboratory to moving frame. Thus have the relation

$$\text{“}B(d\boldsymbol{\beta})\text{”} = B(\beta)B(\boldsymbol{a}_{\text{comoving}}\, d\tau)B^{-1}(\boldsymbol{\beta}). \tag{41.41}$$

The equation for the determination of the Thomas precession now reads

$$B(\boldsymbol{\beta} + d\boldsymbol{\beta})R(\boldsymbol{\omega}\, dt) = B(\boldsymbol{\beta})B(\boldsymbol{a}_{\text{comoving}}\, d\tau) \tag{41.42}$$

or, with all unknowns put on the left,

$$R(\boldsymbol{\omega}\, dt)B^{-1}(\boldsymbol{a}_{\text{comoving}}\, d\tau) = B^{-1}(\boldsymbol{\beta} + d\boldsymbol{\beta})B(\boldsymbol{\beta}). \tag{41.43}$$

The first task, to replace the erroneous value of the velocity change ($d\boldsymbol{\beta}$) by a correct value ($\boldsymbol{a}_{\text{comoving}}\, d\tau$), is now made part of the problem along with the evaluation of the Thomas precession itself.

Principles settled, the calculation proceeds by inserting the appropriate expressions for all four factors in (41.43), and evaluating both sides of the equation to the first order of small quantities, as follows:

$$\begin{aligned} 1 - (i\, dt\boldsymbol{\omega} + d\tau\boldsymbol{a})\cdot\boldsymbol{\sigma}/2 = {} & [\cosh(\alpha'/2) - (\boldsymbol{n}_{\alpha'}\cdot\boldsymbol{\sigma})\sinh(\alpha'/2)] \\ & \times [\cosh(\alpha/2) + (\boldsymbol{n}_{\alpha}\cdot\boldsymbol{\sigma})\sinh(\alpha/2)]. \end{aligned} \tag{41.44}$$

Here α and $\boldsymbol{n}_\alpha$ are the velocity parameter and unit vector that go with the velocity $\boldsymbol{\beta}$; $\alpha' = \alpha + d\alpha$, and $\boldsymbol{n}_{\alpha'} = \boldsymbol{n}_\alpha + d\boldsymbol{n}_\alpha$, go with $\boldsymbol{\beta} + d\boldsymbol{\beta}$. Develop the righthand side of (41.44) by the methods of calculus, writing $\alpha' = \alpha + d\alpha$ and $\boldsymbol{n}_{\alpha'} = \boldsymbol{n}_\alpha + d\boldsymbol{n}_\alpha$, and applying the rule for the differentiation of a product. Equate coefficients of $-\boldsymbol{\sigma}/2$ and $-i\boldsymbol{\sigma}/2$ on both sides of the equation. Thus find

$$\boldsymbol{a}_{\text{comoving}}\, d\tau = (d\alpha)\boldsymbol{n}_\alpha + (\sinh\alpha)\, d\boldsymbol{n}_\alpha \tag{41.45}$$

and

$$\boldsymbol{\omega}\, dt = [2\sinh^2(\alpha/2)]\, d\boldsymbol{n}_\alpha \times \boldsymbol{n}_\alpha. \tag{41.46}$$

Angular velocity of Thomas precession

The one expression gives the change of velocity as seen in a comoving inertial frame. The other gives the precession as seen in the laboratory frame. For low velocities the expression for the Thomas precession reduces to

$$\boldsymbol{\omega} = \boldsymbol{a} \times \boldsymbol{\beta}/2. \tag{41.47}$$

Here $\boldsymbol{a}$ is the acceleration. Only the component perpendicular to the velocity $\boldsymbol{\beta}$ is relevant for the precession.

For an elementary account of the importance of the Thomas precession in atomic physics, see, for example, Ruark and Urey (1930).

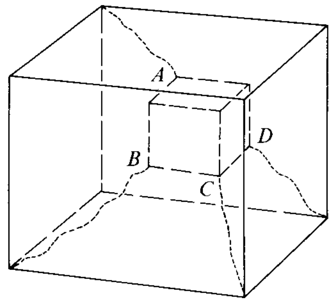


Figure 41.5.
"Orientation-entanglement relation" between a cube and the walls of a room. A 360° rotation of the cube entangles the threads. A 720° rotation might be thought to entangle them still more—but instead makes it possible completely to disentangle them.

§41.5. SPINORS

Orientation-entanglement relation

Paint each face of a cube a different color. Then connect each corner of the cube to the corresponding corner of the room with an elastic thread (Figure 41.5). Now rotate the cube through $2\pi = 360°$. The threads become tangled. Nothing one can do will untangle them. It is impossible for every thread to proceed on its way in a straight line. Now rotate the cube about the same axis by a further 2π. The threads become still more tangled. However, a little work now completely straightens out the tangle (Figure 41.6). Every thread runs as it did in the beginning in a straight line from its corner of the cube to the corresponding corner of the room. More generally, rotations by $0, \pm 4\pi, \pm 8\pi, \ldots$, leave the cube in its standard "orientation-entanglement relation" with its surroundings, whereas rotations by $\pm 2\pi, \pm 6\pi, \pm 10\pi, \ldots$, restore to the cube only its orientation, not its orientation-entanglement relation with its surroundings. Evidently there is something about the geometry of orientation that is not fully taken into account in the usual concept of orientation; hence the concept of "orientation-entanglement relation" or (briefer term!) "version" (Latin *versor,* turn). Whether there is also a detectable difference in the physics (contact potential between a metallic object and its metallic surroundings, for example) for two inequivalent versions of an object is not known [Aharonov and Susskind (1967)].

In keeping with the distinction between the two inequivalent versions of an object, the spin matrix associated with a rotation,

$$R = \cos(\theta/2) - i(\boldsymbol{n} \cdot \boldsymbol{\sigma}) \sin(\theta/2), \tag{41.48}$$

reverses sign on a rotation through an odd multiple of 2π. This sign change never shows up in the law of transformation of a vector, as summarized in the formula

$$X \longrightarrow X' = RXR^* \tag{41.49}$$

Spinor defined

(two factors R; sign change in each!). The sign change does show up when one turns from a vector to a 2-component quantity that transforms according to the law

$$\xi \longrightarrow \xi' = R\xi. \tag{41.50}$$

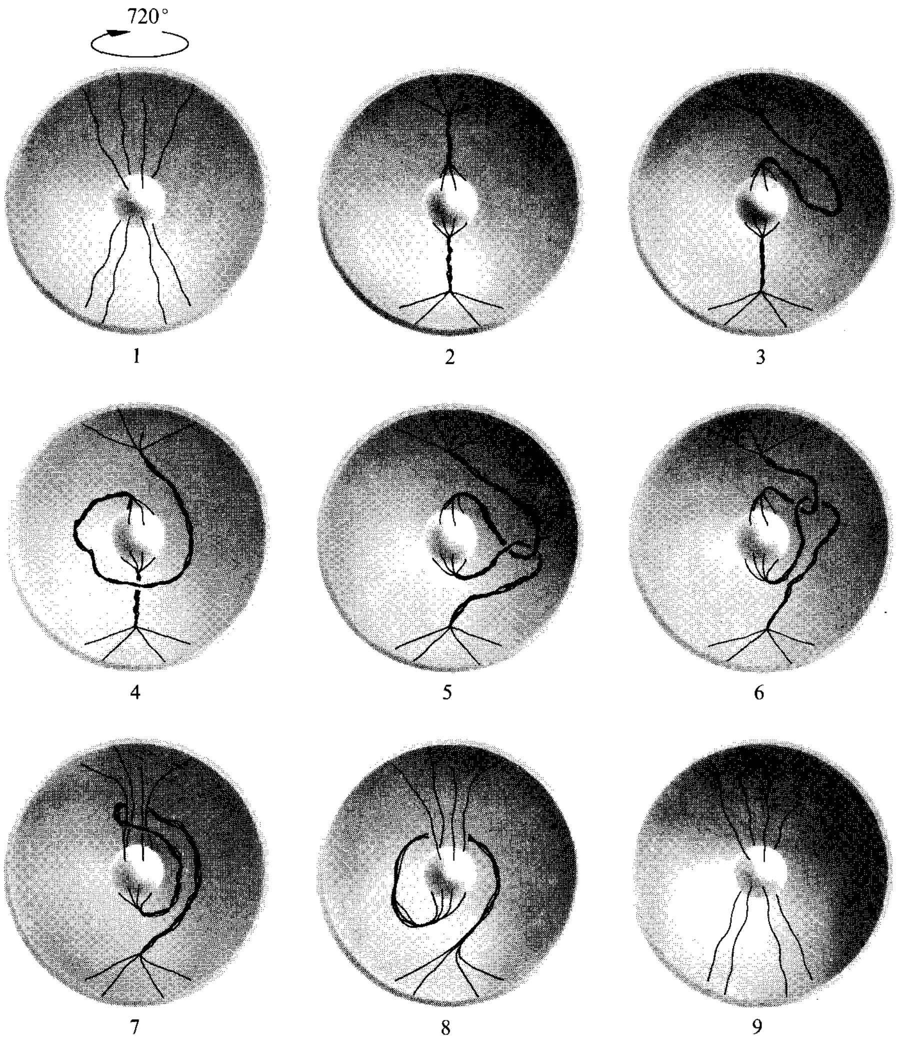


Figure 41.6.
An object is connected to its surroundings by elastic threads as in Figure 41.5. (Eight are shown here; any number could be used.) Rotating the object through 720° and then following the procedure outlined (Edward McDonald) in frames 2–8 (with the object remaining fixed), one finds that the connecting threads are left disentangled, as in frame 9 (lower right).

Such a quantity is known as a spinor. A spinor reverses sign on a 360° rotation. It therefore provides a reasonable means to keep track of the difference between the two inequivalent versions of the cube. More generally, with each orientation-entanglement relation between the cube and its surroundings one can associate a different value of the spinor ξ. Moreover, there is nothing that limits the usefulness of the spinor concept to rotations. Also, for the general combination of boost and rotation, one can write

Lorentz transformation of a spinor

$$\xi \longrightarrow \xi' = L\xi. \tag{41.51}$$

When the boost and rotation are both infinitesimal, the explicit form of this transformation is simple:

$$\xi' = [1 - (i\,d\theta/2)(\boldsymbol{n}\cdot\boldsymbol{\sigma}) + (d\boldsymbol{\beta}/2)\cdot\boldsymbol{\sigma}]\xi,$$

or, according to (41.1),

$$\begin{pmatrix}\xi'^1\\ \xi'^2\end{pmatrix} = \begin{Vmatrix} 1 + \frac{1}{2}(-i\theta_{xy} + \beta_z) & \frac{1}{2}(-i\theta_{yz} - \theta_{zx} + \beta_x - i\beta_y) \\ \frac{1}{2}(-i\theta_{yz} + \theta_{zx} + \beta_x + i\beta_y) & 1 + \frac{1}{2}(i\theta_{xy} - \beta_z)\end{Vmatrix}\begin{pmatrix}\xi^1\\ \xi^2\end{pmatrix} \tag{41.52}$$

For any combination of a boost in the z-direction of any magnitude and a finite rotation about the z-axis, one has

$$\begin{pmatrix}\xi'^1\\ \xi'^2\end{pmatrix} = \begin{Vmatrix} e^{-1/2i\theta_{xy}+1/2\beta_z} & 0 \\ 0 & e^{1/2i\theta_{xy}-1/2\beta_z}\end{Vmatrix}\begin{pmatrix}\xi^1\\ \xi^2\end{pmatrix}. \tag{41.53}$$

To keep track of the two components of the spinor, it is convenient and customary to introduce a label (capital Roman letter near beginning of alphabet) that takes on the values 1 and 2; thus (41.51) becomes

$$\xi'^A = L^A{}_B\xi^B. \tag{41.54}$$

The spinor has acquired a significance of its own through one's having pulled out half of the transformation formula

$$X' = LXL^*. \tag{41.55}$$

Second type of spinor

To be able to recover this formula, one requires the other half as well. It contains the conjugate complex of the Lorentz transformation. Therefore introduce another spinor η that transforms according to the law

$$\eta'^{\dot{U}} = \bar{L}^{\dot{U}}{}_{\dot{V}}\eta^{\dot{V}} \tag{41.56}$$

[$\dot{U} = \dot{1},\dot{2}$; $\dot{V} = \dot{1},\dot{2}$; dots and capital letters near the end of the alphabet are used to distinguish components that transform according to the conjugate complex (no transpose!) of the Lorentz spin matrix].

§41.6. CORRESPONDENCE BETWEEN VECTORS AND SPINORS

Vector regarded as a Hermitian second-rank spinor

To go back from spinors to vectors, note that the spin matrix X in (41.55) has the form

$$X = t + (\boldsymbol{x}\cdot\boldsymbol{\sigma}) = \begin{Vmatrix}(t+z) & (x-iy)\\ (x+iy) & (t-z)\end{Vmatrix} = \begin{Vmatrix}X^{1\dot{1}} & X^{1\dot{2}}\\ X^{2\dot{1}} & X^{2\dot{2}}\end{Vmatrix}, \tag{41.57}$$

where the labels receive dots or no dots according as they are coupled in (41.55) to L^* or to L. That equation of transformation becomes

$$X'^{A\dot{U}} = L^A{}_B \bar{L}^{\dot{U}}{}_{\dot{V}} X^{B\dot{V}} \tag{41.58}$$

(transpose obtained automatically by ordering of indices; thus $\bar{L}^{\dot{U}}{}_{\dot{V}}$, not $L^{*\dot{U}}{}_{\dot{V}}$). The coefficients in this transformation are identical with the coefficients in the law for the transformation of a "second-rank spinor with one index undotted and the other dotted:"

$$\xi'^A \eta'^{\dot{U}} = L^A{}_B \bar{L}^{\dot{U}}{}_{\dot{V}} \xi^\beta \eta^{\dot{V}}. \tag{41.59}$$

In this sense one can say that "a 4-vector transforms like a second-rank spinor." To be completely explicit about this connection between a 4-vector and a second-rank spinor, note from (41.57) the relations

$$\begin{aligned} X^{1\dot{1}} &= x^0 + x^3, \\ X^{1\dot{2}} &= x^1 - ix^2, \\ X^{2\dot{1}} &= x^1 + ix^2, \\ X^{2\dot{2}} &= x^0 - x^3. \end{aligned} \tag{41.60}$$

In a more compact form, one has

$$X^{A\dot{U}} = [t + (\boldsymbol{x} \cdot \boldsymbol{\sigma})]^{A\dot{U}} = x^\mu \sigma_\mu{}^{A\dot{U}} \tag{41.61}$$

where σ_0 is the unit matrix. This equation tells immediately how to go from the components of a 4-vector, or "1-index tensor," to the components of the corresponding "1,1-spinor" (one undotted and one dotted index).

With each real 4-vector x^α is associated a 1,1-spinor that is *Hermitian* in the sense that

$$X^{A\dot{U}} = \overline{X^{U\dot{A}}}. \tag{41.62}$$

An example of a Hermitian 1,1-spinor is provided by (41.61). The concept of Hermiticity can be stated in other words, and more generally. Associated with any N,N-spinor Φ with components $\Phi^{A_1 \cdots A_N \dot{U}_1 \cdots \dot{U}_N}$ is the *conjugate complex spinor* $\bar{\Phi}$ with

N,N-spinors and Hermiticity

$$(\bar{\Phi})^{A_1 \cdots A_N \dot{U}_1 \cdots \dot{U}_N} = \overline{(\Phi^{U_1 \cdots U_N \dot{A}_1 \cdots \dot{A}_N})} \tag{41.63}$$

An N,N-spinor is said to be Hermitian when it is equal to its conjugate complex.

§41.7. SPINOR ALGEBRA

Equation (41.53) showed the component ξ'^1 of a spinor rising exponentially with a boost in proportion to the factor $e^{1/2\beta_z}$, and the other component, ξ'^2 falling exponentially. If from two spinors ξ and ζ, there is to be any quantity constructed which is unaffected in value by the boost, it must be formed out of such products

Spinor algebra:

as $\xi^1\zeta^2$ and $\xi^2\zeta^1$. One can restate this product prescription in other language. Introduce the alternating symbols ε^{AB} and ε_{AB} such that $\varepsilon^{12} = \varepsilon_{12} = 1$ and

(1) ε^{AB}, ε_{AB} defined

$$\varepsilon^{AB} = -\varepsilon^{BA}, \qquad \varepsilon_{AB} = -\varepsilon_{BA}, \tag{41.64}$$

the only other nonvanishing components being $\varepsilon^{21} = \varepsilon_{21} = -1$. Define the lower-label spinor ξ_A in terms of the upper-label spinor ξ^A by the equation

(2) raising and lowering spinor indices

$$\xi_A = \xi^B \varepsilon_{BA}, \tag{41.65}$$

with the inverse

$$\xi^B = \varepsilon^{BC}\xi_C. \tag{41.66}$$

Then the scalar product of one spinor by another is defined to be

(3) scalar products of spinors

$$\xi_A \zeta^A. \tag{41.67}$$

The value of this scalar product is unaffected by any boost or rotation or combination thereof:

$$\begin{aligned} \xi'_A \zeta'^A &= \xi'^B \varepsilon_{BA} \zeta'^A \\ &= (L^B{}_D \xi^D)\varepsilon_{BA}(L^A{}_C \zeta^C) \\ &= (\det L)\xi^D \varepsilon_{DC} \zeta^C \\ &= \xi_C \zeta^C. \end{aligned} \tag{41.68}$$

The proof uses the fact that the expression $L^B{}_D \varepsilon_{BA} L^A{}_C$ (1) vanishes when $D = C$, and (2) reduces to the determinant of L (unity!) or its negative when $D = 1$, $C = 2$, or $D = 2$, $C = 1$. Note that the scalar product $\xi^A \zeta_A$ is the negative of the scalar product $\xi_A \zeta^A$. The value of the scalar product of a spinor with itself is automatically zero ("built-in null character of a spinor").

(4) the mapping between vectors and 1,1-spinors

The components of a vector with upper index have been expressed in terms of the components of a 1,1-spinor with upper indices

$$X^{A\dot{U}} = x^\mu \sigma_\mu{}^{A\dot{U}}, \tag{41.69}$$

and a similar correlation holds between vector and 1,1-spinor with lower indices; thus,

$$X_{A\dot{U}} = x_\mu \sigma^\mu{}_{A\dot{U}}. \tag{41.70}$$

(5) σ^μ defined and related to σ_μ

Here the "associated basic spin matrices" have the components

$$\sigma^\mu{}_{A\dot{U}} = \eta^{\mu\nu}\sigma_\nu{}^{B\dot{V}}\varepsilon_{BA}\varepsilon_{\dot{V}\dot{U}}, \tag{41.71}$$

or, explicitly,

$$\begin{Vmatrix} \sigma^{\mu}{}_{1\dot{1}} & \sigma^{\mu}{}_{1\dot{2}} \\ \sigma^{\mu}{}_{2\dot{1}} & \sigma^{\mu}{}_{2\dot{2}} \end{Vmatrix} = \begin{cases} -\begin{Vmatrix} 1 & 0 \\ 0 & 1 \end{Vmatrix} \text{ for } \mu = 0, \\ -\begin{Vmatrix} 0 & 1 \\ 1 & 0 \end{Vmatrix} \text{ for } \mu = 1, \\ +\begin{Vmatrix} 0 & -i \\ i & 0 \end{Vmatrix} \text{ for } \mu = 2, \\ -\begin{Vmatrix} 1 & 0 \\ 0 & -1 \end{Vmatrix} \text{ for } \mu = 3. \end{cases} \tag{41.72}$$

The same type of multiplication law holds for these matrices, $(\sigma^x)^2 = (\sigma^y)^2 = (\sigma^z)^2 = 1$, $\sigma^x\sigma^y = -\sigma^y\sigma^x = i\sigma^z$, etc., as for the matrices σ_x, σ_y, σ_z of (41.2). Between the "basic spin matrices," σ_μ, and the "associated basic spin matrices," σ^μ, the following orthogonality and normalization relations obtain:

$$\sigma_\mu{}^{A\dot{U}}\sigma^\mu{}_{B\dot{V}} = -2\,\delta^A_B\,\delta^{\dot{U}}_{\dot{V}} \tag{41.73}$$

and

$$\sigma_\mu{}^{A\dot{U}}\sigma^\nu{}_{A\dot{U}} = -2\,\delta^\nu_\mu. \tag{41.74}$$

One can use these relations to "go back from a quantity expressed as a 1,1-spinor ('spinor equivalent of a vector') to the same quantity expressed directly as a vector (first-rank tensor)." Thus, multiply through (41.61) on both sides by $-\frac{1}{2}\sigma^\nu{}_{A\dot{U}}$, sum over the spinor indices, and employ (41.74) to find the contravariant components of the vector,

$$x^\nu = -\frac{1}{2}\sigma^\nu{}_{A\dot{U}}X^{A\dot{U}}. \tag{41.75}$$

Similarly from (41.70) and (41.73) one finds the covariant components,

$$x_\nu = -\frac{1}{2}\sigma_\nu{}^{A\dot{U}}X_{A\dot{U}}. \tag{41.76}$$

An N-index tensor T lets itself be expressed in spinor language ("spinor equivalent of the tensor") by a generalization of (41.61) or (41.70); thus, for a mixed tensor of third order, one has

(6) the mapping between rank-N tensors and N,N-spinors

$$T_{A\dot{U}}{}^{B\dot{V}C\dot{W}} = \sigma^\alpha{}_{A\dot{U}}\sigma_\beta{}^{B\dot{V}}\sigma_\gamma{}^{C\dot{W}}T_\alpha{}^{\beta\gamma} \tag{41.77}$$

and the converse relation

$$T_\alpha{}^{\beta\gamma} = \left(-\frac{1}{2}\right)^3\sigma_\alpha{}^{A\dot{U}}\sigma^\beta{}_{B\dot{V}}\sigma^\gamma{}_{C\dot{W}}T_{A\dot{U}}{}^{B\dot{V}C\dot{W}}. \tag{41.78}$$

Box 41.1 gives the spinor representation of several simple tensors.

Box 41.1 SPINOR REPRESENTATION OF CERTAIN SIMPLE TENSORS IN THE CONTEXT OF A LOCAL LORENTZ FRAME

Quantity	*Tensor language*	*Spinor language*
General 4-vector	x^α (four complex numbers)	$X^{A\dot{U}}$ (4 complex numbers)
Real 4-vector (example: 4-momentum)	$x^\alpha = \bar{x}^\alpha$ (four real numbers)	$X^{A\dot{U}} = \overline{(X^{U\dot{A}})}$ (2 real components, 1 distinct complex component)
Null 4-vector	$\eta_{\alpha\beta}x^\alpha x^\beta = 0$	det $X^{A\dot{U}} = 0$ [see (41.57)]; hence there exist two spinors ξ^A and η^U such that $X^{A\dot{U}} = \xi^A\eta^{\dot{U}}$.
Future-pointing real null 4-vector (such as 4-momentum of a photon)	$x^\alpha = \bar{x}^\alpha$ $\eta_{\alpha\beta}x^\alpha x^\beta = 0$ $x^0 > 0$	There exists a spinor ξ^A (two complex numbers, unique up to a common multiplicative phase factor $e^{i\delta}$) such that $X^{A\dot{U}} = \xi^A(\bar{\xi})^{\dot{U}}$
Past-pointing real null 4-vector	$x^0 < 0$	$X^{A\dot{U}} = -\xi^A(\bar{\xi})^{\dot{U}}$.
Real bivector or 2-form (such as Maxwell field)	$F_{[\alpha\beta]}$ (subscript implying $F_{\alpha\beta} = -F_{\beta\alpha}$; six distinct real components)	There exists a symmetric spinor ϕ_{AB} (three distinct complex components ϕ_{11}, ϕ_{12}, ϕ_{22}) such that $F_{A\dot{U}B\dot{V}} = \phi_{AB}\varepsilon_{\dot{U}\dot{V}} + \varepsilon_{AB}(\bar{\phi})_{\dot{U}\dot{V}}$
Real 2-form dual to foregoing real 2-form	${}^*F_{\alpha\beta} = \frac{1}{2}\varepsilon_{\alpha\beta\gamma\delta}F^{\gamma\delta}$	${}^*F_{A\dot{U}B\dot{V}} = -i\phi_{AB}\varepsilon_{\dot{U}\dot{V}} + i\varepsilon_{AB}(\bar{\phi})_{\dot{U}\dot{V}}$ (duality for 2-form corresponds to multiplication of spinor ϕ_{AB} by $-i$)
Real fourth-order tensor with symmetries of Weyl conformal curvature tensor; that is, with symmetries of Riemann curvature tensor and with additional requirement of vanishing Ricci tensor ("empty space;" "vacuum Riemann tensor")	$C_{\alpha\beta\gamma\delta} = C_{([\alpha\beta][\gamma\delta])}$ (antisymmetric in first two indices; antisymmetric in last two indices; symmetric against interchange of first pair with second pair) $C^\alpha{}_{[\beta\gamma\delta]} = 0$ (20 algebraically distinct components, as for the Riemann tensor, reduced to 10 by the further vacuum condition:) $C^\alpha{}_{\beta\alpha\delta} = 0$	There exists a completely symmetric spinor ψ_{ABCD} with five distinct complex components, ψ_{1111} ψ_{1112} ψ_{1122} ψ_{1222} ψ_{2222} such that $C_{A\dot{U}B\dot{V}C\dot{W}D\dot{X}} = \psi_{ABCD}\varepsilon_{\dot{U}\dot{V}}\varepsilon_{\dot{W}\dot{X}} + \varepsilon_{AB}\varepsilon_{CD}\bar{\psi}_{\dot{U}\dot{V}\dot{W}\dot{X}}$

In some treatises on spinor analysis, the factor $(-\frac{1}{2})^N$ in equations like (41.78) is eliminated by the following double prescription: (1) insert into the matrices σ_μ and σ^μ a factor $1/\sqrt{2}$ not included above; and (2) use for the standard metric not diag $\eta_{\mu\nu} = (-1, 1, 1, 1)$ as above, but $(1, -1, -1, -1)$. This prescription was not adopted here (1) because the introduction of $1/\sqrt{2}$ in the matrices σ_x, σ_y, σ_z would put them out of line with the Pauli matrices as used for many years throughout

Quantity	*Tensor language*	*Spinor language*
Fully developed Riemann curvature tensor (space where matter is present)	$R_{\alpha\beta\gamma\delta} = R_{[\alpha\beta]\gamma\delta} = R_{\alpha\beta[\gamma\delta]} = R_{([\alpha\beta][\gamma\delta])}$ $R_{\alpha[\beta\gamma\delta]} = 0$ (20 algebraically distinct components)	There exists a completely symmetric spinor ψ_{ABCD} ("Weyl" or "conformal" part of curvature, or part of nonlocal origin) and a scalar Λ (measure of trace of part of curvature of local origin) and a spinor $\Phi_{AB\dot{U}\dot{V}} = \Phi_{(AB)(\dot{U}\dot{V})} = (\bar{\Phi})_{AB\dot{U}\dot{V}}$ (measure of trace-free part of curvature of local origin; last of the three irreducible parts of the curvature tensor) such that $R_{A\dot{U}B\dot{V}C\dot{W}D\dot{X}} = \psi_{ABCD}\varepsilon_{\dot{U}\dot{V}}\varepsilon_{\dot{W}\dot{X}}$ $+ \varepsilon_{AB}\varepsilon_{CD}(\bar{\psi})_{\dot{U}\dot{V}\dot{W}\dot{X}}$ $+ 2\Lambda(\varepsilon_{AC}\varepsilon_{BD}\varepsilon_{\dot{U}\dot{V}}\varepsilon_{\dot{W}\dot{X}}$ $+ \varepsilon_{AB}\varepsilon_{CD}\varepsilon_{\dot{U}\dot{X}}\varepsilon_{\dot{V}\dot{W}})$ $+ \varepsilon_{AB}\Phi_{CD\dot{U}\dot{V}}\varepsilon_{\dot{W}\dot{X}}$ $+ \varepsilon_{CD}\Phi_{AB\dot{W}\dot{X}}\varepsilon_{\dot{U}\dot{V}}.$
Each physical quantity is described by a geometric object. Every local physical quantity is described by a mathematical quantity that transforms under a proper local Lorentz transformation as an "irreducible representation of the group $L\uparrow_{+}$ of proper Lorentz transformations."	Each local physical quantity is described by a tensor with its own rank and specific symmetry properties.	In order to provide the required finite irreducible representation of $L\uparrow_{+}$ to represent a local physical quantity, the associated spinor must be completely symmetric in all of its undotted indices, and also completely symmetric in all its dotted indices [Gel'fand (1963)].

atomic and nuclear physics, and (2) because a positive definite metric within a spacelike hypersurface has the advantage of naturalness for the analysis of the initial-value problem of geometrodynamics and for the definition of what one means by a 3-geometry. The price of the factor $(-\frac{1}{2})^N$ is paid here for these advantages. Conventions that avoid this price are preferable for extensive spinor computations; see, e.g., Pirani (1965) or Penrose (1968a).

§41.8. SPIN SPACE AND ITS BASIS SPINORS

Linear independence of spinors

The "space" of elementary spinors is two-dimensional. Therefore it is spanned by any two linearly independent spinors λ_A and μ_A. Moreover, it is easy to diagnose a pair of spinors for possible linear dependence, that is, for existence of a relation of the form $\mu_A = \text{const}\,\lambda_A$. In this event, the scalar product of μ_A with λ^A, like the scalar product of λ_A with λ^A (41.67) automatically vanishes. Therefore a nonvanishing scalar product

$$\lambda_A\mu^A \neq 0 \tag{41.79}$$

is a necessary and sufficient condition for the linear independence of two spinors.

The general 4-vector lets itself be represented as a linear combination of four basis vectors. Similarly the general spinor lets itself be represented as a linear combination of two basis spinors:

$$\omega^A = \lambda\xi^A + \mu\eta^A. \tag{41.80}$$

Basis spinors and spinor mates

Here it is understood that the term "basis spinor" implies that ξ^A and η^A satisfy the normalization condition

$$\xi_A\eta^A = 1\,(= -\eta_A\xi^A). \tag{41.81}$$

From this condition one derives simple expressions for the expansion coefficients in (41.80):

$$\begin{aligned}\lambda &= -\eta_A\omega^A\,(= \omega_B\eta^B),\\ \mu &= \xi_A\omega^A\,(= -\omega_B\xi^B).\end{aligned} \tag{41.82}$$

Inserting these expansion coefficients back into (41.80) will reproduce any arbitrarily chosen spinor ω^A. In other words, the following equation has to be an identity in the components of ω_B:

$$\omega^A = \varepsilon^{AB}\omega_B \equiv (\xi^A\eta^B - \eta^A\xi^B)\omega_B. \tag{41.83}$$

From this circumstance, it follows that the components of the two basic spinors are linked by the equations

$$\xi^A\eta^B - \eta^A\xi^B = \varepsilon^{AB}. \tag{41.84}$$

Given two basis spinors ξ^A and η^A, one can get two equally good basis spinors by writing

$$\begin{aligned}\xi^A{}_{\text{new}} &= \xi^A,\\ \eta^A{}_{\text{new}} &= \eta^A + \alpha\xi^A,\end{aligned} \tag{41.85}$$

with α any real or complex constant, as one checks at once by substitution into (41.81) or (41.84). The most general "spinor mate" to a given spinor ξ^A, satisfying the normalization condition (41.81), has this form (41.85).

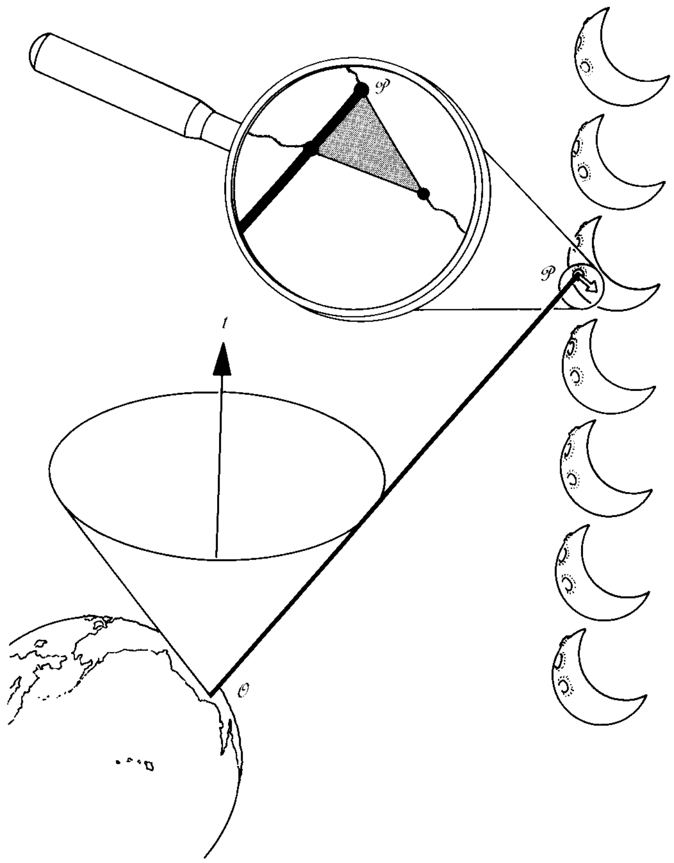


Figure 41.7.
Spinor represented by (1) "flagpole" [Penrose terminology; track of pulse of light; null vector $\mathscr{O}\mathscr{P}$] plus (2) "flag" [arrow ($\mathscr{P} \longrightarrow$) flashed onto surface of moon by laser pulse from earth or, in expanded view in the inset above, a flag itself, substituted for the arrow] plus (3) the orientation-entanglement relation between the flag and its surroundings [symbolized by strings drawn from corners of flag to surroundings]. When the spinor itself is multiplied by a factor $\rho e^{i\sigma}$, the components of the null vector (flagpole) are multiplied by the factor ρ^2 and the flag is rotated through the angle 2σ about the flagpole.

§41.9. SPINOR VIEWED AS FLAGPOLE PLUS FLAG PLUS ORIENTATION-ENTANGLEMENT RELATION

How can one visualize a spinor? Aim the laser, pull the trigger, and send a megajoule pulse from the here and now (event $\mathscr{O}$) to the there and then (event $\mathscr{P}$: center of the crater Aristarchus, 400,000 km from $\mathscr{O}$ in space, and 400,000 km from $\mathscr{O}$ in light-travel time). The laser has been designed to produce, not a mere spot of light, but an illuminated arrow. Following Roger Penrose, speak of the null vector $\mathscr{O}\mathscr{P}$ as a "flagpole," and of the illuminated arrow as a "flag." A spinor (Figure 41.7) consists of this combination of (1) null flagpole plus (2) flag plus (3) the orientation-

Geometric representation of a spinor:

entanglement relation between the flag and its surroundings. "Rotate the flag" by repeatedly firing the laser, with a bit of rotation of the laser about its axis between one firing and the next. When the flag has turned through 360° and has come back to its original direction, the spinor has reversed sign. A rotation of the flag about the flagpole through any even multiple of 2π restores the spinor to its original value.

(1) null vector (flagpole), plus

One goes from a spinor ξ, a mathematical object with two complex components ξ^1 and ξ^2, to the geometric object of "flagpole plus flag plus orientation-entanglement relation" in two steps: first the pole, then the flag. Thus, go from the spinor ξ^A to the real null 4-vector of the "pole" by way of the formula

$$x^\alpha \longrightarrow X^{A\dot{U}} = \xi^A(\bar{\xi})^{\dot{U}} \tag{41.86}$$

or

$$\begin{Vmatrix} (t+z) & (x-iy) \\ (x+iy) & (t-z) \end{Vmatrix} = \begin{Vmatrix} \xi^1\bar{\xi}^{\dot{1}} & \xi^1\bar{\xi}^{\dot{2}} \\ \xi^2\bar{\xi}^{\dot{1}} & \xi^2\bar{\xi}^{\dot{2}} \end{Vmatrix}. \tag{41.87}$$

The matrix on the right has its first row identical up to a factor ξ^1/ξ^2 with its second row. Therefore the determinant of the matrix on the right vanishes. So also for the left. Therefore the 4-vector $\mathscr{OP} = (t, x, y, z)$ is a null vector. One "stretches" this vector by a factor ρ^2 when one multiplies the spinor ξ^A by the nonzero complex number $\lambda = \rho e^{i\sigma}$ (ρ, σ real); however, the vector is unchanged in direction. The 4-vector is also unaffected by the choice of the angle σ. In other words, this null 4-vector is uniquely fixed by the spinor; but the spinor is not fixed with all uniqueness by the 4-vector. To a given 4-vector corresponds a whole family of spinors. They differ from one another by a multiplicative phase factor of the form $e^{i\sigma}$ ("flag factor").

(2) bivector (flag) and its orientation-entanglement relation

Looking further to see the influence of the flag factor showing up, turn from a real vector (four components) generated out of the spinor ξ^A to a real bivector (six components) generated out of the same spinor:

$$\begin{gathered} F^{\mu\nu} \longrightarrow F^{AB\dot{U}\dot{V}} = \xi^A\xi^B\epsilon^{\dot{U}\dot{V}} + \epsilon^{AB}(\bar{\xi})^{\dot{U}}(\bar{\xi})^{\dot{V}}, \\ \mu \longrightarrow A\dot{U};\ \nu \longrightarrow B\dot{V}. \end{gathered} \tag{41.88}$$

That this quantity has no more than six distinct components ($F^{\mu\nu} = -F^{\nu\mu}$) follows from interchanging A with B and $\dot{U}$ with $\dot{V}$, and noting the resultant change in sign on the righthand side of (41.88). To unfold the meaning of this bivector, look in (41.88) for every appearance of the alternating factor ε^{AB}. Wherever such a factor appears, insert the expression (41.84) for this factor in terms of the starting spinor ξ^A and insert the additional spinor η^A that is needed, along with ξ^A, to supply a basis for all spinors. In this way, find

$$\begin{aligned} F^{\mu\nu} \longrightarrow F^{AB\dot{U}\dot{V}} &= \xi^A\xi^B(\bar{\xi}^{\dot{U}}\bar{\eta}^{\dot{V}} - \bar{\eta}^{\dot{U}}\bar{\xi}^{\dot{V}}) + (\xi^A\eta^B - \eta^A\xi^B)\bar{\xi}^{\dot{U}}\bar{\xi}^{\dot{V}} \\ &= \xi^A\bar{\xi}^{\dot{U}}(\xi^B\bar{\eta}^{\dot{V}} + \eta^B\bar{\xi}^{\dot{V}}) - (\xi^A\bar{\eta}^{\dot{U}} + \eta^A\bar{\xi}^{\dot{U}})\xi^B\bar{\xi}^{\dot{V}} \\ &= X^{A\dot{U}}Y^{B\dot{V}} - Y^{A\dot{U}}X^{B\dot{V}} \longrightarrow x^\mu y^\nu - y^\mu x^\nu. \end{aligned} \tag{41.89}$$

Thus the 2,2-spinor built from ξ^A represents a bivector constructed out of the two 4-vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Of these, the first is the "real null vector of the flagpole," already seen to be determined uniquely by the spinor ξ^A. The second vector,

$$y^\alpha \longrightarrow Y^{A\dot{U}} = \xi^A \bar{\eta}^{\dot{U}} + \eta^A \bar{\xi}^{\dot{U}}, \tag{41.90}$$

is also determined by ξ^A, but not uniquely, because the "spinor mate," η^A, to ξ^A is not unique. Go from one choice of mate, η^A, to a new choice of mate (equation 41.85),

$$\eta^A{}_{\text{new}} = \eta^A + \alpha \xi^A. \tag{41.91}$$

Then the real 4-vector y^μ goes to the new real 4-vector

$$y^\mu{}_{\text{new}} = y^\mu + (\alpha + \bar{\alpha}) x^\mu. \tag{41.92}$$

Were the 4-vector $\boldsymbol{y}$ unique, there would project out from the flagpole, not a flag but an arrow. The range of values open for the real constant $\alpha + \bar{\alpha}$ makes one arrow into many arrows, all coplanar; hence the flag of Penrose. Otherwise stated, the choice of a spinor ξ^A fixes no individual arrow, but does fix the totality of the collection of arrows, and thus uniquely specifies the flag.

The 4-vector $\boldsymbol{y}$ (and with it $\boldsymbol{y}_{\text{new}}$) is orthogonal to the null 4-vector $\boldsymbol{x}$,

$$\begin{aligned} \boldsymbol{x} \cdot \boldsymbol{y} &= x_\beta y^\beta = -\frac{1}{2} X_{A\dot{U}} Y^{A\dot{U}} \\ &= -\frac{1}{2} \xi_A \bar{\xi}_{\dot{U}} (\xi^A \bar{\eta}^{\dot{U}} + \eta^A \bar{\xi}^{\dot{U}}) = 0, \end{aligned} \tag{41.93}$$

and spacelike,

$$\begin{aligned} \boldsymbol{y} \cdot \boldsymbol{y} &= -\frac{1}{2} (\xi_A \bar{\eta}_{\dot{U}} + \eta_A \bar{\xi}_{\dot{U}})(\xi^A \bar{\eta}^{\dot{U}} + \eta^A \bar{\xi}^{\dot{U}}) \\ &= -\frac{1}{2} (\xi_A \eta^A)(\bar{\eta}_{\dot{U}} \bar{\xi}^{\dot{U}}) - \frac{1}{2} (\eta_A \xi^A)(\bar{\xi}_{\dot{U}} \bar{\eta}^{\dot{U}}) = 1 \end{aligned} \tag{41.94}$$

("unit length of flag").

Rotation of flag about flagpole

Multiplication of the spinor ξ^A by the "flag factor" $e^{i\sigma}$ rotates the flag about the flagpole by the angle 2σ, because the spinor mate, η^A, of ξ^A is multiplied by the factor $e^{-i\sigma}$ [see the normalization condition (41.81)]. These changes alter the vector $\boldsymbol{y}$ to a rotated vector $\boldsymbol{y}_{\text{rot}}$, with

$$\begin{aligned} y^\alpha_{\text{rot}} \longrightarrow Y^{A\dot{U}}_{\text{rot}} &= e^{2i\sigma} \xi^A \bar{\eta}^{\dot{U}} + e^{-2i\sigma} \eta^A \bar{\xi}^{\dot{U}} \\ &= \cos 2\sigma (\xi^A \bar{\eta}^{\dot{U}} + \eta^A \bar{\xi}^{\dot{U}}) + \sin 2\sigma (i \xi^A \bar{\eta}^{\dot{U}} - i \eta^A \bar{\xi}^{\dot{U}}) \\ &\longrightarrow y^\alpha \cos 2\sigma + z^\alpha \sin 2\sigma. \end{aligned} \tag{41.95}$$

Here the 4-vector $\boldsymbol{z}$ shares with the vector $\boldsymbol{y}$ these properties: it is (1) real, (2) spacelike, (3) of unit magnitude, (4) orthogonal to the null 4-vector $\boldsymbol{x}$ of the flagpole, and (5) uniquely specified by the original spinor ξ^A up to the additive real

Equations relating spinor, flagpole, and flag

multiple $(\alpha + \bar{\alpha})$ of $\boldsymbol{x}$. In addition, $\boldsymbol{z}$ and $\boldsymbol{y}$ are orthogonal. Thus $\boldsymbol{y}$ and $\boldsymbol{z}$ provide basis vectors in the two-dimensional space in which—to overpictorialize—the "tip of the flag" undergoes its rotation.

Recapitulate by returning to the laser pulse. Two numbers, such as the familiar polar angles θ (angle with the z-axis) and ϕ (azimuth around z-axis from x-axis) tell the direction of its flight. A third number, r, gives the distance to the moon and also the travel time for light to reach the moon. A fourth number, an angle ψ, tells the azimuth of the illuminated arrow shot onto the surface of the moon, this azimuth to be measured from the $\boldsymbol{e}_\theta$ direction (where $\psi = 0$), around the flagpole in a righthanded sense. Then the spinor associated with the flagpole plus flag (rotated arrow) is

$$\begin{pmatrix} \xi^1 \\ \xi^2 \end{pmatrix} = (2r)^{1/2} \begin{pmatrix} \cos(\theta/2) e^{-i\phi/2 + i\psi/2} \\ \sin(\theta/2) e^{i\phi/2 + i\psi/2} \end{pmatrix} \tag{41.96}$$

according to the conventions adopted here [see (41.87)]. The mate η_A to this spinor, unique up to an additive multiple of ξ^A, is

$$\begin{pmatrix} \eta^1 \\ \eta^2 \end{pmatrix} = (2r)^{-1/2} \begin{pmatrix} -\sin(\theta/2) e^{-i\phi/2 - i\psi/2} \\ \cos(\theta/2) e^{i\phi/2 - i\psi/2} \end{pmatrix}. \tag{41.97}$$

The 4-vector of the flagpole determined by ξ^A is found from (41.87):

$$\begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix} = \begin{pmatrix} r \\ r \sin\theta \cos\phi \\ r \sin\theta \sin\phi \\ r\cos\theta \end{pmatrix}. \tag{41.98}$$

To determine the flag itself, one requires, in addition to x^α, the unit spacelike 4-vector y^α, normal to x^α, and unique up to an additive real multiple of x^α. This vector is evaluated by use of (41.90) and has the form

$$\begin{pmatrix} y^0 \\ y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} 0 \\ \cos\theta \cos\phi \cos\psi + \sin\phi \sin\psi \\ \cos\theta \sin\phi \cos\psi - \cos\phi \sin\psi \\ -\sin\theta\cos\psi \end{pmatrix}. \tag{41.99}$$

From these expressions for x^μ and y^μ, one calculates the components of the bivector ("flag") $F^{\mu\nu} = x^\mu y^\nu - y^\mu x^\nu$ by simple arithmetic.

§41.10. APPEARANCE OF THE NIGHT SKY: AN APPLICATION OF SPINORS

Attention has gone here to extracting all four pieces of information contained in a spinor: separation in time (equal to separation in space), direction in space, and

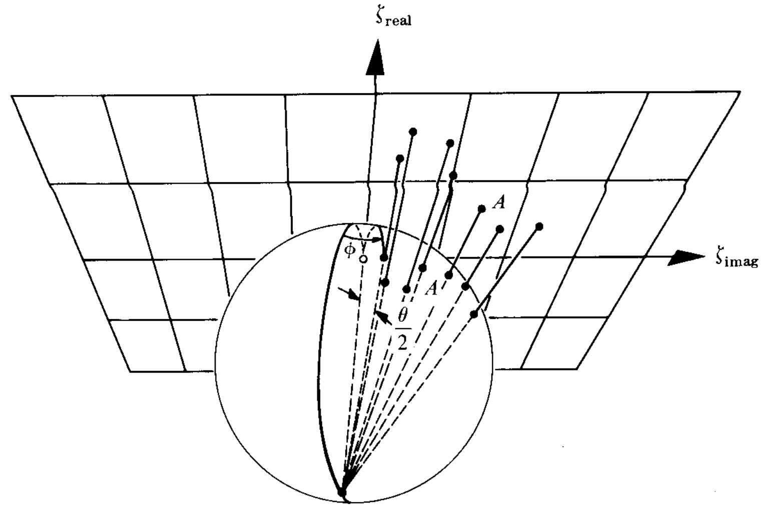


Figure 41.8.
Representation of a direction in space (one of the stars of the Big Dipper, regarded as a point on the celestial sphere) as a point in the complex ζ plane (ζ = ratio ξ^2/ξ^1 of spinor components) by stereographic projection from the South Pole.

Spinors used to analyze "Lorentz transformations" of appearance of night sky

rotation about that direction. Turn now to an application where not all that information is needed. Look at the night sky and ask (1) how to describe its appearance and (2) how to change that appearance. As one way to describe its appearance, give the direction of each star. Abandon any concern about the distance of the star, and any concern about any rotation ψ about the flagpole. In other words, the complex factor

$$(2r)^{1/2}e^{i\psi/2}$$

common to ξ^1 and ξ^2 drops from attention. All that is left as significant is the ratio ζ of these spinor components:

$$\zeta = \xi^2/\xi^1 = \tan(\theta/2)e^{i\phi}. \tag{41.100}$$

To give the one complex number ζ ("stereographic coordinate;" Figure 41.8) for each star in the sky is to catalog the pattern of the stars.

Let the observer change his stance. The celestial sphere appears to rotate. Or let him rocket past his present location in the direction of the North Star with some substantial fraction of the velocity of light. To him all that portion of the celestial sphere is opened out, as if by a magnifying glass. To compensate, the remaining stars are packed into a smaller angular compass. Any such rotation or boost or combination of rotation and boost being described in spinor language by a transformation of the form

$$\xi^A \longrightarrow \xi^A_{\text{new}} = L^A{}_B\xi^B, \tag{41.101}$$

implies a transformation of the complex stereographic coordinate of any given star of the form

$$\zeta \longrightarrow \zeta_{\text{new}} = \frac{\xi^2_{\text{new}}}{\xi^1_{\text{new}}} = \frac{L^2{}_2\zeta + L^2{}_1}{L^1{}_2\zeta + L^1{}_1}. \tag{41.102}$$

In the special case of a boost in the z-direction with velocity parameter α (velocity $\beta = \tanh \alpha$), the off-diagonal components $L^1{}_2$ and $L^2{}_1$ vanish. The magnification of the overhead sky then expresses itself in the simple formula

$$\zeta_{\text{new}} = e^{\alpha}\zeta$$

or

$$\phi_{\text{new}} = \phi,$$
$$\tan(\theta_{\text{new}}/2) = e^{\alpha}\tan(\theta/2). \tag{41.103}$$

Contrary to this prediction and false expectation, no magnification at all is achieved of the regions around the North Star by moving with high velocity in that direction. On the contrary, any photon coming in from a star a little off that direction, with a little transverse momentum, keeps that transverse momentum in the new frame; but its longitudinal momentum against the observer is augmented by his motion. Thus the ratio of the momenta is decreased, and the observed angle relative to the North Star is also decreased. The consequence is not magnification, but diminution ("looking through the wrong end of a telescope"). The correct formula is not (41.103) but

$$\tan(\theta_{\text{new}}/2) = e^{-\alpha}\tan(\theta/2) \tag{41.104}$$

(reversal of the sign of α). The reason for this correction is not far to seek. The spinor analysis so far had dealt with an outgoing light pulse, and a 4-vector with positive time component. That feature was built into the formula adopted to tie the spinor to the 4-vector,

$$r\mathbf{1} + (\boldsymbol{r}\cdot\boldsymbol{\sigma}) \equiv X = \|\xi^A\bar{\xi}^{\dot{U}}\|. \tag{41.105}$$

In contrast the 4-vector that reaches back to the origin of an incoming photon has a time component that is negative (or, alternatively, sign-reversed space components)! For any null 4-vector with negative time component, one employs instead of (41.105) the formula

$$X = -\|\xi^A\bar{\xi}^{\dot{U}}\|. \tag{41.106}$$

It is enough to mention here this point of principle without going through the details that give the altered sign for α in (41.104). From now on, to preserve the previous arithmetic, change the problem. Deal, not with incoming photons, but with outgoing photons. Replace the receiving telescope by the projector of a planetarium. It projects out into space a separate beam of light for each star of the Big Dipper and also one for the North Star itself. Let an observer move in the positive z-direction with velocity parameter α. In his frame of reference the beams actually will be widened out in full accord with (41.103).

"The magnification process changes the size of the Big Dipper but not its shape."

This statement is at the same time true and false. It is true of the Dipper and of any other constellation to the extent that the angular dimensions of that constellation can be idealized to be small compared to the entire compass of the sky. It is false in the sense that any well-rounded projected constellation, however small it may appear to an observer at rest with respect to the earth, can always be so "opened out" by the observer putting on any sufficiently high velocity, the observer still being near the earth, that the constellation encompasses a major fraction of the sky.

Lorentz transformations leave angles on sky unchanged ("conformal invariance")

That the "Lorentz-transformation-induced magnification" of a small object does not change its shape can be seen in three ways. (1) Stereographic projection (Figure 41.8) and "fractional linear transformation" (41.102) are both known to leave all angles unchanged ["conformal invariance;" see for example Penrose (1959)] and known even to turn every old circle into a new circle. (2) Consider a given star, M, in the constellation and immediate neighbors, L and N, just below it and just above it in the count of the members of that constellation. Consider the flagpole pointed at M and the flag pointed first from M to L, then from M to N. The flag has turned about the flagpole through an angle ψ. The two corresponding spinors therefore differ by a phase factor $e^{i\psi/2}$. They differ in no other way. After an arbitrary Lorentz transformation they still differ by the phase factor $e^{i\psi/2}$, and in no other way. The angle between the arcs ML and MN on the celestial sphere therefore remains at its original value ψ after the Lorentz transformation (again conformal invariance of patterns on the celestial sphere!). (3) An even more elementary calculation shows that infinitesimal arc lengths on the unit celestial sphere in the direction of increasing θ and arc lengths in the direction of increasing ϕ are magnified in the same proportion, thus leaving unchanged the angle between arc and arc (conformal invariance). Thus, consider a photon shot out from the planetarium projector to a point on the celestial sphere ("planetarium version of a Big-Dipper star") with inclination θ to the z-axis, as seen by an observer at rest relative to the earth. From the standard laws of transformation of angles in a Lorentz transformation ("aberration"; Box 2.4), one has for the sine of the transformed angle

$$\sin\theta_{\text{new}} = \frac{(1-\beta^2)^{1/2}}{1-\beta\cos\theta}\sin\theta \tag{41.107}$$

and (by differentiating the expression for the cosine of the transformed angle)

$$d\theta_{\text{new}} = \frac{(1-\beta^2)^{1/2}}{1-\beta\cos\theta}\,d\theta. \tag{41.108}$$

From these expressions it follows at once that the inclination, relative to a meridian line, on the transformed celestial sphere is identical to the direction, relative to the same meridian line, on the original celestial sphere:

$$\begin{aligned}\tan\begin{pmatrix}\text{new}\\ \text{inclination}\end{pmatrix} &= \frac{\sin\theta_{\text{new}}\,d\phi_{\text{new}}}{d\theta_{\text{new}}} = \frac{\sin\theta\,d\phi}{d\theta}\\ &= \tan\begin{pmatrix}\text{original}\\ \text{inclination}\end{pmatrix}\end{aligned} \tag{41.109}$$

(again conformal invariance!).

So much for the elementary spinor and what it has to do with a null vector, with a "flagpole" pointed to the celestial sphere, and with rotation of a "flag" about such a flagpole.

§41.11. SPINORS AS A POWERFUL TOOL IN GRAVITATION THEORY

Spinor formalism in curved spacetime

Just as vectors, tensors, and differential forms are easily generalized from flat spacetime to curved, so are spinors.

Each event $\mathscr{P}$ in curved spacetime possesses a tangent space. In that tangent space reside and operate all the vectors, tensors, and forms located at $\mathscr{P}$. The geometry of the tangent space is Lorentzian ("local Lorentz geometry at $\mathscr{P}$"), since the scalar product of any two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ at $\mathscr{P}$, expressed in an orthonormal frame at $\mathscr{P}$, is

$$\boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{g}(\boldsymbol{u}, \boldsymbol{v}) = \eta_{\hat{\alpha}\hat{\beta}} u^{\hat{\alpha}} v^{\hat{\beta}}.$$

Thus, there is no mathematical difference between the tangent space at $\mathscr{P}$ on the one hand, and flat spacetime on the other. Whatever mathematical can be done in the one can also be done in the other. In particular, *the entire formalism of spinors, developed originally in flat spacetime, can be carried over without change to the tangent space at the arbitrary event $\mathscr{P}$ in curved spacetime.*

Let it be done. Now spinors reside at every event in curved spacetime; and at each event one can translate back and forth between spinor language and tensor language, using the equations (valid in orthonormal frames) of §§41.6 and 41.7.

Spinors needed when analyzing fermions in gravitational fields

Spinors in curved spacetime are an indispensible mathematical tool, when one wishes to study the influence of gravity on quantized particles of half-integral spin (neutrinos, electrons, protons, . . .). Consider, for example, Hartle's (1971) proof that a black hole cannot exert any long-range, weak-interaction forces on external matter (i.e., that a black hole has no "weak-interaction hair"). His proof could not function without a spinor description of neutrino fields in curved spacetime. Similarly for Wheeler's (1971b) analysis of the quasibound states of an electron in the gravitational field of a small black hole (gravitational radius $\sim 10^{-13}$ cm): it requires solving the Dirac equation for a spin-$\frac{1}{2}$ particle in the curved spacetime geometry of Schwarzschild. For a detailed discussion of the Dirac equation in curved spacetime see, e.g., Brill and Wheeler (1957).

Equivalence of spinor and tensor formalisms

To use the mathematics of spinors, one need not be dealing with quantum theory or with particles of half-integral spin. The spinor formalism is perfectly applicable in situations where only integral-spin entities (scalars, vectors, tensors) are in view, and where in fact, the spinor formalism is fully equivalent to the tensor formalism that pervades earlier chapters of this book. Equations (41.77) and (41.78) provide the translation from one formalism to the other, once an orthonormal frame has been chosen at each event in spacetime.

Certain types of problems in gravitation theory are far more tractable in the language of spinors than in the language of tensors. Examples are as follows.

2

(1) Geometric Optics (the theory of "null congruences of geodesics")

Applications of spinor formalism in classical gravitation theory

Here spinors make almost trivial the lengthy tensor algebra needed in derivations of the "focusing theorem" [equation (22.37)]; and they yield an elegant, simple formalism for discussing and calculating how, with increasing affine parameter, a bundle of rays alters its size ("expansion"), its shape ("shear"), and its orientation ("rotation"). See, e.g., Sachs (1964), Pirani (1965), or Penrose (1968a) for a review and the original references.

(2) Radiation Theory in Curved Spacetime (both gravitational and electromagnetic)

Spinors provide the most powerful of all formalisms for decomposing radiation fields into spherical harmonics and for manipulating their decomposed components. See, for example, Price's (1972a,b) analysis of how a perturbed Schwarzschild black hole radiates away all its radiatable perturbations, be they electromagnetic perturbations, gravitational perturbations, or perturbations in a fictitious field of spin 17; see, similarly, the analysis by Fackerell and Ipser (1972) and by Ipser (1971) of electromagnetic perturbations of a Kerr black hole, and the analysis by Teukolsky (1972) of gravitational perturbations of a Kerr hole. Spinors also yield an elegant and powerful analysis of the "$1/r$" expansion of a radiation field flowing out from a source into asymptotically flat space. Among its results is a "*peeling theorem,*" which describes the algebraic properties of the coefficients in a $1/r$ expansion of the Riemann tensor. See, e.g., Sachs (1964) or Pirani (1965) for reviews and original references.

(3) Algebraic Properties of Curvature Tensors

The spinor formalism is a more powerful method than any other for deriving the "Petrov-Pirani algebraic classification of the conformal curvature tensor," and for proving theorems about algebraic properties of curvature tensors. See, e.g., Sachs (1964) or Pirani (1965) or Penrose (1968a) for reviews and references.

Of course, the spinor formalism, like any formalism, has its limitations. For example, many of the elementary problems of gravitation theory, and a large fraction of the most difficult ones, would be more difficult in the language of spinors than in the language of tensors! But for certain classes of problems, especially those where null vectors play a central role, spinors are a most valuable tool.

Cartan gave spinors to the world's physics and mathematics. His text (American edition, 1966) is an important reference to the subject.

2

CHAPTER 42

REGGE CALCULUS

§42.1. WHY THE REGGE CALCULUS?

This chapter is entirely Track 2. As preparation for it, Chapter 21 (variational principle and initial-value formalism) is needed. It is not needed as preparation for any later chapter, though it will be helpful in Chapter 43 (dynamics of geometry).

Gravitation theory is entering an era when situations of greater and greater complexity must be analyzed. Before about 1965 the problems of central interest could mostly be handled by idealizations of special symmetry or special simplicity or both. The Schwarzschild geometry and its generalizations, the Friedmann cosmology and its generalizations, the joining together of the Schwarzschild geometry and the Friedmann geometry to describe the collapse of a bounded collection of matter, the vibrations of relativistic stars, weak gravitational waves propagating in an otherwise flat space: all these problems and others were solved by elementary means.

The need for Regge calculus as a computational tool

But today one is pressed to understand situations devoid of symmetry and not amenable to perturbation theory: How do two black holes alter in shape, and how much gravitational radiation do they emit when they collide and coalesce? What are the structures and properties of the singularities at the endpoint of gravitational collapse, predicted by the theorems of Penrose, Hawking, and Geroch? Can a Universe that begins completely chaotic smooth itself out quickly by processes such as inhomogeneous mixmaster oscillations?

To solve such problems, one needs new kinds of mathematical tools—and in response to this need, new tools are being developed. The "global methods" of Chapter 34 provide one set of tools. The Regge Calculus provides another [Regge (1961); see also pp. 467–500 of Wheeler (1964a)].

§42.2. REGGE CALCULUS IN BRIEF

Approximation of smooth geometries by skeleton structures

Consider the geodesic dome that covers a great auditorium, made of a multitude of flat triangles joined edge to edge and vertex to vertex. Similarly envisage spacetime, in the Regge calculus, as made of flat-space "simplexes" (four-dimensional item in this progression: two dimensions, triangle; three dimensions, tetrahedron; four dimensions, simplex) joined face to face, edge to edge, and vertex to vertex. To specify the lengths of the edges is to give the engineer all he needs in order to know the shape of the roof, and the scientist all he needs in order to know the geometry of the spacetime under consideration. A smooth auditorium roof can be approximated arbitrarily closely by a geodesic dome constructed of sufficiently small triangles. A smooth spacetime manifold can be approximated arbitrarily closely by a locked-together assembly of sufficiently small simplexes. Thus the Regge calculus, reaching beyond ordinary algebraic expressions for the metric, provides a way to analyze physical situations deprived, as so many situations are, of spherical symmetry, and systems even altogether lacking in symmetry.

Role of Einstein field equation in fixing the skeleton structure

If the designer can give the roof any shape he pleases, he has more freedom than the analyst who is charting out the geometry of spacetime. Given the geometry of spacetime up to some spacelike slice that, for want of a better name, one may call "now," one has no freedom at all in the geometry from that instant on. Einstein's geometrodynamic law is fully deterministic. Translated into the language of the Regge calculus, it provides a means to calculate the edge lengths of new simplexes from the dimensions of the simplexes that have gone before. Though the geometry is deterministically specified, how it will be approximated is not. The original spacelike hypersurface ("now") is approximated as a collection of tetrahedrons joined together face to face; but how many tetrahedrons there will be and where their vertices will be placed is the option of the analyst. He can endow the skeleton more densely with bones in a region of high curvature than in a region of low curvature to get the most "accuracy profit" from a specified number of points. Some of this freedom of choice for the lengths of the bones remains as one applies the geometrodynamic law in the form given by Regge (1961) to calculate the future from the past. This freedom would be disastrous to any computer program that one tried to write, unless the programmer removed all indefiniteness by adding supplementary conditions of his own choice, either tailored to give good "accuracy profit," or otherwise fixed.

Having determined the lengths of all the bones in the portion of skeletonized spacetime of interest, one can examine any chosen local cluster of bones in and by themselves. In this way one can find out all there is to be learned about the geometry in that region. Of course, the accuracy of one's findings will depend on the fineness with which the skeletonization has been carried out. But in principle that is no limit to the fineness, or therefore to the accuracy, so long as one is working in the context of classical physics. Thus one ends up with a catalog of all the bones, showing the lengths of each. Then one can examine the geometry of whatever spacelike surface one pleases, and look into many other questions besides. For this purpose one has only to pick out the relevant bones and see how they fit together.

§42.3. SIMPLEXES AND DEFICIT ANGLES

Figure 42.1 recalls how a smoothly curved surface can be approximated by flat triangles. All the curvature is concentrated at the vertices. No curvature resides at

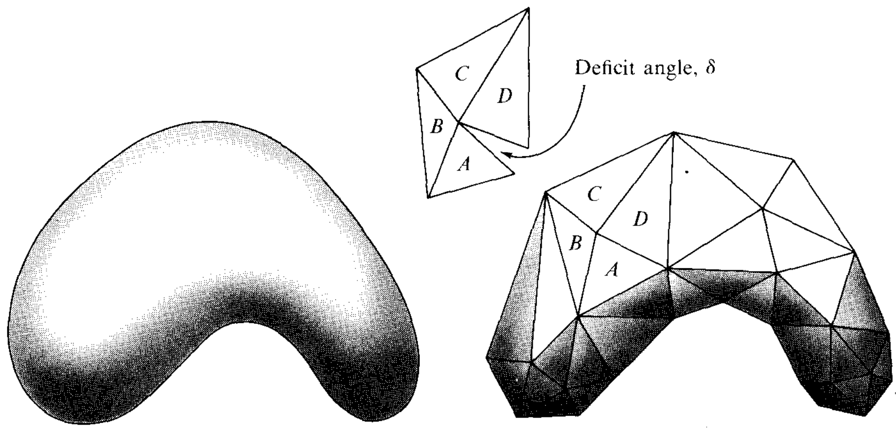


Figure 42.1.
A 2-geometry with continuously varying curvature can be approximated arbitrarily closely by a polyhedron built of triangles, provided only that the number of triangles is made sufficiently great and the size of each sufficiently small. The geometry in each triangle is Euclidean. The curvature of the surface shows up in the amount of deficit angle at each vertex (portion $ABCD$ of polyhedron laid out above on a flat surface).

Deficit angle as a skeletonized measure of curvature:

(1) in two dimensions

the edge between one triangle and the next, despite one's first impression. A vector carried by parallel transport from A through B and C to D, and then carried back by another route through C and B to A returns to its starting point unchanged in direction, as one sees most easily by laying out this complex of triangles on a flat surface. Only if the route is allowed to encircle the vertex common to A, B, C, and D does the vector experience a net rotation. The magnitude of the rotation is equal to the indicated deficit angle, δ, at the vertex. The sum of the deficit angles over all the vertices has the same value, 4π, as does the half-integral of the continuously distributed scalar curvature (${}^{(2)}R = 2/a^2$ for a sphere of radius a) taken over the entirety of the original smooth figure,

$$\sum_{\substack{\text{skeleton}\\ \text{geometry}}} \delta_i = \frac{1}{2} \int_{\substack{\text{actual smooth}\\ \text{geometry}}} {}^{(2)}R \, d(\text{surface}) = 4\pi. \tag{42.1}$$

(2) in n (or four) dimensions

Generalizing from the example of a 2-geometry, Regge calculus approximates a smoothly curved n-dimensional Riemannian manifold as a collection of n-dimensional blocks, each free of any curvature at all, joined by $(n-2)$-dimensional regions in which all the curvature is concentrated (Box 42.1). For the four-dimensional spacetime of general relativity, the "hinge" at which the curvature is concentrated has the shape of a triangle, as indicated schematically in the bottom row of Figure 42.2. In the example illustrated there, ten tetrahedrons have that triangle in common. Between one of these tetrahedrons and the next fits a four-dimensional simplex. Every feature of this simplex is determined by the lengths of its ten edges. One of the features is the angle α between one of the indicated tetrahedrons or "faces" of the simplex and the next. Thus α represents the angle subtended by this simplex

Box 42.1 THE HINGES WHERE THE CURVATURE IS CONCENTRATED IN THE "ANGLE OF RATTLE" BETWEEN BUILDING BLOCKS IN A SKELETON MANIFOLD

Dimensionality of manifold	*2*	*3*	*4*
Elementary flat-space building block:	triangle	tetrahedron	simplex
Edge lengths to define it:	3	4	5
Hinge where cycle of such blocks meet with a deficit angle or "angle of rattle" δ:	vertex	edge	triangle
Dimensionality of hinge:	0	1	2
"Content" of such a hinge:	1	length l	area A
Contribution from all hinges within a given small region to curvature of manifold:	$\sum_{\substack{\text{that}\\ \text{region}}} \delta_i$	$\sum_{\substack{\text{that}\\ \text{region}}} l_i\,\delta_i$	$\sum_{\substack{\text{that}\\ \text{region}}} A_i\,\delta_i$
Continuum limit of this quantity expressed as an integral over the same small region:	$\frac{1}{2}\int {}^{(2)}R({}^{(2)}g)^{1/2}\,d^2x$	$\frac{1}{2}\int {}^{(3)}R({}^{(3)}g)^{1/2}\,d^3x$	$\frac{1}{2}\int {}^{(4)}R(-{}^{(4)}g)^{1/2}\,d^4x$

at the hinge. Summing the angles α for all the simplexes that meet on the given hinge $\mathscr{P}\mathscr{Q}\mathscr{R}$, and subtracting from 2π, one gets the deficit angle associated with that hinge. And by then summing the deficit angles in a given small n-volume with appropriate weighting (Box 42.1), one obtains a number equal to the volume integral of the scalar curvature of the original smooth n-geometry. See Box 42.2.

§42.4. SKELETON FORM OF FIELD EQUATIONS

Einstein-Hilbert variational principle reduced to skeleton form

Rather than translate Einstein's field equations directly into the language of the skeleton calculus, Regge turns to a standard variational principle from which Einstein's law lets itself be derived. It says (see §§21.2 and 43.3) adjust the 4-geometry throughout an extended region of spacetime, subject to certain specified conditions on the boundary, so that the dimensionless integral (action in units of $\hbar$!),

$$I = (c^3/16\pi\hbar G)\int R(-g)^{1/2}\,d^4x, \tag{42.2}$$

is an extremum. This statement applies when space is free of matter and electromag-

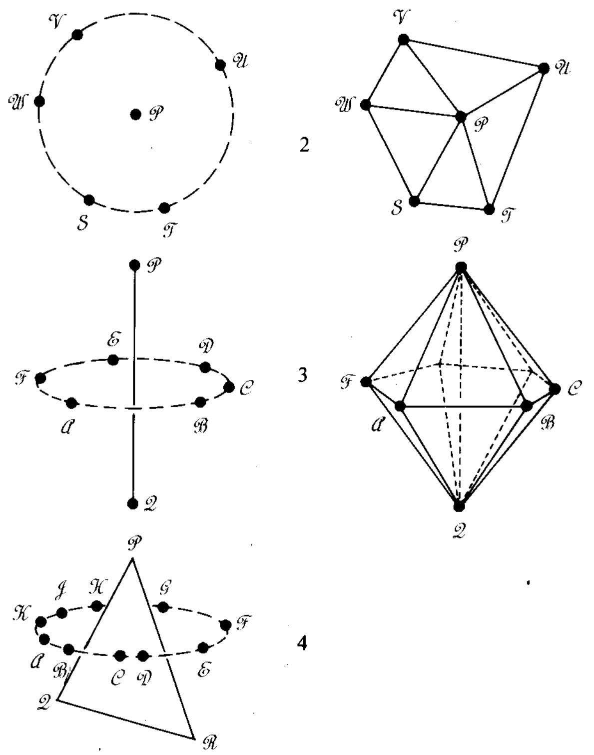


Figure 42.2.
Cycle of building blocks associated with a single hinge. Top row, two dimensions: left, schematic association of vertices $\mathcal{S}$, $\mathcal{T}$, $\mathcal{U}$, $\mathcal{V}$, $\mathcal{W}$ with "hinge" at the vertex $\mathcal{P}$; right, same, but with elementary triangles indicated in full. Middle row, three dimensions: left, schematic; right, perspective representation of the six tetrahedrons that meet on the "hinge" $\mathcal{PQ}$. Bottom row, four dimensions; shown only schematically. The five vertices $\mathcal{PQRCD}$ belong to one simplex, a four-dimensional region throughout the interior of which space is flat. The five vertices $\mathcal{PQRDE}$ belong to the next simplex; and so on around the cycle of simplexes. The two simplexes just named interface at the tetrahedron $\mathcal{PQRD}$, inside which the geometry is also flat. Between that tetrahedron and the next, $\mathcal{PQRE}$, there is a certain hyperdihedral angle α subtended at the "hinge" $\mathcal{PQR}$. The value of this angle is completely fixed by the ten edge lengths of the intervening simplex $\mathcal{PQRDE}$. This dihedral angle, plus the corresponding dihedral angles subtended at the hinge $\mathcal{PQR}$ by the other simplexes of the cycle, do not in general add up to 2π. The deficit, the "angle of rattle" or deficit angle δ, gives the amount of curvature concentrated at the hinge $\mathcal{PQR}$. There is no actual rattle or looseness of fit, unless one tries to imbed the cycle into an over-all flat four-dimensional space (analog of "stamping on" the collection of triangles, and seeing them open out by the amount of the deficit angle, as indicated in inset in Figure 42.1).

netic fields, a simplification that will be made in the subsequent discussion to keep it from becoming too extended. When in addition all lengths are expressed in units of the Planck length

$$L^* = (\hbar G/c^3)^{1/2} = 1.6 \times 10^{-33}\,\text{cm}, \tag{42.3}$$

and the curvature integral is approximated by its expression in terms of deficit angles, Regge shows that the statement $\delta I = 0$ (condition for an extremum!) becomes

$$(1/8\pi)\,\delta \sum_{\substack{\text{hinges}\\ h=1}}^{H} A_h\,\delta_h = 0. \tag{42.4}$$

Box 42.2 FLOW DIAGRAMS FOR REGGE CALCULUS

A skeleton 4-geometry is completely determined by all its edge lengths. From the edge lengths one gets the integrated curvature by pursuing, for each hinge in the 4-geometry, the following flow diagram:

hinge

↓

cycle of blocks swinging on this hinge

↓

one of these blocks

↓

the two tetrahedral "faces" that set this block off from the blocks before and after it in the cycle of blocks

↓

angle α between these two faces fixed by the block's $n(n-1)/2$ edge lengths

↓

deficit angle at the given hinge is $\delta = 2\pi - \sum_{\substack{\text{blocks swinging}\\ \text{on that hinge}}} \alpha$

↓

contribution to integrated curvature (Box 42.1) is δ times area of hinge

One finds it natural to apply this analysis in either of two ways. First, one can probe a given 4-geometry (given set of edge lengths!) in the sense

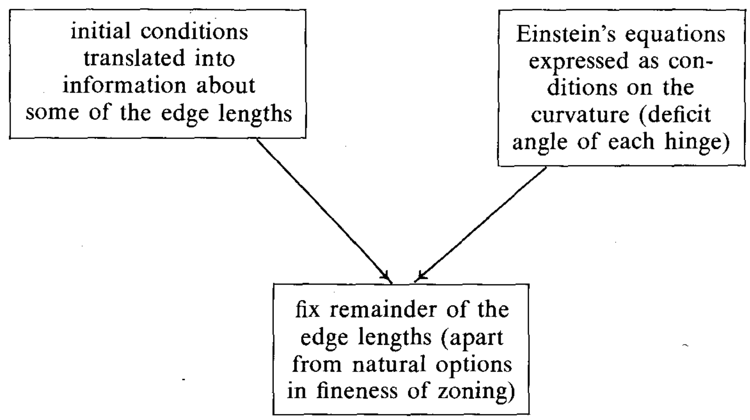
edge lengths

↓

curvature

Box 42.2 (*continued*)

Second—and this is the rationale of Regge calculus—one can use the skeleton calculus to deduce a previously unknown 4-geometry from Einstein's geometrodynamic law, proceeding in the direction



In the changes contemplated in this variational principle, certain edge lengths are thought of as being fixed. They have to do with the conditions specified at the boundaries of the region of spacetime under study. It is not necessary here to enter into the precise formulation of these boundary conditions, fortunately, since some questions of principle still remain to be clarified about the precise formulation of boundary conditions in general relativity (see §21.12). Rather, what is important is the effect of changes in the lengths of the edges of the blocks in the interior of the region being analyzed, as they augment or decrease the deficit angles at the various hinges. In his basic paper on the subject, Regge (1961) notes that the typical deficit angle δ_h depends in a complicated trigonometric way on the values of numerous edge lengths ℓ_p. However, he proves (Appendix of his paper) that "quite remarkably, we can carry out the variation as if the δ_h were constants," thus reducing the variational principle to the form

$$(1/8\pi) \sum_{\substack{\text{hinges} \\ h=1}}^{H} \delta_h \, \delta A_h = 0. \tag{42.5}$$

Here the change in area of the h-th triangle-shaped hinge, according to elementary trigonometry, is

$$\delta A_h = \frac{1}{2} \sum_p \ell_p \, \delta \ell_p \cot \theta_{ph}. \tag{42.6}$$

In this equation θ_{ph} is the angle opposite to the p-th edge in the triangle. Consequently, Einstein's equations in empty space reduce in skeleton geometry to the form

$$\sum_{\substack{\text{hinges that}\\ \text{have the}\\ \text{given edge}\\ p \text{ in common}}} \delta_h \operatorname{cotan} \theta_{ph} = 0, \qquad (p = 1, 2, \ldots), \tag{42.7}$$

Einstein field equation reduced to skeleton form

one equation for each edge length in the interior of the region of spacetime being analyzed.

§42.5. THE CHOICE OF LATTICE STRUCTURE

Two questions arise in the actual application of Regge calculus, and it is not clear that either has yet received the resolution which is most convenient for practical applications of this skeleton analysis: What kind of lattice to use? How best to capitalize on the freedom that exists in the choice of edge lengths? The first question is discussed in this section, the second in the next section.

The choice of lattice structure: (1) avoiding floppiness

It might seem most natural to use a lattice made of small, nearly rectangular blocks, the departure of each from rectangularity being conditioned by the amount and directionality of the local curvature. However, such building blocks are "floppy." One could give them rigidity by specifying certain angles as well as the edge lengths. But then one would lose the cleanness of Regge's prescription: give edge lengths, and give only edge lengths, and give each edge length freely and independently, in order to define a geometry. In addition one would have to rederive the Regge equations, including new equations for the determination of the new angles. Therefore one discards the quasirectangle in favor of the simplex with its $5 \cdot 4/2 = 10$ edge lengths. This decided, one also concludes that even in flat spacetime the simplexes cannot all have identical edge lengths. Two-dimensional flat space can be filled with identical equilateral triangles, but already at three dimensions it ceases to be possible to fill out the manifold with identical equilateral tetrahedrons. One knows that a given carbon atom in diamond is joined to its nearest neighbors with tetrahedral bonds, but a little reflection shows that the cell assignable to the given atom is far from having the shape of an equilateral tetrahedron.

(2) necessity for unequal edge lengths

Synthesis would appear to be a natural way to put together the building blocks: first make one-dimensional structures; assemble these into two-dimensional structures; these, into three-dimensional ones; and these, into the final four-dimensional structure. The one-dimensional structure is made of points, 1, 2, 3, ..., alternating with line segments, 12, 23, 34, ... To start building a two-dimensional structure, pick up a second one-dimensional structure. It might seem natural to label its points $1'$, $2'$, $3'$, ..., etc. However, that labeling would imply a cross-connection between 1 and $1'$, between 2 and $2'$, etc., after the fashion of a ladder. Then the elementary cells would be quasirectangles. They would have the "floppiness" that is to be excluded. Therefore relabel the points of the second one-dimensional structure as $1\frac{1}{2}'$, $2\frac{1}{2}'$, $3\frac{1}{2}'$, etc. The implication is that one cross-connects $2\frac{1}{2}'$ with points 2 and 3 of the original one-dimensional structure, etc. One ends up with something like the

(3) construction of two-dimensional structures

girder structure of a bridge, fully rigid in the context of two dimensions, as desired. The same construction, extended, fills out the plane with triangles. One now has a simple, standard two-dimensional structure. One might mistakenly conclude that one is ready to go ahead to build up a three-dimensional structure: the mistake lies in the tacit assumption that the flat-space topology is necessarily correct.

Let it be the problem, for example, to determine the development in time of a 3-geometry that has the topology of a 3-sphere. This 3-sphere is perhaps strongly deformed from ideality by long-wavelength gravitational waves. A right arrangement of the points is the immediate desideratum. Therefore put aside for the present any consideration of the deformation of the geometry by the waves (alteration of edge lengths from ideality). Ask how to divide a perfect 3-sphere into two-dimensional sheets. Here each sheet is understood to be separated from the next by a certain distance. At this point two alternative approaches suggest themselves that one can call for brevity "blocks" and "spheres."

(4) 3-D structures built from 2-D structures by "method of blocks"

(1) *Blocks.* Note that a 3-sphere lets itself be decomposed into 5 identical, tetrahedron-like solid blocks (5 vertices; 5 ways to leave out any one of these vertices!) Fix on one of these "tetrahedrons." Select one vertex as summit and the face through the other three vertices as base. Give that base the two-dimensional lattice structure already described. Introduce a multitude of additional sheets piled above it as evenly spaced layers reaching to the summit. Each layer has fewer points than the layer before. The decomposition of the 3-geometry inside one "tetrahedron" is thereby accomplished. However, an unresolved question remains; not merely how to join on this layered structure in a regular way to the corresponding structure in the adjacent "tetrahedrons"; but even whether such a regular joinup is at all possible. The same question can be asked about the other two ways to break up the 3-sphere into identical "tetrahedrons" [Coxeter (1948), esp. pp. 292–293: 16 tetrahedrons defined by a total of 8 vertices or 600 tetrahedrons defined by a total of 120 vertices]. One can eliminate the question of joinup of structure in a simple way, but at the price of putting a ceiling on the accuracy attainable: take the stated number of vertices (5 or 8 or 120) as the total number of points that will be employed in the skeletonization of the 3-geometry (no further subdivision required or admitted). Considering the boundedness of the memory capacity of any computer, it is hardly ridiculous to contemplate a limitation to 120 tracer points in exploratory calculations!

(5) 3-D structures from 2-D by "method of spheres"

(2) *Spheres.* An alternative approach to the "atomization" of the 3-sphere begins by introducing on the 3-sphere a North Pole and a South Pole and the hyperspherical angle χ ($\chi = 0$ at the first pole, $\chi = \pi$ at the second, $\chi = \pi/2$ at the equator; see Box 27.2). Let each two-dimensional layer lie on a surface of constant χ (χ equal to some integer times some interval $\Delta\chi$). The structure of this 2-sphere is already to be regarded as skeletonized into elementary triangles ("fully complete Buckminster Fuller geodesic dome"). Therefore the number of "faces" or triangles F, the number of edge lengths E, and the number of vertices V must be connected by the relation of Euler:

$$F - E + V = \begin{pmatrix}\text{a topology-dependent} \\ \text{number or "Euler character"}\end{pmatrix} = \begin{cases} 2 \text{ for 2-sphere,} \\ 0 \text{ for 2-torus.}\end{cases} \tag{42.8}$$

It follows from this relation that it is impossible for each vertex to sit at the center of a hexagon (each vertex the point of convergence of 6 triangles). This being the case, one is not astonished that a close inspection of the pattern of a geodesic dome shows several vertices where only 5 triangles meet. It is enough to have 12 such 5-triangle vertices among what are otherwise all 6-triangle vertices in order to meet the requirements of the Euler relation:

$$\begin{aligned}
n &\quad \text{5-triangle vertices}\\
V - n &\quad \text{6-triangle vertices}\\
F &= (V-n)(6/3) + n(5/3) \text{ triangles}\\
E &= (V-n)(6/2) + n(5/2) \text{ edges}\\
V &= (V-n)(6/6) + n \text{ vertices}\\
2 = F - E + V &= n/6 \qquad \text{Euler characteristic}\\
n &= 12
\end{aligned} \tag{42.9}$$

Among all figures with triangular faces, the icosahedron is the one with the smallest number of faces that meets this condition (5-triangle vertices exclusively!)

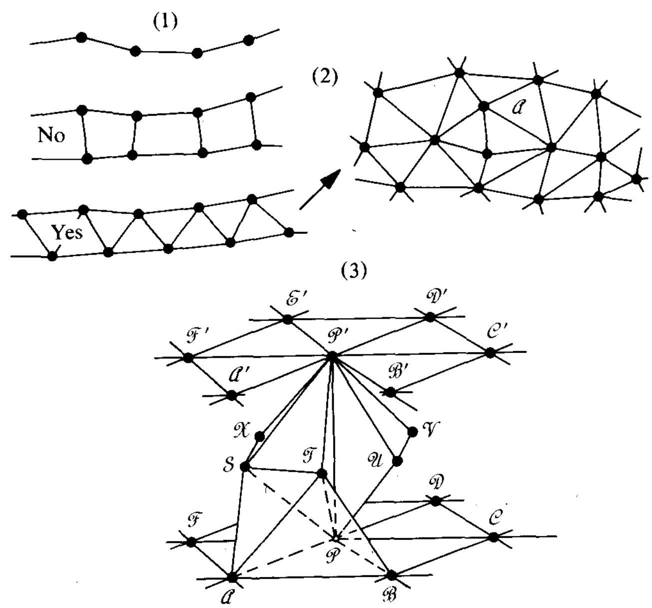
If each 2-surface has the pattern of vertices of a geodesic dome, how is one dome to be joined to the next to make a rigid skeleton 3-geometry? Were the domes imbedded in a flat 3-geometry, rigidity would be no issue. Each dome would already be rigid in and by itself. However, the 3-geometry is not given to be flat. Only by a completely deterministic skeletonization of the space between the two 2-spheres will they be given rigidity in the context of curved space geometry. (1) Not by running a single connector from each vertex in one surface to the corresponding vertex in the next ("floppy structure"!) (2) Not by displacing one surface so each of its vertices comes above, or nearly above, the center of a triangle in the surface "below." First, the numbers of vertices and triangles ordinarily will not agree. Second, even when they do, it will not give the structure the necessary rigidity to connect the vertex of the surface above to the three vertices of the triangle below. The space between will contain some tetrahedrons, but it will not be throughout decomposed into tetrahedrons. (3) A natural and workable approach to the skeletonization of the 3-geometry is to run a connector from each vertex in the one surface to the corresponding vertex in the next, but to flesh out this connection with additional structure that will give rigidity to the 3-geometry: intervening vertices and connectors as illustrated in Box 42.3.

(6) 4-D structures built from 3-D structures

In working up from the skeletonization of a 3-geometry to the skeletonization of a 4-geometry, it is natural to proceed similarly. (1) Use identical patterns of points in the two 3-geometries. (2) Tie corresponding points together by single connectors. (3) Halfway, or approximately half way between the two 3-geometries insert a whole additional pattern of vertices. Each of these supplementary vertices is "dual" to and lies nearly "below" the center of a tetrahedron in the 3-geometry immediately above. (4) Connect each supplementary vertex to the vertices of the tetrahedron immediately above, to the vertices of the tetrahedron immediately below, and to those other supplementary vertices that are its immediate neighbors. (5) In this way get the edge lengths needed to divide the 4-geometry into simplexes, each of rigidly defined dimensions.

2

Box 42.3 SYNTHESIS OF HIGHER-DIMENSIONAL SKELETON GEOMETRIES OUT OF LOWER-DIMENSIONAL SKELETON GEOMETRIES



(1) One-dimensional structure as alternation of points and line segments. (2) Two-dimensional structure (a) "floppy" (unacceptable) and (b) rigidified (angles of triangles fully determined by edge lengths). When this structure is extended, as at right, the "normal" vertex has six triangles hinging on it. However, at least twelve 5-triangle vertices of the type indicated at $\mathcal{A}$ are to be interpolated if the 2-geometry is to be able to close up into a 2-sphere. (3) Skeleton 3-geometry obtained by filling in between the skeleton 2-geometry... $\mathcal{AB}\ldots\mathcal{FPC}\ldots\mathcal{ED}\ldots$ and the similar structure $\ldots\mathcal{A}'\mathcal{B}'\ldots\mathcal{F}'\mathcal{P}'\mathcal{C}'\ldots\mathcal{E}'\mathcal{D}'\ldots$ as follows. (a) Insert direct connectors such as $\mathcal{PP}'$ between corresponding points in the two 2-geometries. (b) Insert an intermediate layer of "supplementary vertices" such as $\mathcal{STUVWX}$.... Each of these supplementary vertices lies roughly halfway between the center of the triangle "above" it and the center of the corresponding triangle "below" it. (c) Connect each such "supplementary vertex" with its immediate neighbors above, below, and in the same plane. (d) Give all edge lengths. (e) Then the skeleton 3-geometry between the two 2-geometries is rigidly specified. It is made up of five types of tetrahedrons, as follows. (1) "Right-through blocks," such as $\mathcal{PP}'\mathcal{ST}$ (six of these hinge on $\mathcal{PP}'$ when $\mathcal{P}$ is a normal vertex; five, when it is a 5-fold vertex, such as indicated by $\mathcal{A}$ at the upper right). (2) "Lower-facing blocks," such as $\mathcal{ABPT}$. (3) "Lower-packing blocks," such as $\mathcal{APST}$. (4, 5) Corresponding "upper-facing blocks" and "upper-packing blocks" (not shown). The number of blocks of each kind is appropriately listed here for the two extreme cases of a 2-geometry that consists (a) of a normal hexagonal lattice extending indefinitely in a plane and (b) of a lattice consisting of the minimum number of 5-fold vertices ("type $\mathcal{A}$ vertices") that will permit close-up into a 2-sphere.

2-geometry of upper (or lower) face	*Hexagonal pattern of triangles*	*Icosahedron made of triangles*
Its topology	Infinite 2-plane	2-sphere
Vertices on upper face	V	12
Nature of these vertices	6-fold	5-fold
Edge lengths on upper face	$3V$	$\frac{5}{2}V = 30$
Triangles on upper face	$2V$	20
Number of "supplementary vertices"	$2V$	20
Outer facing blocks	$2V$	20
Outer packing blocks	$3V$	30
Right through blocks	$6V$	60
Inner packing blocks	$3V$	30
Inner facing blocks	$2V$	20

§42.6. THE CHOICE OF EDGE LENGTHS

The choice of edge lengths:

So much for the lattice structure of the 4-geometry; now for the other issue, the freedom that exists in the choice of edge lengths. Why not make the simplest choice and let all edges be light rays? Because the 4-geometry would not then be fully determined. The geometry $g_{\alpha\beta}(x^\mu)$ differs from the geometry $\lambda(x^\mu)\ g_{\alpha\beta}(x^\mu)$, even though the same points that are connected by light rays in the one geometry are also connected by light rays in the other geometry.

(1) choose some timelike, others spacelike

If none of the edges is null, it is nevertheless natural to take some of the edge lengths to be spacelike and some to be timelike. In consequence the area A of the triangle in some cases will be real, in other cases imaginary. In 3-space the parallelogram (double triangle) spanned by two vectors $\boldsymbol{B}$ and $\boldsymbol{C}$ is described by a vector

$$2\boldsymbol{A} = \boldsymbol{B} \times \boldsymbol{C}$$

perpendicular to the two vectors. One obtains the magnitude of $\boldsymbol{A}$ from the formula

$$4A^2 = \boldsymbol{B}^2\boldsymbol{C}^2 - (\boldsymbol{B} \cdot \boldsymbol{C})^2.$$

In 4-space, let $\boldsymbol{B}$ and $\boldsymbol{C}$ be two edges of the triangle. Then, as in three dimensions, $2\boldsymbol{A}$ is *dual* to the bivector built from $\boldsymbol{B}$ and $\boldsymbol{C}$. In other words, if $\boldsymbol{B}$ goes in the t direction and $\boldsymbol{C}$ in the z direction, then $\boldsymbol{A}$ is a bivector lying in the (x, y) plane. Consequently its magnitude A is to be thought of as a *real* quantity. Therefore the appropriate formula for the area A is (Tullio Regge)

$$4A^2 = (\boldsymbol{B} \cdot \boldsymbol{C})^2 - \boldsymbol{B}^2\boldsymbol{C}^2. \tag{42.10}$$

The quantity A is real when the deficit angle δ is real. Thus the geometrically important product $A\delta$ is also real.

When the hinge lies in the (x, y) plane, on the other hand, the quantity A is purely imaginary. In that instance a test vector taken around the cycle of simplexes that swing on this hinge has undergone change only in its z and t components; that is, it has experienced a Lorentz boost; that is, the deficit angle δ is also purely imaginary. So again the product $A\delta$ is a purely real quantity.

(2) choose timelike lengths comparable to spacelike lengths

Turn now from character of edge lengths to magnitude of edge lengths. It is desirable that the elementary building blocks sample the curvatures of space in different directions on a roughly equal basis. In other words, it is desirable not to have long needle-shaped building blocks nor pancake-shaped tetrahedrons and simplexes. This natural requirement means that the step forward in time should be comparable to the steps "sidewise" in space. The very fact that one should have to state such a requirement brings out one circumstance that should have been obvious before: the "hinge equations"

$$\sum_{\substack{\text{hinges } h \text{ that}\\ \text{have edge } p\\ \text{in common}}} \delta_h \operatorname{cotan} \theta_{ph} = 0 \qquad (p = 1, 2, \ldots), \tag{42.7}$$

(3) why some lengths must be chosen arbitrarily

though they are as numerous as the edges, cannot be regarded as adequate to determine all edge lengths. There are necessarily relations between these equations that keep them from being independent. The equations cannot determine all the details of the necessarily largely arbitrary skeletonization process. They cannot do so any more than the field equations of general relativity can determine the coordinate system. With a given pattern of vertices (four-dimensional generalization of drawings in Box 42.3), one still has (a) the option how close together one will take successive layers of the structure and (b) how one will distribute a given number of points in space on a given layer to achieve the maximum payoff in accuracy (greater density of points in regions of greater curvature). To prepare a practical computer program founded on Regge calculus, one has to supply the machine not only with the hinge equations and initial conditions, but also with definite algorithms to remove all the arbitrariness that resides in options (a) and (b).

Deficit angles in terms of edge lengths

Formulas from solid geometry and four-dimensional geometry, out of which to determine the necessary hyperdihedral angles α and the deficit angles δ in terms of edge lengths and nothing but edge lengths, are summarized by Wheeler (1964a, pp. 469, 470, and 490) and by C. Y. Wong (1971). Regge (1961) also gives a formula for the Riemann curvature tensor itself in terms of deficit angles and number of edges running in a given direction [see also Wheeler (1964a, p. 471)].

§42.7. PAST APPLICATIONS OF REGGE CALCULUS

Past applications of Regge calculus

Wong (1971) has applied Regge calculus to a problem where no time development shows itself, where the geometry can therefore be treated as static, and where in addition it is spherically symmetric. He determined the Schwarzschild and Reissner-Nordstrøm geometries by the method of skeletonization. Consider successive spheres surrounding the center of attraction. Wong approximates each as an icosahedron. The condition

$$^{(3)}R = 16\pi \begin{pmatrix} \text{energy density} \\ \text{on the 3-space} \end{pmatrix}$$

(§21.5) gives a recursion relation that determines the dimension of each icosahedron in terms of the two preceding icosahedra. Errors in the skeleton representation of the exact geometry range from roughly 10 percent to less than 1 percent, depending on the method of analysis, the quantity under analysis, and the fineness of the subdivision.

Partial skeletonization

Skeletonization of geometry is to be distinguished from mere rewriting of partial differential equations as difference equations. One has by now three illustrations that one can capitalize on skeletonization without fragmenting spacetime all the way to the level of individual simplexes. The first illustration is the first part of Wong's work, where the time dimension never explicitly makes an appearance, so that the building blocks are three-dimensional only. The second is an alternative treatment, also given by Wong, that goes beyond the symmetry in t to take account of the symmetry in θ and ϕ. It divides space into spherical shells, in each of which the geometry is "pseudo-flat" in much the same sense that the geometry of a paper cone is flat. The third is the numerical solution for the gravitational collapse of a spherical star by May and White (1966), in which there is symmetry in θ and ϕ, but not in r or t. This zoning takes place exclusively in the r, t-plane. Each zone is a spherical shell. The difference as compared to Regge calculus (flat geometry within each building block) is the adjustable "conicity" given to each shell. The examples show that the decision about skeletonizing the geometry in a calculation is ordinarily not "whether" but "how much."

§42.8. THE FUTURE OF REGGE CALCULUS

Hopes for the future

In summary, Regge's skeleton calculus puts within the reach of computation problems that in practical terms are beyond the power of normal analytical methods. It affords any desired level of accuracy by sufficiently fine subdivision of the spacetime region under consideration. By way of its numbered building blocks, it also offers a practical way to display the results of such calculations. Finally, one can hope that Regge's truly geometric way of formulating general relativity will someday make the content of the Einstein field equations (Cartan's "moment of rotation"; see Chapter 15) stand out sharp and clear, and unveil the geometric significance of the so-called "geometrodynamic field momentum" (analysis of the boundary-value problem associated with the variational problem of general relativity in Regge calculus; see §21.12).

2

CHAPTER 43

SUPERSPACE: ARENA FOR THE DYNAMICS OF GEOMETRY

Traveler, there are no paths.
Paths are made by walking.
ANTONIO MACHADO (1940)

This chapter is entirely Track 2. Chapter 21 (initial-value formalism) is needed as preparation for it. In reading it, one will be helped by Chapter 42 (Regge calculus). It is not needed as preparation for any later chapter, but it will be helpful in Chapter 44 (vision of the future).

§43.1. SPACE, SUPERSPACE, AND SPACETIME DISTINGUISHED

Superspace is the arena for geometrodynamics

Superspace [Wheeler (1964a), pp. 459 ff] is the arena of geometrodynamics. The dynamics of Einstein's curved space geometry runs its course in superspace as the dynamics of a particle unfolds in spacetime. This chapter gives one simple version of superspace, and a little impression of alternative versions of superspace that also have their uses. It describes the classical dynamics of geometry in superspace in terms of the Hamilton-Jacobi principle of Boxes 25.3 and 25.4. No version of mechanics makes any shorter the leap from classical dynamics to quantum. Thus it provides a principle ("Einstein-Hamilton-Jacobi or EHJ equation") for the propagation of wave crests in superspace, and for finding where those wave crests give one the classical equivalent of constructive interference ("envelope formation"). In this way one finds the track of development of 3-geometry with time expressed as a sharp, thin "leaf of history" that slices through superspace. The quantum principle replaces this deterministic account with a fuzzed-out leaf of history of finite thickness. In consequence, quantum fluctuations take place in the geometry of space that dominate the scene at distances of the order of the Planck length, $L^* = (\hbar G/c^3)^{1/2} = 1.6 \times 10^{-33}$ cm, and less. The present survey simplifies by considering only the dynamics of curved empty space. When sources are present and are to be taken into account, supplementary terms are to be added, some of the literature on which is listed.

In all the difficult investigations that led in the course of half a century to some understanding of the dynamics of geometry, both classical and quantum, the most

Box 43.1 GEOMETRODYNAMICS COMPARED WITH PARTICLE DYNAMICS

Concept	*Particle dynamics*	*Geometrodynamics*
Dynamic entity	Particle	Space
Descriptors of momentary configuration	x, t ("event")	${}^{(3)}\mathcal{G}$ ("3-geometry")
Classical history	$x = x(t)$	${}^{(4)}\mathcal{G}$ ("4-geometry")
History is a stockpile of configurations?	Yes. Every point on world line gives a momentary configuration of particle	Yes. Every spacelike slice through ${}^{(4)}\mathcal{G}$ gives a momentary configuration of space
Dynamic arena	Spacetime (totality of all points x, t)	Superspace (totality of all ${}^{(3)}\mathcal{G}$'s)

difficult point was also the simplest: The dynamic object is not spacetime. It is space. The geometric configuration of space changes with time. But it is space, three-dimensional space, that does the changing (see Box 43.1).

3-geometry is the dynamic object

Space will be treated here as "closed" or, in mathematical language, "compact," either because physics adds to Einstein's second-order differential equations the requirement of closure as a necessary and appropriate boundary condition [Einstein (1934, p. 52; 1950); Wheeler (1959; 1964c). Hönl (1962); see also §21.12] or because that requirement simplifies the mathematical analysis, or for both reasons together.

Finite-dimensional "truncated superspace"

One can approximate a smooth, closed 3-geometry by a skeleton 3-geometry built out of tetrahedrons, as indicated schematically in Figure 43.1 (see Chapter 42 on the Regge calculus). Specify the 98 edge-lengths in this example to fix all the features of the geometry; and fix these 98 edge-lengths by giving the location of a single point in a space of 98 dimensions. This 98-dimensional manifold, this "truncated superspace," goes over into superspace [Wheeler (1964a), pp. 453, 459, 463, 495] in the idealization in which the tracer points increase in density of coverage without limit. Accounts of superspace with more mathematical detail are given by DeWitt (1967a,b), Wheeler (1970), and Fischer (1970).

Let the representative point move from one location to a nearby location, either in truncated superspace or in full superspace. Then all edge-lengths alter, and the 3-geometry of Figure 43.1 moves as if alive. No better illustration can one easily supply of what it means to speak of the "dynamics of space."

The term "3-geometry" makes sense as well in quantum geometrodynamics as in classical theory. So does superspace. But spacetime does not. Give a 3-geometry, and give its time rate of change. That is enough, under typical circumstances (see Chapter 21) to fix the whole time-evolution of the geometry; enough in other words, to determine the entire four-dimensional spacetime geometry, provided one is

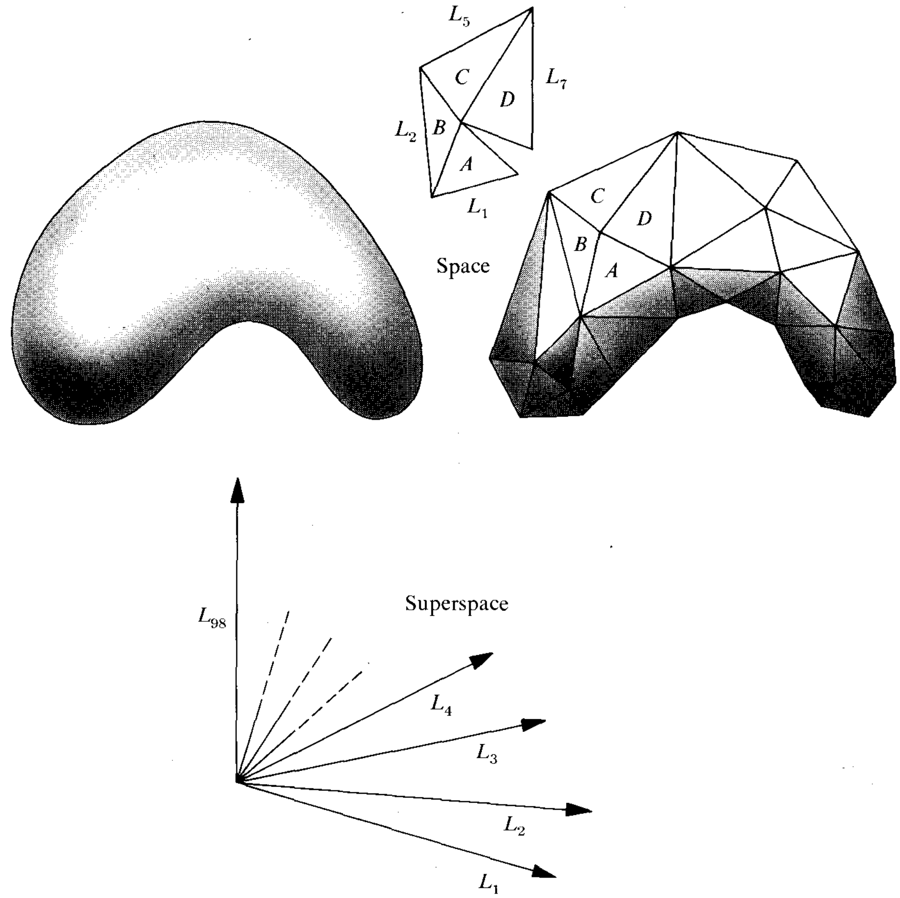


Figure 43.1.
Superspace in the simplicial approximation. Upper left, space (depicted as two-dimensional but actually three-dimensional). Upper right, simplical approximation to space. The approximation can be made arbitrarily good by going to the limit of an arbitrarily fine decomposition. The curvature at a typical location is measured by a deficit angle. This angle is completely determined by the edge lengths (L_1, L_2, . . . L_8 in the figure) of the simplexes that meet at that location. When there are 98 edge lengths altogether in the simplicial representation of the geometry, then this geometry is completely specified by a single point in a 98-dimensional space (lower diagram; "superspace").

The concept of spacetime is incompatible with the quantum principle

considering the problem in the context of classical physics. In the real world of quantum physics, however, one cannot give both a dynamic variable and its time-rate of change. The principle of complementarity forbids. Given the precise 3-geometry at one instant, one cannot also know at that instant the time-rate of change of the 3-geometry. In other words, given the geometrodynamic field coordinate, one cannot know the geometrodynamic field momentum. If one assigns the intrinsic 3-geometry, one cannot also specify the extrinsic curvature.

The uncertainty principle thus deprives one of any way whatsoever to predict, or even to give meaning to, "the deterministic classical history of space evolving

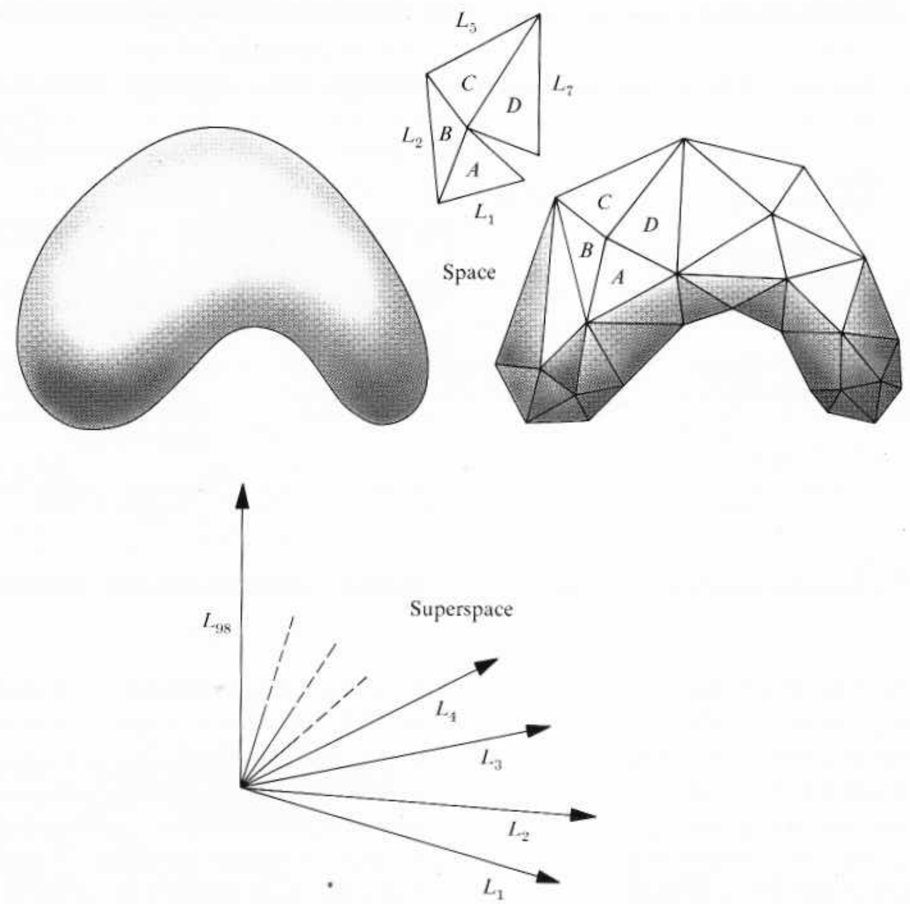


Figure 43.1.
Superspace in the simplicial approximation. Upper left, space (depicted as two-dimensional but actually three-dimensional). Upper right, simplical approximation to space. The approximation can be made arbitrarily good by going to the limit of an arbitrarily fine decomposition. The curvature at a typical location is measured by a deficit angle. This angle is completely determined by the edge lengths (L_1, L_2, . . . L_8 in the figure) of the simplexes that meet at that location. When there are 98 edge lengths altogether in the simplicial representation of the geometry, then this geometry is completely specified by a single point in a 98-dimensional space (lower diagram; "superspace").

The concept of spacetime is incompatible with the quantum principle

considering the problem in the context of classical physics. In the real world of quantum physics, however, one cannot give both a dynamic variable and its time-rate of change. The principle of complementarity forbids. Given the precise 3-geometry at one instant, one cannot also know at that instant the time-rate of change of the 3-geometry. In other words, given the geometrodynamic field coordinate, one cannot know the geometrodynamic field momentum. If one assigns the intrinsic 3-geometry, one cannot also specify the extrinsic curvature.

The uncertainty principle thus deprives one of any way whatsoever to predict, or even to give meaning to, "the deterministic classical history of space evolving in time." *No prediction of spacetime, therefore no meaning for spacetime,* is the verdict of the quantum principle. That object which is central to all of classical general relativity, the four-dimensional spacetime geometry, simply does not exist, except in a classical approximation.

These considerations reveal that the concepts of spacetime and time are not primary but secondary ideas in the structure of physical theory. These concepts are valid in the classical approximation. However, they have neither meaning nor application under circumstances where quantum geometrodynamic effects become important. Then one has to forego that view of nature in which every event, past, present, or future, occupies its preordained position in a grand catalog called "spacetime," with the Einstein interval from each event to its neighbor eternally established. There is no spacetime, there is no time, there is no before, there is no after. The question of what happens "next" is without meaning.

That spacetime is not the right way does not mean there is *no* right way to describe the dynamics of geometry consistent with the quantum principle. Superspace is the key to *one* right way to describe the dynamics (see Figure 43.2).

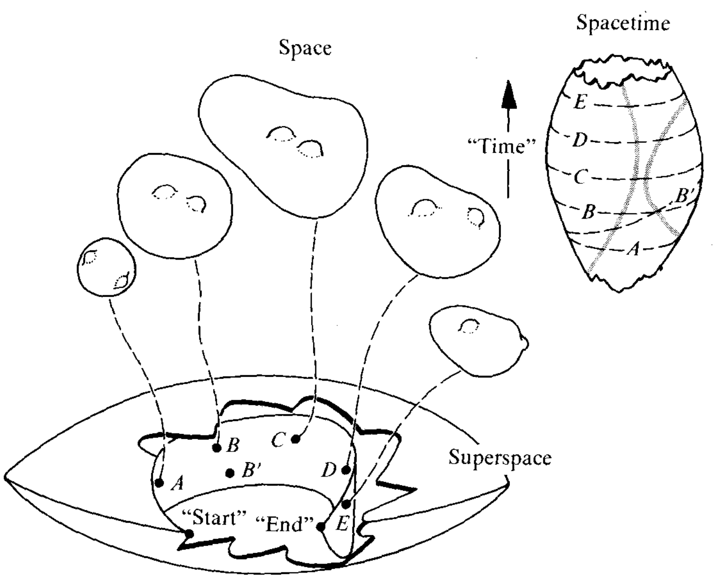


Figure 43.2.
Space, spacetime, and superspace. Upper left: Five sample configurations, A, B, C, D, E, attained by space in the course of its expansion and recontraction. Below: Superspace and these five sample configurations, each represented by a point in superspace. Upper right: Spacetime. A spacelike cut, like A, through spacetime gives a momentary configuration of space. There is no compulsion to limit attention to a one-parameter family of slices, A, B, C, D, E through spacetime. The phrase "many-fingered time" is a slogan telling one not to so limit one's slices, and B' is an example of this freedom in action. The 3-geometries B' and A, B, C, D, E, like all 3-geometries obtained by all spacelike slices whatsoever through the given classical spacetime, lie on a single bent leaf of history, indicated in the diagram, and cutting its thin slice through superspace. A different spacetime, in other words, a different solution of Einstein's field equation, means a different leaf of history (not indicated) slicing through superspace.

2

§43.2. THE DYNAMICS OF GEOMETRY DESCRIBED IN THE LANGUAGE OF THE SUPERSPACE OF THE $^{(3)}\mathcal{G}$'s

Spacetime is a classical leaf of history slicing through superspace

Given a spacetime, one can construct the corresponding leaf of history slicing through superspace. Conversely, given the leaf of history, one can reconstruct the spacetime.

Consider the child's toy commonly known as "Chinese boxes." One opens the outermost box only to reveal another box; when the second box is opened, there is another box, and so on, until eventually there are dozens of boxes scattered over the floor. Or conversely the boxes can be put back together, nested one inside the other, to reconstitute the original package. The packaging of $^{(3)}\mathcal{G}$'s into a $^{(4)}\mathcal{G}$ is much more sophisticated. Nature provides no monotonic ordering of the $^{(3)}\mathcal{G}$'s. Two of the dynamically allowed $^{(3)}\mathcal{G}$'s, taken at random, will often cross each other one or more times. When one shakes the $^{(4)}\mathcal{G}$ apart, one therefore gets enormously more $^{(3)}\mathcal{G}$'s "spread out over the floor" than might have been imagined. Conversely, when one puts back together all of the $^{(3)}\mathcal{G}$'s lying on the leaf of history, one gets a structure with a rigidity that might not otherwise have been foreseen. This rigidity arises from the infinitely rich interleaving and intercrossing of clear-cut, well-defined $^{(3)}\mathcal{G}$'s one with another.

In summary: (1) Classical geometrodynamics in principle constitutes a device, an algorithm, a rule for calculating and constructing a leaf of history that slices through superspace. (2) The $^{(3)}\mathcal{G}$'s that lie on this leaf of history are YES 3-geometries; the vastly more numerous $^{(3)}\mathcal{G}$'s that do not are NO 3-geometries. (3) The YES $^{(3)}\mathcal{G}$'s are the building blocks of the $^{(4)}\mathcal{G}$ that is classical spacetime. (4) The interweaving and interconnections of these building blocks give the $^{(4)}\mathcal{G}$ its existence, its dimensionality, and its structure. (5) In this structure every $^{(3)}\mathcal{G}$ has a rigidly fixed location of its own. (6) In this sense one can say that the "many-fingered time" of each 3-geometry is specified by the very interlocking construction itself. Baierlein, Sharp and Wheeler (1962) say a little more on this concept of "3-geometry as carrier of information about time."

How different from the textbook concept of spacetime! There the geometry of spacetime is conceived as constructed out of elementary objects, or points, known as "events." Here, by contrast, the primary concept is 3-geometry, *in abstracto,* and out of it is derived the idea of event. Thus, (1) the event lies at the intersection of such and such $^{(3)}\mathcal{G}$'s; and (2) it has a timelike relation to (earlier or later than, or synchronous with) some other $^{(3)}\mathcal{G}$, which in turn (3) derives from the intercrossings of all the other $^{(3)}\mathcal{G}$'s.

When one turns from classical theory to quantum theory, one gives up the concept of spacetime, except in the semiclassical approximation. Therefore, one gives up any immediate possibility whatsoever of defining the concept, normally regarded as so elemental, of an "event." The theory itself, however, here as always [Bohr and Rosenfeld (1933)], defines in and by itself, in its own natural way, the procedures-in-principle for measuring all those quantities that are in principle measurable.

Quantum theory upsets the sharp distinction between YES 3-geometries and NO

3-geometries. It assigns to each 3-geometry not a YES or a NO, but a probability amplitude,

Probability amplitude for a 3-geometry

$$\psi = \psi({}^{(3)}\mathcal{G}). \tag{43.1}$$

This probability amplitude is highest near the classically forecast leaf of history and falls off steeply outside a zone of finite thickness extending a little way on either side of the leaf.

Were one to take, instead of a physically relevant probability amplitude function, a *typical* solution of the relevant wave equation, one would have to expect to see not one trace of anything like classical geometrodynamics. The typical probability amplitude function is spread all over superspace. No surprise! Already in classical theory one has to reckon with a Hamilton-Jacobi function,

$$S = S({}^{(3)}\mathcal{G}), \tag{43.2}$$

spread out over superspace. Moreover, this "dynamic phase function" of classical geometrodynamics gives at once the phase of ψ, according to the formula

$$\psi({}^{(3)}\mathcal{G}) = \begin{pmatrix}\text{slowly varying} \\ \text{amplitude function}\end{pmatrix} e^{(i/\hbar)S({}^{(3)}\mathcal{G})}, \tag{43.3}$$

indication enough that ψ and S are both unlocalized.

Dynamics first clearly becomes recognizable when sufficiently many such spread-out probability amplitude functions are superposed to build up a localized wave packet, as in the elementary examples of Boxes 25.3 and 25.4; thus,

Wave packet recovers classical geometrodynamics

$$\psi = c_1\psi_1 + c_2\psi_2 + \cdots. \tag{43.4}$$

Constructive interference occurs where the phases of the several individual waves agree:

$$S_1({}^{(3)}\mathcal{G}) = S_2({}^{(3)}\mathcal{G}) = \cdots. \tag{43.5}$$

This is the condition that distinguishes YES 3-geometries from NO 3-geometries. It is the tool for constructing a leaf of history in superspace. It is the key to the dynamics of geometry. Moreover, it is an equation that says not one word about the quantum principle. It is not surprising that the equation of constructive interference in (43.5) makes the leap from classical theory to quantum theory the shortest.

§43.3. THE EINSTEIN-HAMILTON-JACOBI EQUATION

Should one write down a differential equation for the Hamilton-Jacobi function $S({}^{(3)}\mathcal{G})$, solve it, and then analyze the properties of the solution? The exact opposite is simpler: look at the properties of the solution, and from that inspection find out what equation the dynamic phase or action S must satisfy.

Hilberts' principle of least action reads

$$I_{\text{Hilbert}} = (1/16\pi)\int {}^{(4)}R(-g)^{1/2}\,d^4x = \text{extremum}. \tag{43.6}$$

After one separates off complete derivatives in the integrand, what is left [see equations (21.13) and (21.95)] becomes

$$\begin{aligned}(1/16\pi)I_{\text{ADM}} = I_{\text{true}} = (1/16\pi)\int\Big\{&\pi^{ij}\,\partial g_{ij}/\partial t + Ng^{1/2}R \\ &+ Ng^{-1/2}\Big[\frac{1}{2}(\operatorname{Tr}\boldsymbol{\pi})^2 - \operatorname{Tr}(\boldsymbol{\pi}^2)\Big] + 2N_i\pi^{ij}{}_{|j}\Big\}\,d^4x. \end{aligned}\tag{43.7}$$

In (43.7), but not in (43.6), g stands for the determinant of the three-dimensional metric tensor, g_{ij}, and R for the scalar curvature invariant of the 3-geometry; the suffix $^{(3)}$ is omitted for simplicity. The integral is extended from (1) a spacelike hypersurface on which a 3-geometry is given with metric $g_{ij}'(x, y, z)$ to (2) a spacelike hypersurface on which a 3-geometry is given with metric $g_{ij}''(x, y, z)$. Whatever is adjustable in the chunk of spacetime between is now to be considered as having been adjusted to extremize the integral. Therefore the value of the integral I_{ADM} becomes a functional of the metrics on the two hypersurfaces and nothing more.

Next, holding fixed the metric $g'_{ij}(x, y, z)$ on the earlier hypersurface, change slightly or even more than slightly the metric on the later hypersurface. Solve the new variation problem and get a new value of I_{ADM}. Proceeding further in this way, for each new g_{ij}'' one gets a new value of I_{ADM}. Call the functional I_{ADM} of the metric defined in this way "Hamilton's principal function," or the "action" or the "dynamic path length,*" $S(g_{ij}(x, y, z))$ of the "history of geometry" that connects the two given 3-geometries. The double prime suffix is dropped from g_{ij}'' here and hereafter to simplify the notation. One knows from other branches of mechanics that the quantity defined in this way, $S(g_{ij})$, when it exists, even though it is a special solution, nevertheless is *always* a solution of the Hamilton-Jacobi equation. Jacobi could look for more general solutions, but Hamilton already had one!

For (43.7) to be an extremal with respect to variations of the lapse N and the shift components N_i, it was necessary (see Chapter 21) that the coefficients of these four quantities should vanish; thus,

$$g^{-1/2}\Big[\frac{1}{2}(\operatorname{Tr}\boldsymbol{\pi})^2 - \operatorname{Tr}\boldsymbol{\pi}^2\Big] + g^{1/2}R = 0 \tag{43.8}$$

and

$$\pi^{ij}{}_{|j} = 0. \tag{43.9}$$

In the expression for the extremal value of the action, only one term, the first, is left:

$$S(g(x, y, z)) = I_{\text{ADM, extremal}} = \int_{g'_{ij}}^{g_{ij}} \{\pi^{ij}\,\partial g_{ij}/\partial t\}\,d^4x. \tag{43.10}$$

---

*Actually $S \equiv S_{\text{ADM}} \equiv 16\pi S_{\text{true}} = 16\pi$ (true dynamic path length).

The effect of a slight change, δg_{ij}, in the 3-metric at the upper limit is therefore easy to read off:

$$\delta S = \int \pi^{ij}(x, y, z)\, \delta g_{ij}(x, y, z)\, d^3x. \tag{43.11}$$

The language of "functional derivative" [see, for example, Bogoliubov and Shirkov (1959)] allows one to speak in terms of a derivative rather than an integral:

Geometrodynamic momentum as rate of change of dynamic path length with respect to 3-geometry of terminal hypersurface

$$\frac{\delta S}{\delta g_{ij}} = \pi^{ij}. \tag{43.12}$$

The "field momenta" acquire a simple meaning: they give the rate of change of the action with respect to the continuous infinitude of "field coordinates," $g_{ij}(x, y, z)$. (Here the x, y, z, as well as the i and j, serve as mere labels.)

Action depends on 3-geometry, not on metric coefficients individually

Although the phase function S appears to depend on all six metric coefficients g_{ij} individually, it depends in actuality only on that combination of the g_{ij} which is locked to the 3-geometry. To verify this point, express a particular 3-geometry $^{(3)}\mathscr{G}$ throughout one local coordinate patch in terms of one set of coordinates x^p by one set of metric coefficients g_{pq}. Reexpress the same 3-geometry in terms of coordinates $\bar{x}^p$ shifted by the small amount ξ^p,

$$\bar{x}^p = x^p - \xi^p. \tag{43.13}$$

To keep the 3-geometry the same, that is, to keep unchanged the distance ds from one coordinate-independent point to another, the metric coefficients have to change:

$$\bar{g}_{pq} = g_{pq} + \xi_{p|q} + \xi_{q|p}. \tag{43.14}$$

Let the phase function S (or in quantum mechanics, let the probability amplitude ψ) be considered to be expressed as a functional of the metric coefficients $g_{11}(x)$, $g_{12}(x), \ldots, g_{33}(x)$. Changes $\delta g_{pq}(x)$ in these coefficients alter the H-J phase function and the probability amplitude by the amounts

$$\begin{aligned} \delta S &= \int (\delta S/\delta g_{pq})\, \delta g_{pq}\, d^3x, \\ \delta\psi &= \int (\delta\psi/\delta g_{pq})\, \delta g_{pq}\, d^3x, \end{aligned} \tag{43.15}$$

according to the standard definition of functional derivative. Therefore the coordinate change produces an ostensible change in the dynamic path length or phase S given by

$$\begin{aligned} \delta S &= \int (\delta S/\delta g_{pq})(\xi_{p|q} + \xi_{q|p})\, d^3x \\ &= -2\int (\delta S/\delta g_{pq})_{|q}\xi_p\, d^3x. \end{aligned} \tag{43.16}$$

This change must vanish if S is to depend on the 3-geometry alone, and not on the coordinates in terms of which that 3-geometry is expressed; and must vanish, moreover, for arbitrary choice of the ξ_p. From this condition, one concludes

$$\left(\frac{\delta S}{\delta g_{pq}}\right)_{|q} = 0. \tag{43.17}$$

Likewise, one finds three equations on the wave function ψ itself, as distinguished from its phase $S/\hbar$; thus,

$$\left(\frac{\delta \psi}{\delta g_{pq}}\right)_{|q} = 0. \tag{43.18}$$

But (43.17), by virtue of (43.12), is identical with (43.9). In this sense (43.9) merely verifies what one already knew had to be true: the classical Hamilton-Jacobi function S (like the probability amplitude function ψ of quantum theory) depends on 3-geometry, not on individual metric coefficients, and not on choice of coordinates.

All the dynamic content of geometrodynamics is summarized in the sole remaining equation (43.8), which takes the form

Law of propagation of wave crests in superspace

$$g^{-1/2}\left[\frac{1}{2} g_{pq} g_{rs} - g_{pr} g_{qs}\right] \frac{\delta S}{\delta g_{pq}} \frac{\delta S}{\delta g_{rs}} + g^{1/2} R = 0. \tag{43.19}$$

This is the Einstein-Hamilton-Jacobi equation, first given explicitly in the literature by Peres (1962) on the foundation of earlier work by himself and others on the Hamiltonian formulation of geometrodynamics. This equation tells how fronts of constant S ("wave crests") propagate in superspace.

That the one EHJ equation (43.19) contains as much information as all ten components of Einstein's field equation has been demonstrated by Gerlach (1969). The central point in the analysis is the principle of constructive interference, and the main requirement for a proper treatment of this point is the concept of a completely parametrized solution of the EHJ equation.

The problem of a particle moving in two-dimensional space, as treated by the Hamilton-Jacobi method in Boxes 25.3 and 25.4, required for complete analysis a solution containing two distinct and independently adjustable parameters, the energy per unit mass, $\tilde{E}$, and angular momentum per unit mass, $\tilde{L}$; thus

$$S(r, \theta, t;\ \tilde{E}, \tilde{L}) = -\tilde{E}t + \tilde{L}\theta$$
$$+ \int^r [\tilde{E}^2 - (1 - 2M/r)(1 + \tilde{L}^2/r^2)]^{1/2} \frac{dr}{(1 - 2M/r)} + \delta(\tilde{E}, \tilde{L}). \tag{43.20}$$

Here the additive phase $\delta(\tilde{E}, \tilde{L})$ is required if one is to be able to arrange for the particle to arrive at a given r-value at a specified t value and at a specified value of θ. One thinks of superposing four probability amplitudes, as in (43.4), with dynamic phases, S, given by (43.20) and the parameters taking on, respectively, the following four sets of values: $(\tilde{E}, \tilde{L})$; $(\tilde{E} + \Delta\tilde{E}, \tilde{L})$; $(\tilde{E}, \tilde{L} + \Delta\tilde{L})$; and $(\tilde{E} + \Delta\tilde{E}, \tilde{L} + \Delta\tilde{L})$. The principle of constructive interference leads to the conditions

$$\begin{aligned} \partial S/\partial \tilde{E} &= 0, \\ \partial S/\partial \tilde{L} &= 0. \end{aligned} \tag{43.21}$$

The points in the spacetime (r, θ, t) that satisfy these conditions are the YES points; they lie on the world line. The ones that don't are the NO points.

The desired solution of the EHJ equation (43.19) contains not two parameters (plus an additive phase, δ, depending on these two parameters), but an infinity of parameters, and even a continuous infinity of parameters. Thus the parameters are not to be designated as $\alpha_1, \alpha_2, \ldots; \beta_1, \beta_2, \ldots$ (parameters labeled by a discrete index), but as

$$\alpha(u, v, w)$$

and

$$\beta(u, v, w)$$

(two parameters "labeled" by three continuous indices u, v, w). Accidentally omit one of this infinitude of parameters? How could one ever hope to know that what purported to be a complete solution of the EHJ equation was not in actuality *complete? Happily Gerlach provides a procedure to test the parameters for com*pleteness.

Condition of constructive interference gives classical "leaf of history" or spacetime

Granted completeness, Gerlach goes on to show that the "leaf of history in superspace" or collection of 3-geometries that satisfy the conditions of constructive interference,

$$\begin{aligned} \frac{\delta S({}^{(3)}\mathfrak{G};\ \alpha(u, v, w),\ \beta(u, v, w))}{\delta \alpha} &= 0, \\ \frac{\delta S({}^{(3)}\mathfrak{G};\ \alpha(u, v, w),\ \beta(u, v, w))}{\delta \beta} &= 0, \end{aligned} \tag{43.22}$$

is identical with the leaf of history, or equivalent 4-geometry, given by the ten components of Einstein's geometrodynamic law.

From the Hamilton-Jacobi equation for a problem in elementary mechanics, it is a short step to the corresponding Schroedinger equation; similarly in geometrodynamics. No one has done more than Bryce DeWitt to explore the meaning and consequences of this "Einstein-Schroedinger equation" [DeWitt (1967a,b)]. One of the most interesting consequences is the existence of a conserved current in superspace, analogous to the conserved current

$$j_\mu = \frac{\hbar}{2im}\left(\psi^* \frac{\partial \psi}{\partial x^\mu} - \psi \frac{\partial \psi^*}{\partial x^\mu}\right)$$

that one encounters in the Klein-Gordon wave equation for a particle of spin zero.

It is an unhappy feature of this Einstein-Schroedinger wave equation that it contains second derivatives; so one has to specify both the probability amplitude, and the normal derivative of the probability amplitude, on the appropriate "superhypersurface" in superspace, in order to be able to predict the evolution of this state function elsewhere in superspace. One suggested way out of this situation—it is at least an inconvenience, possibly a real difficulty—has been proposed by Leutwyler (1968): impose a natural boundary condition that reduces the number of independent solutions from the number characteristic of a second-order equation to the number characteristic of a first-order equation. Another way out is to formulate the dynamics quite differently, in the way proposed by Kuchař (see Chapter 21), in which the resulting equation is of first order in the variable analogous to time.

The exploration of quantum geometrodynamics is simplified when one treats most of the degrees of freedom of the geometry as frozen out, by imposition of a high degree of symmetry. Then one is left with one, two, or three degrees of freedom (see Chapter 30, on mixmaster cosmology), or even infinitely many, and is led to manageable problems of quantum mechanics [Misner (1972a, 1973)].

§43.4. FLUCTUATIONS IN GEOMETRY

Of all the remarkable developments of physics since World War II, none is more impressive than the prediction and verification of the effects of the vacuum fluctuations in the electromagnetic field on the motion of the electron in the hydrogen atom (Figure 43.3). That development made it impossible to overlook the effects of such fluctuations throughout all physics and, not least, in the geometry of spacetime itself.

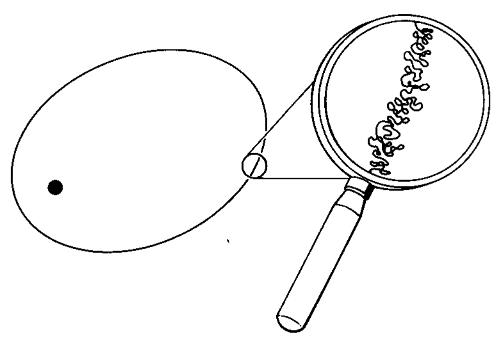


Figure 43.3.
Symbolic representation of motion of electron in hydrogen atom as affected by fluctuations in electric field in vacuum ("vacuum" or "ground state" or "zero-point" fluctuations). The electric field associated with the fluctuation, $E_x(t) = \int E_x(\omega)e^{-i\omega t}\,d\omega$, adds to the static electric field provided by the nucleus itself. The additional field brings about in the most elementary approximation the displacement $\Delta x = \int (e/m\omega^2)E_x(\omega)e^{-i\omega t}\,d\omega$. The average vanishes but the root mean square $\langle(\Delta x)^2\rangle$ does not. In consequence the electron feels an effective atomic potential altered from the expected value $V(x, y, z)$ by the amount

$$\Delta V(x, y, z) = \frac{1}{2}\langle(\Delta x)^2\rangle\,\nabla^2 V(x, y, z).$$

The average of this perturbation over the unperturbed motion accounts for the major part of the observed Lamb-Retherford shift $\Delta E = \langle \Delta V(x, y, z)\rangle$ in the energy level. Conversely, the observation of the expected shift makes the reality of the vacuum fluctuations inescapably evident.

2

From the zero-point fluctuations of a single oscillator to the fluctuations of the electromagnetic field to geometrodynamic fluctuations is a natural order of progression.

Fluctuations for oscillator and for electromagnetic field

A harmonic oscillator in its ground state has a probability amplitude of

$$\psi(x) = \left(\frac{m\omega}{\pi\hbar}\right)^{1/4} e^{-(m\omega/2\hbar)x^2} \tag{43.23}$$

to be displaced by the distance x from its natural classical position of equilibrium. In this sense, it may be said to "resonate" or "fluctuate" between locations in space ranging over a region of extent

$$\Delta x \sim (\hbar/m\omega)^{1/2}. \tag{43.24}$$

The electromagnetic field can be treated as an infinite collection of independent "field oscillators," with amplitudes $\xi_1, \xi_2, \ldots$. When the Maxwell field is in its state of lowest energy, the probability amplitude—for the first oscillator to have amplitude ξ_1, and simultaneously the second oscillator to have amplitude ξ_2, the third ξ_3, and so on—is the product of functions of the form (43.23), one for each oscillator. When the scale of amplitudes for each oscillator is suitably normalized, the resulting infinite product takes the form

$$\psi(\xi_1, \xi_2, \ldots) = N \exp\left[-(\xi_1{}^2 + \xi_2{}^2 + \cdots)\right]. \tag{43.25}$$

This expression gives the probability amplitude ψ for a configuration $\boldsymbol{B}(x,y,z)$ of the magnetic field that is described by the Fourier coefficients $\xi_1, \xi_2, \ldots$. One can forgo any mention of these Fourier coefficients if one desires, however, and rewrite (43.25) directly in terms of the magnetic field configuration itself [Wheeler (1962)]:

$$\psi(\boldsymbol{B}(x,y,z)) = \mathfrak{N} \exp\left(-\int\int \frac{\boldsymbol{B}(\boldsymbol{x}_1)\cdot\boldsymbol{B}(\boldsymbol{x}_2)}{16\pi^3\hbar c r_{12}^2}\, d^3x_1\, d^3x_2\right). \tag{43.26}$$

One no longer speaks of "the" magnetic field, but instead of the probability of this, that, or the other configuration of the magnetic field, even under circumstances, as here, where the electromagnetic field is in its ground state. [See Kuchař (1970) for a similar expression for the "ground state" functional of the linearized gravitational field.]

It is reasonable enough under these circumstances that the configuration of greatest probability is $\boldsymbol{B}(x,y,z) = 0$. Consider for comparison a configuration where the magnetic field is again everywhere zero except in a region of dimension L. There let the field, subject as always to the condition div $\boldsymbol{B} = 0$, be of the order of magnitude ΔB. The probability amplitude for this configuration will be reduced relative to the nil configuration by a factor exp $(-I)$. Here the quantity I in the exponent is of the order $(\Delta B)^2 L^4/\hbar c$. Configurations for which I is large compared to 1 occur with negligible probability. Configurations for which I is small compared to 1 occur with practically the same probability as the nil configuration. In this sense, one can say that the fluctuations in the magnetic field in a region of extension L are of the order of magnitude

$$\Delta B \sim \frac{(\hbar c)^{1/2}}{L^2}. \tag{43.27}$$

In other words, the field "resonates" between one configuration and another with the range of configurations of significance given by (43.27). Moreover, the smaller is the region of space under consideration, the larger are the field magnitudes that occur with appreciable probability.

Still another familiar way of speaking about electromagnetic field fluctuations gives additional insight relevant to geometrodynamics. One considers a measuring device responsive in comparable measure to the magnetic field at all points in a region of dimension L. One asks for the effect on this device of electromagnetic disturbances of various wavelengths. A disturbance of wavelength short compared to L will cause forces to act one way in some parts of the detector, and will give rise to nearly compensating forces in other parts of it. In contrast, a disturbance of a long wavelength λ produces forces everywhere in the same direction, but of a magnitude too low to have much effect. Thus the field, estimated from the equation

$$\begin{pmatrix}\text{energy of electromagnetic}\\ \text{wave of wavelength } \lambda \text{ in a}\\ \text{domain of volume } \lambda^3\end{pmatrix} \sim \begin{pmatrix}\text{energy of one quantum}\\ \text{of wavelength } \lambda\end{pmatrix}$$

or

$$B^2\lambda^3 \sim \frac{\hbar c}{\lambda}$$

or

$$B \sim \frac{(\hbar c)^{1/2}}{\lambda^2} \tag{43.28}$$

is very small if λ is large compared to the domain size L. The biggest effect is caused by a disturbance of wavelength λ comparable to L itself. This line of reasoning leads directly from (43.28) to the standard fluctuation formula (43.27).

Fluctuations in geometry dominate at the Planck scale of distances

Similar considerations apply in geometrodynamics. Quantum fluctuations in the geometry are superposed on and coexist with the large-scale, slowly varying curvature predicted by classical deterministic general relativity. Thus, in a region of dimension L, where in a local Lorentz frame the normal values of the metric coefficients will be $-1, 1, 1, 1$, there will occur fluctuations in these coefficients of the order

$$\Delta g \sim \frac{L^*}{L}, \tag{43.29}$$

fluctuations in the first derivatives of the g_{ik}'s of the order

$$\Delta\Gamma \sim \frac{\Delta g}{L} \sim \frac{L^*}{L^2}, \tag{43.30}$$

2

and fluctuations in the curvature of space of the order

$$\Delta R \sim \frac{\Delta g}{L^2} \sim \frac{L^*}{L^3}. \tag{43.31}$$

Here

$$L^* = \left(\frac{\hbar G}{c^3}\right)^{1/2} = 1.6 \times 10^{-33}\ \text{cm} \tag{43.32}$$

is the so-called Planck length [Planck (1899)].

It is appropriate to look at orders of magnitude. The curvature of space within and near the earth, according to classical Einstein theory, is of the order

$$\begin{aligned} R &\sim \left(\frac{G}{c^2}\right)\rho \sim (0.7 \times 10^{-28}\ \text{cm/g})(5\ \text{g/cm}^3) \\ &\sim 4 \times 10^{-28}\ \text{cm}^{-2}. \end{aligned} \tag{43.33}$$

This quantity has a very direct physical significance. It measures the "tide-producing component of the gravitational field" as sensed, for example, in a freely falling elevator or in a spaceship in free orbit around the earth. By comparison, the quantum fluctuations in the curvature of space are only

$$\Delta R \sim 10^{-33}\ \text{cm}^{-2}, \tag{43.34}$$

even in a domain of observation as small as 1 cm in extent. Thus the quantum fluctuations in the geometry of space are completely negligible under everyday circumstances.

Even in atomic and nuclear physics the fluctuations in the metric,

$$\Delta g \sim \frac{10^{-33}\ \text{cm}}{10^{-8}\ \text{cm}} \sim 10^{-25}$$

and

$$\Delta g \sim \frac{10^{-33}\ \text{cm}}{10^{-13}\ \text{cm}} \sim 10^{-20}, \tag{43.35}$$

are so small that it is completely in order to idealize the physics as taking place in a flat Lorentzian spacetime manifold.

The quantum fluctuations in the geometry are nevertheless inescapable, if one is to believe the quantum principle and Einstein's theory. They coexist with the geometrodynamic development predicted by classical general relativity. The fluctuations widen the narrow swathe cut through superspace by the classical history of the geometry. In other words, the geometry is not deterministic, even though it looks so at the everyday scale of observation. Instead, at a submicroscopic scale it "resonates" between one configuration and another and another. This terminology means no more and no less than the following: (1) Each configuration ${}^{(3)}\mathcal{G}$ has its own probability amplitude $\psi = \psi({}^{(3)}\mathcal{G})$. (2) These probability amplitudes have comparable magnitudes for a whole range of 3-geometries included within the limits (43.29) on either side of the classical swathe through superspace. (3) This range of 3-geometries is far too variegated on the submicroscopic sale to fit into any one 4-geometry, or any one classical geometrodynamic history. (4) Only when one overlooks these small-scale fluctuations ($\sim 10^{-33}$ cm) and examines the larger-scale features of the 3-geometries do they appear to fit into a single space-time manifold, such as comports with the classical field equations.

These small-scale fluctuations tell one that something like gravitational collapse is taking place everywhere in space and all the time; that gravitational collapse is in effect perpetually being done and undone; that in addition to the gravitational collapse of the universe, and of a star, one has also to deal with a third and, because it is constantly being undone, most significant level of gravitational collapse at the Planck scale of distances.

EXERCISES

Exercise 43.1. THE ACTION PRINCIPLE FOR A FREE PARTICLE IN NONRELATIVISTIC MECHANICS

Taking as action principle $I = \int L\, dt =$ extremum, with specified x', t' and x'', t'' at the two limits, and with $L = \frac{1}{2}m(dx/dt)^2$, find (1) the extremizing history $x = x(t)$ and (2) the dynamical path length or action $S(x'', t''; x', t') = I_{\text{extremum}}$ in its dependence on the end points. Also (3) write down the Hamilton-Jacobi equation for this problem, and (4) verify that $S(x, t; x', t')$ satisfies this equation. Finally, imagining the Hamilton-Jacobi equation not to be known, (5) derive it from the already known properties of the function S itself.

Exercise 43.2. THE ACTION FOR THE HARMONIC OSCILLATOR

The kinetic energy is $\frac{1}{2}m(dx/dt)^2$ and the potential energy is $\frac{1}{2}m\omega^2x^2$. Carry through the analysis of parts (1), (2), (3), (4) of the preceding exercise. Partial answer:

$$S = \frac{m\omega}{2}\frac{(x^2 + x'^2)\cos\omega(t - t') - 2xx'}{\sin\omega(t - t')}.$$

Verify that $\partial S/\partial x$ gives momentum and $-\partial S/\partial t$ gives energy.

Exercise 43.3. QUANTUM PROPAGATOR FOR HARMONIC OSCILLATOR

Show that the probability amplitude for a simple harmonic oscillator to transit from x', t' to x'', t'' is

$$\langle x'', t''; x', t'\rangle = \left(\frac{m\omega}{2\pi i\hbar \sin\omega(t'' - t')}\right)^{1/2} \times \exp\frac{im\omega[(x''^2 + x'^2)\cos\omega(t'' - t') - 2x''x']}{2\hbar\sin\omega(t'' - t')},$$

and that it reduces for the case of a free particle to

$$\langle x'', t''; x', t'\rangle = \left(\frac{m}{2\pi i\hbar(t'' - t')}\right)^{1/2} \exp\frac{im(x'' - x')^2}{2\hbar(t'' - t')}.$$

Note that one can derive all the harmonic-oscillator wave functions from the solution by use of the formula

$$\langle x'', t''; x', t'\rangle = \sum_n u_n(x'')u_n^*(x')\exp iE_n(t' - t'')/\hbar.$$

Ex[redacted]TUM PROPAGATOR FOR FREE ELECTROMAGNETIC FIELD

In flat spacetime, one is given on the spacelike hypersurface $t = t'$ the divergence-free magnetic field $B'(x, y, z)$ and on the spacelike hypersurface $t = t''$ the divergence-free magnetic field $B''(x, y, z)$. By Fourier analysis (reducing this problem to the preceding problem) or otherwise, find the probability amplitude to transit from B' at t' to B'' at t''.

Exercise 43.5. HAMILTON-JACOBI FORMULATION OF MAXWELL ELECTRODYNAMICS

Regard the four components A_μ of the electromagnetic 4-potential as the primary quantities; split them into a space part A_i and a scalar potential ϕ. (1) Derive from the action principle (in flat spacetime)

$$I = (1/8\pi) \int (E^2 - B^2)\, d^4x,$$

by splitting off an appropriate divergence, an expression qualitatively similar in character to (43.7). (2) Show that the appropriate quantity to be fixed on the initial and final spacelike hypersurface is not really A_i itself, but the magnetic field, defined by $\boldsymbol{B} = \operatorname{curl} \boldsymbol{A}$. (3) Derive the Hamilton-Jacobi equation for the dynamic phase or action $S(\boldsymbol{B}, \mathcal{S})$ in its dependence on the choice of hypersurface $\mathcal{S}$, and the choice of magnetic field $\boldsymbol{B}$ on this hypersurface,

$$-\frac{\delta S}{\delta \Omega} = \frac{1}{8\pi} \boldsymbol{B}^2 + \frac{(4\pi)^2}{8\pi} \left(\frac{\delta S}{\delta \boldsymbol{A}}\right)^2.$$

The quantity on the left is Tomonaga's "bubble time" derivative [Tomonaga (1946); see also Box 21.1].

2

CHAPTER 44

BEYOND THE END OF TIME

"Heaven wheels above you
Displaying to you her eternal glories
And still your eyes are on the ground"

DANTE

The world "stands before us as a great, eternal riddle"

EINSTEIN (1949a)

§44.1. GRAVITATIONAL COLLAPSE AS THE GREATEST CRISIS IN PHYSICS OF ALL TIME

This chapter is entirely Track 2. No previous Track-2 material is needed as preparation for it, but Chapter 43 will be helpful.

The universe starts with a big bang, expands to a maximum dimension, then recontracts and collapses: no more awe-inspiring prediction was ever made. It is preposterous. Einstein himself could not believe his own prediction. It took Hubble's observations to force him and the scientific community to abandon the concept of a universe that endures from everlasting to everlasting.

Later work of Tolman (1934a), Avez (1960), Geroch (1967), and Hawking and Penrose (1969) generalizes the conclusion. A model universe that is closed, that obeys Einstein's geometrodynamic law, and that contains a nowhere negative density of mass-energy, inevitably develops a singularity. No one sees any escape from the density of mass-energy rising without limit. A computing machine calculating ahead step by step the dynamical evolution of the geometry comes to the point where it can not go on. Smoke, figuratively speaking, starts to pour out of the computer. Yet physics surely continues to go on if for no other reason than this: Physics is by definition that which does go on its eternal way despite all the shadowy changes in the surface appearance of reality.

The Marchon lecture given by J. A. W. at the University of Newcastle-upon-Tyne, May 18, 1971, and the Nuffield lecture given at Cambridge University July 19, 1971, were based on the material presented in this chapter.

Some day a door will surely open and expose the glittering central mechanism of the world in its beauty and simplicity. Toward the arrival of that day, no development holds out more hope than the paradox of gravitational collapse. Why paradox? Because Einstein's equation says "this is the end" and physics says "there is no end." Why hope? Because among all paradigms for probing a puzzle, physics proffers none with more promise than a paradox.

The paradox of collapse: physics stops, but physics must go on

No previous period of physics brought a greater paradox than 1911 (Box 44.1). Rutherford had just been forced to conclude that matter is made up of localized positive and negative charges. Matter as so constituted should undergo electric collapse in $\sim 10^{-17}$ sec, according to theory. Observation equally clearly proclaimed that matter is stable. No one took the paradox more seriously than Bohr. No one worked around and around the central mystery with more energy wherever work was possible. No one brought to bear a more judicious combination of daring and conservativeness, nor a deeper feel for the harmony of physics. The direct opposite

The 1911 crisis of electric collapse

Box 44.1 COLLAPSE OF UNIVERSE PREDICTED BY CLASSICAL THEORY, COMPARED AND CONTRASTED WITH CLASSICALLY PREDICTED COLLAPSE OF ATOM

System	*Atom (1911)*	*Universe (1970's)*
Dynamic entity	System of electrons	Geometry of space
Nature of classically predicted collapse	Electron headed toward point-center of attraction is driven in a finite time to infinite energy	Not only matter but space itself arrives in a finite proper time at a condition of infinite compaction
One rejected "way out"	Give up Coulomb law of force	Give up Einstein's field equation
Another proposal for a "cheap way out" that has to be rejected	"Accelerated charge need not radiate"	"Matter cannot be compressed beyond a certain density by any pressure, however high
How this proposal violates principle of causality	Coulomb field of point-charge cannot readjust itself with infinite speed out to indefinitely great distances to sudden changes in velocity of charge	Speed of sound cannot exceed speed of light; pressure cannot exceed density of mass-energy
A major new consideration introduced by recognizing quantum principle as overarching organizing principle of physics	Uncertainty principle; binding too close to center of attraction makes zero-point kinetic energy outbalance potential energy; consequent existence of a lowest quantum state; can't radiate because no lower state available to drop to	Uncertainty principle; propagation of representative wave packet in superspace does not lead deterministically to a singular configuration for the geometry of space; expect rather a probability distribution of outcomes, each outcome describing a universe with a different size, a different set of particle masses, a different number of particles, and a different length of time required for its expansion and recontraction.

of harmony, cacophony, is the impression that comes from sampling the literature of the 'teens on the structure of the atom. (1) Change the Coulomb law of force between electric charges? (2) Give up the principle that an accelerated charge must radiate? There was little inhibition against this and still wilder abandonings of the well-established. In contrast, Bohr held fast to these two principles. At the same time he insisted on the importance of a third principle, firmly established by Planck in quite another domain of physics, the quantum principle. With that key he opened the door to the world of the atom.

Great as was the crisis of 1911, today gravitational collapse confronts physics with its greatest crisis ever. At issue is the fate, not of matter alone, but of the universe itself. The dynamics of collapse, or rather of its reverse, expansion, is evidenced not by theory alone, but also by observation; and not by one observation, but by observations many in number and carried out by astronomers of unsurpassed ability and integrity. Collapse, moreover, is not unique to the large-scale dynamics of the universe. A white dwarf star or a neutron star of more than critical mass is predicted to undergo gravitational collapse to a black hole (Chapters 32 and 33). Sufficiently many stars falling sufficiently close together at the center of the nucleus of a galaxy are expected to collapse to a black hole many powers of ten more massive than the sun. An active search is under way for evidence for a black hole in this Galaxy (Box 33.3). The process that makes a black hole is predicted to provide an experimental model for the gravitational collapse of the universe itself, with one difference. For collapse to a black hole, the observer has his choice whether (1) to observe from a safe distance, in which case his observations will reveal nothing of what goes on inside the horizon; or (2) to follow the falling matter on in, in which case he sees the final stages of the collapse, not only of the matter itself, but of the geometry surrounding the matter, to indefinitely high compaction, but only at the cost of his own early demise. For the gravitational collapse of a closed model universe, no such choice is available to the observer. His fate is sealed. So too is the fate of matter and elementary particles, driven up to indefinitely high densities. The stakes in the crisis of collapse are hard to match: the dynamics of the largest object, space, and the smallest object, an elementary particle—and how both began.

§44.2. ASSESSMENT OF THE THEORY THAT PREDICTS COLLAPSE

No one reflecting on the paradox of collapse ("collapse ends physics"; "collapse cannot end physics") can fail to ask, "What are the limits of validity of Einstein's geometric theory of gravity?" A similar question posed itself in the crisis of 1911. The Coulomb law for the force acting between two charges had been tested at distances of meters and millimeters, but what warrant was there to believe that it holds down to the 10^{-8} cm of atomic dimensions? Of course, in the end it proves to hold not only at the level of the atom, and at the 10^{-13} cm level of the nucleus, but even down to 5×10^{-15} cm [Barber, Gittelman, O'Neill, and Richter, and Bailey *et al.* (1968), as reviewed by Farley (1969) and Brodsky and Drell (1970)], a striking example of what Wigner (1960) calls the "unreasonable effectiveness of mathematics in the natural sciences."

No theory more resembles Maxwell's electrodynamics in its simplicity, beauty, and scope than Einstein's geometrodynamics. Few principles in physics are more firmly established than those on which it rests: the local validity of special relativity (Chapters 2–7), the equivalence principle (Chapter 16), the conservation of momentum and energy (Chapters 5, 15 and 16), and the prevalence of second-order field equations throughout physics (Chapter 17). Those principles and the demand for no "extraneous fields" (e.g., Dicke's scalar field) and "no prior geometry" (§17.6) lead to the conclusion that the geometry of spacetime must be Riemannian and the geometrodynamic law must be Einstein's.

Battle-tested theory of gravitation

To say that the geometry is Riemannian is to say that the interval between any two nearby events C and D, anywhere in spacetime, stated in terms of the interval AB between two nearby fiducial events, at quite another point in spacetime, has a value CD/AB independent of the route of intercomparison (Chapter 13 and Box 16.4). There are a thousand routes. By this hydraheaded prediction, Einstein's theory thus exposes itself to destruction in a thousand ways (Box 16.4).

Geometrodynamics lends itself to being disproven in other ways as well. The geometry has no option about the control it exerts on the dynamics of particles and fields (Chapter 20). The theory makes predictions about the equilibrium configurations and pulsations of compact stars (Chapters 23–26). It gives formulas (Chapters 27–29) for the deceleration of the expansion of the universe, for the density of mass-energy, and for the magnifying power of the curvature of space, the tests of which are not far off. It predicts gravitational collapse, and the existence of black holes, and a wealth of physics associated with these objects (Chapters 31–34). It predicts gravitational waves (Chapters 35–37). In the appropriate approximation, it encompasses all the well-tested predictions of the Newtonian theory of gravity for the dynamics of the solar system, and predicts testable post-Newtonian corrections besides, including several already verified effects (Chapters 38–40).

No inconsistency of principle has ever been found in Einstein's geometric theory of gravity. No purported observational evidence against the theory has ever stood the test of time. No other acceptable account of physics of comparable simplicity and scope has ever been put forward.

Continue this assessment of general relativity a little further before returning to the central issue, the limits of validity of the theory and their bearing on the issue of gravitational collapse. What has Einstein's geometrodynamics contributed to the understanding of physics?

New view of nature flowing from Einstein's geometrodynamics

First, it has dethroned spacetime from a post of preordained perfection high above the battles of matter and energy, and marked it as a new dynamic entity participating actively in this combat.

Second, by tying energy and momentum to the curvature of spacetime, Einstein's theory has recognized the law of conservation of momentum and energy as an automatic consequence of the geometric identity that the boundary of a boundary is zero (Chapters 15 and 17).

Third, it has recognized gravitation as a manifestation of the curvature of the geometry of spacetime rather than as something foreign and "physical" imbedded in spacetime.

Fourth, general relativity has reinforced the view that "physics is local"; that the analysis of physics becomes simple when it connects quantities at a given event with quantities at immediately adjacent events.

Fifth, obedient to the quantum principle, it recognizes that spacetime and time itself are ideas valid only at the classical level of approximation; that the proper arena for the Einstein dynamics of geometry is not spacetime, but superspace; and that this dynamics is described in accordance with the quantum principle by the propagation of a probability amplitude through superspace (Chapter 43). In consequence, the geometry of space is subject to quantum fluctuations in metric coefficients of the order

$$\delta g \sim \frac{(\text{Planck length, } L^* = (\hbar G/c^3)^{1/2} = 1.6 \times 10^{-33}\,\text{cm})}{(\text{linear extension of region under study})}.$$

Electric charge as lines of force trapped in the topology of space

Sixth, standard Einstein geometrodynamics is partial as little to Euclidean topology as to Euclidean geometry. A multiply connected topology provides a natural description for electric charge as electric lines of force trapped in the topology of a multiply connected space (Figure 44.1). Any other description of electricity postulates a breakdown in Maxwell's field equations for the vacuum at a site where charge

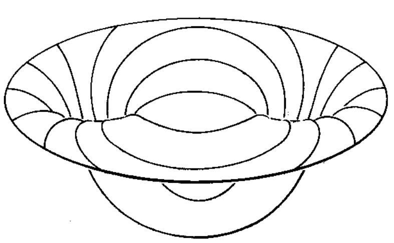


Figure 44.1.
Electric charge viewed as electric lines of force trapped in the topology of a multiply connected space [for the history of this concept see reference 36 of Wheeler (1968a)]. The wormhole or handle is envisaged as connecting two very different regions in the same space. One of the wormhole mouths, viewed by an observer with poor resolving power, appears to be the seat of an electric charge. Out of this region of 3-space he finds lines of force emerging over the whole 4π solid angle. He may construct a boundary around this charge, determine the flux through this boundary, incorrectly apply the theorem of Gauss and "prove" that there is a charge "inside the boundary." It isn't a boundary. Someone caught within it—to speak figuratively—can go into that mouth of the wormhole, through the throat, out the other mouth, and return by way of the surrounding space to look at his "prison" from the outside. Lines of force nowhere end. Maxwell's equations nowhere fail. Nowhere can one place a finger and say, "Here there is some charge." This classical type of electric charge has no direct relation whatsoever to quantized electric charge. There is a freedom of choice about the flux through the wormhole, and a specificity about the connection between one charge and another, which is quite foreign to the charges of elementary particle physics. For ease of visualization the number of space dimensions in the above diagram has been reduced from three to two. The third dimension, measured off the surface, has no physical meaning—it only provides an extra dimension in which to imbed the surface for more convenient diagrammatic representation. [For more detail see Misner and Wheeler (1957), reprinted in Wheeler (1962)].

is located, or postulates the existence of some foreign and "physical" electric jelly imbedded in space, or both. No one has ever found a way to describe electricity free of these unhappy features except to say that the quantum fluctuations in the geometry of space are so great at small distances that even the topology fluctuates, makes "wormholes," and traps lines of force. These fluctuations have to be viewed, not as tied to particles, and endowed with the scale of distances associated with particle physics ($\sim 10^{-13}$ cm) but as pervading all space ("foam-like structure of geometry") and characterized by the Planck distance ($\sim 10^{-33}$ cm). Thus a third type of gravitational collapse forces itself on one's attention, a collapse continually being done and being undone everywhere in space: surely a guide to the outcome of collapse at the level of a star and at the level of the universe (Box 44.2).

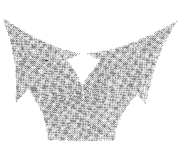
Box 44.2 THREE LEVELS OF GRAVITATIONAL COLLAPSE

1. Universe
2. Black hole
3. Fluctuations at the Planck scale of distances

Recontraction and collapse of the universe is a kind of mirror image of the "big bang," on which one already has so much evidence.

Collapse of matter to form a black hole is most natural at two distinct levels: (a) collapse of the dense white-dwarf core of an individual star (when that core exceeds the critical mass, $\sim 1 M_\odot$ or $2 M_\odot$, at which a neutron star is no longer a possible stable end-point for collapse) and (b) coalescence one by one of the stars in a galactic nucleus to make a black hole of mass up to $10^6 M_\odot$ or even $10^9 M_\odot$.

In either case, no feature of principle about matter falling into the black hole is more interesting than the option that the observer has (symbolized by the branching arrow in the inset). He can go along with the infalling matter, in which case he sees the final stages of collapse, but only at the cost of his own demise. Or he can stay safely outside, in which case even after indefinitely long time he sees only the first part of the collapse, with the infalling matter creeping up more and more slowly to the horizon.



In the final stages of the collapse of a closed model universe, all black holes present are caught up and driven together, amalgamating one by one. No one has any way to look at the event from safely outside; one is inevitably caught up in it oneself.

Collapse at the Planck scale of distances is taking place everywhere and all the time in quantum fluctuations in the geometry and, one believes, the topology of space. In this sense, collapse is continually being done and undone, modeling the undoing of the collapse of the universe itself, summarized in the term, "the reprocessing of the universe" (see text).

§44.3. VACUUM FLUCTUATIONS: THEIR PREVALENCE AND FINAL DOMINANCE

If Einstein's theory thus throws light on the rest of physics, the rest of physics also throws light on geometrodynamics. No point is more central than this, that empty space is not empty. It is the seat of the most violent physics. The electromagnetic field fluctuates (Chapter 43). Virtual pairs of positive and negative electrons, in effect, are continually being created and annihilated, and likewise pairs of mu mesons, pairs of baryons, and pairs of other particles. All these fluctuations coexist with the quantum fluctuations in the geometry and topology of space. Are they additional to those geometrodynamic zero-point disturbances, or are they, in some sense not now well-understood, mere manifestations of them?

Is matter built out of geometry?

Put the question in other words. Recall Clifford, inspired by Riemann, speaking to the Cambridge Philosophical Society on February 21, 1870, "On the Space Theory of Matter" [Clifford (1879), pp. 244 and 322; and (1882), p. 21], and saying, "I hold in fact (1) That small portions of space *are* in fact of a nature analogous to little hills on a surface which is on the average flat; namely, that the ordinary laws of geometry are not valid in them. (2) That this property of being curved or distorted is continually being passed on from one portion of space to another after the manner of a wave. (3) That this variation of the curvature of space is what really happens in that phenomenon which we call the *motion of matter,* whether ponderable or etherial. (4) That in the physical world nothing else takes place but this variation, subject (possibly) to the law of continuity." Ask if there is a sense in which one can speak of a particle as constructed out of geometry. Or rephrase the question in updated language: "Is a particle a geometrodynamic exciton?" What else is there out of which to build a particle except geometry itself? And what else is there to give discreteness to such an object except the quantum principle?

The Clifford-Einstein space theory of matter has not been forgotten in recent years. "In conclusion," one of the authors wrote a decade ago [Wheeler (1962)], "the vision of Riemann, Clifford, and Einstein, of a purely geometric basis for physics, today has come to a higher state of development, and offers richer prospects—and presents deeper problems—than ever before. The quantum of action adds to this geometrodynamics new features, of which the most striking is the presence of fluctuations of the wormhole type throughout all space. If there is any correspondence at all between this virtual foam-like structure and the physical vacuum as it has come to be known through quantum electrodynamics, then there seems to be no escape from identifying these wormholes with 'undressed electrons.' Completely different from these 'undressed electrons,' according to all available evidence, are the electrons and other particles of experimental physics. For these particles the geometrodynamic picture suggests the model of collective disturbances in a virtual foam-like vacuum, analogous to different kinds of phonons or excitons in a solid.

The richness of the physics of the vacuum

"The enormous factor from nuclear densities $\sim 10^{14}$ g/cm^3 to the density of field fluctuation energy in the vacuum, $\sim 10^{94}$ g/cm^3, argues that elementary particles represent a percentage-wise almost completely negligible change in the locally violent conditions that characterize the vacuum. ['A particle (10^{14} g/cm^3) means as little to the physics of the vacuum (10^{94} g/cm^3) as a cloud (10^{-6} g/cm^3) means to the physics of the sky (10^{-3} g/cm^3).'] In other words, elementary particles do not form a really basic starting point for the description of nature. Instead, they represent a first-order correction to vacuum physics. That vacuum, that zero-order state of affairs, with its enormous densities of virtual photons and virtual positive-negative pairs and virtual wormholes, has to be described properly before one has a fundamental starting point for a proper perturbation-theoretic analysis."

"These conclusions about the energy density of the vacuum, its complicated topological character, and the richness of the physics which goes on in the vacuum, stand in no evident contradiction with what quantum electrodynamics has to say about the vacuum. Instead the conclusions from the 'small distance' analysis (10^{-33} cm)—sketchy as it is—and from 'larger distance' analysis (10^{-11} cm) would seem to [be able] to reinforce each other in a most natural way.

"The most evident shortcoming of the geometrodynamic model as it stands is this, that *it fails to supply any completely natural place* for spin $\frac{1}{2}$ in general and *for the neutrino* in particular."

Attempts to find a natural place for spin $\frac{1}{2}$ in Einstein's standard geometrodynamics (Box 44.3) founder because there is no natural way for a *change* in connectivity to take place within the context of classical differential geometry.

A uranium nucleus undergoing fission starts with one topology and nevertheless ends up with another topology. It makes this transition in a perfectly continuous way, classical differential geometry notwithstanding.

No place in geometrodynamics for change of topology; therefore turn to "pregeometry"

There are reputed to be two kinds of lawyers. One tells the client what not to do. The other listens to what the client has to do and tells him how to do it. From the first lawyer, classical differential geometry, the client goes away disappointed, still searching for a natural way to describe quantum fluctuations in the connectivity of space. Only in this way can he hope to describe electric charge as lines of electric force trapped in the topology of space. Only in this way does he expect to be able to understand and analyze the final stages of gravitational collapse. Pondering his problems, he comes to the office of a second lawyer, with the name "Pregeometry" on the door. Full of hope, he knocks and enters. What is pregeometry to be and say? Born of a combination of hope and need, of philosophy and physics and mathematics and logic, pregeometry will tell a story unfinished at this writing, but full of incidents of evolution so far as it goes.

§44.4. NOT GEOMETRY, BUT PREGEOMETRY, AS THE MAGIC BUILDING MATERIAL

An early survey (Box 44.4) asked whether geometry can be constructed with the help of the quantum principle out of more basic elements, that do not themselves have any specific dimensionality.

The focus of attention in this 1964 discussion was "dimensionality without dimensionality." However, the prime pressures to ponder pregeometry were and remain

(*continued on page 1206*)

Box 44.3 THE DIFFICULTIES WITH ATTEMPTS TO FIND A NATURAL PLACE FOR SPIN $\frac{1}{2}$ IN EINSTEIN'S STANDARD GEOMETRODYNAMICS

"It is impossible" [Wheeler (1962)] "to accept any description of elementary particles that does not have a place for spin $\frac{1}{2}$. What, then, has any purely geometric description to offer in explanation of spin $\frac{1}{2}$ in general? More particularly and more importantly, what place is there in quantum geometrodynamics for the neutrino—the only entity of half-integral spin that is a pure field in its own right, in the sense that it has zero rest mass and moves with the speed of light? No clear or satisfactory answer is known to this question today. Unless and until an answer is forthcoming, *pure geometrodynamics must be judged deficient as a basis for elementary particle physics.*"

A later publication [Wheeler (1968a)] takes up this issue again, noting that, "A new world opens out for analysis in quantum geometrodynamics. The central new concept is space resonating between one foamlike structure and another. For this multiple-connectedness of space at submicroscopic distances no single feature of nature speaks more powerfully than electric charge. Yet at least as impressive as charge is the prevalence of spin $\frac{1}{2}$ throughout the world of elementary particles."

Repeating the statement that "It is impossible to accept any description of elementary particles that does not have a place for spin $\frac{1}{2}$," the article adds to the discussion a new note: "Happily, the concept of spin manifold has come to light, not least through the work of John Milnor [see Lichnerowicz (1961a,b,c) and (1964); Milnor (1962), (1963), and (1965a,b); Hsiang and Anderson (1965); Anderson, Brown, and Peterson (1966a,b); and Penrose (1968a)]. This concept suggests a new and *interesting interpretation of a spinor field* within the context of the resonating microtopology of quantum geometrodynamics, *as the nonclassical two-valuedness* [Pauli's standard term for spin; see, for example, Pauli (1947)] *that attaches to the probability amplitude for otherwise identical 3-geometries endowed with alternative 'spin structures.'*" More specifically: "One does not classify the closed orientable 3-manifold of physics completely when one gives its topology, its differential structures, and its metric. One must tell which spin structure it has." [On a 3-geometry with the topology of a 3-sphere, one can lay down a continuous field of triads (a triad consisting of three orthonormal vectors). Any other continuous field of triads can be deformed into the first field by a continuous sequence of small readjustments. One says that the 3-sphere admits only one "spin structure," a potentially misleading standard word for what could just as well have been called a "triad structure." In contrast, a 3-sphere with n handles or wormholes admits 2^n "spin structures" (continuous fields of triads) inequivalent to one another under any continuous sequence of small readjustments whatsoever, and distinguished from one another in any convenient way by n "descriptors" $w_1, w_2, \ldots,$ $w_k, \ldots w_n$.] It is natural in quantum geometrodynamics to expect "separate probability amplitudes for a 3-geometry with descriptor $w_k = +1$ and for an otherwise identical 3-geometry with descriptor $w_k = -1$. Does this circumstance imply that quantum geometrodynamics supplies all the machinery one needs to describe fields of spin $\frac{1}{2}$ in general and the neutrino field in particular? . . . That is the only way that has ever turned up within the framework of Einstein's general relativity and Planck's quantum principle. Is this the right path? It is difficult to name any question more decisive than this in one's assessment of 'everything as geometry.'"

Why not spell out these concepts, reduce them to practice, and compare them with what one knows about the behavior of fields of spin $\frac{1}{2}$? There is a central difficulty in this enterprise. It assumes and demands on physical grounds that the topology of the 3-geometry shall be free to change from one connectivity to another. In contrast, classical differential geometry says, in effect, "Once one topology, always that topology." Try a question like this, "When a new handle develops and the number of descriptors rises by one, what boundary condition in superspace connects the probability amplitude ψ for 3-geometries of the original topology with the probability amplitudes ψ_+ and ψ_- for the two spin structures of the new topology?" Classical differential geometry not only gives one no help in answering this question; it even forbids one to ask it. In other words, one cannot even get the enterprise "on the road" for want of a natural mathematical way to describe the required change in topology. The idea is therefore abandoned here and now that 3-geometry is "the magic building material of the universe." In contrast, pregeometry (see text), far from being endowed with any frozen topology, is to be viewed as not even possessing any dimensionality.

Box 44.4 "A BUCKET OF DUST"—AN EARLY ATTEMPT TO FORMULATE THE CONCEPT OF PREGEOMETRY [Wheeler (1964a)]

"What line of thought could ever be imagined as *leading* to four dimensions—or any dimensionality at all—out of more primitive considerations? In the case of atoms one derives the yellow color of the sodium D-lines by analyzing the quantum dynamics of a system, no part of which is ever endowed with anything remotely resembling the attribute of color. Likewise any derivation of the four-dimensionality of spacetime can hardly *start* with the idea of dimensionality."

"Recall the notion of a Borel set. Loosely speaking, a Borel set is a collection of points ("bucket of dust") which have not yet been assembled into a manifold of any particular dimensionality. . . . Recalling the universal sway of the quantum principle, one can imagine probability amplitudes for the points in a Borel set to be assembled into points with this, that, and the other dimensionality. . . . More conditions have to be imposed on a given number of points—as to which has which for a nearest neighbor—when the points are put together in a five-dimensional array than when these same points are arranged in a two-dimensional pattern. Thus one can think of each dimensionality as having a much higher statistical weight than the next *higher* dimensionality. On the other hand, for manifolds with one, two, and three dimensions, the geometry is too rudimentary—one can suppose—to give anything interesting. Thus Einstein's field equations, applied to a manifold of dimensionality so low, demand flat space; only when the dimensionality is as high as four do really interesting possibilities arise. Can four, therefore, be considered to be the unique dimensionality which is at the same time high enough to give any real physics and yet low enough to have great statistical weight?

"It is too much to imagine that one has yet made enough mistakes in this domain of thought to explore such ideas with any degree of good judgment."

Consider a handle on the geometry. Let it thin halfway along its length to a point. In other words, let the handle dissolve into two bent prongs that touch at a point. Let these prongs separate and shorten. In this process two points part company that were once immediate neighbors. "However sudden the change is in classical theory, in quantum theory there is a probability amplitude function which falls off in the classically forbidden domain. In other words, there is some residual connection between points which are ostensibly very far away (travel from one 'tip' down one prong, then through the larger space to which these prongs are attached, and then up the other prong to the other tip). But there is nothing distinctive in principle about the two points that have happened to come into discussion. Thus it would seem that there must be a connection . . . between *every* point and every other point. Under these conditions the concept of nearest neighbor would appear no longer to make sense. Thus the tool disappears with the help of which one might otherwise try to speak [un]ambiguously about dimensionality."

two features of nature, spin $\frac{1}{2}$ and charge, that speak out powerfully from every part of elementary particle physics.

A fresh perspective on pregeometry comes from a fresh assessment of general relativity. "Geometrodynamics is neither as important or as simple as it looks. Do not make it the point of departure in searching for underlying simplicity. Look deeper, at elementary particle physics." This is the tenor of interesting new considerations put forward by Sakharov [*the* Sakharov] (1967) and summarized under the heading, "Gravitation as the metric elasticity of space," in Box 17.2. In brief, as elasticity is to atomic physics, so—in Sakharov's view—gravitation is to elementary particle physics. The energy of an elastic deformation is nothing but energy put into the bonds between atom and atom by the deformation. The energy that it takes to curve space is nothing but perturbation in the vacuum energy of fields plus particles brought about by that curvature, according to Sakharov. The energy required for the deformation is governed in the one case by two elastic constants and in the other case by one elastic constant (the Newtonian constant of gravity) but in both cases, he reasons, the constants arise by combination of a multitude of complicated individual effects, not by a brave clean stroke on an empty slate.

Sakharov: gravitation is the "metric elasticity of space"

One gives all the more favorable reception to Sakhorov's view of gravity because one knows today, as one did not in 1915, how opulent in physics the vacuum is. In Einstein's day one had come in a single decade from the ideal God-given Lorentz perfection of flat spacetime to curved spacetime. It took courage to assign even one physical constant to that world of geometry that had always stood so far above physics. The vacuum looked for long as innocent of structure as a sheet of glass emerging from a rolling mill. With the discovery of the positive electron [Anderson (1933)], one came to recognize a little of the life that heat can unfreeze in "empty" space. Each new particle and radiation that was discovered brought a new accretion to the recognized richness of the vacuum. Macadam looks smooth, but a bulldozer has only to cut a single furrow through the roadway to disclose all the complications beneath the surface.

The stratification of space

Think of a particle as built out of the geometry of space; think of a particle as a "geometrodynamic exciton"? No model—it would seem to follow from Sakharov's assessment—could be less in harmony with nature, except to think of an atom as built out of elasticity! Elasticity did not explain atoms. Atoms explained elasticity. If, likewise, particles fix the constant in Einstein's geometrodynamic law (Sakharov), must it not be unreasonable to think of the geometrodynamic law as explaining particles?

Comparison with everyday elasticity

Carry the comparison between geometry and elasticity one stage deeper (Fig. 44.2). In a mixed solid there are hundreds of distinct bonds, all of which contribute to the elastic constants; some of them arise from Van der Waal's forces, some from ionic coupling, some from homopolar linkage; they have the greatest variety of strengths; but all have their origin in something so fantastically simple as a system of positively and negatively charged masses moving in accordance with the laws of quantum mechanics. In no way was it required or right to meet each complication of the chemistry and physics of a myriad of bonds with a corresponding complication of principle. By going to a level of analysis deeper than bond strengths, one had

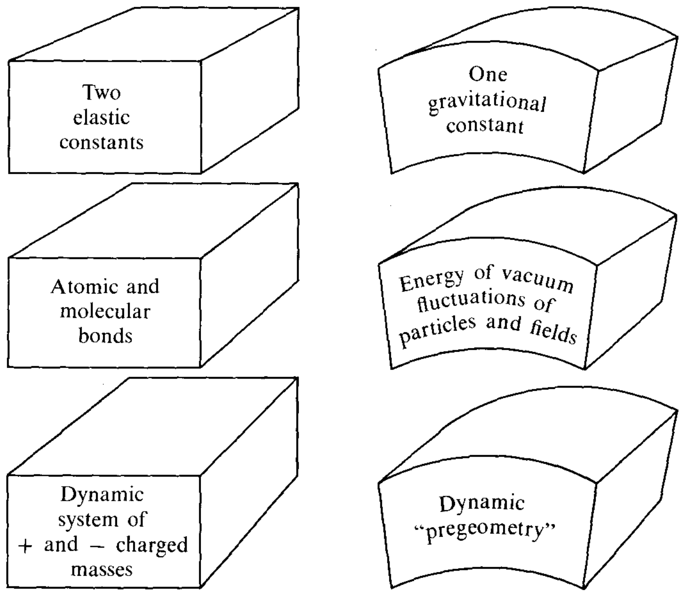


Figure 44.2.
Elasticity and geometrodynamics, as viewed at three levels of analysis. A hundred years of the study of elasticity did not reveal the existence of molecules, and a hundred years of the study of molecular chemistry did not reveal Schrödinger's equation. Revelation moved upward in the diagram, not downward.

emerged into a world of light, where nothing but simplicity and unity was to be seen.

Compare with geometry. The vacuum is animated with the zero-point activity of distinct fields and scores of distinct particles, all of which, according to Sakharov, contribute to the Newtonian G, the "elastic constant of the metric." Some interact via weak forces, some by way of electromagnetic forces, and some through strong forces. These interactions have the greatest variety of strengths. But must not all these particles and interactions have their origin in something fantastically simple? And must not this something, this "pregeometry," be as far removed from geometry as the quantum mechanics of electrons is far removed from elasticity?

If one once thought of general relativity as a guide to the discovery of pregeometry, nothing might seem more dismaying than this comparison with an older realm of physics. No one would dream of studying the laws of elasticity to uncover the principles of quantum mechanics. Neither would anyone investigate the work-hardening of a metal to learn about atomic physics. The order of understanding ran not

$$\text{Work-hardening (1 cm)} \longrightarrow \text{dislocations } (10^{-4}\text{ cm}) \longrightarrow \text{atoms } (10^{-8}\text{ cm}),$$

but the direct opposite,

$$\text{Atoms } (10^{-8}\text{ cm}) \longrightarrow \text{dislocations } (10^{-4}\text{ cm}) \longrightarrow \text{work-hardening (1 cm)}$$

One had to know about atoms to conceive of dislocations, and had to know about dislocations to understand work-hardening. Is it not likewise hopeless to go from the "elasticity of geometry" to an understanding of particle physics, and from particle physics to the uncovering of pregeometry? Must not the order of progress again be the direct opposite? And is not the source of any dismay the apparent loss of guidance that one experiences in giving up geometrodynamics—and not only geometrodynamics, but geometry itself—as a crutch to lean on as one hobbles forward? Yet there is so much chance that this view of nature is right that one must take it seriously and explore its consequences. Never more than today does one have the incentive to explore pregeometry.

§44.5. PREGEOMETRY AS THE CALCULUS OF PROPOSITIONS

Paper in white the floor of the room, and rule it off in one-foot squares. Down on one's hands and knees, write in the first square a set of equations conceived as able to govern the physics of the universe. Think more overnight. Next day put a better set of equations into square two. Invite one's most respected colleagues to contribute to other squares. At the end of these labors, one has worked oneself out into the door way. Stand up, look back on all those equations, some perhaps more hopeful than others, raise one's finger commandingly, and give the order "Fly!" Not one of those equations will put on wings, take off, or fly. Yet the universe "flies."

Search for the central principle of pregeometry

Some principle uniquely right and uniquely simple must, when one knows it, be also so compelling that it is clear the universe is built, and must be built, in such and such a way, and that it could not possibly be otherwise. But how can one discover that principle? If it was hopeless to learn atomic physics by studying work-hardening and dislocations, it may be equally hopeless to learn the basic operating principle of the universe, call it pregeometry or call it what one will, by any amount of work in general relativity and particle physics.

Thomas Mann (1937), in his essay on Freud, utters what Niels Bohr would surely have called a great truth ("A great truth is a truth whose opposite is also a great truth") when he says, "Science never makes an advance until philosophy authorizes and encourages it to do so." If the equivalence principle (Chapter 16) and Mach's principle (§21.9) were the philosophical godfathers of general relativity, it is also true that what those principles do mean, and ought to mean, only becomes clear by study and restudy of Einstein's theory itself. Therefore it would seem reasonable to expect the primary guidance in the search for pregeometry to come from a principle both philosophical and powerful, but one also perhaps not destined to be wholly clear in its contents or its implications until some later day.

Among all the principles that one can name out of the world of science, it is difficult to think of one more compelling than *simplicity;* and among all the simplicities of dynamics and life and movement, none is starker [Werner (1969)] than the *binary choice* yes–no or true–false. It in no way proves that this choice for a starting principle is correct, but it at least gives one some comfort in the choice, that Pauli's "nonclassical two-valuedness" or "spin" so dominates the world of particle physics.

It is one thing to have a start, a tentative construction of pregeometry; but how does one go on? How not to go on is illustrated by Figure 44.3. The "sewing machine" builds objects of one or another definite dimensionality, or of mixed dimensionalities, according to the instructions that it receives on the input tape in yes–no binary code. Some of the difficulties of building up structure on the binary element according to this model, or any one of a dozen other models, stand out at once. (1) Why $N = 10{,}000$ building units? Why not a different N? And if one feeds in one such arbitrary number at the start, why not fix more features "by hand?" No natural stopping point is evident, nor any principle that would fix such a stopping point. Such arbitrariness contradicts the principle of simplicity and rules out the model. (2) Quantum mechanics is added from outside, not generated from inside (from the model itself). On this point too the principle of simplicity speaks against the model. (3) The passage from pregeometry to geometry is made in a too-literal-minded way, with no appreciation of the need for particles and fields to appear along the way. The model, in the words used by Bohr on another occasion, is "crazy, but not crazy enough to be right."

A first try at a pregeometry built on the principle of binary choice

Noting these difficulties, and fruitlessly trying model after model of pregeometry to see if it might be free of them, one suddenly realizes that a machinery for the combination of yes–no or true–false elements does not have to be invented. It already exists. What else can pregeometry be, one asks oneself, than the calculus of propositions? (Box 44.5.)

A more reasonable picture: pregeometry is the calculus of propositions

§44.6. THE BLACK BOX: THE REPROCESSING OF THE UNIVERSE

No amount of searching has ever disclosed a "cheap way" out of gravitational collapse, any more than earlier it revealed a cheap way out of the collapse of the atom. Physicists in that earlier crisis found themselves in the end confronted with a revolutionary pistol, "Understand nothing—or accept the quantum principle." Today's crisis can hardly force a lesser revolution. One sees no alternative except to say that geometry fails and pregeometry has to take its place to ferry physics through the final stages of gravitational collapse and on into what happens next. No guide is evident on this uncharted way except the principle of simplicity, applied to drastic lengths.

Whether the whole universe is squeezed down to the Planck dimension, or more or less, before reexpansion can begin and dynamics can return to normal, may be irrelevant for some of the questions one wants to consider. Physics has long used the "black box" to symbolize situations where one wishes to concentrate on what goes in and what goes out, disregarding what takes place in between.

The role of the black box in physics

At the beginning of the crisis of electric collapse one conceived of the electron as headed on a deterministic path toward a point-center of attraction, and unhappily destined to arrive at a condition of infinite kinetic energy in a finite time. After the advent of quantum mechanics, one learned to summarize the interaction between

(*continued on page 1213*)

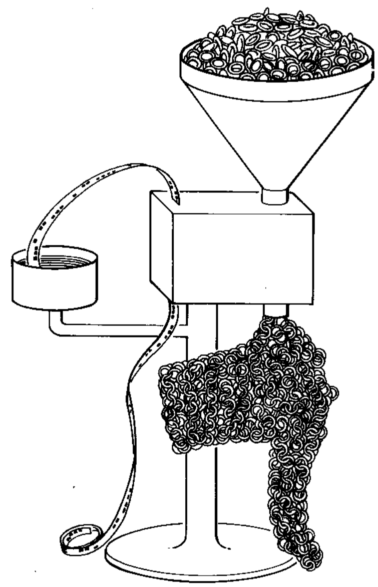


Figure 44.3.
"Ten thousand rings"; or an example of a way to think of the connection between pregeometry and geometry, wrong because it is too literal-minded, and for other reasons spelled out in the text. The vizier [story by Wheeler, as alluded to by Kilmister (1971)*] speaks: "Take $N = 10{,}000$ brass rings. Take an automatic fastening device that will cut open a ring, loop it through another ring, and resolder the joint. Pour the brass rings into the hopper that feeds this machine. Take a strip of instruction paper that is long enough to contain $N(N-1)/2$ binary digits. Look at the instruction in the (jk)-th location on this instruction tape ($j,k = 1, 2, \ldots, N$; $j < k$). When the binary digit at that location is 0, it is a signal to leave the j-th ring disconnected from the k-th ring. When it is 1, it is an instruction to connect that particular pair of rings. Thread the tape into the machine and press the start button. The clatter begins. Out comes a chain of rings 10,000 links long. It falls on the table and the machine stops. Pour in another 10,000 rings, feed in a new instruction tape, and push the button again. This time it is not a one-dimensional structure that emerges, but a two-dimensional one: a Crusader's coat of mail, complete with neck opening and sleeves. Take still another tape from the library of tapes and repeat. Onto the table thuds a smaller version of the suit of mail, this time filled out internally with a solid network of rings, a three-dimensional structure. Now forego the library and make one's own instruction tape, a random string of 0's and 1's. Guided by it, the fastener produces a "Christmas tree ornament," a collection of segments of one-dimensional chain, two-dimensional surfaces, and three-, four-, five-, and higher-dimensional entities, some joined together, some free-floating. Now turn from a structure deterministically fixed by a tape to a probability amplitude, a complex number,

$$\psi(\text{tape}) = \psi(n_{12}, n_{13}, n_{14}, \ldots\ldots, n_{N-1,N}) \qquad (n_{ij} = 0, 1), \tag{1}$$

defined over the entire range of possibilities for structures built of 10,000 rings. Let these probability amplitudes *not* be assigned randomly. Instead, couple together amplitudes, for structures that differ from each other by the breaking of a single ring, by linear formulas that treat all rings on the same footing. The separate ψ's, no longer entirely independent, will still give non-zero probability amplitudes for "Christmas tree ornaments." Of greater immediate interest than these "unruly" parts of the structures are the following questions about the smoother parts: (1) In what kinds of structures is the bulk of the probability concentrated? (2) What is the dominant dimensionality of these structures in an appropriate correspondence principle limit? (3) In this semiclassical limit, what is the form taken by the dynamic law of evolution of the geometry?" No principle more clearly rules out this model for pregeometry than the principle of simplicity (see text).

*Wheeler's story about the vizier and what the vizier had to say about superspace was told at the May 18, 1970, Gwatt Seminar on the Bearings of Topology upon General Relativity. Kilmister's (1971) published article alludes to the unpublished story, but does not actually contain it.

Box 44.5 "PREGEOMETRY AS THE CALCULUS OF PROPOSITIONS"

A sample proposition taken out of a standard text on logic selected almost at random reads [Kneebone (1963), p. 40]

$$[X \longrightarrow ((X \longrightarrow X) \longrightarrow Y)] \,\&\, (\bar{X} \longrightarrow Z) \text{ eq } (\bar{X} \vee Y \vee Z) \,\&\, (\bar{X} \vee Y \vee \bar{Z}) \,\&\, (X \vee Y \vee Z) \,\&\, (X \vee \bar{Y} \vee Z).$$

The symbols have the following meaning:

$\bar{A}$,	Not A;
$A \vee B$,	A or B or both ("A *vel* B");
$A \,\&\, B$,	A and B;
$A \longrightarrow B$,	A implies B ("if A, then B");
$A \longleftrightarrow B$,	B is equivalent to A ("B if and only if A").

Propositional formula $\mathfrak{A}$ is said to be equivalent ("eq") to propositional formula $\mathfrak{B}$ if and only if $\mathfrak{A} \longleftrightarrow \mathfrak{B}$ is a tautology. The letters A, B, etc., serve as connectors to "wire together" one proposition with another. Proceeding in this way, one can construct propositions of indefinitely great length.

A switching circuit [see, for example, Shannon (1938) or Hohn (1966)] is isomorphic to a proposition.

Compare a short proposition or an elementary switching circuit to a molecular collision. No idea seemed more preposterous than that of Daniel Bernoulli (1733), that heat is a manifestation of molecular collisions. Moreover, a three-body encounter is difficult to treat, a four-body collision is more difficult, and a five- or more molecule system is essentially intractable. Nevertheless, mechanics acquires new elements of simplicity in the limit in which the number of molecules is very great and in which one can use the concept of density in phase space. Out of statistical mechanics in this limit come such concepts as temperature and entropy. When the temperature is well-defined, the energy of the system is not a well-defined idea; and when the energy is well-defined, the temperature is not. This complementarity is built inescapably into the principles of the subject. Thrust the finger into the flame of a match and experience a sensation like nothing else on heaven or earth; yet what happens is all a consequence of molecular collisions, early critics notwithstanding.

Any individual proposition is difficult for the mind to apprehend when it is long; and still more difficult to grasp is the content of a cluster of propositions. Nevertheless, make a statistical analysis of the calculus of propositions in the limit where the number of propositions is great and most of them are long. Ask if parameters force themselves on one's attention in this analysis (1) analogous in some small measure to the temperature and entropy of statistical mechanics but (2) so much

Box 44.5 (*continued*)

more numerous, and everyday dynamic in character, that they reproduce the continuum of everyday physics.

Nothing could seem so preposterous at first glance as the thought that nature is built on a foundation as ethereal as the calculus of propositions. Yet, beyond the push to look in this direction provided by the principle of simplicity, there are two pulls. First, bare-bones quantum mechanics lends itself in a marvelously natural way to formulation in the language of the calculus of propositions, as witnesses not least the book of Jauch (1968). If the quantum principle were not in this way already automatically contained in one's proposed model for pregeometry, and if in contrast it had to be introduced from outside, by that very token one would conclude that the model violated the principle of simplicity, and would have to reject it. Second, the pursuit of reality seems always to take one away from reality. Who would have imagined describing something so much a part of the here and now as gravitation in terms of curvature of the geometry of spacetime? And when later this geometry came to be recognized as dynamic, who would have dreamed that geometrodynamics unfolds in an arena so ethereal as superspace? Little astonishment there should be, therefore, if the description of nature carries one in the end to logic, the ethereal eyrie at the center of mathematics. If, as one believes, all mathematics reduces to the mathematics of logic, and all physics reduces to mathematics, what alternative is there but for all physics to reduce to the mathematics of logic? Logic is the only branch of mathematics that can "think about itself."

"An issue of logic having nothing to do with physics" was the assessment by many of a controversy of old about the axiom, "parallel lines never meet." Does it follow from the other axioms of Euclidean geometry or is it independent? "Independent," Bolyai and Lobachevsky proved. With this and the work of Gauss as a start, Riemann went on to create Riemannian geometry. Study nature, not Euclid, to find out about geometry, he advised; and Einstein went on to take that advice and to make geometry a part of physics.

"An issue of logic having nothing to do with physics" is one's natural first assessment of the startling limitations on logic discovered by Gödel (1931), Cohen (1966), and others [for a review, see, for example, Kac and Ulam (1968)]. The exact opposite must be one's assessment if the real pregeometry of the real physical world indeed turns out to be identical with the calculus of propositions.

"Physics as manifestation of logic" or "pregeometry as the calculus of propositions" is as yet [Wheeler (1971a)] not an idea, but an idea for an idea. It is put forward here only to make it a little clearer what it means to suggest that the order of progress may not be

$$\text{physics} \longrightarrow \text{pregeometry}$$

but

$$\text{pregeometry} \longrightarrow \text{physics.}$$

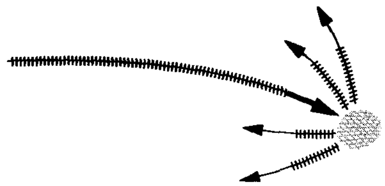


Figure 44.4.
The "black-box model" applied (1) to the scattering of an electron by a center of attraction and (2) to the collapse of the universe itself. The deterministic electron world line of classical theory is replaced in quantum theory by a probability amplitude, the wave crests of which are illustrated schematically in the diagram. The catastrophe of classical theory is replaced in quantum theory by a probability distribution of outputs. The same diagram illustrates the "black-box account" of gravitational collapse mentioned in the text. The arena of the diagram is no longer spacetime, but superspace. The incident arrow marks no longer a classical world line of an electron through spacetime, but a classical "leaf of history of geometry" slicing through superspace (Chapter 43). The wave crests symbolize no longer the electron wave function propagating through spacetime, but the geometrodynamic wave function propagating through superspace. The cross-hatched region is no longer the region where the one-body potential goes to infinity, but the region of gravitational collapse where the curvature of space goes to infinity. The outgoing waves describe no longer alternative directions for the new course of the scattered electron, but the beginnings of alternative new histories for the universe itself after collapse and "reprocessing" end the present cycle.

center of attraction and electron in a "black box:" fire in a wave-train of electrons traveling in one direction, and get electrons coming out in this, that, and the other direction with this, that, and the other well-determined probability amplitude (Figure 44.4). Moreover, to predict these probability amplitudes quantitatively and correctly, it was enough to translate the Hamiltonian of classical theory into the language of wave mechanics and solve the resulting wave equation, the key to the "black box."

Probability distribution of the outcomes of collapse

A similar "black box" view of gravitational collapse leads one to expect a "probability distribution of outcomes." Here, however, one outcome is distinguished from another, one must anticipate, not by a single parameter, such as the angle of scattering of the electron, but by many. They govern, one foresees, such quantities as the size of the system at its maximum of expansion, the time from the start of this new cycle to the moment it ends in collapse, the number of particles present, and a thousand other features. The "probabilities" of these outcomes will be governed by a dynamic law, analogous to (1) the Schrödinger wave equation for the electron, or, to cite another black box problem, (2) the Maxwell equations that couple together, at a wave-guide junction, electromagnetic waves running in otherwise separate wave guides. However, it is hardly reasonable to expect the necessary dynamic law to spring forth as soon as one translates the Hamilton-Jacobi equation of general relativity (Chapter 43) into a Schrödinger equation, simply because geometrodynamics, in both its classical and its quantum version, is built on standard differential geometry. That standard geometry leaves no room for any of those quantum fluctuations in connectivity that seem inescapable at small distances and therefore also inescapable in the final stages of gravitational collapse. Not geometry, but pregeometry, must fill the black box of gravitational collapse.

2

"Reprocessing" the universe

Little as one knows the internal machinery of the black box, one sees no escape from this picture of what goes on: the universe transforms, or transmutes, or transits, or is *reprocessed* probabilistically from one cycle of history to another in the era of collapse.

However straightforwardly and inescapably this picture of the reprocessing of the universe would seem to follow from the leading features of general relativity and the quantum principle, the two overarching principles of twentieth-century physics, it is nevertheless fantastic to contemplate. How can the dynamics of a system so incredibly gigantic be switched, and switched at the whim of probability, from one cycle that has lasted 10^{11} years to one that will last only 10^{6} years? At first, only the circumstance that the system gets squeezed down in the course of this dynamics to incredibly small distances reconciles one to a transformation otherwise so unbelievable. Then one looks at the upended strata of a mountain slope, or a bird not seen before, and marvels that the whole universe is incredible:

mutation of a species,
metamorphosis of a rock,
chemical transformation,
spontaneous transformation of a nucleus,
radioactive decay of a particle,
reprocessing of the universe itself.

If it cast a new light on geology to know that rocks can be raised and lowered thousands of meters and hundreds of degrees, what does it mean for physics to think of the universe as being from time to time "squeezed through a knothole," drastically "reprocessed," and started out on a fresh dynamic cycle? Three considerations above all press themselves on one's attention, prefigured in these compressed phrases:

destruction of all constants of motion in collapse;
particles, and the physical "constants" themselves, as the
  "frozen-in part of the meteorology of collapse;"
"the biological selection of physical constants."

All conservation laws transcended in the collapse of the universe

The gravitational collapse of a star, or a collection of stars, to a black hole extinguishes all details of the system (see Chapters 32 and 33) except mass and charge and angular momentum. Whether made of matter or antimatter or radiation, whether endowed with much entropy or little entropy, whether in smooth motion or chaotic turbulence, the collapsing system ends up as seen from outside, according to all indications, in the same standard state. The laws of conservation of baryon number and lepton number are transcended [Chapter 33; also Wheeler (1971b)]. No known means whatsoever will distinguish between black holes of the most different provenance if only they have the same mass, charge, and angular momentum. But for a closed universe, even these constants vanish from the scene. Total charge is automatically zero because lines of force have nowhere to end except upon charge. Total mass and total angular momentum have absolutely no definable meaning whatsoever for a closed universe. This conclusion follows not least because there is no asymptotically flat space outside where one can put a test particle into Keplerian orbit to determine period and precession.

Of all principles of physics, the laws of conservation of charge, lepton number, baryon number, mass, and angular momentum are among the most firmly established. Yet with gravitational collapse the content of these conservation laws also collapses. The established is disestablished. No determinant of motion does one see left that could continue unchanged in value from cycle to cycle of the universe. Moreover, if particles are dynamic in construction, and if the spectrum of particle masses is therefore dynamic in origin, no option would seem left except to conclude that the mass spectrum is itself reprocessed at the time when "the universe is squeezed through a knot hole." A molecule in this piece of paper is a "fossil" from photochemical synthesis in a tree a few years ago. A nucleus of the oxygen in this air is a fossil from thermonuclear combustion at a much higher temperature in a star a few 10^9 years ago. What else can a particle be but a fossil from the most violent event of all, gravitational collapse?

Three hierarchies of fossils: molecules, nuclei, particles

That one geological stratum has one many-miles long slope, with marvelous linearity of structure, and another stratum has another slope, is either an everyday triteness, taken as for granted by every passerby, or a miracle, until one understands the mechanism. That an electron here has the same mass as an electron there is also a triviality or a miracle. It is a triviality in quantum electrodynamics because it is assumed rather than derived. However, it is a miracle on any view that regards the universe as being from time to time "reprocessed." How can electrons at different times and places in the present cycle of the universe have the same mass if the spectrum of particle masses differs between one cycle of the universe and another?

Inspect the interior of a particle of one type, and magnify it up enormously, and in that interior see one view of the whole universe [compare the concept of monad of Leibniz (1714), "The monads have no window through which anything can enter or depart"]; and do likewise for another particle of the same type. Are particles of the same pattern identical in any one cycle of the universe because they give identically patterned views of the same universe? No acceptable explanation for the miraculous identity of particles of the same type has ever been put forward. That identity must be regarded, not as a triviality, but as a central mystery of physics.

Reason for identity in mass of particles of the same species?

Not the spectrum of particle masses alone, but the physical "constants" themselves, would seem most reasonably regarded as reprocessed from one cycle to another. Reprocessed relative to what? Relative, for example, to the Planck system of units,

Reprocessing of physical constants

$$L^* = (\hbar G/c^3)^{1/2} = 1.6 \times 10^{-33}\,\mathrm{cm},$$

$$T^* = (\hbar G/c^5)^{1/2} = 5.4 \times 10^{-44}\,\mathrm{sec},$$

$$M^* = (\hbar c/G)^{1/2} = 2.2 \times 10^{-5}\,\mathrm{g},$$

the only system of units, Planck (1899) pointed out, free, like black-body radiation itself, of all complications of solid-state physics, molecular binding, atomic constitution, and elementary particle structure, and drawing for its background only on the simplest and most universal principles of physics, the laws of gravitation and black-body radiation. Relative to the Planck units, every constant in every other part of physics is expressed as a pure number.

No pure numbers in physics are more impressive than $\hbar c/e^2 = 137.0360$ and the so-called "big numbers" [Eddington (1931, 1936, 1946); Dirac (1937, 1938); Jordan (1955, 1959); Dicke (1959b, 1961, 1964b); Hayakawa (1965a,b); Carter (1968b)]:

$$\sim 10^{80} \text{ particles in the universe,}^*$$

$$\sim 10^{40} \sim \frac{10^{28}\text{ cm}}{10^{-12}\text{ cm}} \sim \frac{\begin{pmatrix}\text{radius of universe at}\\ \text{maximum expansion}\end{pmatrix}^*}{\begin{pmatrix}\text{"size" of an elementary}\\ \text{particle}\end{pmatrix}},$$

$$\sim 10^{40} \sim \frac{e^2}{GmM} \sim \frac{(\text{electric forces})}{(\text{gravitational forces})},$$

$$\sim 10^{20} \sim \frac{e^2/mc^2}{(\hbar G/c^3)^{1/2}} \sim \frac{\begin{pmatrix}\text{"size" of an elementary}\\ \text{particle}\end{pmatrix}}{(\text{Planck length})},$$

$$\sim 10^{10} \sim \frac{\begin{pmatrix}\text{number of photons}\\ \text{in universe}\end{pmatrix}}{\begin{pmatrix}\text{number of baryons}\\ \text{in universe}\end{pmatrix}}.$$

Some understanding of the relationships between these numbers has been won [Carter (1968b)]. Never has any explanation appeared for their enormous magnitude, nor will there ever, if the view is correct that reprocessing the universe reprocesses also the physical constants. These constants on that view are not part of the laws of physics. They are part of the initial-value data. Such numbers are freshly given for each fresh cycle of expansion of the universe. To look for a physical explanation for the "big numbers" would thus seem to be looking for the right answer to the wrong question.

In the week between one storm and the next, most features of the weather are ever-changing, but some special patterns of the wind last the week. If the term "frozen features of the meteorology" is appropriate for them, much more so would it seem appropriate for the big numbers, the physical constants and the spectrum of particle masses in the cycle between one reprocessing of the universe and another.

Values of physical constants as related to the possibilities for life

A per cent or so change one way in one of the "constants," $\hbar c/e^2$, will cause all stars to be red stars; and a comparable change the other way will make all stars be blue stars, according to Carter (1968b). In neither case will any star like the sun be possible. He raises the question whether life could have developed if the determinants of the physical constants had differed substantially from those that characterize this cycle of the universe.

Dicke (1961) has pointed out that the right order of ideas may not be, here is the universe, so what must man be; but here is man, so what must the universe

*Values based on the "typical cosmological model" of Box 27.4; subject to much uncertainty, in the present state of astrophysical distance determinations, not least because the latitude in these numbers is even enough to be compatible with an open universe.

be? In other words: (1) What good is a universe without awareness of that universe? But: (2) Awareness demands life. (3) Life demands the presence of elements heavier than hydrogen. (4) The production of heavy elements demands thermonuclear combustion. (5) Thermonuclear combustion normally requires several 10^9 years of cooking time in a star. (6) Several 10^9 years of time will not and cannot be available in a closed universe, according to general relativity, unless the radius-at-maximum-expansion of that universe is several 10^9 light years or more. So why on this view is the universe as big as it is? Because only so can man be here!

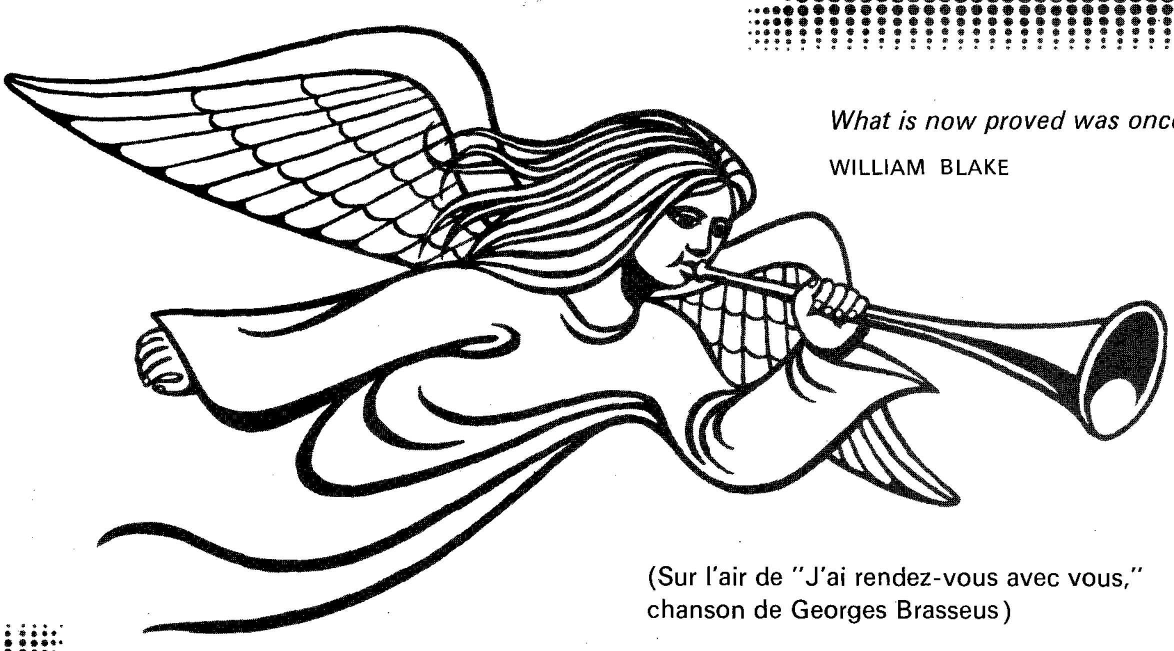
In brief, the considerations of Carter and Dicke would seem to raise the idea of the "biological selection of physical constants." However, to "select" is impossible unless there are options to select between. Exactly such options would seem for the first time to be held out by the only over-all picture of the gravitational collapse of the universe that one sees how to put forward today, the *pregeometry black-box model of the reprocessing of the universe.*

Proceeding with all caution into uncharted territory, one must nevertheless be aware that the conclusions one is reaching and the questions one is asking at a given stage of the analysis may be only stepping stones on the way to still more penetrating questions and an even more remarkable picture. To speak of "reprocessing and selection" may only be a halfway point on the road toward thinking of the universe as Leibniz did, as a world of relationships, not a world of machinery. Far from being brought into its present condition by "reprocessing" from earlier cycles, may the universe in some strange sense be "brought into being" by the participation of those who participate? On this view the concept of "cycles" would even seem to be altogether wrong. Instead the vital act is the act of participation. "Participator" is the incontrovertible new concept given by quantum mechanics; it strikes down the term "observer" of classical theory, the man who stands safely behind the thick glass wall and watches what goes on without taking part. It can't be done, quantum mechanics says. Even with the lowly electron one must participate before one can give any meaning whatsoever to its position or its momentum. Is this firmly established result the tiny tip of a giant iceberg? Does the universe also derive its meaning from "participation"? Are we destined to return to the great concept of Leibniz, of "preestablished harmony" ("Leibniz logic loop"), before we can make the next great advance?

Black hole as "laboratory" model for collapse of universe

Rich prospects stand open for investigation in gravitation physics, from neutron stars to cosmology and from post-Newtonian celestial mechanics to gravitational waves. Einstein's geometrodynamics exposes itself to destruction on a dozen fronts and by a thousand predictions. No predictions subject to early test are more entrancing than those on the formation and properties of a black hole, "laboratory model" for some of what is predicted for the universe itself. No field is more pregnant with the future than gravitational collapse. No more revolutionary views of man and the universe has one ever been driven to consider seriously than those that come out of pondering the paradox of collapse, the greatest crisis of physics of all time.

> *All of these endeavors are based on the belief that existence should have a completely harmonious structure. Today we have less ground than ever before for allowing ourselves to be forced away from this wonderful belief.*
>
> EINSTEIN (1934)



What is now proved was once only imagin'd.

WILLIAM BLAKE

We will first understand
How simple the universe is
When we realize
How strange it is.

ANON.

To some one who could grasp the
universe from a unified standpoint,
the entire creation would appear
as a unique truth and necessity.

J. D'ALEMBERT

Yo ho, it's hot . . . the sun is not
A place where we could live
But here on earth there'd be no life
Without the light it gives

H. ZARET

Probable-Possible, my black hen,
She lays eggs in the Relative When.
She doesn't lay eggs in the Positive Now
Because she's unable to postulate How.

F. WINSOR

(Sur l'air de "J'ai rendez-vous avec vous,"
chanson de Georges Brasseus)

Le Rayonnement dipolaire
On sait qu'il n'est pas pour nous
C'est pour Maxwell, oui mais Maxwell on s'en fout
Tout est r'latif après tout

Un argument qu'on révère
Celui de Synge pour dire le tout
Nous promet le quadrupolaire
Tout est r'latif après tout

Les sources quasi stellaires
Disparaissent comme dans un trou
Dans le Schwarzschild, oui mais Schwarzschild on s'en fout
Tout est r'latif après tout

Aux solutions singulières
On préfère et de beaucoup
Une métrique partout régulière
Tout est r'latif après tout

Les physiciens nucléaires
Comme ils nous aiment pas beaucoup
Y gardent tout l'fric, oui mais le fric on s'en fout
Tout est r'latif après tout

Les expériences de Weber
Le gyroscope, ça coûtent des sous
Celles de pensées sont moins chères
Tout est r'latif après tout

M. A. TONNELAT

"Omnibus ex nihil ducendis sufficit unum!"
(One suffices to create Everything of nothing!)

GOTTFRIED WILHELM VON LEIBNIZ

(Sur l'air de Auprès de ma blonde)

Dans les jardins d'Asnières
La science a refleuri
Tous les savants du monde
Apportent leurs écrits

Refrain:

Auprès de nos ondes
Qu'il fait bon, fait bon, fait bon
Auprès de nos ondes
Qu'il fait bon rêver

Tous les savants du monde
Apportent leurs écrits
Loi gravitationnelle
Sans tenseur d'énergie

Loi gravitationnelle
Sans tenseur d'énergie
De ravissants modèles
Pour la cosmologie

De ravissants modèles
Pour la cosmologie
Pour moi ne m'en faut guère
Car j'en ai un joli

Pour moi ne m'en faut guère
Car j'en ai un joli
Il est dans ma cervelle
Voici mon manuscrit

Le champ laisse des plumes
Aux bosses de l'espace-temps
En prendrons quelques unes
Pour décrire le mouvement

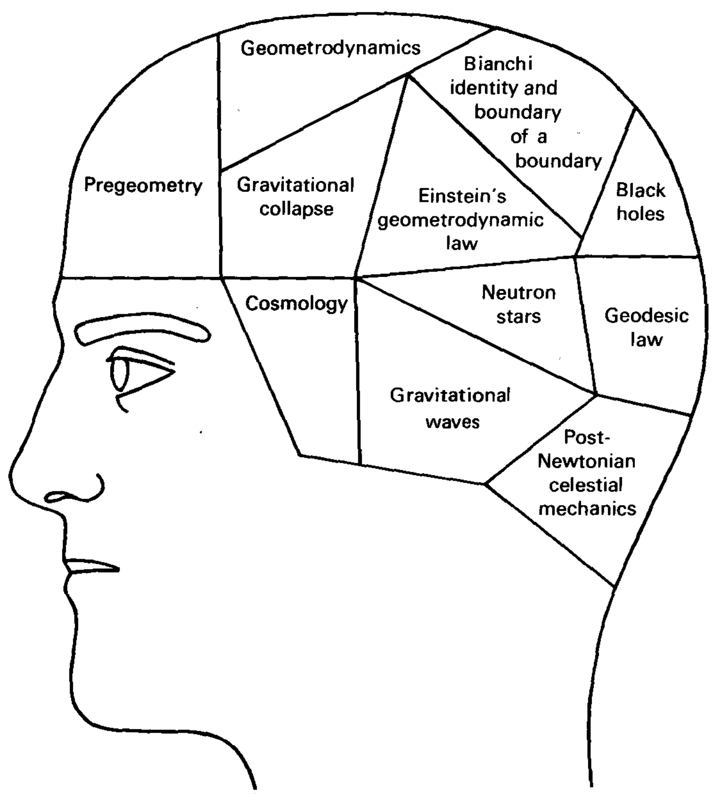
C. CATTANEO, J. GÉHÉNIAU
M. MAVRIDES, and M. A. TONNELAT

When Arthur Evans began this excavation
neither he nor anyone knew that he
would uncover an unknown world.

PAT (Mrs. Hypatia Vourloumis
at Knossos (1971))

And as imagination bodies forth
The form of things unknown, the poet's pen
Turns them to shapes, and gives to airy nothing
A local habitation and a name.

SHAKESPEARE



Appreciation and farewell to our patient reader.

Charles W. Misner Kip S. Thorne John A. Wheeler

BIBLIOGRAPHY AND INDEX OF NAMES

This bibliography, like that of Synge (1960b), merely provides points of entry into the literature, which is far too extensive for a comprehensive listing. For more extensive bibliographies of certain segments of the literature, see, for example, Lecat (1924), Boni, Russ, and Laurence (1960), Chick (1960, 1964), and Combridge (1965). Citations are sometimes not to the earliest publication, but to a later and more accessible source. For ease of reference, conference and summer-school proceedings are cited both under editors' names and under the more familiar place names (Brandeis, Les Houches, Varenna, etc.). Persons mentioned in the text without explicit bibliographical reference are also indexed here. Most doctoral dissertations in the United States are available from University Microfilms, Inc., Ann Arbor, Mich. 48106. Appreciation is expressed to Gregory Cherlin for preparing the initial version of this bibliography, to Nigel Coote for numerous subsequent amendments, and to colleagues without whose help some of the most elusive, and most important, of these references would have escaped capture.

Adler, R., M. Bazin, and M. Schiffer, 1965, *Introduction to General Relativity,* McGraw-Hill, New York. [endpapers]

Aharonov, Y., and L. Susskind, 1967, "Observability of the sign change of spinors under 2π rotations," *Phys. Rev. 158,* 1237–1238. [§41.5]

Ahrens, L. H., F. Press, and S. K. Runcorn, eds., 1964, *Physics and Chemistry of the Earth,* Pergamon Press, London. *Cited under* Bolt 1964.

al Bīrūnī, Abū Raihān, *cited by name in* Box 27.7.

Aldrin, E. E., Jr., *cited by name in* Fig. 40.4.

Alfvén, H., 1971, "Plasma physics applied to cosmology," *Phys. Today 24* (Feb. 1971), 28–33. [§§27.10, 27.11, 28.4]

Alfvén, H., and O. Klein, 1962, "Matter-antimatter annihilation and cosmology," *Ark. Fys. 23,* 187–194. [§§27.10, 27.11, 28.4, Box 27.7]

al Khazini, 1115, *Book of the Balance of Wisdom,* partially edited and translated by N. Khanikoff in *J. Am. Oriental Soc. 6* (1860), 25–38. [§1.7]

Allen, C. W., 1963, *Astrophysical Quantities,* University of London, Athlone Press, London. [§36.6, Box 25.1, endpapers]

Allendoerfer, C., and A. Weil, 1943, "The Gauss-Bonnet theorem for Riemannian polyhedra," *Trans. Am. Math. Soc. 53,* 101–129. [Box 15.3]

Alley, C. O., *cited by name in* Fig. 40.4. *See also* Bender *et al.* 1971.

Alpher, R. A., *cited by name in* Box 27.7.

Alpher, R. A., and R. C. Herman 1948a, "Evolution of the universe," *Nature 162,* 774–775. [§28.1]

Alpher, R. A., and R. C. Herman, 1948b, "On the relative abundance of the elements," *Phys. Rev. 74,* 1737–1742. [§28.1]

Alpher, R. A., and R. C. Herman, 1950, "Theory of the origin and relative-abundance distribution of the elements," *Rev. Mod. Phys. 22,* 153–213. [§28.1]

American Ephemeris and Nautical Almanac for 1964, 1962, U.S. Government Printing Office, Washington, D.C. [Fig. 25.1]

Ames, W. L., and K. S. Thorne, 1968, "The optical appearance of a star that is collapsing through its gravitational radius," *Astrophys. J. 151,* 659–670. [§32.3]

Anderson, C. D., 1933, "The positive electron," *Phys. Rev. 43,* 491–494. [§44.4]

Anderson, D. W., E. H. Brown, Jr., and F. P. Peterson, 1966a, "Spin cobordism," *Bull. Am. Math. Soc. 72,* 256–260. [Box 44.3]

Anderson, D. W., E. H. Brown, Jr., and F. P. Peterson, 1966b, "SU-cobordism, KO-characteristic numbers, and the Kervaire invariant," *Ann. Math. 83,* 54. [Box 44.3]

Anderson, J. D., 1973, "Lectures on physical and technical problems posed by precision radio tracking," in Bertotti 1973. [§40.1]

Anderson, J. D., P. B. Esposito, W. L. Martin, and D. O. Muhleman, 1971, "Measurements of the general relativistic time delay with data from Mariners 6 and 7," in Davies 1971, pp. 111–135. [Box 40.2]

Anderson, J. L., 1967, *Principles of Relativity Physics,* Academic Press, New York. [§17.6, endpapers]

Anderson, J. L., 1970, "Relativistic Boltzmann theory and the Grad method of moments," in Carmeli, Fickler, and Witten 1970, pp. 109–124. [Ex. 22.7]

Anonymous, source of quotation at end of Chapter 44: found on beach at Rincon, Puerto Rico, March 21, 1973.

Aristotle, *ca.* 330 B.C., *Physics,* ed. by W. D. Ross, Clarendon Press, Oxford (1950). [Box 1.2] *See also* Guthrie 1939, Warrington 1956.

Armstrong, N. A., *cited by name in* Fig. 40.4.

Arnett, W. D., 1966, "Gravitational collapse and weak interactions," *Can. J. Phys. 44,* 2553–2594. [§32.7]

Arnett, W. D., 1967, "Mass dependence in gravitational collapse of stellar cores," *Can. J. Phys. 45,* 1621–1641. [§32.7]

Arnold, V. I., and A. Avez. 1968, *Ergodic Problems of Classical Mechanics,* W. A. Benjamin, New York. [§8.2]

Arnowitt, R., *cited by name in* §§21.1, 35.8, Box 17.2.

Arnowitt, R., S. Deser, and C. W. Misner, 1962, "The dynamics of general relativity," in Witten 1962, pp. 227–265. [§§21.1, 21.4, 21.5, 21.6, 21.7, 21.8, 35.4]

Arp, H., 1971, "Observational paradoxes in extragalactic astronomy," *Science 174,* 1189–1200. [Box 28.1]

Ash, M. E. *See* Shapiro, Ash, *et. al.* 1971, Shapiro, Smith, *et. al.* 1971, Shapiro *et. al.* 1972.

Auslander, L., and R. E. MacKenzie, 1963, *Introduction to Differentiable Manifolds,* McGraw-Hill, New York. [Box 8.1]

Auslander, L., and L. Markus, 1959, *Flat Lorentz 3-Manifolds,* American Mathematical Society, Providence, R.I. [Box 27.2]

Avez, A., 1960, "Propriétés globales des espace-temps périodiques clos," *Acad. des Sci., Paris, Comptes Rend. 250,* 3583–3587. [§44.1]

Avez, A., 1964, "Le ds^2 de Schwarzschild parmi les ds^2 stationnaires," *Ann. Inst. Henri Poincaré A1,* 291–300. [§23.1]

Avez, A. *See also* Arnold and Avez 1968.

Avicenna, *cited by name in* Box 27.7.

Baade, W., 1942, "The Crab Nebula," *Astrophys. J. 96,* 188–198. [Box 24.1]

Baade, W., 1952, Report to Commission 28 of I.A.U., Rome, September 1952. [Box 27.1]

Baade, W., 1956. "The period-luminosity relation of the cepheids," *Pub. Astron. Soc. Pacific, 28,* 5–16. [Box 27.1]

Baade, W., *cited by name in* Box 27.7.

Baade, W., and F. Zwicky, 1934a, "Cosmic rays from supernovae," *Proc. Nat. Acad. Sci. U.S. 20,* 259–263. [Box 24.1]

Baade, W., and F. Zwicky, 1934b, "On supernovae," *Proc. Nat. Acad. Sci. U.S. 20,* 254–259. [Box 24.1]

Baade, W., and F. Zwicky, 1934c. "Supernovae and cosmic rays," *Phys. Rev. 45,* 138. Abstract of paper presented at Stanford meeting of Amer. Phys. Soc., Dec. 15–16, 1933. [Box 24.1]

Bacon, R., 1268, *Opus Majus,* J. H. Bridges, ed., Clarendon Press, Oxford, Eng. Vols. 1 and 2, 1897; vol. 3, 1900. [Box 27.7]

Bahcall, J. N., W. L. W. Sargent, and M. Schmidt, 1967, "An analysis of the absorption spectrum of 3C191," *Astrophys. J. 149,* 11–15. [§38.6]

Bahcall, J. N., and M. Schmidt, 1967, "Does the fine-structure constant vary with cosmic time?" *Phys. Rev. Lett. 19,* 1294–1295. [§38.6]

Baierlein, R. F., 1967, "Testing general relativity with laser ranging to the moon," *Phys. Rev. 162,* 1275–1288. [§40.6]

Baierlein, R. F., *cited by name in* §21.1. *See also* Krogh and Baierlein 1968.

Baierlein, R. F., D. H. Sharp, and J. A. Wheeler, 1962, "Three-dimensional geometry as carrier of information about time," *Phys. Rev. 126,* 1864–1865. [§§21.9, 43.2]

Bailey, J. *See* Farley *et al.* 1966.

Bailey, J., W. Bartl, G. von Bochmann, R. C. A. Brown, F. J. M. Farley, H. Jostlein, E. Picasso, and R. W. Williams, 1968, "Precision measurement of the anomalous magnetic moment of the muon," *Phys. Lett. 28B,* 287–289. [§44.2]

Bardeen, J. M., 1968, "General relativistic collapse of charged dust (abstract)," *Bull. Am. Phys. Soc. 13,* 41. [§34.6, Fig. 34.4]

Bardeen, J. M., 1970a, "Kerr metric black holes," *Nature 226,* 64–65. [§33.8, Boxes 33.3, 33.5]

Bardeen, J. M., 1970b, "A variational principle for rotating stars in general relativity," *Astrophys. J. 162,* 71–95. [Ex. 33.3]

Bardeen, J. M. *See also* Press and Bardeen 1971.

Bardeen, J. M., B. Carter, and S. W. Hawking, 1973, "The four laws of black-hole mechanics," *Commun. Math. Phys.,* in press. [Box 33.4]

Bardeen, J. M., and W. H. Press, 1972, "Radiation fields in the Schwarzschild background," *J. Math. Phys. 14,* 7–19. [§35.14, Ex. 32.11]

Bardeen, J. M., W. H. Press, and S. A. Teukolsky, 1972, "Rotating black holes: locally nonrotating frames, energy extraction, and scalar synchrotron radiation," *Astrophys. J. 178,* 347–369. [§33.7, Ex. 33.3, 33.14]

Bardeen, J. M., K. S. Thorne, and D. W. Meltzer, 1966, "A catalogue of methods for studying the normal modes of radial pulsation of general-relativistic stellar models," *Astrophys. J. 145,* 505–513. [Box 26.1]

Bardeen, J. M., and R. V. Wagoner, 1971, "Relativistic disks, I: Uniform rotation," *Astrophys. J. 167,* 359–423. [Box 24.1]

Barger, R. L. *See* Evenson *et. al.* 1972.

Bargmann, V., and E. P. Wigner, 1948, "Group theoretical discussion of relativistic wave equations," *Proc. Nat. Acad. Sci. U.S. 34,* 211–222. [Box 32.2]

Barnes, J. A., 1971, "A non-mathematical discussion of some basic concepts of precise time measurement," in *Tracor on Frequency 2,* No. 2 (Tracor Industrial Instruments, Austin, Texas). [Box 1.4]

Barschall, H. H. and W. Haeberli, 1971. "The Madison convention," pp. xxv–xxix in *Polarization Phenomena in Nuclear Reactions,* Univ. of Wisconsin Press, Madison. [§37.7]

Bartl, W. *See* Bailey *et al.* 1968.

Barton, D., S. R. Bourne, and J. P. Fitch, 1970, "An algebra system," *Comput. J. 13,* 32–39. [Box 14.3]

Barton, D., and J. P. Fitch, 1971, "General relativity and the application of algebraic manipulative systems," *Comm. Assoc. Comput. Machinery 14,* 542–547. [Box 14.3]

Battelle Rencontres, Seattle 1967. *See* DeWitt and Wheeler 1968.

Baym, G., 1969, *Lectures on Quantum Mechanics,* W. A. Benjamin, New York. [Box 22.4]

Baym, G., H. A. Bethe, and C. J. Pethick, 1971, "Neutron star matter," *Nuclear Phys A. 175,* 225–271. [§24.2]

Bazin, M. *See* Adler, Bazin, and Schiffer 1965.

Beall, E. F., 1970, "Measuring the gravitational interaction of elementary particles," *Phys. Rev. D 1,* 961–978. [§1.3, Box 1.2]

Beams, J. W., 1971, "Finding a better value for G," *Phys. Today 24* (May 1971), 35–40. [§40.8]

Beams, J. W. *See also* Rose *et al.* 1969.

Beckedorff, D. L., and C. W. Misner, 1962, unpublished; see D. L. Beckedorff's A. B. Senior Thesis, Princeton University. [§32.4]

Bekenstein, J., 1972a, "Nonexistence of baryon number for static black holes," *Phys. Rev. D 5,* 1239–1246. [Boxes 25.2, 32.2, 33.1]

Bekenstein, J., 1972b, "Nonexistence of baryon number for static black holes, II," *Phys. Rev. D 5,* 2403–2412. [Boxes 25.2, 32.2, 33.1]

Bel, L., 1958, "Sur la radiation gravitationelle," *Acad. Sci., Paris, Comptes Rend. 247,* 1094–1096. [Ex. 15.2]

Bel, L., 1959, "Introduction d'un tenseur du quatrième ordre," *Acad. Sci., Paris, Comptes Rend. 248,* 1297–1300. [Ex. 15.2]

Bel, L., 1962, "La radiation gravitationelle," in CNRS 1962, pp. 119–126. [Ex. 15.2]

Belasco, E. P., and H. C. Ohanian, 1969, "Initial conditions in general relativity: Lapse and shift formulation," *J. Math. Phys. 10,* 1503–1507. [§21.9]

Belinfante, F., 1940, "On the current and the density of the electric charge, the energy, the linear momentum, and the angular momentum of arbitrary fields," *Physica 7,* 449–474. [§21.3]

Belinsky, V. A. *See also* Ryan 1972b.

Belinsky, V. A., and I. M. Khalatnikov, 1969a, "On the nature of the singularities in the general solution of the gravitational equations," *Zh. Eksp. & Teor. Fiz. 56,* 1700–1712. English translation in *Sov. Phys.—JETP 29,* 911–917. [§34.6]

Belinsky, V. A., and I. M. Khalatnikov, 1969b, "General solution of the gravitational equations with a physical singularity," *Zh. Eksp. & Teor. Fiz. 57,* 2163–2175. English translation in *Sov. Phys.—JETP 30,* 1174–1180 (1970). [§34.6]

Belinsky, V. A., and I. M. Khalatnikov, 1970, "General solution of the gravitational equations with a physical oscillatory singularity," *Zh. Eksp. & Teor. Fiz. 59,* 314–321. English translation in *Sov. Phys.—JETP 32,* 169–172. [§§30.6, 34.6]

Belinsky, V. A., I. M. Khalatnikov, and E. M. Lifshitz, 1970, "Oscillatory approach to a singular point in the relativistic cosmology," *Usp. Fiz. Nauk 102,* 463–500. English translation in *Advances in Physics 19,* 525–573. [§30.6, Box 30.1]

Belinsky, V. A., E. M. Lifshitz, and I. M. Khalatnikov, 1971, "Oscillatory mode of approach to a singularity in homogeneous cosmological models with rotating axes," *Zh. Eksp. & Teor. Fiz. 60,* 1969–1979. English translation in *Sov. Phys.—JETP 33,* 1061–1066 (1971). [§§30.7, 34.6]

Bell, C. C., R. L. Forward, L. R. Miller, T. D. Beard, and T. M. Baran, 1969, "Static gravitational gradient field detection with a rotating torsional gradiometer," in NASL 1969. [Box 16.5]

Bell, C. C., R. L. Forward, and H. P. Williams, 1970, "Simulated terrain mapping with the rotating gravity gradiometer, in W. T. Kattner, ed., *Advances in Dynamic Gravimetry,* Instrument Society of America, Pittsburgh, 115–129. [Box 16.5]

Bell, S. J., *cited by name in* Box 27.7. *See also* Hewish *et al.* 1968.

Beltran-Lopez, V. B. *See* Hughes, Robinson, and Beltran-Lopez 1960.

Bender, P. L., R. H. Dicke, D. T. Wilkinson, C. O. Alley, D. G. Currie, J. E. Faller, J. D. Mulholland, E. C. Silverberg, H. E. Plotkin, W. M. Kaulla, and G. J. F. McDonald, 1971, "The lunar laser-ranging experiment," in Davies 1971. [§40.9]

Bennett, A. S., *cited by name in* Box 27.7.

Bentley, R., *cited by name in* Box 27.7.

Berger, B. K., D. M. Chitre, V. E. Moncrief, and Y. Nutku, 1972, "Hamiltonian formulation of spherically symmetric gravitational fields," *Phys. Rev. D 5,* 2467–2470. [§21.10]

Berger, M., 1966, "Sur les spectres d'une variété Riemannienne," *Acad. Sci Paris, Comptes Rendus 263,* 13–16. [Box 17.2 part 6]

Berger, M., P. Gauduchon, and E. Mazet, 1971, *Le Spectre d'une variété riemannienne,* Springer, Berlin and New York. [Box 17.2 part 6]

Bergmann, P. G., 1942, *Introduction to the Theory of Relativity,* Prentice-Hall, New York. [endpapers]

Bergmann, P. G., 1968, "Comments on the scalar-tensor theory," *Int. J. Theor. Phys. 1,* 25–36. [§38.2]

Bern, Switzerland, 1955, Jubilee of Relativity Theory, conference held July 11–16, 1955; *see* Mercier and Kervâire 1956.

Bernoulli, D., 1733, *Hydrodynamica,* chapter 10, Archives of Academy of Science, Leningrad; summarized in C. C. Gillispie, ed., *Dictionary of Scientific Biography,* Scribners, New York, vol. 2, 1970. [Box 44.5]

Bers, L., 1952, "Singularities of minimal surfaces," in *Proceedings of the International Congress of Mathematicians, Cambridge, Mass., 1950,* American Mathematical Society, Providence, R.I. [§33.2]

Bers, L., *cited by name in* Box 8.5.

Bertotti, B., 1959, "Uniform electromagnetic field in the theory of general relativity," *Phys. Rev. 116,* 1331–1333. [Ex. 32.1]

Bertotti, B., ed., 1973, proceedings of Course 56 of the International School of Physics Enrico Fermi, dealing with experimental tests of relativistic gravitation theories, Academic Press, New York, in press.

Bertotti, B., D. R. Brill, and R. Krotkov, 1962, "Experiments on gravitation," in Witten 1962. [§38.5, Box 40.1]

Bessel, F. W., *cited by name in* Box 27.7.

Bethe, H. A., *cited by name in* Box 27.7. *See also* Baym, Bethe, and Pethick 1971.

Birkhoff, G. D., 1923, *Relativity and Modern Physics,* Harvard University Press, Cambridge, Mass. [§32.2]

Birkhoff, G. D., 1943, "Matter, electricity and gravitation in flat spacetime," *Proc. Nat. Acad. Sci. U.S. 29,* 231–239. [§39.1]

Bishop, R. L., and S. I. Goldberg, 1968, *Tensor Analysis on Manifolds,* Macmillan, New York. [Box 8.1, Ex. 21.8]

Bisnovatyi-Kogan, G. S., and K. S. Thorne, 1970, "Relativistic gas spheres and clusters of point masses with arbitrarily large central redshifts: Can they be stable?" *Astrophys. J. 160,* 875–885. [Ex. 25.28]

Bisnovatyi-Kogan, G. S., and Ya. B. Zel'dovich, 1969, "Models of clusters of point masses with great central redshift," *Astrofizika 5,* 223–234. English translation in *Astrophys. 5,* 105–109. [Ex. 25.28]

Blackman, R. B., and J. W. Tukey, 1959, *The Measurement of Power Spectra, from the Point of View of Communications Engineering,* Dover, New York. [§37.8]

Blake, W., *quoted at end of* Chapter 44

Blaton, J., 1950, "On a geometrical interpretation of energy and momentum conservation in atomic collisions and disintegration processes," *Kgl. Danske Videnskab. Sels., Mat.-fys. Medd. 24,* no. 20. [§2.10]

Blatt, J. M., and V. F. Weisskopf, 1952, *Theoretical Nuclear Physics,* Wiley, New York. [§37.7]

Blumenthal, L. M., 1953, *Theory and Applications of Distance Geometry,* Clarendon Press, Oxford, England. [Box 13.1]

Bochmann, G. von. *See* Bailey *et al.* 1968.

Bogoliubov, N. N., and D. V. Shirkov, 1959, *Introduction to the Theory of Quantized Fields,* Interscience (now a division of Wiley), New York. [§43.3]

Bogorodskii, A. F., 1962, *Uravneniya Polya Einsteina i ikh Primeneniye v Astronomii* (*Einstein's Field Equations and Their Application in Astronomy*), Kiev Univ. Press, USSR. [Ex. 25.22]

Bohr, A., and B. R. Mottleson, 1969, *Nuclear Structure, Vol. I: Single-Particle Motion,* W. A. Benjamin, New York. [§23.4]

Bohr, N., 1918, "On the quantum theory of line spectra," *Mém. Acad. R. Sci. & Lett. Danemark 4,* no. 1, parts 1–2; part 1 reprinted in Van der Waerden 1967, pp. 95–137. [Box 17.1]

Bohr, N., *cited by name in* Preface, §§44.1, 44.5, Box 17.1

Bohr, N., and L. Rosenfeld, 1933, "Zur Frage der Messbarkeit der elektromagnetischen Feldgrössen," *Kgl. Danske Videnshab. Sels. Mat.-fys. Medd. 12,* no. 8. [§43.2]

Boissonade, J. F. *See* Westerman and Boissonade 1878.

Bolt, B. A., 1964, "Recent information of the earth's interior from studies of mantle waves and eigenvibrations," in Ahrens, Press, and Runcorn 1964, pp. 55–119. [Ex. 37.14]

Bolton, J. G., *cited by name in* Box 27.7.

Bondi, H., 1947, "Spherically Symmetrical Models in General Relativity," *Mon. Not. R. Astron. Soc. 107,* 410–425. [§30.6]

Bondi, H., 1957, "Plane gravitational waves in general relativity," *Nature 179,* 1072–1073. [Ex. 18.5, Fig. 18.1]

Bondi, H., 1961, *Cosmology,* 2d ed., Cambridge Univ. Press, Cambridge, Eng. [Box 27.1]

Bondi, H., 1965, "Some special solutions of the Einstein equations," in Trautman, Pirani, and Bondi 1965, pp. 375–459, esp. p. 431. [Ex. 18.5]

Bondi, H., *cited by name in* §21.10, Boxes 27.7, 34.2. *See also* Trautman, Pirani, and Bondi 1965.

Bondi, H., and T. Gold, 1948, "The steady-state theory of the expanding universe," *Mon. Not. R. Astron. Soc. 108,* 252–270. [§28.4]

Bondi, H., and W. H. McCrea, 1960, "Energy transfer by gravitation in Newtonian theory," *Proc. Cambridge Phil. Soc. 56,* 410–413. [Ex. 18.5]

Bondi, H., F. A. E. Pirani, and I. Robinson, 1959, "Gravitational waves in general relativity, III: Exact plane waves," *Proc. R. Soc. London A 251,* 519–533. [§35.9]

Boni, N., M. Russ, and D. H. Laurence, 1960, *A Bibliographical Check List and Index to the Published Writings of Albert Einstein,* Pageant Books, Paterson, N.J., cited in preface of present bibliography; based on earlier bibliography of M. C. Shields in Schilpp 1949.

Born, M., 1971, *The Born-Einstein letters: Correspondence Between Albert Einstein and Max and Hedwig Born from 1916 to 1955 with Commentaries by Max Born,* translated by Irene Born, Walker, New York. [Box 1.11]

Born, M., *cited by name in* Box 17.2

Boughn, S. P., D. M. Fram, and R. B. Partridge, 1971, "Isotropy of the microwave background at 8-millimeter wavelength," *Astrophys. J. 165,* 439–444. [§§27.1, 27.2, 30.8]

Bourbaki, N., *cited by name in* Box 8.2; pseudonym for an organization of French mathematicians who, since 1939, have nearly completed the first ten *Books* of a projected larger work: *Éléments de Mathématique,* Hermann, Paris.

Bourne, S. R. *See* Barton, Bourne, and Fitch 1970.

Boyer, R. H., and R. W. Lindquist, 1967, "Maximal analytic extension of the Kerr metric," *J. Math. Phys. 8,* 265–281. [§33.2, Box 33.2, Ex. 33.8]

Braginsky, V. B., 1965, "Gravitational radiation and the prospect of its experimental discovery," *Usp. Fiz. Nauk 86,* 433–446. English translation in *Sov. Phys.—Uspekhi 8,* 513–521. [Box 36.3]

Braginsky, V. B., 1970, *Fizicheskie Eksperimenti s Probnimi Telami* (*Physics Experiments With Test Bodies*), Nauka, Moscow, Russia. English translation published as NASA-TT F-762, available from National Technical Information Service, Springfield, Va. [§37.8]

Braginsky, V. B., A. B. Manukin, and M. Yu. Tikhonov, 1970, "Investigation of dissipative ponderomotive effects of electromagnetic radiation," *Zh. Eksp. & Teor. Fiz. 58,* 1549–1552, English translation in *Sov. Phys.—JETP 31,* 829–830. [§37.8]

Braginsky, V. B., and M. B. Menskii, 1971, "High-frequency detection of gravitational waves," *Zh. Eksp. & Teor. Fiz. Pis'ma 13,* 585. English translation in *Sov. Phys.—JETP Lett. 13,* 417–419. [§37.9, Box 37.6]

Braginsky, V. B., and V. S. Nazarenko, 1971, "On the detection of gravitational radiation from some astronomical sources with the heterodyne detector," in Davies 1971, pp. 45–46. [§37.3]

Braginsky, V. B., and V. I. Panov, 1971, "Verification of the equivalence of inertial and gravitational mass," *Zh. Eksp. & Teor. Fiz. 61,* 873–879. English translation in *Sov. Phys.—JETP 34,* 464–466. [§§38.3, 40.9, Box 1.2]

Braginsky, V. B., Ya. B. Zel'dovich, and V. N. Rudenko, 1969, "Reception of gravitational radiation of extraterrestrial origin," *Zh. Eksp. & Teor. Fiz. Pis'ma 10,* 437–441. English translation in *Sov. Phys.—JETP Lett. 10,* 280–283. [§37.3, Fig. 37.2]

Brandeis Summer Institute in Theoretical Physics: 1964, *see* Trautman, Pirani, and Bondi 1965; 1968, *see* Chrétian, Deser, and Goldstein 1969.

Brans, C., *cited by name in* §§40.1, 40.6, 40.9.

Brans, C., and R. H. Dicke, 1961, "Mach's principle and a relativistic theory of gravitation," *Phys. Rev. 124,* 925–935. Reprinted in Dicke 1964b. [§§17.6, 28.4, 38.2, 39.2, 40.8, Boxes 38.3, 39.1]

Brasseur, Georges, *quoted in* Fig. 44.5.

Brault, J. W., 1962, *The Gravitational Redshift in the Solar Spectrum,* doctoral dissertation, Princeton Univ. Abstract, "Gravitational Redshift of Solar Lines," in *Bull. Amer. Phys. Soc. 8,* 28 (1963). [§38.5, Fig. 38.2]

Brehme, R. W. *See* DeWitt and Brehme 1960.

Bressanone, Italy: 1970 summer school sponsored by Centro Interzionale Matematico Estivo. *See* Cattaneo 1971.

Brill, D. R., 1959, "On the positive definite mass of the Bondi-Weber-Wheeler time-symmetric gravitational waves," *Ann. Phys.* (*USA*) *7,* 466–483. [§21.10]

Brill, D. R., 1961, *General Relativity: Selected Topics of Current Interest,* report SUI61–4, Physics Department, State University of Iowa, Iowa City, Iowa (multilithed); summarized in DeWitt and DeWitt 1964b, pp. 431–432; published as Brill 1964. [§21.10]

Brill, D. R., 1964, "General relativity: selected topics of current interest," *Suppl. Nuovo Cimento 2,* 1–56. *See* Brill 1961.

Brill, D. R. *See also* Bertotti, Brill, and Krotkov 1962, Cohen and Brill 1967 *and* 1968, *and* Graves and Brill 1960.

Brill, D. R., P. L. Chrzanowski, C. M. Pereira, E. D. Fackerell, and J. R. Ipser, 1972, "Solution of the scalar wave equation in a Kerr background by separation of variables," *Phys. Rev. D 5,* 1913–1915. [Ex. 33.14]

Brill, D. R., and J. M. Cohen, 1966, "Rotating masses and their effect on inertial frames," *Phys. Rev. 143,* 1011–1015. [§§21.12, 40.7, Ex. 26.1]

Brill, D. R., and J. B. Hartle, 1964, "Method of self-consistent field in general relativity and its application to the gravitational geon," *Phys. Rev. B 135,* 271–278. [§§21.12, 35.13]

Brill, D. R., and J. A. Wheeler, 1957, "Interaction of neutrinos and gravitational fields," *Rev. Mod. Phys. 29,* 465–479. [§§27.10, 41.11]

Brillouin, L., 1970, *Relativity Reexamined,* Academic Press, New York. [Ex. 25.21]

Brodetsky, S. *See* Dobson and Brodetsky 1947.

Brodsky, S. J., and S. D. Drell, 1970, "The present status of quantum electrodynamics," in *Ann. Rev. Nuclear Sci. 20,* 147–194. [§44.2]

Bromberg, V. A., 1958, "Relativistic corrections in the theory of the motion of the moon," *Bull. Inst. Theor. Astron., Akad. Nauk SSSR 6,* 733–756. [§40.6]

Bronshtein, M. P., *cited by name in* §29.2.

Brown, E. H., Jr. *See* Anderson, Brown, and Peterson 1966a *and* 1966b.

Brown, R. C. A. *See* Bailey *et al.* 1968 *and* Farley *et al.* 1966.

Bruhat, Y., 1962, "The Cauchy problem," in Witten 1962. [§§21.11, 21.13]

Bruhat, Y. *See also* Choquet-Bruhat, Y., and Fourès-Bruhat, Y.

Bruno, G. *See* Singer 1950.

Bunsen, R. W., *cited by name in* Box 38.2.

Burbidge, E. M., *cited by name in* Box 27.7. *See also* Burbidge and Burbidge 1969a *and* 1969b *and* Hoyle *et al.* 1964.

Burbidge, G. R., *cited by name in* Box 27.7. *See also* Hoyle *et al.* 1964.

Burbidge, G. R., and E. M. Burbidge, 1969a, "Red-shifts of quasi-stellar objects and related extragalactic systems," *Nature 222,* 735–741. [Ex. 27.14]

Burbidge, G. R., and E. M. Burbidge, 1969b, "Quasi-stellar objects—a progress report," *Nature 224,* 21–24. [Ex. 27.14]

Burke, W. L., 1970, "Runaway solutions: remarks on the asymptotic theory of radiation damping," *Phys. Rev. A 2,* 1501–1505. [§20.6]

Burke, W. L., 1971, "Gravitational radiation damping of slowly moving systems calculated using matched asymptotic expansions," *J. Math. Phys. 12,* 402–418. [§36.8]

Burke, W. L., *cited by name in* §22.5.

Burnett, D. S. *See* Wasserburg and Burnett 1968.

Cairns, S. S., ed., 1965, *Differential and Combinatorial Topology,* Princeton Univ. Press, Princeton, N.J.

Cairns, S. S. *See also* Morse and Cairns 1969.

Callaway, J., 1953, "The equations of motion in Einstein's new unified field theory," *Phys. Rev. 92,* 1567–1570. [§20.6]

Callippus, *cited by name in* Box 27.7.

Caltech-JPL conference on experimental tests of gravitational theories, November 1970. *See* Davies 1971.

Cameron, A. G. W., 1970, "Neutron stars," *Ann. Rev. Astron. & Astrophys. 8,* 179–208. [§24.2]

Cameron, A. G. W., *See also* Langer and Cameron 1969, *and* Truran and Cameron 1971.

Cameron, A. G. W., and J. Truran, 1971, *work cited by* Peebles 1971. [Box 33.3]

Campbell, D. B. *See* Shapiro, Ash, *et al.* 1971 *and* Shapiro *et al.* 1972.

Campbell, L., and Garnett, W., 1882, *Life of James Clerk Maxwell,* Macmillan, London. [Box 4.6]

Campbell, W. W., and R. J. Trumpler, 1928, "Observations made with a pair of five-foot cameras on the light-deflections in the sun's gravitational field at the total solar eclipse of September 21, 1922." *Lick Obs. Bull. 13,* 130–160. [Box 40.1]

Carmeli, M., S. I. Fickler, and L. Witten, 1970, *Relativity,* Plenum, New York. [§40.7]

Carroll, L. *See* Dodgson, C. L.

Cartan, É., 1901, "Sur quelques quadratures dont l'élément différentiel contient des fonctions arbitraires," *Bull. Soc. Math. France 29,* 118–130. [Box 8.2]

Cartan, É., 1922a, "Sur les équations de la gravitation de Einstein," *J. Math. Pures Appl. 1,* 141–203. [§21.1]
Cartan, É., 1922b, "Sur une généralisation de la notion de courbure de Riemann et les espaces à torsion," *Acad. Sci., Paris, Comptes Rend. 174,* 593–595. [§§38.2, 39.2]
Cartan, É., 1923, "Sur les variétés à connexion affine et la théorie de la relativité generalisée (premiére partie)," *Ann. École Norm. Sup. 40,* 325–412. [§§8.2, 12.1, 12.2, 38.2, 39.2]
Cartan, É., 1924, "Sur les variétés à connexion affine et la théorie de la relativité generalisée (suite)," *Ann. École Norm. Sup. 41,* 1–25. [§§8.2, 8.3, 12.1]
Cartan, É., 1925, *La Géométrie des Espaces de Riemann,* Memorial des Sciences Mathématiques, Fascicule IX, Gauthier-Villars, Paris, France. [§14.5]
Cartan, É., 1928 and 1946, *Leçons sur la Géométrie des Espaces de Riemann,* Gauthier-Villars, Paris, France. [§§14.5, 15.3, Box 8.2, Ex. 21.4, endpapers]
Cartan, É., 1935, *La Méthode du Repère Mobile, la Théorie des Groupes Continus et les Espaces Generalisées,* Hermann, Paris, France. [§6.4]
Cartan, É., 1945, "Les systèmes différentiels extérieurs et leurs applications géométriques," *Exposés de Géométrie 14,* Hermann, Paris, France. [§4.1]
Cartan, É., 1946. *See* Cartan 1928.
Cartan, É., 1966, *Theory of Spinors,* M.I.T. Press, Cambridge, Mass. [§41.11]
Cartan, É., *cited by name in* §§9.1, 13.1, 13.4, 15.1, Boxes 10.1, 12.3, Ex. 12.7, 12.8.
Cartan, H., 1970, *Formes Différentiels,* Hermann, Paris, France. [§4.1]
Carter, B., 1966a, "Complete analytic extension of the symmetry axis of Kerr's solution of Einstein's equations," *Phys. Rev. 141,* 1242–1247. [Box 33.2]
Carter, B., 1966b, "The complete analytic extension of the Reissner-Nordstrom metric in the special case $e^2 = m^2$," *Phys. Lett. 21,* 423–424. [Fig. 34.4, Ex. 31.8, 34.3]
Carter, B., 1968a, "Global structure of the Kerr family of gravitational fields," *Phys. Rev. 174,* 1559–1571. [§33.5, Box 33.2, Ex. 33.7]
Carter, B., 1968b, "Large numbers in astrophysics and cosmology," unpublished preprint, Institute of Theoretical Astronomy, Cambridge, Eng. [§44.6]
Carter, B., 1968c, "Hamilton-Jacobi and Schrodinger separable solutions of Einstein's equations," *Commun. Math. Phys. 10,* 280–310. [§33.5]
Carter, B., 1970, "An axisymmetric black hole has only two degrees of freedom," *Phys. Rev. Lett. 26,* 331–333. [Box 33.1]
Carter, B., 1973, "Properties of the Kerr metric," *in* DeWitt and DeWitt 1973. [Box 33.1]
Carter, B., *cited by name in* Box 34.3. *See also* Bardeen, Carter, and Hawking 1973.
Cattaneo, C., ed., 1971, *Relativistic Fluid Dynamics,* proceedings of a summer school sponsored by Centro Internazionale Matematico Estivo, Bressanone, Italy, June 1970, Edizione Cremonese, Rome. [§22.3]
Cattaneo, C., *quoted at end of* Chap. 44.
Cauchy, A. L., *cited by name in* Box 17.2.
Cavendish, H., 1798, "Experiments to determine the density of the Earth" (read June 21, 1798), *Phil. Trans. R. Soc. London, Part II* (1798), pp. 469–526. [§40.8]
Caves, C. M., *cited by name in* §26.1.
Chadwick, J., 1932, "Possible existence of a neutron," *Nature 129,* 312. [Box 24.1]
Chandrasekhar, S., 1931a, "The density of white dwarf stars," *Phil. Mag. 11,* 592–596. [Box 24.1]
Chandrasekhar, S., 1931b, "The maximum mass of ideal white dwarfs," *Astrophys. J. 74,* 81–82. [Box 24.1]
Chandrasekhar, S., 1964a. "Dynamical instability of gaseous masses approaching the Schwarzschild limit in general relativity," *Phys. Rev. Lett. 12,* 114–116 and 437–438. [§26.1, Boxes 24.1, 26.2]
Chandrasekhar, S., 1964b, "The dynamical instability of gaseous masses approaching the Schwarzschild limit in general relativity," *Astrophys. J. 140,* 417–433. [§26.1, Boxes 24.1, 26.2]
Chandrasekhar, S., 1965a, "The post-Newtonian equations of hydrodynamics in general relativity," *Astrophys. J. 142,* 1488–1512. [§39.3, Box 39.3, Ex. 39.14]
Chandrasekhar, S., 1965b, "The post-Newtonian effects of general relativity on the equilibrium of uniformly rotating bodies, I: The Maclaurin spheroids and the virial theorem," *Astrophys. J. 142,* 1513–1518. [§39.3, Box 39.3]
Chandrasekhar, S., 1965c, "The stability of gaseous masses for radial and non-radial oscillations in the post-Newtonian approximation of general relativity," *Astrophys. J. 142,* 1519–1540. [§39.3, Box 26.2]
Chandrasekhar, S., 1968, *Hydrodynamic and Hydromagnetic Stability,* Clarendon Press, Oxford. See especially §108. [Box 32.3]
Chandrasekhar, S., and G. Contopolous, 1967, "On a post-Galilean transformation appropriate to the post-Newtonian theory of Einstein, Infeld, and Hoffmann," *Proc. R. Soc. London A 298,* 123–141. [§39.3, 39.9]
Chandrasekhar, S., and F. P. Esposito, 1970, "The $2\frac{1}{2}$-post-Newtonian equations of hydrodynamics and radiation reaction in general relativity," *Astrophys. J. 160,* 153–180. [§36.8, Box 39.3]
Chandrasekhar, S., and N. R. Lebovitz, 1968, "The pulsations and the dynamical stability of gaseous masses in uniform rotation," *Astrophys. J., 152,* 267–291. [Ex. 39.7]
Chandrasekhar, S., and Y. Nutku, 1969, "The second post-Newtonian equations of hydrodynamics in general relativity," *Astrophys. J. 158,* 55–80. [Box 39.3]
Chapel Hill 1957, Conference on the Role of Gravitation in Physics, Jan. 18–23, 1957. *See* DeWitt 1957.
Chase, D. M., 1954, "The equations of motion of charged test particles in general relativity," *Phys. Rev. 95,* 243–246. [§20.6]
Cherlin, G., *cited by name in introduction to this bibliography.*
Chern, S.-S., 1955, "On curvature and characteristic classes of a Riemann manifold," *Hamburg Abh. 20,* 117–126. [Box 15.3]
Chern, S.-S., 1960, *Geometric Structures on Manifolds,* colloquium lectures, Amer. Math. Soc., Providence, R.I. [§2.2]
Chern, S.-S., 1962, "Geometry of a quadratic differential

form," *J. Soc. Indust. & Appl. Math. 10,* 751–755, also in *H. P. Robertson January 27, 1903–August 26, 1961, In Memoriam,* Soc. Indust. and Applied Math., Philadelphia 1963, pp. 13–17. [Box 15.3]

Chern, S.-S., 1966, "The geometry of G-structures," *Bull. Am. Math. Soc. 72,* 167–219. [§2.2]

Chern, S.-S., 1971, "Differential geometry: its past and its future," in *Proceedings of the International Congress of Mathematicians, Nice, September 1970,* Gauthier-Villars, Paris, pp. 41–53. [§2.2]

Chern, S.-S., and C. Chevalley, 1952, "Élie Cartan and his mathematical work," *Bull. Am. Math. Soc. 58,* 217–250. [Box 8.2]

Chevalley, C. *See* Chern and Chevalley 1952.

Chick, H. J., 1960, *Theory of Gravitation 1920–Present (Bibliography),* Report LMS-2328 of Los Alamos Scientific Laboratory, Los Alamos, New Mexico, cited in preface of present bibliography.

Chick, H. J., 1964, *Theory of Gravitation 1959–1963 (Bibliography),* Report LMS-3099 of Los Alamos Scientific Laboratory, Los Alamos, New Mexico, cited in preface of present bibliography.

Chinnapared, R. *See* Newman *et al.* 1965.

Chitre, D. M., 1972a, *Investigations of the Vanishing of a Horizon for Bianchi Type IX (Mixmaster) Universe,* doctoral dissertation, Univ. of Maryland. [§30.8]

Chitre, D. M., 1972b, "High-frequency sound waves to eliminate a horizon in the mixmaster universe," *Phys. Rev. D 6,* 3390–3396. [§30.8]

Chitre, D. M. *See also* Berger *et al.* 1972.

Chiu, H.-Y., and W. F. Hoffman, eds., 1964, *Gravitation and Relativity,* W. A. Benjamin, New York.

Choquard, P., 1955, "Traitement semi-classique des forces générales dans la représentation de Feynman," *Helv. Phys. Acta 28,* 89–157. [Box 21.1, Fig. 13.2]

Choquet-Bruhat, Y., 1968a, *Géométrie différentielle et systèmes extérieurs,* Monographies Universitaires de Mathématiques no. 28, Dunod, Paris, France. [§4.1]

Choquet-Bruhat, Y., 1968b, "Espace-temps Einsteiniens généraux, chocs gravitationels," *Ann. Inst. Henri Poincaré 8,* 327–338. [§21.13]

Choquet-Bruhat, Y., 1969, "Construction de solutions radiatives approchées des équations d'Einstein," *Commun. Math. Phys. 12,* 16–35. [§35.13]

Choquet-Bruhat, Y., 1972, "Problème des contraintes sur une variété compacte," *Acad. Sci. Paris, Comptes Rend. A 274,* 682–684; fuller account in "Global solutions of the equations of constraint in general relativity, on open and closed manifolds," *Gen. Rel. & Grav.* 1973, in press. [§21.1]

Choquet-Bruhat, Y. *See also* Bruhat, Y., *and* Fourès-Bruhat, Y.

Chrétien, M., S. Deser, and J. Goldstein, eds., 1969, *Brandeis Summer Institute 1968, Astrophysics and General Relativity,* Gordon and Breach, New York.

Christodolou, D., 1970, "Reversible and irreversible transformations in black-hole physics," *Phys. Rev. Lett. 25,* 1596–1597. [§33.8, Box 33.4]

Christodolou, D., 1971, *Investigations in Gravitational Collapse and the Physics of Black Holes,* doctoral dissertation, Princeton Univ. [§35.8, Boxes 33.3, 33.4, Fig. 33.3]

Christodolou, D., and R. Ruffini, 1971, "Reversible transformations of a charged black hole," *Phys. Rev. D 4,* 3552–3555. [§33.8, Boxes 33.4, 33.5]

Chrzanowski, P. L., and C. W. Misner, 1973, "Geodesic synchrotron radiation in the Kerr geometry by the method of asymptotically factorized Green's functions," technical report of the Dept. of Physics and Astronomy, University of Maryland. [§33.8]

Chrzanowski, P. L. *See also* Brill *et al.* 1972.

Cincinnati 1969, Relativity Conference in The Midwest, June 2–6, 1969, Cincinnati, Ohio. *See* Carmeli, Fickler, and Witten 1970.

Clark, G. *See* Eddington and Clark 1938.

Clarke, C. J. S., 1970, "On the global isometric embedding of pseudo-Riemannian manifolds," *Proc. Roy. Soc. London A 314,* 417–428. [§23.8]

Clayton, D., 1968, *Principles of Stellar Evolution and Nucleosynthesis,* McGraw-Hill, New York. [§28.1]

Clemence, G. M., 1943, "The motion of Mercury, 1765–1937," *Astron. Papers Am. Ephemeris 11,* 1–221. [Box 40.3]

Clemence, G. M., 1947, "The relativity effect in planetary motions," *Rev. Mod. Phys. 19,* 361–364. [Box 40.3]

Clemens, R. *See also* Fletcher *et al.* 1967.

Clemens, R., and R. Matzner, 1967, Technical Report no. 635 of the Univ. of Maryland Department of Physics and Astronomy. [Box 14.3]

Clifford, W. K., 1879, *Lectures and Essays,* L. Stephen and F. Pollock, eds., Macmillan, London, I, 244 and 322. [§44.3]

Clifford, W. K., 1882, *Mathematical Papers,* R. Tucker, ed., Macmillan, London. [§44.3]

CNRS, 1962, *Les Théories relativistes de la gravitation,* Éditions du Centre National de la Recherche Scientifique, Paris, France. *Same as* Royaumont 1959.

Cocke, W. J., 1965, "A maximum-entropy principle in general relativity and the stability of fluid spheres," *Ann. Inst. Henri Poincaré A 2,* 283–306. [§26.1]

Cocke, W. J., H. J. Disney and D. J. Taylor, 1969, "Discovery of optical signals from pulsar NP0532," *Nature 221,* 525–527. [Box 24.1]

Codazzi, D., *cited by name in* §21.5.

Cohen, J. M., 1968, "Angular momentum and the Kerr metric," *J. Math. Phys. 9,* 905–906. [Ex. 26.1]

Cohen, J. M. *See also* Brill and Cohen 1966.

Cohen, J. M., and D. Brill, 1966, "Rotating masses and their effect on inertial frames," *Phys. Rev. 143,* 1011–1015. [§21.12]

Cohen, J. M., and D. R. Brill, 1968, "Further examples of 'Machian effects' of rotating bodies in general relativity," *Nuovo Cim. 56B,* 209–219. [Ex. 26.1]

Cohen, J. M., and R. M. Wald, 1971, "Point charge in the vicinity of a Schwarzschild black hole," *J. Math. Phys. 12,* 1845–1849. [Box 33.1]

Cohen, P., 1966, *Set Theory and the Continuum Hypothesis,* W. A. Benjamin, New York. [Box 44.5]

Cole, J., 1968, *Perturbation Methods in Applied Mathematics,* Ginn-Blaisdell, Waltham, Mass. [§22.5]

Coleman, C. J., 1971, "A simple relativistic theory of gravitation," *J. Phys. A 4,* 611–616. [§40.5]

Coleman, J. A., 1967, *Early Theories of the Universe,* New American Library, New York. [§1.7]

Colgate, S. A., 1967, "Stellar coalescence and the multiple-supernova interpretation of quasi-stellar sources," *Astrophys. J. 150,* 163–192. [§24.5, Box 33.3]

Colgate, S. A., and R. H. White, 1966, "The hydrodynamic behavior of supernovae explosions," *Astrophys. J. 143,* 626–681. [§32.7]

Collins, C. B., and S. W. Hawking, 1973, "Why is the universe isotropic?" *Astrophys. J. 180,* 317–334. [Box 34.3]

Collins, R. A., *cited by name in* Box 27.7. *See also* Hewish *et al.* 1968.

Combridge, J. T., 1965, *Bibliography of Relativity and Gravitation Theory 1921–1937,* Kings College, London, cited in preface of present bibliography.

Comte, A., 1835, *Cours de philosophie positive,* Paris. [Box 27.7].

Conklin, E. K., 1969, "Velocity of the Earth with respect to the cosmic background radiation," *Nature 222,* 971–972. [§30.8]

Contopolous, G. *See* Chandrasekhar and Contopolous 1967.

Coote, N., *cited by name in introduction to this bibliography.*

Copenhagen 1971, Sixth International Conference on General Relativity and Gravitation, held in Copenhagen, July 1971; portions of the proceedings appear as articles in *Gen. Rel. Grav.,* especially vol. 3, 1972.

Copernicus, N. *See* Dobson and Brodetsky 1947.

Corben, H. C., and P. Stehle, 1960, *Classical Mechanics,* 2d ed., Wiley, New York. [Ex. 33.7]

Cornell University, 1965, Summer Seminar on Applied Mathematics, arranged by the American Mathematical Society. Proceedings (vols. 8–10 in Lectures in Applied Mathematics series) published as Ehlers 1967.

Cornford, F. M., 1937, *Plato's Cosmology: The Timaeus of Plato,* Humanities Press, New York. [Box 27.7]

Corson, E. M., 1953, *Introduction to Tensors, Spinors, and Relativistic Wave-Equations,* Hafner, New York. [§21.3]

Couch, W. E. *See also* Newman *et al.* 1965.

Couch, W. E., and E. T. Newman, 1972, "Generalized Lienard-Wiechart fields," *J. Math. Phys. 13,* 929–931. [§36.1]

Couch, W. E., R. J. Torrence, A. I. Janis, and E. T. Newman, 1968, "Tail of a gravitational wave," *J. Math. Phys. 9,* 484–496. [§§35.8, 35.14]

Courant, R., 1937, "Plateau's problem and Dirichlet's principle," *Ann. Math. U.S. 38,* 679–724. [§33.2]

Coxeter, H. S. M., 1963, *Regular Polytopes,* 2d ed., Macmillan, New York. [§42.5, Box 27.2]

Curott, D. R., 1966, "Earth deceleration from ancient solar eclipses," *Astron. J. 71,* 264–269. [Fig. 1.8]

Currie, D. G., *cited by name in* Fig. 40.4. *See also* Bender *et al.* 1971.

Curtis, H. D., *cited by name in* Box 27.7.

Cusanus, N. *See* Heron 1954.

Dabbs, J. W. T., J. A. Harvey, D. Paya, and H. Horstmann, 1965, "Gravitational acceleration of free neutrons," *Phys. Rev. B 139,* 756–760. [§1.3]

D'Alembert, J., *quoted at end of* Chap. 44.

Danielson, B. L. *See* Evenson *et al.* 1972.

Dante Alighieri, *Paradiso,* translated by Philip H. Wicksteed, 1921, Dent, London, England. [Epigraph to Chap. 44]

Danziger, I. J., 1970, "The cosmic abundance of helium," *Ann. Rev. Astron. Astrophys., 8,* 161–178. [§28.1]

Darboux, G., 1941, *Leçons sur la Théorie Générale des Surfaces, Part I,* Gauthier-Villars, Paris, France. [§33.2]

Darmois, G., 1927, *Les équations de la gravitation einsteinienne,* Mémorial des sciences mathematiques XXV, Gauthier-Villars, Paris. [§21.13]

Darwin, C., 1959, "The gravity field of a particle, I," *Proc. R. Soc. London A 249,* 180–194. [§25.5, Fig. 25.7]

Darwin, C., 1961, "The gravity field of a particle, II," *Proc. R. Soc. London A 263,* 39–50. [§25.5, Fig. 25.7]

Datt, B., 1938, "Über eine Klasse von Lösungen der Gravitationsgleichungen der Relativität," *Z. Physik 108,* 314–321. [§30.6, Ex. 32.8]

Davenport, W. B., and W. L. Root, 1958, *An Introduction to the Theory of Random Signals and Noise.* McGraw Hill, New York, [§37.8]

Davies, R. W., ed., 1971, *Proceedings of the Conference on Experimental Tests of Gravitational Theories, November 11–13, 1970, California Institute of Technology,* J.P.L. Technical Memorandum, 33–499.

Davis, M., R. Ruffini, W. H. Press, and R. H. Price, 1971, "Gravitational radiation from a particle falling radially into a Schwarzschild black hole," *Phys. Rev. Lett. 27,* 1466–1469. [§§33.7, 36.5, Box 33.3, Fig. 36.2]

Davis, W. R., 1970, *Classical Theory of Particles and Fields and the Theory of Relativity,* Gordon and Breach, New York. [§21.3, endpapers]

Day, G. W. *See* Evenson *et al.* 1972.

de Broglie, M. *See* Einstein 1912.

Debye, P. J. W., *cited by name in* §22.5.

De Chéseaux, P. L., 1744, *Traité de la comète qui a paru en décembre 1743,* Paris; the relevant material is in the Appendix. [Box 27.7]

De Felice, F., 1968, "Equatorial geodesic motion in the gravitational field of a rotating source," *Nuovo Cimento B 57,* 351–388. [Box 33.5]

Defoe, D., 1722, *The Fortunes and Misfortunes of the Famous Moll Flanders,* reprinted by Fine Editions Press, Cleveland, 1953. [epigraph to Part X]

De Groot, S. R. *See* Kluitenberg and de Groot 1954.

De Hoffmann, F., and E. Teller, 1950, "Magneto-hydrodynamic shocks," *Phys. Rev. 80,* 692–703. [§22.2, Box 22.2]

De la Cruz, V., and W. Israel, 1967, "Gravitational bounce," *Nuovo Cimento A 51,* 744–760. [§34.6, Fig. 34.4]

De Lange, F. G. *See* Refsdal, Stabell, and de Lange 1967.

Dempster, J. R. H., 1963, *Feynman Quantization in Field Theory: Topology of Commutators and Functional Representation of States,* doctoral dissertation, Princeton University. [Box 21.1]

De Rham, G., 1955, *Variétés Différentiables,* Herman, Paris. [§§4.1, 22.4]

Deser, S., 1970, "Self-interaction and gauge invariance," *Gen. Rel. & Grav. 1,* 9–18. [§§7.1, 18.1, Box 17.2 part 5]

Deser, S., *cited by name in* §§21.1, 35.8, Box 17.2, part 2′. *See also* Arnowitt, Deser, and Misner 1962 *and* Chrétien, Deser, and Goldstein 1969.

Deser, S., and K. W. Ford, eds., 1965, *Lectures on General Relativity; 1964 Brandeis Summer Institute in Theoretical Physics I,* Prentice-Hall, Englewood Cliffs, N.J. *Same as* Trautman, Pirani, and Bondi 1965.

De Sitter, W., 1916, "On Einstein's theory of gravitation and its astronomical consequences," *Mon. Not. R. Astron. Soc. 77,* 155–184 and 481. [§§40.6, 40.7]

De Sitter, W., 1917a, "On the relativity of inertia: Remarks concerning Einstein's latest hypothesis," *Proc. Kon. Ned. Akad. Wet. 19,* 1217–1225. [§27.11]

De Sitter, W., 1917b, "On the curvature of space," *Proc. Kon. Ned. Akad. Wet. 20,* 229–243. [§27.11]

De Sitter, W., *cited by name in* §30.3, Box 27.7.

Deutsch, A. J., and W. B. Klemperer, eds., 1962, *Space Age Astronomy,* Academic Press, New York.

De Valera, E., *cited by name in* §41.1.

De Vaucouleurs, G., 1971, "The large-scale distribution of galaxies and clusters of galaxies," *Astron. Soc. Pacific Publ. 83,* 113–143. [§27.11]

DeWitt, B. S., 1953, "Pair production by a curved metric," *Phys. Rev. 90,* 357. [§30.5]

DeWitt, B. S., 1962, "The quantization of geometry," in Witten 1962. [Ex. 40.8]

DeWitt, B. S., 1967a, "Quantum theory of gravity, I," *Phys. Rev. 160,* 1113–1148. [§§21.1, 43.1, 43.3]

DeWitt, B. S., 1967b. "Quantum theory of gravity, II; The manifestly covariant theory." *Phys. Rev. 162,* 1195–1239. [§§43.1, 43.3]

DeWitt, B. S., 1967c, "Quantum theory of gravity, III; Application of the covariant theory," *Phys. Rev. 162,* 1239–1256. [Box 17.2 part 6h]

DeWitt, B. S., 1971, *Lectures on Relativity—Stanford, Fall 1971* (unpublished manuscript). [§36.7]

DeWitt, B. S., *cited by name in* Ex. 35.20. *See also* DeWitt and DeWitt 1964a, 1964b, and 1973.

DeWitt, B. S., and R. W. Brehme, 1960, "Radiation damping in a gravitational field," *Ann. Phys. (U.S.A.) 9,* 220–259. [§35.8, Ex. 35.14]

DeWitt, C., ed. 1957, *Conference on the Role of Gravitation in Physics, Proceedings,* W.A.D.C. Technical Report 57-216; ASTIA Document No. AD118180, Wright Air Development Center, Wright-Patterson Air Force Base, Ohio.

DeWitt, C. *See also* Morette, C.

DeWitt, C., and B. S. DeWitt, 1964a, "Falling charges," *Physics 1,* 3–20. [§35.8]

DeWitt, C., and B. S. DeWitt, eds., 1964b, *Relativity, Groups, and Topology,* Gordon and Breach, New York. [§38.2]

DeWitt, C., and B. S. DeWitt, eds., 1973, *Black Holes,* Proceedings of 1972 session of Ecole d'été de physique theorique, Gordon and Breach, New York. [§33.1]

DeWitt, C., E. Schatzman, and P. Veron, eds., 1967, *High-Energy Astrophysics, Vol. III,* Gordon and Breach, New York. [§24.4]

DeWitt, C., and J. A. Wheeler, eds., 1968, *Battelle Rencontres: 1967 Lectures in Mathematics and Physics,* W. A. Benjamin, New York.

Dicke, R. H., 1954, "Coherence in spontaneous radiation processes," *Phys. Rev. 93,* 99–110. [Box 38.3]

Dicke, R. H., 1957, "Coherence and the quantum," *J. Opt. Soc. Am. 47,* 527. [Box 38.3]

Dicke, R. H., 1959a, "Dirac's cosmology and the dating of meteorites," *Nature 183,* 170–171. [§38.6]

Dicke, R. H., 1959b, "New research on old gravitation," *Science 129,* 621–624. [§§38.6, 40.9, 44.6]

Dicke, R. H., 1961, "Dirac's cosmology and Mach's principle," *Nature 192,* 440–441. [§§38.6, 44.6, Box 38.3]

Dicke, R. H., 1962, "Mach's principle and invariance under transformation of units," *Phys. Rev. 125,* 2163–2167. Reprinted in Dicke 1964b. [Box 39.1]

Dicke, R. H., 1964a, "The sun's rotation and relativity," *Nature 202,* 432–435. [Box 38.3]

Dicke, R. H., 1964b, *The Theoretical Significance of Experimental Relativity,* Gordon and Breach, New York; also in DeWitt and DeWitt 1964a. [§§38.2, 38.7, 44.6, Box 40.1]

Dicke, R. H., 1964c, "Remarks on the observational basis of general relativity," in Chiu and Hoffman 1964, pp. 1–16. [§40.9]

Dicke, R. H., 1968, "Scalar-tensor gravitation and the cosmic fireball," *Astrophys. J., 152,* 1–24. [§28.4]

Dicke, R. H., 1969, *Gravitation and the Universe,* American Philosophical Society, Philadelphia. [§§38.6, 40.9]

Dicke, R. H., *cited by name in* §§10.1, 38.3, 38.5, 40.1, 40.6, 40.9, Box 27.7. *See also* Bender *et al.* 1971, Brans and Dicke 1961, *and* Roll, Krotkov, and Dicke 1964.

Dicke, R. H., and H. M. Goldenberg, 1967, "Solar oblateness and general relativity," *Phys. Rev. Lett. 18,* 313–316. [Boxes 38.3, 40.3]

Dicke, R. H., and P. J. E. Peebles, 1962a, "Significance of spatial isotropy," *Phys. Rev. 127,* 629–631. Reprinted in Dicke 1964b. [§38.7]

Dicke, R. H., and P. J. E. Peebles, 1962b, "Cosmology and the radioactive decay ages of terrestrial rocks and meteorites," *Phys. Rev. 128,* 2006–2011. [§38.6, Box 38.3]

Dicke, R. H., and P. J. E. Peebles, 1965, "Gravitation and space science," *Space Sci. Rev. 4,* 419–460. [§40.8]

Dicke, R. H., P. J. E. Peebles, P. G. Roll, and D. T. Wilkinson, 1965, "Cosmic-black-body radiation," *Astrophys. J. 142,* 414–419. [§§27.2, 28.1]

Dieter, N. H., E. E. Epstein, A. E. Lilley, and M. S. Roberts, 1962, "A radio and optical investigation of extragalactic redshifts," *Astron. J. 67,* 270–271. [§29.2]

Dieudonné, J., 1960, *Foundations of Modern Analysis,* Academic Press, New York. [§2.6]

Digges, T., *cited by name in* Box 27.7.

d'Inverno, R. A., 1969, "ALAM—Atlas Lisp algebraic manipulator," *Comput. J. 12,* 124–127. [Box 14.3]

Dirac, P. A. M., 1926, "On the theory of quantum mechanics," *Proc. R. Soc. London A 112,* 661–677. [Box 24.1]

Dirac, P. A. M., 1936, "Relativistic wave equations," *Proc. R. Soc. London A 155,* 447–459. [Box 32.2]

Dirac, P. A. M., 1937, "The cosmological constants," *Nature 139,* 323. [§§38.6, 44.6]

Dirac, P. A. M., 1938, "New basis for cosmology," *Proc. R. Soc. London A 165,* 199–208. [§§38.6, 44.6]

Dirac, P. A. M., 1959, "Fixation of coordinates in the Hamiltonian theory of gravitation," *Phys. Rev. 114,* 924–930. [§21.7]

Dirac, P. A. M. 1964, *Lectures on Quantum Mechanics,* Belfer Graduate School of Science Monograph Series Number Two, Yeshiva University, New York. [§21.7]

Dirac, Paul A. M., *Festschrift* on the occasion of his 70th birthday, Aug. 8, 1972. *See* Salam and Wigner 1972.

Dirac, P. A. M., *cited by name in* §§20.6, 21.1, Box 4.4.

Dirac, P. A. M., V. A. Fock, and B. Podolsky, 1932, "On quantum electrodynamics," *Phys. Z. Sowjetunion 2,* 468–479. [§21.8]

Dobson, J. F., and S. Brodetsky, 1947, *Nicolaus Copernicus's De Revolutionibus* (translation of Preface and Book I), Royal Astronomical Society, London. [Box 27.7].

Dodgson, C. L., 1871, *Through the Looking Glass,* Macmillan, London. [Epigraph to Chap. 32]

Doroshkevich, A. G., 1966, "Change of the angular momentum of a star during accretion," *Astron. Zhur. 43,* 105–107. Terrible English translation in *Sov. Astron.—A. J. 10,* 83–84. [Box 33.3]

Doroshkevich, A. G., V. N. Lukash, and I. D. Novikov, 1971, "Impossibility of mixing in the Bianchi type IX cosmological model," *Zh. Eksp. & Teor. Fiz. 60,* 1201–1205. English translation in *Sov. Phys.—JETP 33,* 649–651. [§30.8]

Doroshkevich, A. G., Ya. B. Zel'dovich, and I. D. Novikov, 1965, "Gravitational collapse of nonsymmetric and rotating masses," *Zh. Eksp. & Teor. Fiz. 49,* 170–181. English translation in *Sov. Phys.—JETP 22,* 122–130. [Box 33.3, Ex. 26.1]

Douglas, K. N., I. Robinson, A. Schild, E. L. Schücking, J. A. Wheeler, and N. J. Woolf, eds., 1969, *Quasars and High-Energy Astronomy,* Proceedings of the Second Texas Symposium on Relativistic Astrophysics, Gordon and Breach, New York.

Douglass, D. H., Jr., 1971, "Properties of 'hollow square' gravitational wave detectors," in Davies 1971. [§37.3, Fig. 37.2]

Douglass, D. H., Jr., *cited by name in* Ex. 37.13.

Douglass, D. H., Jr., and J. A. Tyson, 1971, "New class of gravitational-wave detectors," *Nature 229,* 34–36. [§37.3]

Drabkin, I. E., and S. Drake, 1960, *Galileo Galilei on Motion and on Mechanics,* translations with introductions and notes, University of Wisconsin Press, Madison, Wisc. [Box 1.2]

Drake, S. *See* Drabkin and Drake 1960.

Drell, S. D. *See* Brodsky and Drell 1970.

Drever, R. W., 1961, "A search for anisotropy of inertial mass using a free precession technique," *Phil. Mag. 6,* 683–687. [§38.7]

Dunford, N. and J. T. Schwartz, 1958, *Linear operators,* pp. 467–471, Interscience (now a division of Wiley), New York. *Cited under* Lefschetz 1949.

Durbin, R. P., H. H. Loar, and W. W. Havens, Jr., 1952, "The lifetimes of the pi-plus and pi-minus mesons," *Phys. Rev. 88,* 179–183. [§38.4]

Duyvendak, J. J. L., 1942, "Further data bearing on the identification of the Crab Nebula with the supernova of A.D. 1054, Part I: The ancient oriental chronicles," *Proc. Astron. Soc. Pacific 54,* 91–94. [Box 24.1]

Dyce, R. B. *See* Shapiro, Ash, *et al.* 1971 and Shapiro *et al.* 1972.

Dyson, F. J., 1954, *Advanced Quantum Mechanics,* multilithographed lecture notes, Physics Department, Cornell University. [Box 17.2 part 6]

Dyson, F. J., 1967, "Time variation of the charge of the proton," *Phys. Rev. Lett. 19,* 1291–1293. [§37.3, Fig. 37.2]

Dyson, F. J., 1969, "Seismic response of the Earth to a gravitational wave in the 1-Hz band," *Astrophys. J. 156,* 529–540. [§37.3, Fig. 37.2]

Dyson, F. J., 1972, "The fundamental constants and their time variation," in Salam and Wigner 1972. [§38.6]

Eardley, D., E. Liang, and R. Sachs, 1972, "Velocity-dominated singularities in irrotational dust cosmologies," *J. Math. Phys. 13,* 99–107. [§30.7]

Eckart, C., 1940, "The thermodynamics of irreversible processes, III: Relativistic theory of the simple fluid," *Phys. Rev., 58,* 919–924. [Ex. 22.7]

Ecphantus, *cited by name in* Box 27.7.

Eddington, A. S., 1922, *The Mathematical Theory of Relativity,* Cambridge Univ. Press, Cambridge, Eng. [§§38.2, 39.3]

Eddington, A. S., 1924, "A comparison of Whitehead's and Einstein's formulas," *Nature 113,* 192. [Boxes 31.1, 31.2]

Eddington, A. S., 1931, "Preliminary note on the masses of the electron, the proton, and the universe," *Proc. Cambridge Phil. Soc. 27,* 15–19. [§44.6]

Eddington, A. S., 1936, *Relativity Theory of Protons and Electrons,* Cambridge Univ. Press, Cambridge, Eng. [§44.6]

Eddington, A. S., 1946, *Fundamental Theory,* Cambridge Univ. Press, Cambridge, Eng. [§44.6]

Eddington, A. S., *cited by name in* Box 27.7.

Eddington, A., and G. Clark, 1938, "The problem of *n* bodies in general relativity theory," *Proc. Roy. Soc. A 166,* 465–475. [Ex. 39.15]

Ehlers, J., 1961, "Contributions to the relativistic mechanics of continuous media," *Akad. Wiss. Lit. Mainz Abh. Math.-Nat. Kl.,* 793–837. [§22.3]

Ehlers, J., ed., 1967, *Relativity Theory and Astrophysics: I, Relativity and Cosmology; II, Galactic Structure; III, Stellar Structure,* American Mathematical Society, Providence, R.I.

Ehlers, J., 1971, "General relativity and kinetic theory," in Sachs 1971. [§22.6, Ex. 22.7, endpapers]

Ehlers, J. *See also* Jordan, Ehlers, and Sachs 1961.

Ehlers, J., P. Geren, and R. K. Sachs, 1968, "Isotropic solutions of the Einstein-Liouville equations," *J. Math. Phys. 9,* 1344–1349. [§27.2]

Ehlers, J., and W. Kundt, 1962, "Exact solutions of the gravitational field equations," in Witten 1962. [§§33.6, 35.9, Ex. 35.8, 35.9]

Ehlers, J., F. A. E. Pirani, and A. Schild, 1972, "The geometry of free-fall and light propagation," *in* O'Raifeartaigh 1972, pp. 63–84. [Boxes 10.2, 16.4]

Ehrenfest, P., 1916, "On the adiabatic changes of a system in connection with the quantum theory," *Amsterdam Acad., Proc. 19,* 576–597. [Box 22.4]

Einstein, A., 1905, "Zur elektrodynamik bewegter Körper," *Ann. Phys. (Germany) 17,* 891–921. [Box 1.5]

Einstein, A., 1908, "Über das Relativitätsprinzip und die aus demselben gezogenen Folgerungen," *Jarb. Radioakt. 4,* 411–462. [§§7.4, 17.7]

Einstein, A., 1911, "Über den Einfluss der Schwerkraft auf die Ausbreitung des Lichtes," *Ann. Phys. (Germany) 35,* 898–908, English translation in Lorentz *et al.* 1923. [§§1.3, 7.2, 7.4, 10.1, 16.2]

Einstein, A. 1912, "Si l'on modifie la longueur du pendule . . . ," p. 450 in P. Langevin and M. de Broglie, eds., *La théorie du rayonnement et les quanta,* Gauthier-Villars, Paris, 1912. [Box 22.4]

Einstein, A., 1913a, "Zum gegenwärtigen Stand des Gravitationsproblems," *Phys. Z. 14,* 1249–1262, discussion 1262–1266. [Epigraph to Chap. 35]

Einstein, A., 1913b, letter to Ernst Mach dated June 26. [§21.12, Fig. 21.5]

Einstein, A., 1915a, "Zür allgemeinen Relativitätstheorie," *Preuss. Akad. Wiss. Berlin, Sitzber.,* 778–786 (published November 11). [§17.7]

Einstein, A. 1915b, "Zür allgemeinen Relativitätstheorie (Nachtrag)," *Preuss. Akad. Wiss. Berlin, Sitzber.,* 799–801 (published November 18). [§17.7, Box 27.8]

Einstein, A., 1915c. "Erklärung der Perihelbewegung des Merkur aus der allgemeinen Relativitätstheorie," *Preuss. Akad. Wiss. Berlin, Sitzber. 47,* 831–839 (published November 25). [§17.7]

Einstein, A., 1915d, "Die Feldgleichungen der Gravitation," *Preuss. Akad. Wiss. Berlin, Sitzber.,* 844–847 (published December 2). [§17.7, Box 27.7]

Einstein, A., 1916a, "Ernst Mach," Phys. Z. Sowjetunion 17, 101–104. [Fig. 21.5]

Einstein, A., 1916b, "Näherungsweise Integration der Feldgleichungen der Gravitation," *Preuss, Akad. Wiss. Berlin, Sitzber.* 688–696. [§17.7]

Einstein, A., 1916c, "Hamiltonsches Prinzip und allgemeine Relativitätstheorie," *Preuss. Akad. Wiss. Berlin, Sitzber.,* 1111–1116. [§17.7]

Einstein, A., 1917, "Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie," *Preuss. Akad. Wiss. Berlin, Sitzber.* 142–152, reprinted in English in Lorentz *et al.* 1923. [§17.3, Boxes 27.1, 27.7]

Einstein, A., 1921, "Geometry and Experience," address to Prussian Academy of Sciences, Berlin, Jan. 27, 1921; published in *Sidelights on Relativity,* Mathuen, London, 1922. [Box 1.11]

Einstein, A., 1931, "Zum kosmologischen Problem der allgemeinen Relativitätstheorie," *Preuss, Akad. Wiss. Berlin, Sitzber,* 235–237. [§27.11]

Einstein, A., 1934, *Essays in Science,* Philosophical Library, New York. Translated from *Mein Weltbild,* Querido Verlag, Amsterdam, 1933. [Epigraph to Chap. 17, §§21.12, 27.11, 27.12, 43.1, 44.6, Box 8.5]

Einstein, A., 1939, "On a stationary system with spherical symmetry consisting of many gravitating masses," *Ann. Math. (U.S.A.) 40,* 922–936. [Ex. 25.29]

Einstein, Albert, *Festschrift,* in commemoration of his seventieth birthday, March 14, 1949; *Rev. Mod. Phys. 21* (July 1949), 343–540.

Einstein, A., 1949a, introduction to Schilpp 1949. [Preface, Epigraph to Chap. 44, §1.2]

Einstein, A., 1949b, "Reply to criticisms," in Schilpp 1949. [Epigraph to Chap. 27]

Einstein, A., 1950, *The Meaning of Relativity,* 3d ed., Princeton Univ. Press, Princeton, N.J. [§21.12, 29.5, 43.1, Box 27.1, endpapers]

Einstein, A., 1954, preface to *Concepts of Space,* M. Jammer, ed., Harper, New York, p. xv. [§1.3]

Einstein, A., 1970, quoted by G. Gamow in *My World Line,* Viking Press, New York (1970), p. 44. [§§17.3, 27.8, Box 27.1]

Einstein, A., *cited by name in* §§1.1, 1.6, 7.1, 7.2, 7.4, 10.1, 12.1, 12.2, 17.3, 20.6, Boxes 12.2, 12.3, 27.7. *See also* Lorentz *et al.* 1923, Born 1971, Holton 1971, Klein 1971.

Einstein, A., and A. D. Fokker, 1914, *Ann. Phys. (Germany), 44,* 321. [§17.6]

Einstein, A., and J. Grommer, 1927, "Allgemeine Relativitätstheorie und Bewegungsgesetze," *Sitzber. Deut. Akad. Wiss. Berlin, Kl. Math. Phys. Tech.* 2–13, 235–245. [§20.6]

Einstein, A., and M. Grossmann, 1913, "Entwerf einer verallgemeinerten Relativitätstheorie und einer Theorie der Gravitation," *Zeit. Math. Phys. 62,* 225–261. [§17.7]

Einstein, A., and M. Grossmann, 1914, "Kovarianzeigenschaften der Feldgleichungen der auf die verallgemeinerte Relativitätstheorie gegründeten Gravitationstheorie," *Zeit. Math. Phys. 63,* 215–225. [§17.7]

Einstein, A., L. Infeld, and B. Hoffmann, 1938, "The gravitational equations and the problem of motion," *Ann. Math.* 39, 65–100. [Ex. 39.15]

Eisenhart, L. P., 1926, *Riemannian Geometry.* Princeton Univ. Press, Princeton, N.J. [§§21.5, 21.12, Ex. 27.4, endpapers]

Ekers, R. D. *See* Muhleman, Ekers, and Fomalont 1970.

Eliot, T. S., 1920, "Tradition and individual talent," in *The Sacred Wood; Essays on Poetry and Criticism,* Methuen, London. [Box 27.7]

Eliot, T. S., 1942, *Little Gidding,* Faber and Faber, London. [§1.7]

Ellis, G. F. R., 1971, "Relativistic cosmology," in Sachs 1971. [§22.3, Ex. 22.6, 29.5]

Ellis., G. F. R., *cited by name in* Boxes 27.7, 34.3, Ex. 32.1. *See also* Hawking and Ellis 1968 *and* 1973, *and* Stewart and Ellis 1968.

Eötvös, R. V., 1889, "Über die Anziehung der Erde auf verschiedene Substanzen," *Math. Naturw. Ber. aus Ungarn 8,* 65–68. [Box 1.2]

Eötvös, R. V., *cited by name in* §§10.1, 38.3, 38.5, 40.9.

Eötvös, R. V., D. Pekár, and E. Fekete, 1922, "Beiträge zum Gesetze der Proportionalität von Trägheit und Gravität," *Ann. Phys. (Germany) 68,* 11–66. [§§10.1, 38.3, 40.9, Boxes 1.2, 38.2]

Epstein, E. E. *See* Dieter *et al.* 1962.

Ernst, F. J., 1968, "7090 FORMAC subroutines: Exterior calculus," report number PORS IIT-9, Illinois Institute of Technology. [Box 14.3]

Esposito, F. P. *See* Chandrasekhar and Esposito 1970.
Esposito, P. B. *See* Anderson *et al.* 1971.
ESRO Conference on Space Physics, Interlaken, Switzerland, September 4, 1969. *See* Ruffini and Wheeler 1971b.
Estabrook, F. B., 1969, "Post-Newtonian *n*-body equations of the Brans-Dicke theory," *Astrophys. J. 158,* 81–83. [Ex. 39.15]
Estermann, I., O. C. Simpson, and O. Stern, 1938, "The free fall of molecules," *Phys. Rev. 53,* 947–948. [§1.3]
Eudoxus, *cited by name in* Box 27.7.
Euler, H. *See* Heisenberg and Euler 1936.
Euler, L., 1760, "Demonstratio theorematis Fermatiani omnem numerum esse summam quatuor quadratorum," *Novi Comment, Petrop. 5, 93, Opera 12,* 369ff. [§41.1]
Evenson, K. M., J. S. Wells, F. R. Petersen, B. L. Danielson, G. W. Day, R. L. Barger, and J. L. Hall, 1972, "Speed of light from direct frequency and wavelength measurements of the methane-stabilized laser", *Phys. Rev. Letters 29,* 1346–1349. [§1.5]
Everitt, C. W. F., 1970, "Maxwell as a student," preprint, Stanford University, Stanford, Calif. [Box 14.6]
Everitt, C. W. F., W. M. Fairbank, and W. O. Hamilton, 1970, "General relativity experiments using low-temperature techniques," in Carmeli, Fickler, and Witten 1970, pp. 145–159. [§40.7]
Exton, A. *See* Newman *et al.* 1965.
Fackerell, E. D., 1966, *Relativistic Stellar Dynamics,* unpublished Ph.D. thesis, University of Sydney. [Box 25.9]
Fackerell, E. D., 1968, "Relativistic stellar dynamics," *Astrophys. J. 153,* 643–660. [Box 25.8]
Fackerell, E. D. *See also* Brill *et al.* 1972.
Fackerell, E. D., and J. R. Ipser, 1972, "Weak electromagnetic fields around a rotating black hole," *Phys. Rev. D5,* 2455–2458. [§41.11]
Fackerell, E. D., J. R. Ipser, and K. S. Thorne, 1969, "Relativistic star clusters," *Comments Astrophys. & Space Phys. 1,* 134–139. [§24.6, Box 25.9]
Fadeev, L. D., 1971, "Symplectic structure and quantization of the Einstein gravitation theory," in *Actes du Congrés International des Mathématiciens, Nice 1–10 Septembre 1970,* Gauthier-Villars, Paris, III, 35–40. [Box 21.1]
Fairbank, W. M. *cited by name in* Ex. 37.13. *See also* Everitt, Fairbank, and Hamilton 1970, Witteborn and Fairbank 1967, *and* Witteborn, Knight, and Fairbank 1967.
Faller, J. E. *See* Bender *et al.* 1971.
Faraday, M., *cited by name in* §5.6.
Farley, F. J. M., 1969, "The status of quantum electrodynamics," *Nuovo Cimento 1,* Numero Speciale, 59–86. [§44.2]
Farley, F. J. M. *See also* Bailey *et al.* 1968.
Farley, F. J. M., J. Bailey, R. C. A. Brown, M. Giesch, H. Jöstlein, S. van der Meer, E. Picasso, and M. Tannenbaum, 1966, "The anomolous magnetic moment of the negative muon," *Nuovo Cimento 45,* 281–286. [§38.4]
Farnsworth, D., J. Fink, J. Porter, and A. Thomson, eds., 1972, *Methods of Local and Global Differential Geometry in General Relativity,* Springer, Berlin.
Faulkner, J., and J. R. Gribbin, 1968, "Stability and radial vibration periods of the Hamada-Salpeter white-dwarf models," *Nature 218,* 734–735. [§24.2, Boxes 24.1, 26.1]
Fejér, L., and S. Mikola, eds., 1918, "Báró Eötvös Loránd Füzet," *Matem. Fiz. Lapok 27,* Nos. 6–7, Oct.–Nov., kindly translated into English for use here by Rev. Stanley Jaki. [Box 38.2]
Fekete, E. *See* Eötvös, Pekar, and Fekete 1922.
Fermi, E., 1922, "Sopra i fenomeni che avvengono in vicinanza di una linea oraria," *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. & Nat. 31,* 184–187, 306–309. [§6.3]
Fermi, E., 1932, "Quantum theory of radiation," *Rev. Mod. Phys. 4,* 87–132. [§21.12]
Fermi, E., 1949, "On the origin of the cosmic radiation," *Phys. Rev. 75,* 1169–1174. [Box 27.7].
Fermi, E., and A. Turkevich, 1950, unpublished work described in Alpher and Herman 1950, pp. 194ff. [§28.1]
Ferrari, A., and R. Ruffini, 1969, "Theoretical implications of the second time derivative of the period of the pulsar NP 0532," *Astrophys. J. Letters 158,* L71–L75. [Ex. 36.4]
Feynman, R. P., 1942. *The Principle of Least Action in Quantum Mechanics,* doctoral dissertation, Princeton University. [Boxes 17.2, 21.1]
Feynman, R. P., 1948, "Spacetime approach to nonrelativistic quantum mechanics," *Rev. Mod. Phys. 20,* 367–387. [Box 21.1]
Feynman, R. P., 1949, "Spacetime approach to quantum electrodynamics," *Phys. Rev. 76,* 769–789. [Box 21.1]
Feynman, R. P., 1951, "The concept of probability in quantum mechanics," in Neyman 1951, p. 533. [Box 21.1]
Feynman, R. P., 1955, "Slow electrons in polar crystals," *Phys. Rev. 97,* 660–665. [Box 21.1]
Feynman, R. P., 1963, *Lectures on Gravitation,* unpublished lecture notes prepared by F. B. Morinigo and W. G. Wagner, Calif. Institute of Technology. [§7.1, Box 7.1, §18.1]
Feynman, R. P., 1964, unpublished work cited in Fowler 1964. [Box 24.1]
Feynman, R. P. *See also* Wheeler and Feynman 1945 *and* Wheeler and Feynman 1949.
Feynman, R. P., and A. R. Hibbs, 1965, *Quantum Mechanics and Path Integrals,* McGraw-Hill, New York. [§21.1, Fig. 13.3, Boxes 17.2, 21.1]
Feynman, R. P., R. B. Leighton and M. Sands, 1964, *The Feynman Lectures on Physics,* Addison-Wesley, Reading, Mass. [§21.3, Box 16.3]
Fickler, S. I. *See* Carmeli, Fickler, and Witten 1970.
Field, G., 1973, "The origin of galaxies," in A. and M. Sandage, eds., *Stars and Stellar Systems, IX: Galaxies and the Universe,* Univ. of Chicago Press, Chicago (in press). [§30.1]
Fierz, M., and W. Pauli, 1939, "Relativistic wave equations for particles of arbitrary spin in an electromagnetic field," *Proc. R. Soc. London A 173,* 211–232. [§§18.1, 18.3, 39.1, Box 7.1]
Fierz, M., and V. F. Weisskopf, 1960, *Theoretical Physics in the 20th Century: A Memorial Volume to Wolfgang Pauli,* Interscience, New York.
Fink, J. *See* Farnsworth *et al.* 1972.
Finkelstein, D., 1958, "Past-future asymmetry of the gravitational field of a point particle," *Phys. Rev. 110,* 965–967. [Box 31.2]

Fischer, A. E., and J. B. Marsden, 1972, "The Einstein equations of evolution—A geometric approach," *J. Math. Phys. 13,* 546–568. [§§21.5, 21.6]

Fischer, A. E., 1970, "The theory of superspace," in Carmeli, Fickler, and Witten 1970. [§43.1]

Fitch, J. P. *See* Barton and Fitch 1971 *and* Barton, Bourne, and Fitch 1970.

Flamm, L., 1916, "Beitrage zur Einsteinschen Gravitationstheorie," *Physik. Z. 17,* 448–454. [§23.8]

Flanders, H., 1953, "Development of an extended exterior differential calculus," *Trans. Am. Math. Soc. 75,* 311–326. [§14.5, Ex. 14.9, 14.10]

Flanders, H., 1963, *Differential Forms With Applications to the Physical Sciences,* Academic Press, New York. [§4.1, Boxes 4.4, 8.1]

Fletcher, J. G., 1966, "GRAD Assistant—a program for symbolic algebraic manipulation and differentiation," *Commun. Assoc. Comput. Machinery 9,* 552–553. [Box 14.3]

Fletcher, J. G., R. Clemens, R. Matzner, K. S. Thorne, and B. A. Zimmerman, 1967, "Computer programs for calculating general-relativistic curvature tensors," *Astrophys. J. Lett. 148,* L91–L92. [Box 14.3]

Fock, V. A., 1959, *The Theory of Space, Time, and Gravitation,* Pergamon, New York, translated by N. Kemmer. [Box 39.3, Ex. 39.15, Epigraph to Chap. 21, endpapers]

Fock, V. A., chief editor, 1968, *Abstracts, Fifth International Conference on Gravitation and the Theory of Relativity,* Publishing House of Tbilisi University, Tbilisi, USSR.

Fock, V. A. *See also* Dirac, Fock, and Podolsky 1932.

Fokker, A. D., 1965, *Time and Space, Weight and Inertia,* Pergamon Press, London. [endpapers]

Fokker, A. D. *See* Einstein and Fokker 1914.

Fomalont, E. B. *See* Muhleman, Ekers, and Fomalont 1970.

Ford, K. W. *See* Deser and Ford 1967.

Ford, K. W., and J. A. Wheeler, 1959a, "Application of semiclassical scattering analysis," *Ann. Phys. (U.S.A.) 7,* 287–322. [§25.6]

Ford, K. W., and J. A. Wheeler, 1959b, "Semiclassical description of scattering," *Ann. Phys. (U.S.A.) 7,* 259–286. [§25.6]

Forman, F. *See* Gittelman *et al.* 1968.

Forward, R. L., 1972, "Geodesy with orbiting gravity gradiometers," in *The Use of Artificial Satellites for Geodesy,* Geophysical Monograph Series, vol. 15, American Geophysical Union, Washington, 239–243. [Box 16.5]

Fourès-Bruhat, Y. *See* Bruhat, Y., *and* Choquet-Bruhat, Y.

Fowler, R. H., 1926, "Dense matter," *Mon. Not. R. Astron. Soc. 87,* 114–122. [Box 24.1]

Fowler, W. A., 1964, "Massive stars, relativistic polytropes, and gravitational radiation," *Rev. Mod. Phys. 36,* 545–555 and 1104. [Boxes 24.1, 26.2]

Fowler, W. A., 1966, "The stability of supermassive stars," *Astrophys. J. 144,* 180–200. [§24.4, Box 26.2]

Fowler, W. A., 1967, *Nuclear Astrophysics,* American Philosophical Society, Philadelphia, Penn. [§28.1]

Fowler, W. A., 1972, "New observations and old nucleocosmochronologies," in Reines 1972. [§29.7]

Fowler, W. A. *Cited by name in* Box 27.7. *See also* Hoyle and Fowler 1963a *and* 1963b, Hoyle *et al.* 1964, *and* Wagoner, Fowler, and Hoyle 1967.

Fowler, W. A., and F. Hoyle, 1964, "Neutrino processes and pair formation in massive stars and supernovae," *Astrophys. J. Supp. 9,* 201–319. [§24.4]

Fram, D. M. *See* Boughn, Fram, and Partridge 1971.

Friedmann, A., 1922, "Über die Krümmung des Raumes," *Z. Phys. 10,* 377–386. [§§14.1, 27.9, 29.2, Boxes 14.5, 27.6]

Friedmann, A., *cited by name in* §§30.3, 42.1, Boxes 27.6, 27.7.

Frisch, D. H., and J. H. Smith, 1963, "Measurement of the relativistic time dilation using mu-mesons," *Am. J. Phys. 31,* 342–355. [§38.4]

Fröman, N., and P. O. Fröman, *JWKB approximation: contributions to the theory,* North-Holland, Amsterdam, 1965. [§22.5]

Fröman, P. O. *See* Fröman and Fröman 1965.

Fronsdal, C., 1959, "Completion and embedding of the Schwarzschild solution," *Phys. Rev. 116,* 778–781. [§23.8, Box 31.1]

Fujiwara, I., 1962, "The correspondence principle, I," in *Proceedings of the Physics Seminar in Trondheim,* no. 3, Nordita, Copenhagen, Denmark. [Box 21.1]

Fuller, R. W., and J. A. Wheeler, 1962, "Causality and multiply-connected space-time," *Phys. Rev. 128,* 919–929. [§31.6, Fig. 31.5]

Galilei, G., 1592, *De Motu,* unpublished work written at the University of Pisa, 1589–1592; translation in Drabkin and Drake 1961, p. 48. [Box 1.2]

Galilei, G., 1610, *Sidereus Nuncius,* Translated by E. S. Carlos, *The Sidereal messenger,* Rivingtons, London, 1880. [Boxes 1.9, 27.7]

Galilei, G., 1623, *Il Saggiatore,* Accademia dei Lincei, Rome. [epigraph, Chap. 13]

Galilei, G., 1632, *Dialogo dei due massimi sistemi del mundo,* Landini, Florence. Translation by S. Drake, *Galileo Galilei—Dialogue Concerning the Two Chief World Systems—Ptolemaic and Copernican,* Univ. of California Press, Berkeley and Los Angeles, 1953. [Box 1.9]

Galilei, G., 1638, *Discorsi e dimostriazioni matematiche intorno a due nuove scienze,* Elzevir, Leiden. Translation by H. Crew and A. de Salvio, *Dialogues Concerning Two New Sciences,* Macmillan, New York, 1914; reprinted by Dover, New York, 1954. [Boxes 1.2, 1.9, 27.7]

Galilei, G., *cited by name in* §12.2, *quoted in* epigraph of Chap. 38.

Gamow, G., 1948, "The evolution of the universe," *Nature 162,* 680–682. [§28.1]

Gamow, G., 1967, "Electricity, gravity, and cosmology," *Phys. Rev. Lett. 19,* 759–761. [§38.6]

Gamow, G., 1970, *My World Line,* Viking Press, New York. [Box 27.7]

Gamow, George, March 4, 1904–August 20, 1968, Memorial volume. *See* Reines 1972.

Gaposchkin, S., 1958, "The eclipsing binaries," *Handbuch der Physik 50,* 225–242. [Box 36.3]

Gårding, L., 1945, "Relativistic wave equations for zero rest mass," *Proc. Cambridge Phil. Soc. 41,* 49–56. [Box 32.2]

Garnett, W. *See* Campbell and Garnett 1882.

Gauduchon, P. *See* Berger, Gauduchon, and Mazet 1971.

Gasiorowizc, S., 1966, *Elementary Particle Physics,* Wiley, New York. [Box 7.1]

Gauss, K. F., 1817, letter to Olbers in his *Werke,* VIII, 177. [Epigraph to Chap. 8]

Gauss, K. F., 1830, letter to Bessel in his *Werke,* VIII, 201. [Epigraph to Chap. 8]

Gauss, K. F., 1827, "Disquisitiones generales circa superficies curvas," in *Karl Friedrich Gauss Werke,* IV, 217–258; English translation by J. C. Morehead and A. M. Hiltebeitel, *General Investigations of Curved Surfaces of 1827 and 1825,* reprinted by Raven Press, New York, 1965. [§§5.9, 22.2, Boxes 5.3, 5.4, 14.1]

Gauss, K. F., *cited by name in* §41.1.

Géhéniau, J., *quoted at end of* Chap. 44.

Geren, P. *See* Ehlers, Geren, and Sachs 1968.

Gerlach, U., 1968, "The equation of state at supranuclear densities and the existence of a third family of superdense stars," *Phys. Rev. 172,* 1325–1330. [§24.2]

Gerlach, U., 1969, "Derivation of the ten Einstein field equations from the semiclassical approximation to quantum geometrodynamics," *Phys. Rev. 177,* 1929–1941. [§43.3, Box 17.2 part 4]

Geroch, R. P., 1967, *Singularities in the Spacetime of General Relativity: Their Definition, Existence, and Local Characterization,* doctoral dissertation, Princeton University. [§§32.7, 44.1, Epigraph to Chap. 34]

Geroch, R. P., 1968, "What is a singularity in general relativity?" *Ann. Phys. (U.S.A.) 48,* 526–540. [§34.6]

Geroch, R. P., 1971, "Spacetime structure from a global viewpoint," in Sachs 1971. [§34.1]

Geroch, R. P., *cited by name in* §§32.7, 42.1, Box 34.1.

Giacconi, R., *cited by name in* Box 27.7.

Gibbons, J. J., *cited by name in* Box 34.3.

Giesch, M. *See* Farley *et al.* 1966.

Gilbert, R. P., and R. Newton, 1970, *Analytic Methods in Mathematical Physics,* Gordon and Breach, New York. [Box 8.5]

Gillispie, C. C. *See* Bernoulli 1733.

Ginzel, F. K., 1882, "Astronomische Untersuchungen über Finsternisse, I: Über die zwischen 26 und 103 nach Chr. stattgefundenen Sonnenfinsternisse im Allgemeinen and die Finsterniss des Plutarch ins besonders," *Sitzber. Kais. Akad. Wiss. Wien Math.-Nat. Kl. 85II,* 663–747. [Fig. 1.8]

Ginzel, F. K., 1883, "Astronomische Untersuchungen über Finsternisse, II: Grundlagen aus historischen Sonnenfinsternissen zur Ableitung empirischer Correctionen der Mondbahn," *Sitzber. Kais. Akad. Wiss. Wien Math.-Nat. Kl. 88II,* 629–755. [Fig. 1.8]

Ginzel, F. K., 1884, "Astronomische Untersuchungen über Finsternisse, III: Ermittlung empirischer Correction der Bahn des Mondes," *Sitzber. Kais. Akad. Wiss. Wien Math.-Nat. Kl. 89II.* [Fig. 1.8]

Ginzel, F. K., 1899, *Spezieller Kanon der Sonnen- und Mondfinsternisse für das Landergebiet der Klassischen Altertumwissenschaften und den Zeitraum von 900 vor Chr. bis 600 nach Chr.,* Mayer, Berlin. [Fig. 1.8]

Godfrey, B. B., 1970a, "Mach's principle, the Kerr metric, and black-hole physics," *Phys. Rev. D 1,* 2721–2725. [Box 33.3]

Godfrey, B. B., 1970b, *A Survey of the Mathematical Properties of Static, Axisymmetric Vacuum Spacetimes,* doctoral dissertation, Princeton University. [§34.1]

Goebel, C. J., 1972, "Comments on the 'vibrations' of a black hole," *Astrophys. J. 172,* L95–L96. [Box 33.3]

Gödel, K., 1931, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandte Systeme I," *Monatsh. Math. Phys. 38,* 173–198, English translation by B. Meltzer, *On Formally Undecidable Propositions,* with an introduction by R. O. Braithwaite, Basic Books, New York, 1962; outlined and discussed in Nagel and Newman 1958. [Box 44.5]

Gold, T., 1968, "Rotating neutron stars as the origin of the pulsating radio sources," *Nature 218,* 731–732. [Box 24.1]

Gold, T., 1972, oral remark to J. A. Wheeler. [Epigraph to Chap. 21]

Gold, T., *cited by name in* Box 27.7. *See also* Bondi and Gold 1948.

Goldberg, S. I. *See* Bishop and Goldberg 1968.

Goldenberg, H. M. *See* Dicke and Goldenberg 1967.

Goldreich, P. and W. H. Julian, 1968, "Pulsar electrodynamics," *Astrophys. J. 157,* 869–880. [§24.3, Box 33.3]

Goldstein, H., 1959, *Classical Mechanics,* Addison-Wesley, Reading, Mass. [§§2.1, 9.7, 33.5, Ex. 33.7]

Goldstein, J. *See* Chrétien, Deser, and Goldstein 1969.

Gott, J. R., III, and J. E. Gunn, 1971, "The Coma Cluster as an X-ray source: Some cosmological implications," *Astrophys. J. Lett. 169,* 13–15. [§29.6]

Gowdy, R. H., 1971, "Gravitational waves in closed universes," *Phys. Rev. Lett. 27,* 826–829, and erratum p. 1102. [§30.6]

Gowdy, R. H., 1973, "Vacuum spacetimes with two-parameter spacelike isometry groups and compact invariant hypersurfaces: Topologies and Boundary conditions," University of Maryland technical report 73-089 (unpublished). [§§21.10, 30.6]

Gratton, L., ed., 1966, *High-Energy Astrophysics,* Academic Press, New York.

Graves, J. C., and D. R. Brill, 1960, "Oscillatory character of Reissner-Nordstrom metric for an ideal charged wormhole," *Phys. Rev. 120,* 1507–1513. [Fig. 34.4, Ex. 31.8]

Greenstein, G. S., 1968a, "Brans-Dicke Cosmology," *Astrophys. Letters, 1,* 139–141. [§28.4]

Greenstein, G. S., 1968b, "Brans-Dicke Cosmology, II," *Ap. Space Sci., 2,* 157–165. [§28.4]

Gribbin, J. R. *See* Faulkner and Gribbin 1968.

GR*n, International Conferences on General Relativity and Gravitation. See* Berne 1955 (GR0), Chapel Hill 1957 (GR1), Royaumont 1959 (GR2), Jablonna 1962 (GR3), London 1965 (GR4), Tbilisi 1968 (GR5), Copenhagen 1971 (GR6).

Grommer, J. *See* Einstein and Grommer 1927.

Grossmann, M. *See* Einstein and Grossman 1913 *and* 1914.

Gunn, J. E., 1967, "On the propagation of light in inhomogeneous cosmologies: I, Mean effects," *Astrophys. J. 150,* 737–753. [§29.5]

Gunn, J. E., 1971, "On the distances of the quasistellar objects," *Astrophys. J. Letters, 164,* L133–L118. [Box 28.1]

Gunn, J. E. *Cited by name in* Box 27.7. *See also* Gott and Gunn 1971, Ostriker and Gunn 1969.

Gupta, S. N., 1954, "Gravitation and electromagnetism," *Phys. Rev. 96,* 1683–1685. [§§7.1, 18.1]

Gupta, S. N., 1957, "Einstein's and other theories of gravitation," *Rev. Mod. Phys. 29,* 337–350. [§§7.1, 18.1]

Gupta, S. N., 1962, "Quantum theory of gravitation," in *Recent Developments in General Relativity,* Pergamon, New York, pp. 251–258. [§§7.1, 18.1]

Gurovich, V. Ts., 1965, "The gravitational field of rotating superdense configurations," *Astron. Zhur. 42,* 974–976. English translation in *Sov. Astron.—A.J. 9,* 750–751. [Ex. 26.1]

Gursky, H., *cited by name in* Box 27.7.

Guseynov, O. H. *See* Zel'dovich and Guseynov 1965.

Guthrie, W. K. C., trans., 1939, *Aristotle's On the Heavens,* Harvard Univ. Press, Cambridge, Mass. [Box 27.7]

Haeberli, W. *See* Barschall and Haeberli 1971.

Hagedorn, R., 1964, *Relativistic Kinematics,* W. A. Benjamin, New York. [§2.10]

Hagihara, Y., 1931, "Theory of the relativistic trajectories in a gravitational field of Schwarzschild," *Japan. J. Astron. Geophys., 8,* 67–175. [§25.5]

Hall, D. B. *See* Rossi and Hall 1941.

Hall, J. L., 1972. *See* Evenson *et al.* 1972.

Halley, E., 1693, "Emendationes ac notal in vetustas Ablatenii observationes astronomicas cum restitutione tabularum lunisolarum," *Roy. Soc., London, Phil. Trans. 17,* 913. [§1.5]

Halley, E. 1720, "Of the infinity of the sphere of fix'd stars" and "Of the number, order, and light of the fix'd stars," *Roy. Soc. London, Phil. Trans. 31,* 22–24 and 24–26. [Box 27.7]

Hamilton, W. O., *cited by name in* Ex. 37.13. *See also* Everitt, Fairbank, and Hamilton 1970.

Hamilton, W. R., *cited by name in* §41.1 Boxes 17.2, 25.3.

Hamoui, A. *See* Papapetrou and Hamoui 1967 *and* 1968.

Hanni, R. S., and R. Ruffini, 1973, paper in preparation. [Box 33.2]

Hardy, E. *See* Sandage, Tamman, and Hardy 1972.

Harrison, B. K., 1959, "Exact three-variable solutions of the field equations of general relativity," *Phys. Rev. 116,* 1285–1296. [§14.1]

Harrison, B. K., K. S. Thorne, M. Wakano, and J. A. Wheeler, 1965, *Gravitation Theory and Gravitational Collapse,* Univ. of Chicago Press, Chicago. [§26.1, Fig. 24.2]

Harrison, B. K., and J. A. Wheeler, 1958, "The Harrison-Wheeler equation of state," first presented in *Onzième Conseil de Physique Solvay: La Structure et l'évolution de l'univers,* Editions Stoops, Brussels, 1959; presented in greater detail in Harrison *et al.* 1965. [§24.2, Fig. 24.2]

Harrison, G. H., 1970, "A Compact Method for Symbolic Computation of the Riemann Tensor," M. S. thesis, Tech. Report, Dept. of Physics and Astronomy, Univ. of Maryland, College Park. [Box 14.3]

Hartle, J. B., 1967, "Slowly rotating relativistic stars, I: Equations of structure," *Astrophys. J. 150,* 1005–1029. [Ex. 26.1]

Hartle, J. B., 1970, "Slowly rotating relativistic stars, IV: Rotational energy and moment of inertia for stars in differential rotation," *Astrophys. J. 161,* 111–118. [§20.5, Ex. 33.13]

Hartle, J. B., 1971, "Long-range neutrino forces exerted by Kerr black holes," *Phys. Rev. D 3,* 2938–2940. [§41.11, Boxes 25.2, 32.2, 33.1]

Hartle, J. B., 1972, "Can a Schwarzschild black hole exert long-range neutrino forces?" in Klauder 1972. [Boxes 25.2, 32.2, 33.1]

Hartle, J. B. *See also* Brill and Hartle 1964, Hawking and Hartle 1972.

Hartle, J. B., and D. H. Sharp 1965, "Variational principle for the hydrostatic equilibrium of a relativistic, rotating fluid," *Phys. Rev. Lett. 15,* 909–911. [Ex. 26.1]

Hartle, J. B., and K. S. Thorne, 1968, "Slowly rotating relativistic stars, II: Models for neutron stars and supermassive stars," *Astrophys. J. 153,* 807–834. [Box 36.1, Fig. 24.2]

Hartle, J. B., and K. S. Thorne, 1974, "Slowly rotating relativistic stars, VII: Emission of gravitational waves due to quasiradial pulsation," *Astrophys. J.,* in preparation. [Box 14.3]

Harvey, J. A. *See* Dabbs *et al.* 1965.

Hauser, W., 1970, "On the fundamental equations of electromagnetism," *Am. J. Phys. 38,* 80–85. [§4.1]

Havens, W. W., Jr. *See* Durbin, Loar, and Havens 1952.

Hawking, S. W., 1966a, "Singularities and the geometry of space-time" Adams Prize Essay, Cambridge University, Cambridge, England. [Box 34.3]

Hawking, S. W., 1966b, "The occurrence of singularities in cosmology," *Proc. R. Soc. London A 294,* 511–521. [Box 34.3]

Hawking, S. W., 1969, "On the rotation of the Universe," *Observatory 89,* 38–39. [Box 34.3]

Hawking, S. W., 1971a, "Gravitationally collapsed objects of very low mass," *Mon. Not. R. Astron. Soc. 152,* 75–78. [Box 33.3]

Hawking, S. W., 1971b, "Gravitational radiation from colliding black holes," *Phys. Rev. Lett. 26,* 1344–1346. [§34.5, Boxes 33.1, 33.4, Ex. 34.4]

Hawking, S. W., 1972a, "Black holes in general relativity," *Commun. Math. Phys. 25,* 152–166. [§34.5, Boxes 33.1, 33.3, 33.4, 34.3, Ex. 34.4]

Hawking, S. W., 1972b, "Gravitational Radiation: The Theoretical Aspect," *Contemp. Phys. 13,* 273–282. [Box 34.3]

Hawking, S. W., 1973, "The Event Horizon," in DeWitt and DeWitt 1973. [§§34.1, 34.5]

Hawking, S. W., *cited by name in* §§32.7, 42.1, Box 27.7. *See also* Bardeen, Carter, and Hawking 1973 *and* Collins and Hawking 1973.

Hawking, S. W., and G. F. R. Ellis, 1968, "The cosmic black-body radiation and the existence of singularities in our universe," *Astrophys. J. 152,* 25–36. [§28.3, Box 34.3]

Hawking, S. W., and G. F. R. Ellis, 1973, *The Large Scale Structure of Space-time,* Cambridge Univ. Press, Cambridge, Eng. [§§34.1, 34.6, Box 33.1, endpapers]

Hawking, S. W., and J. B. Hartle, 1972, "Energy and angular momentum flow into a black hole," *Commun. Math. Phys. 27,* 283–290. [Box 33.3]

Hawking, S. W., and R. Penrose, 1969, "The singularities of gravitational collapse and cosmology., *Proc. R. Soc. London A 314,* 529–548. [§§34.6, 44.1]

Hayakawa, S., 1965a, "Cosmological interpretation of the weak forces," *Prog. Theor. Phys. 33,* 538–539. [§44.6]
Hayakawa, S., 1965b, "Atomism and cosmology," *Prog. Theor. Phys. Supp.,* special number, Yukawa thirtieth anniversary issue, 532–541. [§44.6]
Hazard, C., *cited by name in* Box 27.7.
Hearn, A. C., 1970, "A program for symbolic algebraic computation," in *Proceedings of SHARE XXXIV,* 4.470–4.481. [Box 14.3]
Heckmann, O. *See* Schücking and Heckmann 1958.
Heisenberg, W., and H. Euler, 1936, "Folgerungen aus der Diracschen Theorie des Positrons," *Z. Phys. 98,* 714–732. [Ex. 40.8]
Helmholtz, H. L. F., *cited by name in* Box 38.2.
Henderson, T., *cited by name in* Box 27.7.
Hepp, K., 1969, *Théorie de la renormalisation,* Springer, Berlin. [Box 17.2 part 6]
Herman, R. C. *See* Alpher and Herman 1948a,b *and* 1950.
Hermann, R., 1966, *Lie Groups for Physicists,* W. A. Benjamin, New York. [§8.2]
Hernandez, W. C., Jr., 1967, "Material sources for the Kerr metric," *Phys. Rev. 159,* 1070–1072. [§33.3]
Heron (Hero of Alexandria), 1899, *Heronis Alexandrini Opera Omnia,* Teubner, Leipzig; *also in* G. Sarton, *A History of Science,* Harvard Univ. Press, Cambridge, Mass., 1952, I, 73. [Box 13.1]
Heron, G., trans., 1954, *Nicolas Cusanus's Of Learned Ignorance,* Yale Univ. Press, New Haven, Conn. [Box 27.7]
Hertz, G. L., *cited by name in* Box 17.3.
Hewish, A., *cited by name in* Box 27.7.
Hewish, A., S. J. Bell, J. D. H. Pilkington, P. F. Scott, and R. A. Collins, 1968, "Observation of a rapidly pulsating radio source," *Nature 217,* 709–713. [Box 24.1]
Hey, J. S., *cited by name in* Box 27.7.
Hibbs, A. R., 1951, "Quantum mechanics," in M. Kac, *Probability and Related Topics in Physical Science,* Interscience, New York, pp. 205–221. [Box 21.1]
Hibbs, A. R. *See also* Feynman and Hibbs 1965.
Hicks, N. J., 1965, *Notes on Differential Geometry,* Van Nostrand, Princeton, N.J. [Box 8.1, endpapers]
Hilbert, D., 1915, "Die Grundlagen der Physik," *Konigl. Gesell. d. Wiss. Göttingen, Nachr., Math.-Phys. Kl.,* 395–407; see also *Math. Annalen 92* (1924), 1–32. [§§17.7, 21.2, Box 17.2]
Hilbert, D., 1917, "Die Grundlagen der Physik (Zweite Mitteilung)," *Konigl. Gesell. d. Wiss. Göttingen, Nachr., Math.-Phys. Kl.,* 53–76. [§17.7]
Hilbert, D., *cited by name in* §21.1.
Hill, H. A. *See* Turner and Hill 1964.
Hill, J. M., 1971, "A Measurement of the gravitational deflection of radio waves by the sun," *Mon. Not. R. Astron. Soc., 154,* 7P-11P. [Box 40.1]
Hill, T., 1967, unpublished work referred to in DeWitt 1967c. [Box 17.2 part 6]
Hiltebeitel, A. M. *See* Gauss 1827
Hipparchus, *cited by name in* Box 16.1.
Hlavatý, Václav, *Festschrift* on the occasion of his 70th birthday, Jan. 27, 1964. *See* Hoffmann 1966.
Hönl, H., 1962, "Allgemeine Relativitätstheorie und Machsches Prinzip," in *Physikertagung Wien,* Physik Verlag, Mosbach/Baden, Germany. [§§21.12, 43.1]
Hoffmann, B., ed., 1966, *Perspectives in Geometry and Relativity: Essays in Honor of Václav Hlavatý,* Indiana Univ. Press, Bloomington.
Hoffman, B. *See also* Einstein, Infeld, and Hoffman 1938.
Hoffman, W. F. *See* Chiu and Hoffman 1964.
Hohn F. E., 1966, *Applied Boolean Algebra: An Elementary Introduction,* 2d ed., Macmillan, New York. [§44.5]
Holden, D. J., 1966, "An investigation of the clustering of radio sources," *Mon. Not. R. Astron. Soc., 133,* 225–228. [§27.1]
Holdridge, D. B. *See* Ohandley *et al.,* 1969.
Holmes, A., *cited by name in* Box 27.7.
Holton, G., 1965, personal communication, London. [Preface]
Holton, G., 1971, *Einstein. The Life and Times, by R. W. Clark,* book review, *New York Times Book Review,* September 5, 1-20. [Box 1.11]
Holton, G. *See also* Kollros 1956.
Horstmann, H. *See* Dabbs *et al.* 1965
Hoyle, F., 1948, "A new model for the expanding universe," *Mon. Not. R. Astron. Soc. 108,* 372–382. [§28.4]
Hoyle, F., *cited by name in* Box 27.7. *See also* Fowler and Hoyle 1964 *and* Wagoner, Fowler, and Hoyle 1967.
Hoyle, F., and W. A. Fowler, 1963a, "On the nature of strong radio sources," *Mon. Not. R. Astron. Soc. 125,* 169–176. [§24.5, Box 24.1]
Hoyle, F., and W. A. Fowler, 1963b, "Nature of strong radio sources," *Nature 197,* 533–535. [§24.5, Box 24.1]
Hoyle, F., W. A. Fowler, G. R. Burbidge, and E. M. Burbidge, 1964, "On relativistic astrophysics," *Astrophys. J. 139,* 909–928. [Box 33.3]
Hsiang, W.-C., and B. J. Sanderson, 1965, "Twist-spinning spheres in spheres," *Illinois J. Math. 9,* 651–659. [Box 44.3]
Hubble, E. P., 1920, "Photographic investigations of faint nebulae" (Ph.D. thesis), *Pub. Yerkes Obs., 4,* 69–85. [Box 29.5]
Hubble, E. P., 1925, "NGC 6822, a remote stellar system," *Astrophys. J. 62,* 409–433. [Box 29.5]
Hubble, E. P., 1926a, "A spiral nebula as a stellar system: Messier 33," *Astrophys. J. 63,* 236–274. [Box 29.5]
Hubble, E. P., 1926b, "Extra-galactic nebulae," *Astrophys. J. 64,* 321–369. [Box 29.5]
Hubble, E. P., 1929, "A relation between distance and radial velocity among extragalactic nebulae," *Proc. Nat. Acad. Sci. U.S. 15,* 169–173. [§§27.9, 29.2, Boxes 27.1, 29.5]
Hubble, E. P., 1934a, *Red-shifts in the spectra of nebulae* (Halley Lecture), Clarendon Press, Oxford, Eng. [Box 29.5]
Hubble, E. P., 1934b, "The distribution of extra-galactic nebulae," *Astrophys. J. 79,* 8–76. [§27.1, Box 29.5]
Hubble, E. P., 1936, "Effects of redshifts on the distribution of nebulae," *Astrophys. J. 84,* 517–554. [§27.1]
Hubble, E. P., *cited by name in* §17.3, Boxes 27.7, 29.5. *See also* Mayall 1970.
Hubble, E. P., and M. L. Humason, 1931, "The velocity-distance relation among extra-galactic nebulae," *Astrophys. J. 74,* 43–80. [Box 29.5]

Hughes, R. G., and M. S. Longair, 1967, "Evidence of the isotropy of faint radio sources," *Mon. Not. R. Astron. Soc. 135,* 131-137. [§27.1]

Hughes, V. W., H. G. Robinson, and V. B. Beltran-Lopez, 1960, "Upper limit for the anisotropy of inertial mass from nuclear resonance experiments," *Phys. Rev. Lett. 4,* 342-344. [§38.7]

Hughston, L., 1969, "Multifluid cosmologies," *Astrophys. J. 158,* 987-989. [§27.9]

Humason, M. L. *See* Hubble and Humason 1931.

Huneke, J. C. *See* Wasserburg, Schramm, and Huneke 1969.

Hunter, C., 1967, "Fragmentation," in Ehlers 1967, pp. 169-194. [Boxes 32.2, 32.3]

Huygens, C., 1698, *Cosmotheoros,* The Hague. English translation, *The Celestial Worlds Discovered or, Conjectures Concerning the Inhabitants, Plants and Productions of the Worlds in the Planets,* London, also 1698. [Box 27.7]

Imshennik, V. S. *See* Ivanova, Imshennik, and Nadezhin 1969.

Infeld, Leopold, 1962, volume dedicated to him in connection with his 60th birthday, Feb. 8, 1958: *Recent Developments in General Relativity,* Pergamon Press, Oxford, and PWN—Polish Scientific Publishers, Warsaw.

Infeld, L., ed., 1964, *Conférence Internationale sur les théories relativistes de la gravitation,* Gauthier-Villars, Paris, France.

Infeld, L. *See also* Einstein, Infeld, and Hoffman 1938.

Infeld, L., and J. Plebanski, 1960, *Motion and Relativity,* Pergamon, New York. [§20.6, endpapers]

Infeld, L., and A. Schild, 1949, "On the motion of test particles in general relativity," *Rev. Mod. Phys. 21,* 408-413. [§20.6]

Ingalls, R. P. *See* Shapiro, Ash *et al.* 1971, Shapiro, Smith *et al.* 1971, Shapiro *et al.* 1972.

Ipser, J. R., 1969, "Relativistic, spherically symmetric star clusters, III: Stability of compact isotropic models," *Astrophys. J. 158,* 17-43. [Boxes 24.1, 25.9]

Ipser, J. R., 1970, "Gravitational radiation from slowly rotating, fully relativistic stars," *Astrophys. J. 166,* 175-196. [Ex. 36.4]

Ipser, J. R., 1971, "Electromagnetic test fields around a Kerr-metric black hole," *Phys. Rev. Lett. 27,* 529-531. [§41.11, Ex. 33.14, Box 33.3]

Ipser, J. R. *See also* Brill *et al.* 1972, Fackerell and Ipser 1972, *and* Fackerell, Ipser, and Thorne 1969.

Ipser, J. R., and K. S. Thorne, 1968, "Relativistic, spherically symmetric star clusters, I: Stability theory for radial perturbations," *Astrophys. J. 154,* 251-270. [Box 24.1]

Isaacson, R. A., 1968a, "Gravitational radiation in the limit of high frequency, I: The linear approximation and geometrical optics," *Phys. Rev. 166,* 1263-1271. [§35.13]

Isaacson, R. A., 1968b, "Gravitational radiation in the limit of high frequency, II: Nonlinear terms and the effective stress tensor," *Phys. Rev. 166,* 1272-1280. [§35.13, Ex. 35.14]

Isenberg, J. A., *cited by name in* Figure 25.7.

Israel, W., 1960, "Relativistic theory of shock waves," *Proc. Roy. Soc. A 259,* 129-143. [§22.2, Box 22.2]

Israel, W., 1966, "Singular hypersurfaces and thin shells in general relativity," *Nuovo Cimento 44 B,* 1-14, and correction in *48 B,* 463. [§§21.5, 21.13]

Israel, W., 1967a, "Event horizons in static vacuum space-times," *Phys. Rev. 164,* 1776-1779. [Box 33.1]

Israel, W., 1967b, "Gravitational collapse and causality," *Phys. Rev. 153,* 1388-1393. [Ex. 21.27]

Israel, W., 1968, "Event horizons in static electrovac space-times," *Commun. Math. Phys. 8,* 245-260. [Box 33.1]

Israel, W., 1970, *Differential Forms in General Relativity,* Communications of the Dublin Institute for Advanced Studies, series A, no. 17. [§4.1]

Israel, W., 1971, personal communication. [Box 33.4]

Israel, W., 1972, "The relativistic Boltzman Equation," in O'Raifeartaigh 1972, pp. 201-241. [§22.6]

Israel, W., *cited by name in* Box 34.3. *See also* de la Cruz and Israel 1967.

Ithaca 1965. *See* Cornell University 1965 *or* Ehlers 1967.

Ivanova, L. N., V. S. Imshennik, and D. K. Nadezhin, 1969, "Investigations into the dynamics of supernova explosions," *Sci. Inf. Astron. Council USSR Acad. Sci. 13,* 3-93. [§32.7]

Jablonna 1962, *Relativistic Theories of Gravitation,* proceedings of a conference held in Warsaw and Jablonna, July 1962, edited by L. Infeld, Pergamon Press, Oxford, 1964.

Jackson, J. D., 1962, *Classical Electrodynamics,* Wiley, New York. [§2.1]

Jacobi, K. G. J., *cited by name in* Box 17.2.

Jaffe, J., 1969, "Collapsing objects and the backward emission of light," *Ann. Phys. U.S.A. 55,* 374-390. [§32.3]

Janis, A. I. *See* Couch *et al.* 1968.

Jauch, J., 1968, *Foundations of Quantum Mechanics,* Addison-Wesley, Reading, Mass. [Box 44.5]

Jeans, J., *cited by name in* Box 27.7.

Jones, B. T. J., and P. J. E. Peebles, 1972, "Chaos in Cosmology," *Comm. Astrophys. Space Sci. 4,* 121-128. [§30.1]

Jöstlein, H. *See* Bailey *et al.* 1968 *and* Farley *et al.* 1966.

Johnson, F. R., 1937, *Astronomical Thought in Renaissance England,* John Hopkins Press, Baltimore, [Box 27.7]

Jordan, P., 1955, *Schwerkraft und Weltall,* Vieweg und Sohn, Braunschweig. [§§38.6, 44.6]

Jordan, P., 1959, "Zum gegenwärtigen Stand der Diracschen kosmologischen Hypothesen," *Z. Phys. 157,* 112-121. [§§38.2, 38.6, 44.6, Box 39.1]

Jordan, P., *cited by name in* §§40.1, 40.6, 40.9.

Jordan, P., J. Ehlers, and R. Sachs, 1961, "Exact solutions of the field equations of general relativity, II: Contributions to the theory of pure gravitational radiation," (in German) *Akad. Wiss. Lit. Mainz Abh. Math.-Nat. Kl. 1,* 3-60. [§22.5]

Julian, W. H. *See* Goldreich and Julian 1968.

Jurgens, R. F. *See* Shapiro, Ash *et al.* 1971.

Kac, M., and S. Ulam, 1968, *Mathematics and Logic: Retrospect and Prospects,* Praeger, New York; abridged paperback edition by Pelican Books, Harmondsworth, Middlesex, 1971. [Box 44.5]

Kant, I., 1754, "Whether the earth has undergone an alteration of its axial rotation," *Wöchentliche Frag- und Anzeigungs-Nachrichten,* Kanter, Königsberg, mid-June issue [§1.5]

Kant, I., 1755, *Allgemeine Naturgeschichte und Theories des Himmels, oder Versuch von der Verfassung und dem Mechanischen Ursprunge des ganzen Welt gebäudes nach Newtonischen Grundsätzen Abgehandelt,* Petersen, Konigsberg and Leipzig. Translation by W. Hastie, with introduction by M. K. Munitz, *Universal Natural History and Theory of the Heavens,* Univ. of Michigan Press, Ann Arbor, 1969. [Box 27.7]

Kaplan, S. A., 1949a, "Sverkhplotnye' zvezdy ('Superdense' stars)," *Naukovy Zapiski* (*Sci. Notes Univ. Lwow*) *15,* 109–115. [Box 24.1]

Kaplan, S. A., 1949b, "O krugovykh orbitakh v teorii tyagoteniya Einshteina (On circular orbits in Einstein's theory of gravitation)," *Zh. Eksp. & Teor. Fiz. 19,* 951–952. [Box 24.1]

Kasner, E., 1921a, "Geometrical theorems on Einstein's cosmological equations," *Am. J. Math. 43,* 217–221. [§30.2]

Kasner, E., 1921b, "Finite representation of the solar gravitational field in a space of six dimensions," *Am. J. Math 43,* 130–133. [§23.8]

Kasner, E., 1921c, "The impossibility of Einstein fields immersed in flat space of five dimensions," *Am. J. Math 43,* 126–129. [§23.8]

Kasner, E., *cited by name in* Boxes 27.7, 30.1.

Kaula, W. M. *See* Bender *et al.* 1971.

Keller, J. B., R. M. Lewis, and B. D. Seckler, 1956, "Asymptotic solution of some diffraction problems," *Commun. Pure & Appl. Math. 9,* 207–265. [§22.5]

Keller, L. V., *cited by name in* Box 27.6.

Kelley, J. L., 1955, *General Topology,* Van Nostrand, New York. [Box 34.2]

Kelvin. *See* Thomson, W., Lord Kelvin.

Kemble, E. C., 1937, *The Fundamental Principles of Quantum Mechanics,* McGraw-Hill, New York. Later editions available in paperback. [Box 25.3]

Keres, P. *See* Unt and Keres 1972.

Kepler, J., 1609, *Astronomia nova αιτιολογητος, seu Physica coelestis tradita comentariis de motibus stellae Martis,* Prague. [Box 27.7]

Kepler, J., 1619, *De Harmonice Mundi,* Augsburg. [Box 27.7]

Kerr, R. P., 1963, "Gravitational field of a spinning mass as an example of algebraically special metrics," *Phys. Rev. Lett. 11,* 237–238. [§§33.2, 33.6, Box 33.2]

Kerr, R. P., and A. Schild, 1965, "A new class of vacuum solutions of the Einstein field equations," in *Proceedings of the Galileo Galilei Centenary Meeting on General Relativity, Problems of Energy and Gravitational Waves,* G. Barbera, ed., Comitato Nazionale per le Manifestazione Celebrative, Florence, pp. 222–233. [§33.6]

Khalatnikov, I. M., *cited by name in* Box 27.7. *See also* Belinsky and Khalatnikov 1969a, 1969b, *and* 1970, Belinsky, Khalatnikov, and Lifshitz 1970, Belinsky, Lifshitz, and Khalatnikov 1971, Lifshitz and Khalatnikov 1963a *and* 1963b, *and* Ryan 1972b.

Khalatnikov, I. M., and E. M. Lifshitz, 1970, "General cosmological solutions of the gravitational equations with a singularity in time," *Phys. Rev. Lett. 24,* 76–79. [§34.6, Box 30.1]

Khuri, N., 1957, unpublished calculation quoted in Lindquist and Wheeler 1957. [§25.5]

Killing, W., 1892, "Über die Grundlagen der Geometrie," *J. Reine Angew. Math. 109,* 121–186. [§25.2]

Kilmister, C. W., 1971, "More about the King and the Vizier," *Gen. Relativity & Grav. 2,* 35–36. [Fig. 44.3]

Kirchhoff, G.-R., *cited by name in* Box 38.2.

Kittel, C., 1958, *Elementary Statistical Physics,* Wiley, New York. [§23.4]

Klauder, J., ed., 1972, *Magic Without Magic: John Archibald Wheeler,* W. H. Freeman, San Francisco.

Klauder, L. T., Jr., M. Wakano, J. A. Wheeler, and R. Willey, 1959, "Kinematics of a spherical isotropic uniform universe," in *Onzième Conseil de Physique Solvay: La Structure et l'évolution de l'univers,* Éditions Stoops, Brussels, Belgium. [§29.5]

Klein, F., 1872, "Erlangerprogram: Vergleichende Betrachtungen über neuere geometrischen Forschungen," translation in *Bull. Amer. Math. Soc. 2,* 215–249, 1893. [§2.2]

Klein, M. J., 1971, *Einstein: The Life and Times, by R. W. Clark,* book review, *Science, 174,* 1315–1316. [Box 1.11]

Klein, O., 1966, "Boundary conditions and general relativity" in A. de-Shalit, H. Feshbach, and L. Van Hove, eds., *Preludes in Theoretical Physics in Honor of V. F. Weisskopf,* North Holland, Amsterdam, pp. 23–26. [Box 27.7]

Klein, O., 1971, "Arguments concerning relativity and cosmology," *Science 171,* 339–345. [§§27.10, 27.11, Box 27.7]

Klein, O. *See also* Alfvén and Klein 1962.

Klemperer, W. B. *See* Deutsch and Klemperer 1962.

Kline, M., 1954, "Asymptotic solution of linear hyperbolic partial-differential equations," *J. Rat. Mech. & Anal. 3,* 315–342. [§22.5]

Klüber, H. V., 1960, "The determination of Einstein's light-deflection in the gravitational field of the sun," *Vistas in Astronomy 3,* 47–77, London, Pergamon Press. [Box 40.1]

Kluitenberg, G. A., and S. R. deGroot, 1954, "Relativistic thermodynamics of irreversible processes, III," *Physica 20,* 199–209. [§22.2]

Kneebone, G. T., 1963, *Mathematical Logic and the Foundations of Mathematics: An Introductory Survey,* Van Nostrand, Princeton, N. J. [Box 44.5]

Knight, L. V. *See* Witteborn, Knight, and Fairbank 1967.

Kollros, L., 1956, *in* Carl Seelig, ed. *Helle Zeit- Dunkle Zeit—In Memoriam Albert Einstein,* Europa, Zurich. We thank Professor G. Holton for this quotation. [preface]

Kovács, S., *cited by name in* ex. 35.20.

Kraichnan, R. H., 1955, "Special-relativistic derivation of generally covariant gravitation theory," *Phys. Rev. 55,* 1118–1122. [§§7.1, 18.1]

Kramers, H. A., 1957, *Quantum Mechanics,* North Holland Pub. Co., Amsterdam, Holland. Translated from the German by D. ter Haar. [§21.1]

Krefetz, E., 1967, "The equilibrium of slowly rotating configurations in the post-Newtonian approximation: Corrections to Clairaut's equations," *Astrophys. J. 148,* 589–612. [Ex. 26.1]

Kretschmann, E., 1917, "Über den physikalischen Sinn der Relativitätspostulate, A. Einsteins neue und seine ursprungliche Relativitätstheorie," *Ann. Phys.* (*Germany*) *53,* 575–614. [§§12.5, 17.6]

Kreuzer, L. B., 1968, "Experimental measurement of the equivalence of active and passive gravitational mass," *Phys. Rev. 169,* 1007–1012. [§40.8]

Krogh, C., and R. Baierlein, 1968, "Lunar laser ranging and the Brans-Dicke theory," *Phys. Rev. 175,* 1576–1579. [§40.6]

Krotkov, R. *See* Bertotti, Brill, and Krotkov 1962 *and* Roll, Krotkov, and Dicke 1964.

Kruskal, M. D., 1960, "Maximal extension of Schwarzschild metric," *Phys. Rev. 119,* 1743–1745. [§31.4, Box 31.1]

Kuchař, K., 1968, "Charged shells in general relativity and their gravitational collapse," *Czech J. Phys. B 18,* 435–463. [§21.13]

Kuchař, K., 1970, "Ground state functional of the linearized gravitational field," *J. Math. Phys. 11,* 3322–3334. [§43.4]

Kuchař, K., 1971a, "Canonical quantization of cylindrical gravitational waves," *Phys. Rev. D 4,* 955–986. [§§21.1, 21.10]

Kuchař, K., 1971b, "Gravitational action functional reexpressed by means of intrinsic and extrinsic curvature," unpublished lecture given in the course *Relativity* at Princeton University. [§§21.6, 21.11, Ex. 21.10]

Kuchař, K., 1972, "A bubble-time canonical formalism for geometrodynamics," *J. Math. Phys. 13,* 768–781. [§§21.1, 21.10, 21.11, Ex. 21.24]

Kuchař, K., *cited by name in* §21.1, Ex. 13.7, 21.10.

Kuhlthau, A. R. *See* Rose *et al.* 1969.

Kundt, W., 1971, "Cosmological density fluctuations during Hadron stage," in Sachs 1971, pp. 365–372. [§28.1]

Kundt, W. *See also* Ehlers and Kundt 1962.

Kundt, W., and E. T. Newman, 1968, "Hyperbolic differential equations in two dimensions," *J. Math. Phys. 9,* 2193–2210. [§§35.8, 38.8]

Künzle, H. P., 1971, "On the Spherical Symmetry of a Static Perfect Fluid," *Commun. Math. Phys. 20,* 85–100. [§23.1]

Kuper, C. G., and A. Peres, eds., 1971, *Relativity and Gravitation,* Gordon and Breach, New York.

Kursunoglu, B., 1962, *Modern Quantum Theory,* W. H. Freeman, San Francisco. [Box 21.1]

Kustaanheimo, P., 1966, "Route dependence of the gravitational red shift," *Phys. Lett. 23,* 75–77. [§39.1]

Lagrange, J. L., *cited by name in* Box 17.2.

Lamb, W. E., Jr., and R. C. Retherford, 1947, "Fine structure of the hydrogen atom by a microwave method," *Phys. Rev. 72,* 241–243. [Fig. 43.3, Box 17.2 part 6]

Lambert, J. H., 1761, *Kosmologischen Briefe,* Augsburg; see his *The System of the World,* translated, with modifications, from the French by J. Jacque, London, 1800; reprinted in part in Munitz 1957. [Box 27.7]

Lanczos, K., 1922, "Bemerkung zur de Sitterschen Welt," *Phys. Z. 23,* 539–543. [§21.13]

Lanczos, K., 1924, "Flächenhafte Verteilung der Materie in der Einsteinschen Gravitationstheorie," *Ann. Phys. (Germany) 74,* 518–540. [§21.13]

Lanczos, K., 1970, *The Variational Principles of Mechanics,* Univ. of Toronto Press, Toronto, 4th ed. [Box 21.1]

Landau, L. D., 1932, "On the theory of stars," *Phys. Z. Sowjetunion 1,* 285–288. [Box 24.1]

Landau, L. D., and E. M. Lifshitz, 1959, *Fluid Mechanics,* Pergamon Press, London. [Ex. 22.7]

Landau, L. D., and E. M. Lifshitz, 1960, *Mechanics,* Addison-Wesley, Reading, Mass. [Box 22.4]

Landau, L. D., and E. M. Lifshitz, 1962, *The Classical Theory of Fields,* translated by M. Hamermesh, Addison-Wesley, Reading, Mass. [§§20.3, 20.5, 36.1, endpapers]

Landau, L. D. and E. M. Lifshitz, 1971, *The Classical Theory of Fields,* 3d rev. English ed., Addison-Wesley, Reading, Massachusetts, and Pergamon, London. [endpapers]

Langer, W. D., and A. G. W. Cameron, 1969, "Effects of hyperons on the vibrations of neutron stars," *Astrophys. & Space Sci. 5,* 213–253. [§24.3]

Lao-tzu, 3d century B.C. *Tao Tê Ching,* Poem 1. Translated as *The Way of Life,* R. B. Blakney, New American Library, 1955; translation revised by Janet Ten Broeck. [Epigraph to Chap. 18]

Langevin, P. *See* Einstein 1912.

Laplace, P. S., 1795, *Le Système du monde,* vol. II, Paris. Published in English as *The System of the World,* W. Flint, London, 1809. [Fig. 24.1, Box 24.1, Epigraph to chap. 33]

Laplace, P. S., *cited by name in* Box 17.2.

Lascoux, J., 1968, "Perturbation theory in quantum field theory and homology," in DeWitt and Wheeler 1968, 354–419. [Ex. 4.5]

Laurence, D. H. *See* Boni, Russ, and Laurence 1960.

Lawson, R. W., *cited by name in* Box 27.7.

Leavitt, H., *cited by name in* Box 27.7.

Lebovitz, N. R. *See* Chandrasekhar and Lebovitz 1968.

Lecat, M., 1924, *"Bibliographie de la relativité,"* Lamertine, Bruxelles, cited in preface of present bibliography.

Ledoux, P., and Th. Walraven, 1958, "Variable stars," *Handbuch der Physik 51,* S, Flügge ed., Springer-Verlag, Berlin, pp. 353–604. [Box 24.2]

Lefschetz, S. 1949, *Introduction to Topology,* Princeton Univ. Press, Princeton, N.J. For a fuller account of fixed-point theorems, see Dunford and Schwartz 1958. [§36.1]

Leibniz, G. W., 1714, *La Monadologie.* English translation available in several books; parts included in P. P. Wiener, ed., *Leibniz Selections,* Scribners, New York, 1951. [§44.6]

Leibniz, G. W., 1961, untitled essay in his *Philosophischen Schriften, vol. 7,* ed. by C. J. Gerhardt, Olms Verlagsbuchhandlung, Hildesheim, p. 200. [Epigraph to Chap. 21]

Leibniz, G. W., *quoted at end of* Chap. 44.

Leighton, R. B., 1959, *Principles of Modern Physics,* McGraw-Hill, New York. [§2.1]

Leighton, R. B. *See also* Feynman, Leighton, and Sands 1964.

Lemaître, G., 1933a, "Formation of nebulae in the expanding universe," *Acad. Sci., Paris, Comptes Rend. 196,* 1085–1087. [§30.6]

Lemaître, G., 1933b, "Spherical condensations in the expanding universe," *Acad. Sci., Paris, Comptes Rend. 196,* 903–904. [§30.6]

Lemaître, G., 1933c, "L'univers en expansion," *Ann. Soc. Sci. Bruxelles I A53,* 51–85. [Box 31.1]

Lemaître, G., 1949, cited in Schilpp 1949, p. 440. [Epigraph to Chap. 36]

Lemaître, G., *cited by name in* Box 27.7.
Lenin, V. I., 1927, *Materialism and Empiriocriticism: Critical Comments on a Reactionary Philosophy,* International Publications, New York, 2d English ed. of *Materializm' i Empiriokrititsizm': Kriticheskia Zametki ob' Odnoi Reaktsionnoi Filosofii,* Zveno Press, Moscow, 1909. [§21.12]
Lense, J. *See* Thirring and Lense 1918.
Les Houches, France, Ecole d'ete de physique theorique: 1963, *see* DeWitt and DeWitt 1964b; 1966, *see* DeWitt, Schatzman, and Veron 1967 (third volume of three); 1972, *see* DeWitt and DeWitt 1973.
Leutwyler, H., 1968, "Boundary conditions for the state functional in the quantum theory of gravity," in DeWitt and Wheeler 1968, pp. 309–317. [§43.3]
Lewis, R. M., 1958, "Asymptotic expansion of steady-state solutions of symmetric hyperbolic linear differential equations," *J. Rat. Mech. & Math. 7,* 593–641. [§22.5]
Lewis, R. M. *See also* Keller, Lewis, and Seckler 1956.
Liang, E. *See* Eardley, Liang, and Sachs 1972.
Lichnerowicz, A., 1944, "L'intégration des équations de la gravitation relativiste et le problème des *n* corps," *J. Math. Pures & Appl. 23,* 37–63. [§21.11]
Lichnerowicz, A., 1955, *Théories relativistes de la gravitation et de l'électromagnétisme,* Masson, Paris. [endpapers]
Lichnerowicz, A., 1961a, "Anticommutateur du champ spinoriel en relativité générale," *Acad. Sci., Paris, Comptes Rend. 252,* 3742–3744. [Box 44.3]
Lichnerowicz, A., 1961b, "Anticommutateur pour le champ correspondant au spin 3/2 sur un espace d'Einstein," *Acad. Sci., Paris, Comptes Rend. 253,* 940–942. [Box 44.3]
Lichnerowicz, A., 1961c, "Théorie de Petiau-Duffin-Kemmeren," *Acad. Sci., Paris, Comptes Rend. 253,* 983–985. [Box 44.3]
Lichnerowicz, A., 1962, "Radiations en relativité générale," especially "Le tenseur de superénergie," in CNRS 1962, pp. 93–106. [Ex. 15.2]
Lichnerowicz, A., 1964, "Propagateurs, commutateurs, et anticommutateurs en relativité générale," in DeWitt and DeWitt 1964. [Box 44.3, Ex. 15.2]
Lichnerowicz, A., 1967, *Relativistic Hydrodynamics and Magnetohydrodynamics; Lectures on the Existence of Solutions,* W. A. Benjamin, New York. [§§22.2, 22.3, Box 22.2]
Lichnerowicz, A., 1971, "Ondes des choc, ondes infinitesimales et rayons en hydrodynamique et magnetohydrodynamique relativistes," in Cattaneo 1971, pp. 88–203. [§22.2, Box 22.2]
Lichtenberg, D. B., 1965, *Meson and Baryon Spectroscopy,* Springer-Verlag, New York. [§38.4]
Lifshitz, E. M., *cited by name in* Box 27.7. *See also* Belinsky, Khalatnikov, and Lifshitz 1970, Belinsky, Lifshitz, and Khalatnikov 1971, Khalatnikov and Lifshitz 1970, *and* Landau and Lifshitz, 1959, 1960, 1962, and 1971.
Lifshitz, E. M., and I. M. Khalatnikov, 1963a, "Investigations in relativistic cosmology," in *Advances in Physics 12,* 185–249; translated from the Russian by J. L. Beeby. [§§30.6, 34.6]
Lifshitz, E. M., and I. M. Khalatnikov, 1963b, "Problems of relativistic cosmology," *Usp. Fiz. Nauk 80,* 391–438. English translation in *Sov. Phys.—Uspekhi 6,* 495–522, 1964. [§§30.6, 34.6]
Lifshitz, E. M., and I. M. Khalatnikov, 1970, "Oscillatory approach to singular point in the open cosmological model," *Zh. Eksp. & Teor. Fiz. Pis'ma., 11,* 200–203. English translation in *Sov. Phys.—JETP Lett. 11,* 123–125 (1971).
Lighthill, M. J., 1958, *Introduction to Fourier Analysis and Generalized Functions,* Cambridge University Press, Cambridge, England. [§4.7]
Lilley, A. E. *See* Dieter *et al.* 1962.
Lin, C. C., L. Mestel, and F. H. Shu, 1965, "The gravitational collapse of a uniform spheroid," *Astrophys. J. 142,* 1431–1446. [Boxes 32.2, 32.3]
Lindquist, R. W., 1960, *The Two-Body Problem in Geometrodynamics,* doctoral dissertation, Princeton University, Princeton, N.J. [Ex. 32.1]
Lindquist, R. W., 1966, "Relativistic transport theory," *Ann. Phys.* (*USA*) *37,* 487–518. [§22.6]
Lindquist, R. W. *See also* Boyer and Lindquist 1967.
Lindquist, R. W., and J. A. Wheeler, 1957, "Dynamics of a lattice universe by the Schwarzschild cell method," *Rev. Mod. Phys. 29,* 432–433. [§27.10, Box 25.2, Fig. 25.3, 27.3]
Liouville, J., 1837, "Second mémoire sur le développement des fonctions ou parties des fonctions en séries . . . . ," Liouville's *J. de Math. Pures et Appliquées 2,* 16–35. [§22.5]
Loar, H. H. *See* Durbin, Loar, and Havens 1952.
London 1965, International Conference on Relativistic Theories of Gravitation, King's College, University of London, July 1965. Proceedings multilithed in 2 volumes by King's College.
Longair, M. S. *See* Hughes and Longair 1967.
Lorentz, H. A., *cited by name in* §20.6 *and* Box 16.4.
Lorentz, H. A., A. Einstein, H. Minkowski, and H. Weyl, 1923, *The Principle of Relativity: A Collection of Original Memoirs,* Methuen, London. Paperback reprint, Dover, New York.
Lowry, R. A. *See* Rose *et al.* 1969.
Lukash, V. N. *See* Doroshkevich, Lukash, and Novikov 1971.
Lurie, D., and N. Mukunda, eds., 1963, *Proceedings of Seminar on Unified Theories of Elementary Particles,* University of Rochester report URPA-11, Rochester, N.Y.
Lynden-Bell, D., 1967, "Statistical mechanics of violent relaxation in stellar systems," *Mon. Not. R. Astron. Soc. 136,* 101–121. [Box 33.3]
Lynden-Bell, D., 1969, "Galactic nuclei as collapsed old quasars," *Nature 223,* 690–694. [Box 33.3]
Lynden-Bell, D., and M. Rees, 1971, "On quasars, dust, and the galactic center," *Mon. Not. R. Astron. Soc. 152,* 461–475. [Box 33.3]
Lynds, C. R., *cited by name in* Box 27.7.
MacCallum, M. A. H. *See* Penrose and MacCallum 1973.
MacDonald, G. J. F. *See* Bender *et al.* 1971 *and* Munk and MacDonald 1960.
Mach, E., 1912, *Die Mechanik in Ihrer Entwicklung Historisch-*

Kritisch Dargestellt, Brockhaus, Leipzig. English translation by T. J. McCormack, published in 1960 as *The Science of Mechanics* with an introduction by Karl Menger, Open Court, La Salle, Ill. [§§21.1, 21.12]

Mach, E., *cited by name in* Box 27.1.

Machado, A., 1940, "Proverbios y cantares XXIX," in his *Obras,* Editorial Seneca, Mexico, pp. 235–236. [Epigraph to Chap. 43]

MacKenzie, R. E. *See* Auslander and MacKenzie 1963.

Mackey, M. B., *cited by name in* Box 27.7.

Manasse, F. K., and C. W. Misner, 1963, "Fermi Normal Coordinates and Some Basic Concepts in Differential Geometry," *J. Math. Phys. 4,* 735–745. [§13.6]

Mandelstam, S. *See* Yourgrau and Mandelstam 1968.

Mann, T., 1937, *Freud, Goethe, Wagner,* Knopf, New York. Essay on Freud published originally as *Freud und die Zukunft* (*Vortrag Gehalten in Wien am 8 Mai 1936 zur Feier von Sigmund Freuds 80 Geburtstag*), Bormann-Fischer, Wien (1936). [§44.5]

Manukin, A. B. *See* Braginsky, Manukin, and Tikhonov 1970.

Marconi, G. W., *cited by name in* Preface.

Marinus of Naples, *cited by name in* Fig. 1.8.

Markus, L. *See* Auslander and Markus 1959.

Marle, C., 1969, "Sur l'établissement des équations de l'hydrodynamique des fluides relativistes dissipatifs, I et II," *Ann. Inst. Henri Poincare, 10,* 67–126 and 127–194. [§22.6, Ex. 22.7]

Marsden, J. B. *See* Fischer and Marsden 1972.

Martin, W. L. *See* Anderson *et al.* 1971.

Marzke, R. F., and J. A. Wheeler, 1964, in Chiu and Hoffman 1964. [Box 16.4]

Mathews, J. *See also* Peters and Mathews 1963.

Mathews, J., and R. Walker, 1965, *Mathematical Methods of Physics,* W. A. Benjamin, New York. [Box 26.1]

Matzner, R. A., 1968, "The present temperature of primordial black-body gravitational radiation," *Astrophys. J. 154,* 1123–1125. [§27.10]

Matzner, R. A. *See also* Clemens and Matzner 1967 *and* Fletcher *et al.* 1967.

Matzner, R. A., and C. W. Misner, 1972, "Dissipative effects in the expansion of the universe," *Astrophys. J. 171,* 415–432. [§§28.2, 30.4]

Maupertuis, P. L. M. de, 1746, "Recherche des loix du mouvement," *Acad. R. Sci. Berlin.* [Epigraph to Chap. 21]

Mavrides, M., *quoted at end of* Chap. 44.

Maxwell, J. C., *cited by name in* Preface, Box 4.6.

May, M., and R. H. White, 1966, "Hydrodynamic calculations of general-relativistic collapse," *Phys. Rev. 141,* 1232–1241. [§§32.7, 42.7, Box 33.3, Ex. 32.7]

May, M., and R. H. White, 1967, "Stellar dynamics and gravitational collapse," *Meth. Computat. Phys. 73,* 219–258. [§22.2, Boxes 22.2, 33.3]

Mayall, N. U., 1970, "Edwin Powell Hubble," *Biograph. Mem. Nat. Acad. Sci., 41,* 176–214. [Box 29.5]

Mayall, N. U., and Oort, J. H., 1942, "Further data bearing on the identification of the Crab Nebula with the supernova of 1054 A.D., part II: The astronomical aspects," *Proc. Astron. Soc. Pacific 54,* 95–104. [Box 24.1]

Mazet, E. *See* Berger, Gauduchon, and Mazet 1971.

McCrea, W. H., *cited by name in* Box 27.7. *See also* Bondi and McCrea 1960.

McDonald, E., *cited by name in* Fig. 41.6.

McVittie, G. C., 1956, *General Relativity and Cosmology,* Chapman and Hall, London. [endpapers]

McWhirter, N., and R. McWhirter, 1971, *The Guinness Book of Records,* Guinness Superlatives, Ltd., Enfield, Middlesex [Chap. 12 epigraph]

Meer, S. van der. *See* Farley *et al.* 1966.

Meixner, J., and H. G. Reik, 1959, "Thermodynamik der irreversiblen Prozesse," *Handbuch der Physik, 3, Teil 2,* 413–505. [§22.2]

Meixner, J., and F. W. Schäfke, 1954, *Mathieusche Funktionen und Sphäroidfunktionen,* Springer-Verlag, Berlin, Germany. [Ex. 33.14]

Melbourne, W. G. *See* Ohandley *et al.* 1969.

Meltzer, D. W. *See* Bardeen, Thorne, and Meltzer 1966.

Meltzer, D. W., and K. S. Thorne, 1966, "Normal modes of radial pulsation of stars at the end point of thermonuclear evolution," *Astrophys. J. 145,* 514–543. [Box 26.1]

Menger, Karl, 1949, see Schilpp 1949, pp. 466–467. [Epigraph to Chap. 2, Epigraph to Chap. 9]

Menskii, M. B. *See* Braginsky and Menskii 1971.

Mercier, A., 1953, *Analytical and Canonical Formalism in Physics,* North-Holland, Amsterdam, Holland. Paperback reprint, Dover, New York, 1963. [Box 21.1]

Mercier, A., and M. Kervaire, 1956, *Jubilee of Relativity Theory,* proceedings of a conference in Bern, Switzerland, July 11–16, 1955, Birkhäuser, Bern.

Merzbacher, E., 1961, *Quantum Mechanics,* Wiley, New York. [§16.3]

Messiah, A., 1961, *Quantum Mechanics,* translated from the French by G. M. Temmer, North-Holland, Amsterdam, Holland. [Ex. 3.12]

Mestel, L. *See* Lin, Mestel, and Shu 1965.

Mielnik, B., and J. Plebanski, 1962, "A study of geodesic motion in the field of Schwarzschild's solution," *Acta Phys. Polon. 21,* 239–268. [§25.5]

Mikola, S. *See* Fejér and Mikola 1918.

Milne, E. A., *cited by name in* Box 27.7.

Milnor, J., 1962, "A survey of cobordism theory," *Enseignement Math. 8,* 16–23. [Box 44.3]

Milnor, J., 1963, "Spin structures on manifolds," *Enseignement Math. 9,* 198–203. [Box 44.3]

Milnor, J., 1965a, "On the Stiefel-Whitney numbers of complex manifolds and of spin manifolds," *Topology 3,* 223–230. [Box 44.3]

Milnor, J., 1965b, "Remarks concerning spin manifolds," in Cairns 1965, p. 55. [Box 44.3]

Milton, J., 1665, *Paradise Lost,* Book V, Lines 623–625, in *The Complete Poems of John Milton,* Union Library Association of New York, 1936. [Epigraph to Chap. 25]

Minkowski, H., 1908, address delivered at the 80th Assembly of German Natural Scientists and Physicians, Cologne. English translation, "Space and time," in *The Principle of Relativity,* Dover, New York, 1923. [§1.5]

Minkowski, H. *See also* Lorentz *et al.* 1923.

Minkowski, R., 1942, "The Crab Nebula," *Astrophys. J. 96,* 199–213. [Box 24.1]

Minkowski, R., *cited by name in* Box 27.7.

Minkowski, R., and O. C. Wilson, 1956, "Proportionality of nebular redshifts to wave length," *Astrophys. J. 123,* 373–376. [§38.6]

Misner, C. W., 1957, "Feynman quantization of general relativity," *Rev. Mod. Phys. 29,* 497–509. [Box 21.1]

Misner, C. W., 1964a, "Differential geometry and differential topology," in DeWitt and DeWitt 1964b, pp. 881–929. [Box 8.1]

Misner, C. W., 1964b, "Waves, Newtonian fields, and coordinate functions," in Infeld 1964. [§35.8]

Misner, C. W., 1965, "Relativistic equations for spherical gravitational collapse with escaping neutrinos," *Phys. Rev. B137,* 1360–1364. [Box 23.1]

Misner, C. W., 1968, "The isotropy of the universe," *Astrophys. J. 151,* 431–457. [§§28.2, 30.3]

Misner, C. W., 1969a, "Gravitational collapse," in Chrétien, Deser, and Goldstein 1969, vol I. [§§12.1, 32.5, Boxes 8.1, 31.2, Ex. 31.8, endpapers]

Misner, C. W., 1969b, "Mixmaster universe," *Phys. Rev. Lett. 22,* 1071–1074. [§§21.10, 28.2, 30.6, 34.6, Box 30.1]

Misner, C. W., 1969c, "Absolute zero of time," *Phys. Rev. 186,* 1328–1333. [§§28.3, 34.6]

Misner, C. W., 1970, "Classical and quantum dynamics of a closed universe," in Carmeli, Fickler, and Witten 1970, pp. 55–79. [Box 30.1]

Misner, C. W., 1972a, "Minisuperspace," in Klauder 1972. [§43.3, Box 30.1]

Misner, C. W., 1972b, "Stability of Kerr black holes against scalar perturbations," *Bull. Am. Phys. Soc. (II) 17,* 472. [Ex. 33.14]

Misner, C. W., 1973, "A minsuperspace example: The Gowdy T^3 cosmology," Caltech Orange Aid Preprint. Submitted to *Phys. Rev.* [§43.3]

Misner, C. W., *cited by name in* §§21.1, 35.8, Boxes 17.2 part 2′, 17.3, 27.7. *See also* Arnowitt, Deser, and Misner 1962, Beckedorff and Misner 1962, Chrzanowski and Misner 1973, Manasse and Misner 1963, *and* Matzner and Misner 1972.

Misner, C. W., and D. H. Sharp, 1964, "Relativistic equations for adiabatic, spherically symmetric gravitational collapse," *Phys. Rev. B 136,* 571–576. [§§21.13, 22.2, Ex. 32.7, Box 23.1]

Misner, C. W., and A. H. Taub, 1968, "A singularity-free empty universe," *Zh. Eksp. & Teor. Fiz. 55,* 233. English original in *Sov. Phys.—JETP 28,* 122, 1969. [§§32.5, 34.6]

Misner, C. W., and J. A. Wheeler, 1957, "Classical physics as geometry: Gravitation, electromagnetism, unquantized charge, and mass as properties of curved empty space," *Ann. Phys. (U.S.A.) 2,* 525–603, reprinted in Wheeler 1962. [§§20.6, 21.8, 21.9, Fig. 44.1, Boxes 4.1, 4.4, 15.1, 21.2, Ex. 20.7]

Misner, C. W., and J. A. Wheeler, 1972, "Conservation laws and the boundary of a boundary," in Shelest 1972, pp. 338–351. [§15.3]

Moffet, A., *cited by name in* Box 40.1.

Møller, C., 1952, *The Theory of Relativity,* Oxford Univ. Press, London; 2d ed., 1972. [endpapers]

Møller, C., ed., 1962, *Evidence for Gravitational Theories,* Academic Press, New York.

Moncrief, V. E. *See* Berger *et al.* 1972.

Morduch, G. E. *See* Whitrow and Morduch 1965.

Morehead, J. C. *See* Gauss 1827.

Morette, C., 1951, "On the definition and approximation of Feynman's path integrals," *Phys. Rev. 81,* 848–852. [Box 21.1]

Morette, C. *See also* DeWitt, C.

Morgan, T., and K. S. Thorne, 1973, in preparation. [Box 32.3]

Morganstern, R. E., 1973, "Observational constraints imposed by Brans-Dicke cosmologies," *Phys. Rev. D 7,* 1570–1579. [§28.4]

Moritz, B. K., 1969, *A Study of the General Relativistic Dynamics of the Klein-Alfvén Cosmology,* doctoral dissertation, University of Maryland [§27.11, §28.4]

Morse, M., 1934, *The calculus of variations in the large,* American Mathematical Society, Providence, R.I. [Fig. 13.2]

Morse, M., and S. S. Cairns, 1969, *Critical Point Theory in Global Analysis and Differential Topology: an Introduction,* Academic Press, New York. [Fig. 13.2]

Moser, J. K. *See* Siegel and Moser 1971.

Mottleson, B. R. *See* Bohr and Mottleson 1969.

Muhleman, D. O. *See* Anderson *et al.* 1971.

Muhleman, D. O., R. D. Ekers, and E. B. Fomalont, 1970, "Radio interferometric test of the general relativistic light bending near the sun," *Phys. Rev. Lett. 24,* 1377–1380. [Box 40.1]

Muhlemann, D. O., and P. Reichley, 1964, "Effects of general relativity on planetary radar distance measurements," *JPL Space Programs Summary 4,* no. 37-29, p. 239. [§40.4]

Muhlemann, D. O., and P. Reichley, 1965, "General relativistic effects on radar propagation," *JPL Space Programs Summary 4,* no. 37-31, p. 342. [§40.4]

Mukunda, N. *See* Lurie and Mukunda 1963.

Mulholland, J. D. *See* Bender *et al.* 1971 *and* Ohandley *et al.* 1969.

Munitz, M. K., 1957, *Theories of the Universe,* Free Press, New York. [Box 27.7]

Munk, W. H., and G. J. F. MacDonald, 1960, *The Rotation of the Earth; a Geophysical Discussion,* Cambridge University Press, Cambridge, England. [Fig. 1.8]

Nadezhin, D. K. *See* Ivanova, Imshennik, and Nadezhin 1969.

Nagel, E., and J. R. Newman, 1958, *Gödel's Proof,* New York Univ. Press, New York.

NASL, 1969, *Proceedings of the 1969 Symposium on Unconventional Inertial Sensors (U) Vol. I,* Naval Applied Science Laboratory, Brooklyn, New York.

Nasr, H., 1964, *Three Muslim sages: Avicenna, Suhrawardi, Ibn Arabi,* Harvard University Press, Cambridge, Mass. [Box 27.7]

Navier, L., *cited by name in* Box 17.2 part 6.

Nazarenko, V. S. *See* Braginsky and Nazarenko 1971.

Ne'eman, Y., 1965, "Expansion as an energy source in quasistellar radio sources," *Astrophys. J. 141,* 1303–1305. [§34.3]

Nester, J. M., *cited by name in* §35.1.

Neugebauer, O., 1959, *Greek Horoscopes,* with H. B. van Hösen, American Philosophical Society, Philadelphia, Pa. [Fig. 1.8]

Neugebauer, P. V., 1927, "Zür astronomischen Chronologie," *Astron. Nachr. 230,* 57–64. [Fig. 1.8]

Neugebauer, P. V., 1929, *Astronomische Chronologie,* 2 vols, De Gruyter, Berlin. [Fig. 1.8]

Neugebauer, P. V., 1930, "Schoch Todesanzeige," *Astron. Nachr. 237,* 221–224. [Fig. 1.8]

Neugebauer, P. V., and C. Schoch, 1930, "Zür astronomischen Chronologie, II," *Astronomische Abhandlungen. Erganzungshefte zu den Astron. Nachrichten 8,* no. 2, 24–32. [Fig. 1.8]

Newman, E. T. *See also* Couch and Newman 1972, Couch *et al.* 1968, *and* Kundt and Newman 1968.

Newman, E. T., E. Couch, K. Chinnapared, A. Exton, A. Prakash, and R. Torrence, 1965, "Metric of a rotating, charged mass," *J. Math. Phys. 6,* 918–919. [§33.2]

Newman, E. T., and R. Penrose, 1965, "10 exact gravitationally-conserved quantities," *Phys. Rev. Lett. 15,* 231–233. [Ex. 32.11]

Newman, J. R. *See* Nagel and Newman 1958.

Newton, I., 1687, *Philosophiae naturalis principia mathematica,* 1st ed., Streater, London; final ed., in English translation by A. Motte, 1729, revised by A. Cajori, *Sir Isaac Newton's Mathematical Principles of Natural Philosophy and His System of the World,* Univ. of California Press, Berkeley and Los Angeles, 1934; paperback, 1962. [Boxes 1.10, 21.1, 27.7]

Newton, I., 1730, *Opticks, or a Treatise of the Reflection, Refraction, Inflection, & Colours of Light,* 4th ed.; reprint, ed. by D. H. D. Roller, Dover, New York, 1952. [Box 1.10]

Newton, I., *cited by name in* §§1.1, 1.3, 12.1, 12.2, Boxes 1.2, 12.2, 12.3, 12.6, Ex. 12.7, 12.8, 13.4, 17.2 part 6.

Newton, R. *See* Gilbert and Newton 1970.

Neyman, J., 1951, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability,* Univ. of California Press, Berkeley, Calif.

Ni, W.-T., 1970, work reported by K. S. Thorne at the NASA-ESRO conference on experimental tests of gravitation theories, California Institute of Technology, November 1970; *see* Thorne, Ni, and Will 1971. [§39.2, Box 39.1]

Ni, W.-T., 1972, "Theoretical frameworks for testing relativistic gravity, IV: A compendium of metric theories of gravity and their post-Newtonian limits," *Astrophys. J. 176,* 769–796. [§§17.6, 38.2, 39.2, 40.1, 40.5, 40.8, Boxes 39.1, 39.2, 39.5]

Ni, W.-T. *See also* Thorne, Ni, and Will 1971.

Ni, W.-T., 1973, "Relativistic stellar stability: an empirical approach," *Astrophys. J. 181,* 939–956. [Box 39.5, Ex. 39.7]

Nicetas, *cited by name in* Box 27.7.

Nickerson, H. K., D. C. Spencer, and N. E. Steenrod, 1959, *Advanced Calculus,* Van Nostrand, Princeton, New Jersey. [§4.1]

Nijenhuis, A., 1952, "Theory of the geometric object," dissertation, Amsterdam. [§2.2]

Nizer, L., 1972, quoted in the *New York Times* February 7, 1972. [Epigraph to Chap. 40]

Noerdlinger, P. D., and V. Petrosian, 1971, "The effect of cosmological expansion on self-gravitating ensembles of particles," *Astrophys. J. 168,* 1–9. [§27.5]

Noonan, T. W. *See* Robertson and Noonan 1968.

Nordstrøm, G., 1913, "Zur Theorie der Gravitation vom Standpunkt des Relativitätsprinzips," *Ann. Phys. (Germany) 42,* 533–554. [§17.6]

Nordstrøm, G., 1918, "On the energy of the gravitational field in Einstein's theory," *Proc. Kon. Ned. Akad. Wet. 20,* 1238–1245. [Ex. 31.8]

Nordstrøm, G., *cited by name in* §§38.2, 42.7.

Nordtvedt, K., 1968a, "Equivalence principle for massive bodies, I: Phenomenology," *Phys. Rev. 169,* 1014–1016. [§§38.2, 40.9]

Nordtvedt, K., 1968b, "Equivalence principle for massive bodies, II: Theory," *Phys. Rev. 169,* 1017–1025. [§§38.2, 39.3, 40.9, Box 39.5]

Nordtvedt, K., Jr., 1969, "Equivalence principle for massive bodies including rotational energy and radiation pressure," *Phys. Rev., 180,* 1293–1298. [§§38.2, 39.3]

Nordtvedt, K., 1970, "Solar system Eötvos experiments," *Icarus 12,* 91–100. [§40.8, Box 39.5]

Nordtvedt, K., 1971a, "Equivalence principle for massive bodies, IV: Planetary bodies and modified Eötvos-type experiments," *Phys. Rev. D 3,* 1683–1689. [§§40.8, 40.9, Box 39.5]

Nordtvedt, K., 1971b, in Davies 1971, pp. 32–37. [§§39.5, 40.9, Box 39.5]

Nordtvedt, K., Jr. *See also* Will and Nordtvedt 1972.

Nordtvedt, K., Jr., and C. M. Will, 1972, "Conservation laws and preferred frames in relativistic gravity, II: Experimental evidence to rule out preferred-frame theories of gravity," *Astrophys. J. 177,* 775–792. [§§39.2, 40.5, 40.8, Box 39.5]

North, J. D., 1965, *The Measure of the Universe,* Clarendon Press, Oxford. [Box 27.7]

Novikov, I. D., 1963, doctoral dissertation, Shternberg Astronomical Institute, Moscow. [§31.4]

Novikov, I. D., 1964, "Delayed explosion of a part of the Friedmann universe, and quasars," *Astron. Zh. 41,* 1075–1083. English translation in *Sov. Astron.-Astron. J. 8,* 857–863. [§34.3]

Novikov, I. D., 1966a, "Change of relativistic collapse into anticollapse and kinematics of a charged sphere," *Zh. Eksp. & Teor. Fiz. Pis'ma 3,* 223–227. English translation in *Sov. Phys.—JETP Lett. 3,* 142–144. [§34.6, Fig. 34.4]

Novikov, I. D., 1966b, "The replacement of relativistic gravitational contraction by expansion, and the physical singularities during contraction," *Astron. Zh. 43,* 911–918. English translation in *Sov. Astron.—A. J. 10,* 731–737. [§34.6, Fig. 34.4]

Novikov, I. D. *See also* Doroshkevich, Lukash, and Novikov 1971, Doroshkevich, Zel'dovich, and Novikov 1965, *and* Zel'dovich and Novikov 1966, 1971, *and* 1974.

Novikov, I. D., and K. S. Thorne, 1973, "Astrophysics of black holes," in DeWitt and DeWitt 1973, pp. 343–450. [Box 33.3]

Nutku, Y. *See* Berger *et al.* 1972 *and* Chandrasekhar and Nutku 1969.

O'Connell, R. F., 1972, "Present status of the theory of the relativistic gyroscope experiment," *Gen. Rel. & Grav. 3,* 123–133. [§40.7, Ex. 40.8]

Ogorodnikov, K. F., 1965, *Dynamics of Stellar Systems,* ed. by A. Beer, Macmillan, New York. [§25.7]

Ohandley, D. A., D. B. Holdridge, W. G. Melbourne, and J. D. Mulholland, 1969, *J.P.L. Development Ephemeris No. 69,* J.P.L. Technical report 32-1465, Caltech Jet Propulsion Laboratory, Pasadena. [§40.1]

Ohanian, H. C. *See* Belasco and Ohanian 1969.

Olbers, H. W. M., 1826, "Über die Durchsichtigkcit des Weltraumes" *Bode Jahrbuch, 110;* substantially the same article in *Bibliothèque Universelle* for February 1826, and in *Edinburgh New Philosophical Journal* for April–October 1826, 141. [Box 27.7]

Omnes, R. L., 1969, "Possibility of matter-antimatter separation at high temperature," *Phys. Rev. Lett. 23,* 38–40. [§27.10]

O'Murchadha, N., and J. W. York, Jr., 1973, personal communication; *see also* O'Murchadha, N., "Existence and uniqueness of solutions of the Hamiltonian constraint of general relativity," doctoral dissertation, Princeton Univ., February 5, 1973. [§21.11]

O'Neill, B., 1966, *Elementary Differential Geometry,* Academic Press, New York. [Box 8.1]

Oort, J. H., 1958, "Distribution of galaxies and the density of the universe," in *Onzième Conseil de Physique Solvay: La Structure et l'évolution de l'univers,* Éditions Stoops, Brussels, Belgium. [§29.6]

Oort, J. H., 1970, "Matter far from the galactic plane associated with spiral arms," in *Spiral Structure of our Galaxy, Basel, Switzerland, 29 Aug.–4 Sep. 1969,* D. Reidel, Dordrecht, Netherlands. [§28.2]

Oort, J. H., *cited by name in* Box 27.7. *See also* Mayall and Oort 1942.

Oppenheimer, J. R., and H. Snyder, 1939, "On continued gravitational contraction," *Phys. Rev. 56,* 455–459. [§§32.3, 32.4, 34.6, Fig. 24.1, Box 24.1]

Oppenheimer, J. R., and G. Volkoff, 1939, "On massive neutron cores," *Phys. Rev. 55,* 374–381. [§23.2, Fig. 24.1, Box 24.1, Ex. 23.10]

Oppolzer, Th. V., 1887, "Canon der Finsternisse," *Math.-Naturwiss. Cl. d. Akad. der Wiss. Denkschriften 52,* Wien. [Fig. 1.8]

O'Raifeartaigh, L., ed., 1972, *General Relativity, Papers in Honor of J. L. Synge,* Oxford Univ. Press, London.

Ostriker, J. P., and J. E. Gunn, 1969, "On the nature of pulsars, I: Theory," *Astrophys. J. 157,* 1395–1417. [§24.3, Ex. 33.13]

Pacini, F., 1968, "Rotating neutron stars, pulsars, and supernova remnants," *Nature 219,* 145–146. [§24.3]

Page, L., 1941, lecture at Dec. 17 meeting of the American Institute of Electrical Engineers, New York. [Epigraph to Chap. 3]

Palatini, A., 1919, "Deduzione invariantiva delle equazioni gravitazionali dal principio di Hamilton," *Rend. Circ. Mat. Palermo 43,* 203–212. [§21.2]

Palatini, A., *cited by name in* §§21.1, 21.2.

Panov, V. I. *See* Braginsky and Panov 1971.

Paolini, F., R., *cited by name in* Box 27.7.

Papanastassiou, D. A., and G. J. Wasserburg, 1969, "Initial strontium isotopic abundances and the resolution of small time differences in the formation of planetary objects," *Earth & Planetary Sci. Lett. 5,* 361–376. [Box 29.3]

Papapetrou, A., 1951, "Spinning test-particles in general relativity, I," *Proc. R. Soc. London A 209,* 248–258. [§20.6, Ex. 40.8]

Papapetrou, A., 1954a, "Eine neue Theorie des Gravitationsfeldes, I," *Math. Nachr. 12,* 129–141. [§40.8]

Papapetrou, A., 1954b, "Eine neue Theorie des Gravitationsfeldes, II," *Math. Nachr. 12,* 143–154. [§40.8]

Papapetrou, A., 1954c, "Eine Theorie des Gravitationsfeldes mit einer Feldfunktion," *Z. Phys. 139,* 518–532. [§40.8]

Papapetrou, A., 1958, "Uber zeit abhängige Lösungen der Feldgleichungen der allgemeinen Relativitätstheorie," *Ann. Phys. (Germany) 2,* 87–96. [§35.8]

Papapetrou, A., and A. Hamoui, 1967, "Surfaces caustiques dégénerées dans la solution de Tolman: La singularité physique en relativité générale," *Ann. Inst. Henri Poincaré A 6,* 343–364. [Ex. 32.7]

Papapetrou, A., and A. Hamoui, 1968, "Couches simples de matières en relativité générale," *Ann. Inst. Henri Poincaré A 9,* 179–211. [§21.13]

Papapetrou, A., and H. Treder, 1959, "Das Sprungproblem erster Ordnung in der allgemeinen Relativitätstheorie," *Math. Nachr. 20,* 53–66. [§21.13]

Papapetrou, A., and H. Treder, 1962, "Shock waves in general relativity," in Infeld 1962, pp. 351–360. [§21.13]

Paris 1967, *Fluides et Champs Gravitationnel en relativité générale,* Colloques Internationaux du CNRS, no. 170, organizé au Collège de France, du 19 au 23 juin 1967, Éditions du Centre National de la Recherche Scientifique, 1969.

Parker, H. M. *See* Rose *et al.* 1969.

Parker, L., 1966, *Particle Creation in the Expanding Universe,* Ph.D. thesis, Harvard University, unpublished. [§30.5]

Parker, L., 1969, "Quantized fields and particle creation in expanding universes, I," *Phys. Rev. 183,* 1057–1068. [§30.5]

Parker, L., 1972, "Particle Creation in Isotropic Cosmologies," *Phys. Rev. Lett. 28,* 705–708. [§30.5]

Parsons, S. J., *cited by name in* Box 27.7.

Partridge, R. B., 1969, "The primeval fireball today," *Am. Sci. 57,* 37–74. [§27.2]

Partridge, R. B. *See also* Boughn, Fram, and Partridge 1971.

Pauli, W., 1921, "Relativitätstheorie," in *Enzyklopädie der Mathematischen Wissenschaften,* II, 539–775, Teubner, Leipzig, Germany. [Preface]

Pauli, W., 1934, "Die allgemeinen Prinzipien der Wellenmechanik," *Handbuch der Physik 24,* Part 1, 2d ed. [§16.3]

Pauli, W., 1947, *Exclusion Principle and Quantum Mechanics,* 1945 Nobel prize lecture, Editions Grisson, Neuchatel, Switzerland. Related discussion in Fierz and Weisskopf 1960. [Box 44.3]

Pauli, W., ed., 1955, *Niels Bohr and the Development of Physics,* McGraw-Hill, New York.

Pauli, W., 1958, *Theory of Relativity,* Pergamon Press, London. English translation by Gerard Field of Pauli 1921 with supplementary notes (1956) by the author. [endpapers]

Pauli, Wolfgang, Memorial volume. *See* Fierz and Weisskopf 1960.

Pauli, W., *cited by name in* Preface. *See also* Fierz and Pauli 1939.

Paya, D. *See* Dabbs *et al.* 1965.

Peebles, P. J. E., 1966, "Primordial He abundance and fireball, II," *Astrophys. J. 146,* 542–552. [§28.1]

Peebles, P. J. E., 1969, "Cosmology," *Proc. Roy. Astron. Soc. Canada, 63,* 4–31. [§30.1]

Peebles, P. J. E., 1971, *Physical Cosmology,* Princeton Univ. Press, Princeton, N.J. [§§27.1, 27.2, 28.1, 29.6, Box 27.7]

Peebles, P. J. E., 1972, "Gravitational collapse and related phenomena from an empirical point of view, or, black holes are where you find them," *Gen. Rel. & Grav. 3,* 63–82. [Box 33.3]

Peebles, P. J. E. *Cited by name in* Box 27.7. *See also* Dicke and Peebles 1962a, 1962b, *and* 1965, Dicke *et al.* 1965, *and* Jones and Peebles 1972.

Peebles, P. J. E., and J. T. Yu, 1970, "Primeval adiababic perturbation in an expanding universe," *Astrophys. J. 162,* 815–836. [§30.1]

Pekar, D. *See* Eötvös, Pekar, and Fekete 1922.

Penrose, R., 1959, "The apparent shape of a relativistically moving sphere," *Proc. Cambridge Phil. Soc. 55,* 137–139. [§41.10]

Penrose, R., 1964, "Conformal treatment of infinity," in DeWitt and DeWitt 1964b. [§34.2]

Penrose, R., 1965a, "Zero rest-mass fields including gravitation: Asymptotic behavior," *Proc. R. Soc. London A 284,* 159–203. [§34.2, Box 32.2]

Penrose, R., 1965b, "Gravitational collapse and space-time singularities," *Phys. Rev. Lett. 14,* 57–59. [§§32.7, 34.6, Box 34.2]

Penrose, R., 1966, "An analysis of the structure of spacetime," Adams Prize Essay. [Box 34.2]

Penrose, R., 1968a, "Structure of spacetime," in DeWitt and Wheeler 1968. [§§34.1, 34.4, 41.7, 41.11, Boxes 34.1, 34.2, 44.3]

Penrose, R., 1968b, "Twistor quantization and the curvature of spacetime," *Int. J. Theor. Phys. 1,* 61–99. [Box 34.2]

Penrose, R., 1969, "Gravitational collapse: The role of general relativity," *Nuovo Cimento 1,* special number, 252–276. [§33.7, Boxes 33.2, 33.4, 34.2]

Penrose, R., 1972, "Techniques of differential topology in relativity," in Farnsworth *et al.* (1972). [§34.1]

Penrose, R., *cited by name in* §§32.7, 41.9, 42.1, Box 27.7, Fig. 41.7. *See also* Hawking and Penrose 1969, Infeld 1964, Klauder 1964, Newman and Penrose 1965, *and* Walker and Penrose 1970.

Penrose, R., and M. A. H. MacCallum, 1973, "Twistor Theory: An approach to the quantization of fields and spacetime," *Physics Reports 6C,* no. 4, 242–315. [Box 34.2]

Penzias, A. A., *cited by name in* Box 27.7.

Penzias, A. A., and R. W. Wilson, 1965, "A measurement of excess antenna temperature at 4080 Mc/s," *Astrophys. J. 142,* 419–421. [§27.2]

Pereira, C. M. *See* Brill *et al.* 1972.

Peres, A., 1962, "On the Cauchy problem in general relativity, II," *Nuovo Cimento 26,* 53–62. [§43.3]

Peres, A., and N. Rosen, 1964, "Gravitational radiation damping of nongravitational motion," *Ann. Phys. (U.S.A.) 10,* 94–99. [§§36.1, 36.8]

Peters, P. C., 1964, "Gravitational radiation and the motion of two point masses," *Phys. Rev. B 136,* 1224–1232. [§§36.1, 36.7, Ex. 36.9]

Peters, P. C., and J. Mathews, 1963, "Gravitational radiation from point masses in a Keplerian orbit," *Phys. Rev. 131,* 435–440. [§36.6, Ex. 36.6]

Peterson, B. A., *cited by name in* Box 27.7.

Peterson, F. P. *See* Anderson, Brown, and Peterson 1966a *and* 1966b.

Peterson, F. R. *See* Evenson *et al.* 1972.

Pethick, C. J. *See* Baym, Bethe, and Pethick 1971.

Petrosian, V. *See* Noerdlinger and Petrosian 1971.

Petrov, A. Z., 1954, *Reports of the State University of Kazan 114,* book 8, 55. [§33.6]

Petrov, A. Z., 1969, *Einstein Spaces,* Pergamon, New York. [§33.6]

Petrov, Aleksei Zinoveivich, Memorial volume. *See* Shelest 1972.

Pettengill, G. H. *See* Shapiro, Ash *et al.* 1971, Shapiro, Smith *et al.* 1971, and Shapiro *et al.* 1972.

Philolaus, *cited by name in* Box 27.7.

Phillips, J. W., *cited by name in* Box 27.7.

Picasso, E. *See* Bailey *et al.* 1968 *and* Farley *et al.* 1966.

Piir, I. I., 1971, "Maxwell equations in the Schwarzschild gravitational field" (in Russian), *Izv. Akad. Nauk Estonian S. S. R., Fiz. Mat., 20,* 413–419. [Ex. 32.11]

Pilkington, J. D., *cited by name in* Box 27.7. *See also* Hewish *et al.* 1968.

Pirani, F. A. E., 1956, "On the physical significance of the Riemann tensor," *Acta Phys. Pol. 15,* 389–405. [§20.6, Ex. 40.8]

Pirani, F. A. E., 1957, "Invariant formulation of gravitational radiation theory," *Phys. Rev. 105,* 1089–1099. [§§21.13, 33.6, Ex. 15.2]

Pirani, F. A. E., 1965, "Introduction to gravitational radiation theory," in Trautman, Pirani, and Bondi 1965, pp. 249–373. [§§41.7, 41.11, endpapers]

Pirani, F. A. E. *See also* Bondi, Pirani, and Robinson 1959, Ehlers, Pirani, and Schild 1972, *and* Trautman, Pirani, and Bondi 1965.

Pittsburgh 1970, Regional Conference on Relativity, July 13–17, 1970. *See* Farnsworth *et al.* 1972.

Planck, M., 1899, "Über irreversible Strahlungsvorgänge," *Sitzungsber. Deut. Akad. Wiss. Berlin, Kl. Math.—Phys. Tech.,* 440–480. [§§43.4, 44.6]

Plato. *See* Cornford 1937.

Plebanski, J. *See* Infeld and Plebanski 1960 *and* Mielnik and Plebanski 1962.

Plotkin, H. E. *See* Bender *et al.* 1971.

Plutarch, *cited by name in* Box 27.7.

Podolsky, B. *See* Dirac, Fock, and Podolsky 1932.

Podurets, M. A., 1964, "Asymptotic behavior of the optical luminosity of a star in gravitational collapse in terms of general relativity theory," *Astron. Zhur. 41,* 1090–1096, English translation in *Sov. Astron.—A. J. 8,* 868–873 (1965). [§32.3]

Podurets, M. A., 1966, "The character of the singularity in the gravitational collapse of a star," *Dokl. Akad. Nauk SSSR*

167, 1025–1027. English translation in *Sov. Ph.—Doklady 11,* 275–276. [§34.6, Ex. 32.3]

Podurets, M. A. *See also* Zel'dovich and Podurets 1965.

Poincaré, H. *cited by name in* §§3.1, 12.3.

Poisson, S., *cited by name in* Box 17.2 part 6.

Polievktov-Nikoladze, N. M., 1969, "On the theory of the gravitational field," *Zh. Eksp. & Teor. Fiz. 57,* 2010–2020. English translation in *Sov. Phys.—JETP 30* (1970), 1089–1095. [Box 17.2 part 6]

Polkinghorne, J. C., 1955, "On the Feynman principle," *Proc. R. Soc. London A 230,* 272–276. [Box 21.1]

Polubarinova-Kochina, P. Ya., 1963, "Alexander Alexandrovitch Friedmann (In memory of A. A. Friedmann)," *Usp. Fiz. Nauk, 80,* 345–352. English translation in *Sov. Phys.—Uspekhi, XX,* 467–472 (1964). [Box 27.6]

Pope, A., 1733, *Essay on Man,* London [Epigraph to Chap. 24]

Porter, J. *See* Farnsworth *et al.* 1972.

Pound, R. V., and G. A. Rebka, 1960, "Apparent weight of Photons," *Phys. Rev. Lett. 4,* 337–341. [§§7.2, 7.4, 38.5, Fig. 38.1]

Pound, R. V., and J. L. Snider, 1964, "Effect of gravity on nuclear resonance," *Phys. Rev. Lett. 13,* 539–540. [§§7.2, 7.4]

Pound, R. V., and J. L. Snider, 1965, "Effect of gravity on gamma radiation," *Phys. Rev. B 140,* 788–803. [§§7.2, 38.5, Fig. 38.1]

Power, E. A., and J. A. Wheeler, 1957, "Thermal geons," *Rev. Mod. Phys. 29,* 480–495. [§25.6]

Prakash, A. *See* Newman *et al.* 1965.

Press, F., 1965, "Resonant vibrations of the Earth," *Sci. American 213,* no. 5, 28–37. [Ex. 37.14]

Press, F. *See also* Ahrens, Press, and Runcorn 1964.

Press, W. H., 1970, unpublished work described in part in Press and Thorne 1972. [§37.3, Fig. 37.2, Box 37.2]

Press, W. H., 1971, "Long wave trains of gravitational waves from a vibrating black hole," *Astrophys. J. Lett. 170,* 105–108. [Box 33.3]

Press, W. H., 1972, "Time evolution of a rotating black hole immersed in a static scalar field," *Astrophys. J. 175,* 243–252. [Ex. 33.14, Box 33.3]

Press, W. H. *See also* Bardeen and Press 1972, Bardeen, Press, and Teukolsky 1972, *and* Davis *et al.* 1971.

Press, W. H., and J. M. Bardeen, 1971, "Nonconservation of the Newman-Penrose conserved quantities," *Phys. Rev. Lett. 27,* 1303–1306. [Ex. 32.11]

Press, W. H., and S. A. Teukolsky, 1972, "Floating orbits, superradiant scattering and the black-hole bomb," *Nature 238,* 211–212. [Box 33.3]

Press, W. H., and S. A. Teukolsky, 1973, "Perturbations of a rotating black hole, II: Dynamical stability of the Kerr metric," *Astrophys. J.,* in press. [Box 33.3, Ex. 33.14]

Press, W. H., and K. S. Thorne, 1972, "Gravitational-wave astronomy," *Ann. Rev. Astron. Astrophys. 10,* 335–374. [§§36.5, 37.4, 37.8, Box 37.5]

Price, R. H., 1971, *Nonspherical Perturbations of Relativistic Gravitational Collapse,* doctoral dissertation, California Institute of Technology. [Ex. 32.10, Box 32.2]

Price, R. H., 1972a, "Nonspherical perturbations of relativistic gravitational collapse, I: Scalar and gravitational perturbations," *Phys. Rev., D5,* 2419–2438. [§§35.14, 41.11, Ex. 32.10, Box 32.2]

Price, R. H., 1972b, "Nonspherical perturbations of relativistic gravitational collapse, II: Integer-spin, zero-rest-mass fields," *Phys. Rev., D5,* 2439–2454. [§41.11, Box 32.2]

Price, R. H. *See also* Davis *et al.* 1971.

Pringle, J. E., and M. J. Rees, 1972, "Accretion disc models for compact X-ray sources," *Astron. Astrophys. 21,* 1–9. [Box 33.3]

Proclus of Athens, *cited by name in* Fig. 1.8.

Ptolemy, C. *See* Taliaferro 1952.

Ramsey, N. F., 1972, "History of atomic and molecular standards of frequency and time," *I.E.E.E. Trans. IM-21,* 90–98. [Box 16.3]

Rayleigh, Lord, 1912, "On the propagation of waves through a stratified medium, with special reference to the question of reflection," *Proc. R. Soc. London A 86,* 207–226. [§22.5]

Rayzer, Yu. P. *See* Zel'dovich and Rayzer 1967.

Reber, G., *cited by name in* Box 27.7.

Rebka, G. A. *See* Pound and Rebka 1960.

Rees, M. *See* Lynden-Bell and Rees 1971 *and* Pringle and Rees 1972.

Refsdal, S., 1964, "The gravitational lens effect," *Mon. Not. R. Astron. Soc. 128,* 295–306. [§25.6, Fig. 22.2, Box 33.3]

Refsdal, S., R. Stabell, and F. G. de Lange, 1967, "Numerical calculations on relativistic cosmological models," *Mem. Roy. Astron. Soc., 71,* 143–248. [§29.4, Box 29.4]

Regge, T., 1961, "General relativity without coordinates," *Nuovo Cimento 19,* 558–571. [§§42.1, 42.2, 42.4, 42.6]

Regge, T., and J. A. Wheeler, 1957, "Stability of a Schwarzschild singularity," *Phys. Rev. 108,* 1063–1069. [§25.5, Box 33.3]

Reichley, P. *See* Muhlemann and Reichley 1964 *and* 1965.

Reik, H. G. *See* Meixner and Reik 1959.

Reines, F., ed., 1972, *Cosmology, Fusion, and Other Matters,* Univ. of Colorado Press, Boulder.

Reissner, H., 1916, "Über die Eigengravitation des elektrischen Feldes nach der Einsteinschen Theorie," *Ann. Phys. (Germany) 50,* 106–120. [Ex. 31.8]

Reissner, H., *cited by name in* §42.7.

Retherford, R. C. *See* Lamb and Retherford 1947.

Rhoades, C. E. J., 1971, *Investigations in the Physics of Neutron Stars,* doctoral dissertation, Princeton University. [§24.2]

Richelot, *cited by name in* Box 38.2.

Riemann, G. F. B., *achievements and picture in* Box 8.5; for collected works, see Weber 1953.

Riesz, M., 1949, "L'intégrale de Riemann-Liouville et le problème de Cauchy," *Acta Math. 81,* 1–223. [§35.8]

Rindler, W., 1969, *Essential Relativity: Special, General, and Cosmological,* Van Nostrand, New York. [Box 27.7]

Robb, A. A., 1914, *A Theory of Time and Space,* Cambridge Univ. Press, Cambridge, Eng. [Box 13.1]

Robb, A. A., 1936, *Geometry of Time and Space,* Cambridge Univ. Press, Cambridge, Eng. Rev. ed. of Robb 1914. [Boxes 1.3, 13.1]

Robb, A. A., *quoted,* epigraph of Chap. 4.
Roberts, M. S., 1965, "21-cm. hydrogen measurements of NGC5668, a relatively distant galaxy," *Astrophys. J. 142,* 148–153. [§29.2]
Roberts, M. S. *See also* Dieter *et al.* 1962.
Robertson, H. P., 1935, "Kinematics and world structure," *Astrophys. J. 82,* 248–301. [Ex. 27.5]
Robertson, H. P., 1936, "Kinematics and world structure," *Astrophys. J. 83,* 187–201, 257–271. [Ex. 27.5]
Robertson, H. P., 1962, "Relativity and cosmology," in Deutsch and Klemperer 1962, pp. 228–235. [§§38.2, 39.3]
Robertson, H. P., *cited by name in* §30.3, Box 27.7.
Robertson, H. P., and T. W. Noonan, 1968, *Relativity and Cosmology,* W. B. Saunders, Philadelphia. [§27.11, Ex. 27.4, 29.6, endpapers]
Robinson, H. G. *See* Hughes, Robinson, and Beltran-Lopez 1960.
Robinson, I., 1959a, "A solution of the Maxwell-Einstein equations," *Bull. Acad. Polon. Sci. 7,* 351–352. [Ex. 32.1]
Robinson, I., 1959b, unpublished report at *Colloque Internationale sur les Théories Relativistes de la Gravitation,* Abbaye de Royaumont. [Ex. 15.2]
Robinson, I., 1961, "Null Electromagnetic Fields," *J. Math. Phys., 2,* 290–291. [§22.5]
Robinson, I., A. Schild, and E. Schücking, eds., 1965, *Quasi-stellar Sources and Gravitational Collapse,* The University of Chicago Press, Chicago.
Robinson, I. *See also* Bondi, Pirani, and Robinson 1959, Douglas *et al.* 1969.
Rohrlich, F., 1965, *Classical Charged Particles; Foundations of Their Theory,* Addison-Wesley, Reading, Mass. [§4.7]
Rohrlich, F., *cited by name in* §20.6.
Roll, P. G., *cited by name in* Box 27.7. *See also* Dicke *et al.* 1965.
Roll, P. G., R. Krotkov, and R. H. Dicke, 1964, "The equivalence of inertial and passive gravitational mass," *Ann. Phys.* (*U.S.A.*) *26,* 442–517. [§§1.3, 10.1, 38.3, Fig. 1.6, Box 1.2]
Root, W. L. *See* Davenport and Root 1958.
Rose, R. D., H. M. Parker, R. A. Lowry, A. R. Kuhlthau, and J. W. Beams, 1969, "Determination of the gravitational constant *G, Phys. Rev. Lett. 23,* 655–658. [§40.8]
Rosen, Nathan, *Festschrift* on The occasion of his 60th birthday, March 22, 1969. *See* Kuper and Peres 1971.
Rosen, N. *See also* Peres and Rosen 1964.
Rosenfeld, L., 1940, "Sur le tenseur d'impulsion-énergie," *Mém. Acad. R. Belg. Sci. 18,* no. 6. [§21.3]
Rosenfeld, L. *See also* Bohr and Rosenfeld 1933.
Rossi, B., *cited by name in* Box 27.7.
Rossi, B., and D. B. Hall, 1941, "Variation of the rate of decay of mesotrons with momentum," *Phys. Rev. 59,* 223–228. [§38.4]
Royaumont 1959, *Les Théories relativistes de la gravitation,* Actes du Colloque International, organizé à Royaumont, du 21 au 27 juin 1959, Centre National de la Recherche Scientifique, Paris, 1962.
Ruark, A. E., and H. C. Urey, 1930, *Atoms, Molecules, and Quanta,* McGraw-Hill, New York. [§§40.7, 41.4]
Rudenko, V. N. *See* Braginsky, Zel'dovich, and Rudenko 1969.
Ruderman, M., 1972, "Pulsars: Structure and dynamics," *Ann. Rev. Astron. Astrophys. 10,* 427–476. [§24.3]
Ruffini, R. *See* Christodolou and Ruffini 1971; Davis *et al.* 1971; Ferrari and Ruffini 1969; *and* Hanni and Ruffini 1973.
Ruffini, R., and J. A. Wheeler, 1970, "Collapse of wave to black hole," *Bull. Am. Phys. Soc. 15,* 76. [§35.8]
Ruffini, R., and J. A. Wheeler, 1971a, "Introducing the black hole," *Phys. Today 24,* 30–36. [§33.7]
Ruffini, R., and J. A. Wheeler, 1971b, "Relativistic cosmology and space platforms," in *Proceedings of the Conference on Space Physics,* European Space Research Organization, Paris, France, pp. 45–174. [Boxes 33.2, 33.3, 33.5, 36.3, Fig. 24.3, Ex. 37.13, 37.14]
Runcorn, S. K. *See* Ahrens, Press, and Runcorn 1964.
Runge, J. *See* Sommerfeld and Runge 1911.
Russ, N. *See* Boni, Russ, and Laurence 1960.
Rutherford, E., *cited by name in* Preface, §44.1.
Ruzmaikin, A. A. *See* Ruzmaikina and Ruzmaikin 1969.
Ruzmaikina, T. V., and A. A. Ruzmaikin, 1969, "Quadratic corrections to the Lagrangian density of the gravitational field and the singularity," *Zh. Eksp. & Teor. Fiz. 57,* 680–685. English translation in *Sov. Phys.—JETP 30,* 372–374. [Box 17.2 part 6]
Ryan, M. P., Jr., 1971, "Qualitative cosmology: Diagrammatic solutions for Bianchi type IX universes with expansion, rotation, and shear, II: The general case," *Ann. Phys. 68,* 541–555. [§30.7]
Ryan, M. P., Jr., 1972a, *Hamiltonian Cosmology,* Springer, Berlin, Germany. [§30.7, Box 30.1]
Ryan, M. P., Jr., 1972b, "The oscillatory regime near the singularity in Bianchi type-IX universes," *Ann. Phys.* (*U.S.A.*) *70,* 301–322. (Includes a report of work by V. A. Belinsky and I. M. Khalatnikov. [§30.7, Box 30.1]
Sachs, R. K., 1961, "Gravitational waves in general relativity, VI: The outgoing radiation condition," *Proc. R. Soc. London A 264,* 309–338. [§22.5]
Sachs, R. K., 1964, "Gravitational radiation," in DeWitt and DeWitt 1964. [§§21.5, 21.13, 33.6, 41.11, Ex. 22.14, endpapers]
Sachs, R. K., ed., 1971, *General Relativity and Cosmology,* proceedings of course 47 of the International School of Physics "Enrico Fermi", Academic Press. [§§27.2, 34.1, 40.7]
Sachs, R. K. *See also* Eardley, Liang, and Sachs 1972, Ehlers, Geren, and Sachs 1968, *and* Jordan, Ehlers, and Sachs 1961.
Sakharov, A. D., 1967, "Vacuum quantum fluctuations in curved space and the theory of gravitation," *Doklady Akad. Nauk S.S.S.R. 177,* 70–71. English translation in *Sov. Phys. Doklady 12,* 1040–1041 (1968). [§44.4, Box 17.2 part 6]
Sakharov, A. D., 1969, unpublished work cited by Braginsky and Nazarenko 1971. [§37.3, Fig. 37.2]
Salam, A., and E. P. Wigner, eds., 1972, *Aspects of Quantum Theory,* Cambridge Univ. Press, Cambridge, Eng.
Salpeter, E. E. and R. V. Wagoner, 1971, "Structure of relativistic disks," *Astrophys. J. 164,* 557–568. [Box 24.1]

Sandage, A., 1958, "Current problems in the extragalactic distance scale," *Astrophys. J. 127,* 513–526. [Box 27.1]

Sandage, A., 1968, "The time scale for creation," in L. Woltjer, ed., *Galaxies and the Universe,* Columbia Univ. Press, New York, pp. 75–112. [§29.7]

Sandage, A., 1970, "Main-sequence photometry, color-magnitude diagrams, and ages for the globular clusters M3, M13, M15, and M92," *Astrophys. J. 162,* 841–870. [§29.7]

Sandage, A., 1972a, "Distances to galaxies: the Hubble constant, the Friedmann time, and the edge of the world," in *Proceedings of the Symposium on the Galaxy and the Distance Scale,* Essex, England, in press. [§§27.1, 29.3, 29.4, Box 29.4]

Sandage, A., 1972b, "The redshift-distance relation, I. Angular diameter of first-ranked cluster galaxies as a function of redshift: the aperture correction to magnitudes," *Astrophys. J. 173,* 485-500. [§29.5, Box 29.4]

Sandage, A., 1972c, "The redshift-distance relation, II. The Hubble diagram and its scatter for first-ranked cluster galaxies: a formal value for q_o," *Astrophys. J. 178,* 1–24. [§29.4, Box 29.4]

Sandage, A., 1972d, "The redshift-distance relation, III. Photometry and the Hubble diagram for radio sources and the possible turn-on time for QSO's," *Astrophys. J. 178,* 25–44. [§29.4, Boxes 28.1, 29.4]

Sandage, A., *cited by name in* Boxes 27.7, 29.5.

Sandage, A., G. A. Tamman, and E. Hardy, 1972, "Limits on the local deviation of the universe from a homogeneous model," *Astrophys. J. 172,* 253–263. [§27.1]

Sanders, R., 1970, "The effects of stellar collisions in dense stellar systems," *Astrophys. J. 162,* 791–809. [§§24.5, 24.6, Box 33.3]

Sanderson, B. J. *See* Hsiang and Sanderson 1965.

Sands, M. *See* Feynman, Leighton, and Sands 1964.

Sard, R. D., 1970, *Relativistic Mechanics,* W. A. Benjamin, New York. [§2.10]

Sargent, W. L. W. *See* Bahcall, Sargent, and Schmidt 1967.

Schäfke, F. W. *See* Meixner and Schäfke 1954.

Schatzman, E. *See* DeWitt, Veron, and Schatzman 1967.

Schiff, L. I., 1962, "General relativity: Theory and experiment," *J. Indust. & Appl. Math 10,* 795–801. [§§38.2, 39.3]

Schiff, L. I., 1967, "Comparison of theory and observation in general relativity," in Ehlers 1967. [§§38.2, 39.3]

Schiffer, M. *See* Adler, Bazin, and Schiffer 1965.

Schild, A., 1960, "Time," *Texas Quarterly, 3,* no. 3, 42–62. [§7.3]

Schild, A., 1962, "Gravitational theories of the Whitehead type and the principle of equivalence," in Møller 1962. [§§7.3, 38.5]

Schild, A., 1967, "Lectures on General Relativity Theory," in Ehlers 1967, pp. 1–105. [§7.3, endpapers]

Schild, A., 1970, "Tearing geometry to pieces: More on conformal geometry," unpublished lecture at Jan. 19 Princeton University relativity seminar. [Box 10.2]

Schild, A., *cited by name in* §10.3, Box 17.3, Ex. 10.6. *See also* Douglas *et al.* 1969, Ehlers, Pirani, and Schild 1972, Infeld and Schild 1949, Kerr and Schild 1965, *and* Robinson, Schild, and Schücking 1965.

Schilpp, P. A., ed., 1949, *Albert Einstein: Philosopher-Scientist,* Library of Living Philosophers, Evanston, Ill.; source of quotations from A. Einstein, K. Menger, and M. von Laue. [Box 1.11, epigraphs to §1.2, Chaps. 2, 9]

Schmidt, B. G., 1971, "A new definition of singular points in general relativity," *Gen. Rel. Grav. 1,* 269–280. [§34.6]

Schmidt, M., 1972, "Statistical studies of the evolution of extragalactic radio sources, I. Quasars," *Astrophys. J. 176,* 273–287. [Box 28.1]

Schmidt, M. *Cited by name in* Box 27.7, Ex. 29.7. *See also* Bahcall, Sargent, and Schmidt 1967 *and* Bahcall and Schmidt 1967.

Schoch, C. *See* Neugebauer and Schoch 1930.

Schouten, J. A., 1954, *Ricci-calculus,* Springer, Berlin, Germany. [§21.5, Ex. 17.8, 21.8, endpapers]

Schramm, D. N. *See* Wasserburg, Schramm, and Huneke 1969.

Schrödinger, E., 1950, *Space-Time Structure,* Cambridge Univ. Press, England. [endpapers]

Schücking, E. *See also* Douglas *et al.* 1969 *and* Robinson, Schild, and Schücking 1965.

Schücking, E., and O. Heckmann, 1958, "World models," in *Onzième Conseil de Physique Solvay,* Éditions Stoops, Brussels, pp. 149–158. [§§30.2, 30.3]

Schutz, B., *cited by name in* Ex. 35.20.

Schutz, S., 1966, statement in January 1966 final examination in course in relativity, Princeton University. [Epigraph to Chap. 16]

Schwartz, J. T. *See* Dunford and Schwartz 1958.

Schwartz, L., 1950–1951, *Théorie des distributions,* vols. 1 and 2, Hermann, Paris. [§4.7]

Schwarzschild, K., 1916a, "Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie," *Sitzber. Deut. Akad. Wiss. Berlin, Kl. Math.-Phys. Tech.,* 189–196. [Preface, §§14.1, 23.6]

Schwarzschild, K., 1916b, "Über das Gravitationsfeld einer Kugel aus inkompressibler Flussigkeit nach der Einsteinschen Theorie," *Sitzber. Deut. Akad. Wiss. Berlin, Kl. Math.-Phys. Tech.,* 424–434. [§23.2, Box 23.2]

Schwarzschild, K., *cited by name in* §§42.1, 42.7.

Schwinger, J., 1948, "Quantum electrodynamics, I: A covariant formulation," *Phys. Rev. 74,* 1439–1461. [Box 15.1]

Schwinger, J., ed., 1958, *Selected Papers on Quantum Electrodynamics,* Dover, New York.

Schwinger, J., 1963a, "Energy and momentum density in field theory," *Phys. Rev. 130,* 800–805. [Ex. 40.8]

Schwinger, J., 1963b, "Quantized gravitational field," *Phys. Rev. 130,* 1253–1258. [Ex. 40.8]

Sciama, D. W., 1957, "Inertia," *Sci. American 196,* no. 2, 99–109. [§21.12]

Sciama, D. W., 1961, *The Unity of the Universe,* Anchor, New York. [§21.12]

Sciama, D. W., 1971, *Modern Cosmology,* Cambridge Univ. Press, Cambridge, Eng. [§28.1, Box 27.7]

Sciama, D. W., *cited by name in* Boxes 34.2, 34.3.

Scott, P. F., *cited by name in* Box 27.7. *See also* Hewish *et al.* 1968.

Sears, F. W., 1953, *An Introduction to Thermodynamics, the Kinetic Theory of Gases, and Statistical Mechanics,* Addison-Wesley, Reading, Mass. [§23.4]

Seattle 1967, Battelles Rencontres. *See* DeWitt and Wheeler 1968.

Seckler, B. D. *See* Keller, Lewis, and Seckler 1956.

Seelig, C., ed., 1956. *See* Kollros 1956.

Seielstad, G. A., R. A. Sramek, and K. W. Weiler, 1970, "Measurement of the deflection of 9.602-GHz radiation from 3C279 in the solar gravitational field," *Phys. Rev. Lett. 24,* 1373–1376. [Box 40.1]

Sejnowski, T. J., 1973, personal communication. [Ex. 15.2]

Shakespeare, W., *quoted at end of* Chap. 44.

Shakura, N. I., and R. A. Sunyaev, 1973, "Black holes in binary systems: Observational appearance," *Astron. Astrophys. 24,* 337–355. [Box 33.3]

Shannon, C., 1938, "A symbolic analysis of relay and switching circuits," *Trans. Am. Inst. Elect. Eng. 57,* 713–723. [§44.5]

Shapiro, I. I., 1964, "Fourth test of general relativity," *Phys. Rev. Lett. 13,* 789–791. [§40.4]

Shapiro, I. I., 1967, "New method for the detection of light deflection by solar gravity," *Science 157,* 806–808. [§40.3]

Shapiro, I. I., 1968, "Fourth test of general relativity: preliminary results," *Phys. Rev. Lett. 20,* 1265–1269. [§40.4, Box 40.2]

Shapiro, I. I., 1970, results reported at Third Cambridge Conference on Experimental Relativity, New York City, June 1970. [Box 40.3]

Shapiro, I. I., 1971a. "Testing general relativity: progress, problems, and prospects," in Davies, 1971, pp. 136–147. [Box 40.3]

Shapiro, I. I., 1971b. "Testing general relativity: progress, problems, and prospects," lecture at Sixth International Conference on Gravitation and Relativity, published by NORDITA (Copenhagen) in multilith form for distribution at conference, also published in *Gen. Rel. Grav. 3,* 135–148 (1972). [§39.3, Box 40.3]

Shapiro, I. I., M. E. Ash, D. B. Campbell, R. B. Dyce, R. P. Ingalls, R. F. Jurgens, and G. H. Pettengill, 1971, "Fourth test of general relativity: New radar result," *Phys. Rev. Lett. 26,* 1132–1135. [Box 40.2]

Shapiro, I. I., W. B. Smith, M. E. Ash, R. P. Ingalls, and G. H. Pettengill, 1971, "Gravitational constant: Experimental bound on its time variation," *Phys. Rev. Lett. 26,* 27–30. [§40.8]

Shapiro, I. I., G. H. Pettengill, M. E. Ash, R. P. Ingalls, D. B. Campbell, and R. B. Dyce, 1972, "Mercury's perihelion advance: determination by radar," *Phys. Rev. Lett. 28,* 1594–1597. [Box 40.3]

Shapley, H., *cited by name in* Box 27.7.

Sharp, D. H., 1960, "One- and two-surface formulations of the boundary-value problem for Einstein-Maxwell theory and for 'already unified' field theory," Physics A. B. senior thesis, Princeton University, unpublished; deposited in Fine Library, Princeton University. [§21.9]

Sharp, D. H., *cited by name in* §21.1. *See also* Baierlein, Sharp, and Wheeler 1962, Hartle and Sharp 1965, *and* Misner and Sharp 1964.

Sheldon, S. E., 1971, "A complete redshift-magnitude formula," *Astron. J. 76,* 751–755. [§29.4]

Shelest, V. P., ed., 1972, *Gravitatsiya: Problemi i Perspektivi: pamyati Alekseya Zinovievicha Petrova posvashaetsya,* Naukova Dumka, Kiev.

Shields, M. C. *See* Boni, Russ, and Laurence 1960.

Shimmins, A. J., *cited by name in* Box 27.7.

Shirkov, D. V. *See* Bogoliubov, N. N.

Shu, F. H. *See* Lin, Mestel, and Shu 1965.

Siegel, C. L., 1956, *Vorlesungen über Himmelsmechanik,* Springer, Berlin. [Box 25.4]

Siegel, C. L., and J. K. Moser, 1971, *Lectures on Classical Mechanics,* translated by C. I. Kalme, Springer, Berlin and New York, 1971. [Box 25.4]

Silverberg, E. C. *See* Bender *et al.* 1971.

Simpson, O. C. *See* Estermann, Simpson, and Stern 1938.

Singer, D. W., 1950, *Giordano Bruno, His Life and Thought with Annotated Translation of His Work, On the Infinite Universe and Worlds,* Abelard-Schuman, New York. [Box 27.7]

Slee, O. B., *cited by name in* Box 27.7.

Slipher, V. M., *cited by name in* Box 27.7.

Smith, J. H. *See* Frisch and Smith 1963.

Smith, W. B. *See* Shapiro, Smith *et al.* 1971.

Snider, J. L. *See* Pound and Snider 1964 *and* 1965.

Snyder, H. *See* Oppenheimer and Snyder 1939.

Söding, P., J. Bartels, A. Barbaro-Galtieri, J. E. Enstrom, T. A. Lasinski, A. Rittenberg, A. H. Rosenfeld, T. G. Trippe, N. Barash-Schmidt, C. Bricman, V. Chaloupka, and M. Roos, 1972, "Review of particle properties," *Physics Letters, 39B,* 13. [endpapers]

Solvay 1958, Onzième Conseil de Physique Solvay, held at the Instituts Solvay, Brussels; proceedings: *La Structure et l'evolution de l'univers,* Éditions Stoops, Brussels, Belgium.

Sommerfeld, A., and J. Runge, 1911, "Anwendung der Vektorrechnung auf die Grundlagen der geometrischen Optik," *Ann. Phys. (Germany) 35,* 277–298. [§22.5]

Spencer, D. C. *See* Nickerson, Spencer, and Steenrod 1959.

Spitzer, L., Jr., 1971, "Dynamical evolution of dense spherical star systems," in *Semaine d'étude sur les noyeaux des galaxies 1970 (Pontificae Academiae Scientarum Scripta Varia no. 35).* [§§24.5, 24.6, Box 33.3]

Spivak, M., 1965, *Calculus on Manifolds,* W. A. Benjamin, New York. [§4.1, Box 4.1]

Sramek, R. A., 1971, "A measurement of the gravitational deflection of microwave radiation near the sun, 1970 October," *Astrophys. J. Lett. 167,* L55–L60. [Box 40.1]

Sramek, R. A. *See also* Seielstad, Sramek, and Weiler 1970.

St. Patrick. *See* Stokes 1887.

Stabell, R. *See* Refsdal, Stabell, and de Lange 1967.

Stanley, G. J., *cited by name in* Box 27.7.

Starobinsky, A. A. *See* Zel'dovich and Starobinsky 1971.

Steenrod, N. E. *See* Nickerson, Spencer, and Steenrod 1959.

Stehle, P. *See* Corben and Stehle 1960.

Steigman, G., 1971, "Antimatter and cosmology," in Sachs 1971, pp. 373–382. [§§27.11, 28.4]

Stephenson, F. R., *contributed* Fig. 1.8.

Stern, O. *See* Estermann, Simpson, and Stern 1938.

Sternberg, S., 1969, *Celestial Mechanics,* W. A. Benjamin, New York. [§8.2]

Steuckelberg, E. C. G., and G. Wanders, 1953, "Thermodynamique en relativité général," *Helv. Phys. Act., 26,* 307–316. [§22.2]

Stewart, J. M., 1971, *Non-Equilibrium Relativistic Kinetic Theory,* lecture notes in Physics 10, Springer-Verlag, Berlin. [§22.6, Ex. 22.7]

Stewart, J. M., *cited by name in* §22.3.

Stewart, J. M., and G. F. R. Ellis, 1968, "Solutions of Einstein's equations for a fluid which exhibits local rotational symmetry," *J. Math. Phys. 9,* 1072–1082. [Ex. 32.1]

Stockton, A. N., *cited by name in* Box 27.7.

Stokes, G. C., *cited by name in* Box 4.6.

Stokes, W., 1887, *The Tripartite Life of St. Patrick With Other Documents Relating to That Saint,* Her Majesty's Stationery Office, London. [§19.4]

Stratimyer, H., *cited by name in* Box 16.3.

Struve, F. G. W., *cited by name in* Box 27.7.

Sunyaev, R. A. *See* Shakura and Sunyaev 1973.

Susskind, L. *See* Aharonov and Susskind 1967.

Symanzik, K., 1963, "Aspects of Euclidean field theory and Feynman path integrals," in Lurie and Mukunda 1963. [Box 21.1]

Synge, J. L., 1950, "The gravitational field of a particle," *Proc. R. Irish Acad. A 53,* 83–114. [§31.5, Box 31.1]

Synge, J. L., 1960a, "On certain identities connected with the Einstein tensor," *Proc. R. Irish Acad. A 61,* 29–36. [Ex. 35.14]

Synge, J. L., 1960b, *Relativity: The General Theory,* North-Holland, Amsterdam. [endpapers]

Synge, J. L., 1962, "Tensorial integral conservation laws in general relativity," in CNRS 1962, pp. 75–83. [Box 15.3]

Synge, John Lighton, *Festschrift* on the occasion of his 75th birthday, March 23, 1972. *See* O'Raifeartaigh 1972.

Szekeres, G., 1960, "On the singularities of a Riemannian manifold," *Publ. Mat. Debrecen 7,* 285–301. [§31.4, Box 31.1]

Taliaferro, R. C., trans. 1952, *Ptolemy, The Almagest,* in *Great Books of the Western World,* Encyclopedia Brittanica, Inc., Chicago, Ill., Volume 16. [Box 27.7]

Tamman, G. A., *cited by name in* Box 29.4. *See also* Sandage, Tamman, and Hardy 1972.

Tannenbaum, M. *See* Farley *et al.* 1966.

Tartaglia, N. F., 1560, *General Trattato dei Numeri et Misure,* Cited in *Enzyklopädie der Mathematischen Wissenschaften,* vol. III, part i, p. 1056, Leipzig 1914. [Box 13.1]

Taub, A. H., 1948, "Relativistic Rankine-Hugonoit equations," *Phys. Rev., 74,* 328–334. [§22.2, Box 22.2]

Taub, A. H., 1965, "Equations of motion of test particles," in London 1965. [§40.9]

Taub, A. H., 1971, "Relativistic Hydrodynamics," *Studies in Applied Mathematics—Math. Assoc. Amer. Studies in Math. 7,* 150–180. [§22.3]

Taub, A. H. *See also* Misner and Taub 1968.

Tauber, G. E., and J. W. Weinberg, 1961, "Internal state of a gravitating gas," *Phys. Rev. 122,* 1342–1365. [§22.6]

Taylor, D. J., *cited by name in* Box 34.3. See also Cocke and Taylor 1969.

Taylor, E. F., and J. A. Wheeler, 1966, *Spacetime Physics,* W. H. Freeman, San Francisco, Calif. [§2.1, Fig. 2.7, Ex. 6.9]

Tbilisi, USSR, 1968. Fifth International Conference on Gravitation and the Theory of Relativity, Tbilisi State University, August 1968. *See* Fock 1968.

Technion 1969, *International Seminar on Relativity and Gravitation* held at Technion City, Israel, in July 1969, in honor of the 60th birthday of Professor Nathan Rosen. *See* Kuper and Peres 1971.

Teitelboim, C., 1970, "Splitting of the Maxwell tensor: radiation reaction without advanced fields," *Phys. Rev. D 1,* 1572–1582. [§20.6, Ex. 20.8]

Teitelboim, C., 1971, "Radiation reaction as a retarded self-interaction," *Phys. Rev. D 4,* 345–347. [§20.6]

Teitelboim, C., 1972a, "Nonmeasurability of the baryon number of a black hole," *Nuovo Cimento Lett., II, no. 3,* 326–328. [Box 25.2, Boxes 32.2, 33.1]

Teitelboim, C., 1972b, "Nonmeasurability of the lepton number of a black hole," *Nuovo Cimento, II 3,* 397–400. [Boxes 25.2, 33.1]

Teitelboim, C., 1972c, "Nonmeasurability of the Quantum Numbers of a Black Hole," *Phys. Rev. D 5,* 2941–2954 [Boxes 25.2, 32.2, 33.1]

Teitelboim, C., *cited by name in* §21.1, Fig. 4.6.

Teller, E., 1948, "On the change of physical constants," *Phys. Rev. 73,* 801–802. [§38.6]

Teller, E., *cited by name in* Box 38.2. *See also* De Hoffman and Teller 1950.

Tempier, É., *cited by name in* Box 27.7.

Teukolsky, S. A., 1972. "Rotating black holes: Separable wave equations for gravitational and electromagnetic perturbations," *Phys. Rev. Lett. 29,* 1114–1118. [§41.11, Box 33.3, Ex. 33.14]

Teukolsky, S. A., 1973, "Perturbations of a rotating black hole, I: Fundamental equations for gravitational, electromagnetic, and neutrino-field perturbations," *Astrophys. J.,* in press. [Ex. 33.14]

Teukolsky, S. A. *See also* Bardeen, Press, and Teukolsky 1972 *and* Press and Teukolsky 1972 *and* 1973.

Texas Symposia on Relativistic Astrophysics: 1963, *see* Robinson, Schild, and Schücking 1965 *and* Harrison *et al.* 1965; 1964, *see* Douglas *et al.* 1969; 1972, *see Trans. N.Y. Acad. Sci.*

Thirring, H., 1918, "Über die Wirkung rotierender ferner Massen in der Einsteinschen Gravitationstheorie," *Phys. Z. 19,* 33–39. [§21.12]

Thirring, H., 1921, "Berichtigung zu meiner Arbeit: "Über die Wirkung rotierender ferner Massen in der Einsteinschen Gravitationstheorie," *Phys. A. 22,* 29–30. [§21.12]

Thirring, H., and J. Lense, 1918, "Über den Einfluss der Eigenrotation der Zentralkörper auf die Bewegung der Planeten und Monde nach der Einsteinschen Gravitationstheorie," *Phys. Z. 19,* 156–163. [§§21.12, 40.7]

Thirring, W. E., 1961, "An alternative approach to the theory of gravitation," *Ann. Phys. (U.S.A.) 16,* 96–117. [§§7.1, 18.1]

Thomas, L. H., 1927, "The kinematics of an electron with an axis," *Phil. Mag. 3,* 1–22. [§§40.7, 41.4, Ex. 6.9]

Thomson, A. *See* Farnsworth *et al.* 1972.

Thomson, J. J., 1907, *Electricity and Matter,* Archibald Constable, London, Chapter III. [Fig. 4.6]

Thomson, J. J., *cited by name in* Fig. 4.6.

Thomson, W., Lord Kelvin, 1889, *Popular Lectures and Addresses, Vol. I,* Macmillan, London. [Epigraph to Chap. 37]

Thomson, W., Lord Kelvin, 1904, *Baltimore Lectures on Mo-*

lecular Dynamics and the Wave Theory of Light, C. J. Clay, London. [Preface]

Thomson, W., Lord Kelvin, *cited by name in* Box 17.2, *quoted in* epigraph of Chap. 20.

Thorne, K. S., 1967, "The general relativistic theory of stellar structure and dynamics," in DeWitt, Schatzman, and Veron 1967, pp. 259–441. [§24.4, endpapers]

Thorne, K. S., 1969a, "Nonradial pulsation of general-relativistic stellar models, III: Analytic and numerical results for neutron stars," *Astrophys. J. 158,* 1–16. [§24.3, Box 36.1]

Thorne, K. S., 1969b, "Nonradial pulsation of general-relativistic stellar models, IV: The weak-field limit," *Astrophys. J. 158,* 997–1019. [§36.8, Box 36.1]

Thorne, K. S., 1971, "Relativistic stars, black holes, and gravitational waves," in Sachs 1971, pp. 238–283. [§40.7]

Thorne, K. S., 1972, "Nonspherical gravitational collapse, a short review," in Klauder 1972, pp. 231–258. [Box 32.3, Ex. 32.10]

Thorne, K. S., 1973a. "Relativistic shocks: The Taub adiabat," *Astrophys. J. 179,* 897–907. [§22.2, Box 22.2]

Thorne, K. S., 1973b, "Black-hole models for compact x-ray sources," *Astrophys. J.,* in press. [Box 33.3]

Thorne, K. S. *See also* Ames and Thorne 1968, Bardeen, Thorne, and Meltzer 1966, Bisnovatyi-Kogan and Thorne 1970, Fackerell, Ipser, and Thorne 1969, Fletcher *et al.* 1967, Harrison *et al.* 1965, Hartle and Thorne 1968 *and* 1974, Ipser and Thorne 1968, Meltzer and Thorne 1966, Morgan and Thorne 1973, Novikov and Thorne 1973, Press and Thorne 1972, *and* Trimble and Thorne 1969.

Thorne, K. S., W. T. Ni, and C. M. Will, 1971, "Theoretical frameworks for testing relativistic gravity: A review," in Davies 1971, pp. 10–31. [§17.6, Box 39.5]

Thorne, K. S., and C. M. Will, 1971, "Theoretical frameworks for texting relativistic gravity, I: Foundations," *Astrophys. J. 163,* 595–610. [§§38.2, 39.2]

Tikhonov, M. Yu. *See* Braginsky, Manukin, and Tikhonov 1970.

Tobey, R. G., *et al.* 1967, *PL/1 Formac Interpreter, Users Reference Manual* IBM Controlled Program Library 360D 03.3004, Hawthorne, N.Y. [Box 14.3]

Tolman, R. C., 1934a, *Relativity, Thermodynamics, and Cosmology,* Clarendon Press, Oxford. [§44.1, Ex. 22.7, endpapers]

Tolman, R. C., 1934b, "Effect of inhomogeneity on cosmological models," *Proc. Nat. Acad. Sci. U.S. 20,* 169–176. [§30.6, Ex. 32.8]

Tolman, R. C., 1939, "Static solutions of Einstein's field equations for spheres of fluid," *Phys. Rev. 55,* 364–373. [§23.2, Ex. 23.9]

Tomonaga, S., 1946, "On a relativistically invariant formulation of the quantum theory of wave fields," *Prog. Theor. Phys. 1,* 27–42, reprinted in Schwinger 1958. [Boxes 15.1, 21.1, Ex. 43.5]

Tonnelat, M. A., *quoted at end of* Chap. 44.

Torrence, R. J. *See* Couch *et al.* 1968 *and* Newman *et al.* 1965.

Trautman, A., 1965, "Foundations and current problems of general relativity theory," in Trautman, Pirani, and Bondi 1965. [§12.1, Boxes 8.1, 12.4, endpapers]

Trautman, A., 1972, "On the Einstein-Cartan equations, I," *Bulletin de l'Academie Polonaise des Sciences (math., astr., phys.) 20,* 185–190. [§§38.2, 39.2]

Trautman, A., F. A. E. Pirani, and H. Bondi, 1965, *Lectures on General Relativity,* Brandeis 1964 Summer Institute on Theoretical Physics, vol. I, Prentice-Hall, Englewood Cliffs, N.J.

Treder, H. J., 1962, *Gravitative Stosswellen, Nichtanalytischen Wellenlösungen der Einsteinschen Gravitationsgleichungen,* Akademie-Verlag, Berlin. [§21.13]

Treder, H. J. *See also* Papapetrou and Treder 1959 *and* 1962.

Trimble, V. L., and K. S. Thorne, 1969, "Spectroscopic binaries and collapsed stars," *Astrophys. J. 156,* 1013–1019. [Box 33.3]

Trumpler, R. J. *See* Campbell and Trumpler 1928.

Truran, J. *See* Cameron and Truran 1971.

Truran, J., and A. G. W. Cameron, 1971, "Evolutionary models of nucleosynthesis in the galaxy," *Astrophys. & Space Sci. 14,* 179–222. [Box 33.3]

Tukey, J. W. *See* Blackman and Tukey 1959.

Turkevich, A. *See* Fermi and Turkevich 1950.

Turner, K. C., and H. A. Hill, 1964, "New experimental limit on velocity-dependent interactions of clocks and distant matter," *Phys. Rev. B 134,* 252–256. [§38.7, Fig. 38.3]

Tyson, J. A., *cited by name in* Ex. 37.13. *See also* Douglass and Tyson 1971.

Ulam, S. *See* Kac and Ulam 1968.

Unruh, W. G., 1971, "Dirac particles and geometrodynamical charge in curved geometries," doctoral dissertation, physics, Princeton University. [§21.9]

Unt, V., and P. Keres, 1972. "Tails of waves in the Schwarzschild field" (in Russian), *Izv. Akad Nauk Estonian S.S.R., Fiz. Mat., 21,* 17–30. [§35.14]

Urey, H. C. *See* Ruark and Urey 1930.

Van de Kamp, P., 1958, "Visual binaries," *Ann. Phys. (U.S.A.) 50,* 187–224. [Box 36.3]

Van der Waerden, B. L., 1967, *Sources of Quantum Mechanics,* North-Holland, Amsterdam, Holland. Reprinted in 1968 by Dover, New York.

Varenna, International School of Physics "Enrico Fermi," organized at Varenna on Lake Como by the Italian Physical Society; 1961, *see* Møller 1962; 1965, *see* Gratton 1966; 1969, *see* Sachs 1971; 1972, *see* Bertotti 1973.

Vatican 1970, Study Week on the Nuclei of Galaxies, April 13–18, 1970, at the Vatican; proceedings published in *Semain d'etude sur les noyeaux des galaxies, Pontificae Academiae Scientarum Script Varia no. 35,* 1971.

Veblen, O., and J. H. C. Whitehead, 1932, *The Foundations of Differential Geometry,* Cambridge Univ. Press, Cambridge, Eng. Reprinted 1953. [§2.2]

Veron, P. *See* DeWitt, Schatzman, and Veron 1967.

Vishveshwara, C. V., 1970, "Stability of the Schwarzschild Metric," *Phys. Rev. D 1,* 2870–2879. [Box 33.3]

Volkoff, G. *See* Oppenheimer and Volkoff 1939.

Voltaire, François Marie Arouet de, 1738, *Éléments de Philosophie de Newton,* Pt. 3, chap. III; reprinted in vol. 5 of *Oevres de Voltaire,* Bacquenois, Paris, 1836. [epigraph to §1.1]

von Karman, T., *cited by name in* Box 38.2.

von Klüber, H., 1960, "The determination of Einstein's light deflection in the gravitational field of the sun," in Beer 1960, pp. 47–77. [Box 40.1]

von Oppolzer, T. R., 1887, "Canon der Finsternisse," *Sitzber. Kais. Akad. Wiss. Wien Math.—Nat. Kl. 52,* i-xviii, 1-321. [Fig. 1.8]

von Neumann, J., *cited by name in* Box 38.2.

Vourloumis, Hypatia, 1971, statement made at Knossos, May 1971, while serving as guide on Swan's Hellenic tour. *Quoted at end of* Chap. 44.

Vyasa, about 400 A.D., *The Mahabharata,* translated from the Sanskrit by P. Lal, Writers Workshop, Calcutta. [Epigraph to §1.5]

Waage, H., *cited by name in* Fig. 1.7.

Wagoner, R. V., 1967, "Some effects of an intervening galaxy on the radiation from very distant objects," *Astrophys. J. 149,* 465–480. [Ex. 29.8]

Wagoner, R. V., *cited by name in* Box 27.7. *See also* Bardeen and Wagoner 1971 *and* Salpeter and Wagoner 1971.

Wagoner, R. V., W. A. Fowler, and F. Hoyle, 1967, "On the synthesis of elements at very high temperatures," *Astrophys. J. 148,* 3–49. [§28.1]

Wakano, M. *See* Harrison *et al.* 1965 *and* Klauder *et al.* 1959.

Wald, R. M., 1971. *See* Cohen and Wald 1971.

Walker, A. G., 1932, "Relative coordinates," *Proc. R. Soc. Edinburgh 52,* 345–353. [§6.3]

Walker, A. G., 1936, "On Milne's theory of world-structure," *Proc. London Math. Soc. 42,* 90–127. [Ex. 27.5]

Walker, A. G. *cited by name in* §30.3, Box 27.7.

Walker, M., and R. Penrose, 1970, "On quadratic first integrals of the geodesic equations for type [22] spacetimes," *Commun. Math. Phys. 18,* 265–274. [§§33.4, 33.5]

Walker, R. *See* Mathews and Walker 1965.

Wallace, A. H., 1957, *An Introduction to Algebraic Topology,* Pergamon, London. [Box 34.1]

Walraven, Th. *See* Ledoux and Walraven 1958.

Wampler, E. J., *cited by name in* Box 27.7.

Wanders, G. *See* Steuckelberg and Wanders 1953.

Ward, W., 1970, "General-relativistic light deflection for the complete celestial sphere," *Astrophys. J. 162,* 345–348. [§40.3]

Warrington, J., trans. 1956, *Aristotle's Metaphysics,* E. P. Dutton, New York. [Box 27.7]

Wasserburg, G. J. *See* Papanastassiou and Wasserburg 1969.

Wasserburg, G. J., and D. S. Burnett, 1968, "The status of isotopic age determinations on iron and stone meteorites," in P. M. Millman, ed., *Meteorite Research,* Reidel, Dordrecht, Holland, pp. 467–479. [§29.7, Box 29.3]

Wasserburg, G. J., D. N. Schramm, and J. C. Huneke, 1969, "Nuclear chronologies for the galaxy," *Astrophys. J. 157,* L91–L96. [§29.7]

Weber, H., ed., 1953, *B. Riemann: Gesammelte Mathematische Werke,* 2d ed., paperback reprint, Dover, New York. The *Habilitationsvorlesung* of 1854 was translated into English by W. K. Clifford, *Nature, 8,* 14 (1873). [Box 8.5]

Weber, J., 1960, "Detection and generation of gravitational waves," *Phys. Rev. 117,* 307–313. [§37.3]

Weber, J., 1961, *General Relativity and Gravitational Waves,* Wiley-Interscience, New York. [§37.3, endpapers]

Weber, J., 1967, "Gravitational radiation," *Phys. Rev. Lett. 18,* 498–501. [§37.3, Fig. 37.2]

Weber, J., 1969, "Evidence for discovery of gravitational radiation," *Phys. Rev. Lett. 22,* 1320–1324. [§§37.3, 37.10, Fig. 37.2]

Weber, J., 1970a, "Gravitational radiation experiments," *Phys. Rev. Lett. 24,* 276–279. [§§37.3, 37.10]

Weber, J., 1970b, "Anisotropy and polarization in the gravitational-radiation experiments," *Phys. Rev. Lett. 25,* 180–184. [§§37.3, 37.10]

Weber, J., *cited by name in* Preface, Box 17.3.

Weil, A. *See* Allendoerfer and Weil 1943.

Weiler, K. W. *See* Seielstad, Sramek, and Weiler 1970.

Weinberg, J. W. *See* Tauber and Weinberg 1961.

Weinberg, S., 1965, "Photons and gravitons in perturbation theory: Derivation of Maxwell's and Einstein's equations," *Phys. Rev. B 138,* 988–1002. [§§7.1, 18.1]

Weinberg, S., 1972, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity,* Wiley, New York. [endpapers]

Weingarten, J., *cited by name in* §21.5.

Weisskopf, V. F. *See* Blatt and Weisskopf 1952 *and* Fierz and Weisskopf 1960.

Wells, J. S. *See* Evenson *et al.* 1972.

Welton, T. A., 1948, "Some observable effects of the quantum-mechanical fluctuations of the electromagnetic field," *Phys. Rev. 74,* 1157–1167. [Box 17.2 part 6]

Wentzel, G., 1949, *Quantum Theory of Fields,* Interscience, New York. [§21.3, Boxes 7.1, 21.2]

Werner, F. G., 1969, private remark to J. A. Wheeler on June 3 at the Cincinnati, Ohio, Relativity Conference in the Midwest. [§44.5]

Westermann, A. and J. F. Boissonade, eds. 1878, "Marini vita Procli," xxxvii, *in Dioginis Laertii de Clarorum Philosophorum,* II, Paris, p. 169. [Fig. 1.8]

Weyl, H., 1922, *Space-Time-Matter,* translated by H. L. Brose, Methuen, London. [endpapers]

Weyl, H., 1924a, "Massenträgheit und Kosmos: Ein Dialog," *Naturwissenschaften 12,* 197–204. [§21.12]

Weyl, H., 1924b, *Was Ist Materie,* Springer, Berlin, Germany. [Box 15.1]

Weyl, H., 1925, "Theorie der Darstellung kontinuierlichen halb-einfacher Gruppen durch lineare Transformationen, I," *Math. Z. 23,* 271–309. [§2.2, Epigraph to Chap. 2]

Weyl, H., 1937, lectures at Princeton University. [§4.5]

Weyl, H., 1946, *Classical Groups: Their Invariants and Representation,* Princeton Univ. Press, Princeton, N.J., 2d ed. [Box 27.2]

Weyl, H., 1949, *Philosophy of Mathematics and Natural Science,* Princeton Univ. Press, Princeton, N.J., p. 116. [Epigraph to Chap. 6, Box 8.5]

Weyl, H., *cited by name in* §4.5, Boxes 16.4, 27.7, *quoted in* epigraph of Chap. 10. *See also* Lorentz *et al.* 1923.

Wheeler, J. A., 1955, "Geons," *Phys. Rev. 97,* 511–536. [§§21.12, 25.5, Box 15.1, Ex. 32.11]

Wheeler, J. A., 1959, "View that the distribution of mass and energy determines the metric," in *Onzième Conseil de Physique Solvay: La Structure et l'évolution de l'univers,* Éditions Stoops, Brussels, Belgium. [§§21.12, 43.1]

Wheeler, J. A., 1961, "Geometrodynamics and the problem of motion," *Rev. Mod. Phys. 33,* 63–78. [§20.6]

Wheeler, J. A., 1962, *Geometrodynamics,* Academic Press, New York. [§§43.4, 44.3, Fig. 44.1, Boxes 33.3, 36.2, 44.3]

Wheeler, J. A., 1964a, "Geometrodynamics and the issue of the final state," in DeWitt and DeWitt 1964b. [§§27.10, 21.11, 34.6, 35.13, 42.6, 43.1, Boxes 27.7, 44.4, Ex. 21.18]

Wheeler, J. A., 1964b, "Gravitation as geometry, II," in Chiu and Hoffman 1964. [§15.3]

Wheeler, J. A., 1964c, "Mach's principle as boundary condition for Einstein's equations," in Chiu and Hoffman 1964, pp. 303–349. [§§21.10, 21.12, 43.1]

Wheeler, J. A., 1966, "Superdense stars," in *Annual Review Astron. and Astrophys. Vol. 4,* L. Goldberg, ed., Palo Alto, Calif. pp. 393–432. [§24.3, Ex. 36.3]

Wheeler, J. A., 1968a, "Superspace and the nature of quantum geometrodynamics," in DeWitt and Wheeler 1968. [§§21.1, 21.8, Box 44.3]

Wheeler, J. A., 1968b, *Einstein's Vision,* Springer, Berlin. Contained in part in Wheeler 1968a. [§§21.8, 21.12, Ex. 21.23]

Wheeler, J. A., 1970, "Superspace," in Gilbert and Newton 1970, pp. 335–378. [§43.1]

Wheeler, J. A., 1971a, notebook entry, "Pregeometry and the calculus of propositions," 9:10 A.M., April 10; seminar, Department of Mathematics, Kings College, London, May 10; letter to L. Thomas, "Pregeometry and propositions," June 11, unpublished. [Box 44.5]

Wheeler, J. A., 1971b, "Transcending the law of conservation of leptons," in *Atti del Convegno Internazionale sul Tema: The Astrophysical Aspects of the Weak Interactions; Quaderno N. 157,* Accademia Nazionale dei Lincei, Roma, pp. 133–164. [§§41.11, 44.6, Box 25.2]

Wheeler, J. A., 1971c, "From Mendeléev's atom to the collapsing star," 189–233 in *Atti del Convegno Mendeléeviano,* Accademia delle Scienze, Torino; reprinted in *Trans. New York Acad. Sci., 33,* 745–779 (1971). *See also* Wyler 1971. [§28.3, Box 25.2]

Wheeler, J. A., 1971d, "Mechanisms for jets," in *Study week on nuclei of galaxies, Pontificiae Academiae Scientarum Scripta Varia no. 35,* 539–567. [Box 33.3]

Wheeler, John Archibald, *Festschrift* on the occasion of his sixtieth birthday, July 9, 1971. *See* Klauder 1972.

Wheeler, J. A., *cited by name in* §21.1, Box 8.5. *See also* Baierlein, Sharp, and Wheeler 1962, Brill and Wheeler 1957, DeWitt and Wheeler 1968, Douglas *et al.* 1969, Ford and Wheeler 1959a *and* 1959b, Fuller and Wheeler 1962, Harrison *et al.* 1965, Harrison and Wheeler 1958, Klauder *et al.* 1959, Lindquist and Wheeler 1957, Marzke and Wheeler 1964, Misner and Wheeler 1957 *and* 1972, Power and Wheeler 1957, Regge and Wheeler 1957, Ruffini and Wheeler 1970, 1971a, *and* 1971b, *and* Taylor and Wheeler 1966.

Wheeler, J. A., and R. P. Feynman, 1945, "Interaction with the absorber as the mechanism of radiation," *Rev. Mod. Phys. 17,* 157–181. [§4.7, Box 2.1]

Wheeler, J. A., and R. P. Feynman, 1949, "Classical electrodynamics in terms of direct inter-particle action," *Rev. Mod. Phys. 21,* 425–434. [§4.7, Box 2.1]

White, R. H. *See* Colgate and White 1966 *and* May and White 1966 *and* 1967.

Whitehead, A. N., 1922, *The Principle of Relativity,* Cambridge Univ. Press, Cambridge, Eng. [§§38.2, 39.1, 39.3, 39.8, Box 17.3]

Whitehead, A. N., 1929, *The Function of Reason,* Princeton Univ. Press, Princeton, N.J. [Epigraph to Chap. 28]

Whitehead, J. H. C. *See* Veblen and Whitehead 1932.

Whitrow, G. J., and G. E. Morduch, 1965, "Relativistic theories of gravitation, a comparative analysis with particular reference to astronomical tests, in *Vistas in Astronomy 6,* 1–67. [§38.2]

Whittaker, E. R., 1904, "On an expression of the electromagnetic field due to electrons by means of two scalar potential functions," *Proc. R. Soc. London 1,* 367–372. [Ex. 4.10]

Wigner, E. P., 1960, "The unreasonable effectiveness of mathematics in the natural sciences," *Commun. Pure & Appl. Math. 13,* 1–14, reprinted in *Symmetries and Reflections: Scientific Essays of Eugene P. Wigner,* W. J. Moore and M. Scriven, eds., Indiana Univ. Press, Bloomington, Ind., 1967. [§44.2]

Wigner, E. P., *cited by name in* Box 38.2. *See also* Bargmann and Wigner 1948 *and* Salam and Wigner 1972.

Wilkins, D., 1972, "Bound geodesics in the Kerr metric," *Phys. Rev. D 5,* 814–822. [Box 33.5]

Wilkinson, D. T., *cited by name in* Box 27.7. *See also* Bender *et al.* 1971 *and* Dicke *et al.* 1965.

Will, C. M., 1971a, "Relativistic gravity in the solar system, I: Effect of an anisotropic gravitational mass on the earth-moon distance," *Astrophys. J. 165,* 409–412. [Box 39.5]

Will, C. M., 1971b, "Relativistic gravity in the solar system, II: Anisotropy in the Newtonian gravitational constant," *Astrophys. J. 169,* 141–156. [§§38.2, 39.1, 40.8, Boxes 17.3, 39.5]

Will, C. M., 1971c, "Theoretical frameworks for testing relativistic gravity, II: Parametrized post-Newtonian hydrodynamics, and the Nordtvedt effect," *Astrophys. J. 163,* 611–628. [§§38.2, 39.3, 40.9, Box 39.5]

Will, C. M., 1971d, "Theoretical frameworks for testing relativistic gravity, III: Conservation laws, Lorentz invariance, and values of the PPN parameters," *Astrophys. J. 169,* 125–140. [§39.9, Box 39.5, Ex. 39.15]

Will, C. M., 1972. "The theoretical tools of experimental gravitation," Caltech Orange-Aid Preprint 289, to be published in Bertotti 1973. [§§38.2, 39.3, Box 39.5]

Will, C. M., 1973, "Relativistic gravity in the solar system, III: Experimental disproof of a class of linear theories of gravitation," *Astrophys. J.,* in press. [§§39.3, 39.8]

Will, C. M. *See also* Nordtvedt and Will 1972, Thorne, Ni, and Will 1971, *and* Thorne and Will 1971.

Will, C. M., and K. Nordtvedt, Jr., 1972, "Conservation laws and preferred frames in relativistic gravity, I: Preferred-frame theories and an extended PPN formalism," *Astrophys. J. 177,* 757–774. [§§38.2, 39.2, 39.3, 39.9, 40.8, Boxes 39.1, 39.5]

Willey, R. *See* Klauder *et al.* 1959.

Williams, R. W. *See* Bailey *et al.* 1968.

Willmore, T. J., 1959, *An Introduction to Differential Geometry,* Clarendon Press, Oxford, [Box 8.1]

Wilson, O. C. *See* Minkowski and Wilson 1956.

Wilson, R. W., *cited by name in* Box 27.7. *See also* Penzias and Wilson 1965.

Winsor, F., *quoted at end of* Chap. 44.
Wintner, A., 1941, *Analytical Foundations of Celestial Mechanics,* Princeton Univ. Press, Princeton, N.J. [Box 25.4]
Witteborn, F. C., and W. M. Fairbank, 1967, "Experimental comparison of the gravitational force on freely falling electrons and metallic electrons," *Phys. Rev. Lett. 19,* 1049–1052. [§1.3]
Witteborn, F. C., L. V. Knight, and W. M. Fairbank, 1965, "Use of low-temperature techniques to measure gravitational forces on charged particles," in *Proceedings of the 9th International Conference on Low-Temperature Physics, Columbus 1964,* Plenum Press, New York, 1965, part B, pp. 1248–1252. [§1.3]
Witten, L., ed., 1962, *Gravitation: An Introduction to Current Research,* Wiley, New York.
Witten, L. *See also* Carmeli, Fickler, and Witten 1970.
Wong, C.-Y., 1971, "Application of Regge calculus to the Schwarzschild and Reissner-Nordstrom geometries at the moment of time-symmetry," *J. Math. Phys. 12,* 70–78. [§§42.6, 42.7]
Woolf, N. *See* Douglas *et al.* 1969.
Wright, J. P., 1964, "General relativistic instability," *Phys. Rev. B 136,* 288–289. [Box 26.2]
Wright, T., 1750, *An Original Theory or New Hypothesis of the Universe,* H. Chapelle, London. Reprinted in Munitz 1957. [Box 27.7]
Wyler, J. A. (pseudonym), 1971, "Rasputin, science, and the transmogrification of destiny," preprint bearing the Princeton University seal.
York, J. W., Jr., 1971, "Gravitational degrees of freedom and the initial-value problem," *Phys. Rev. Lett. 26,* 1656–1658. [§§21.11, 21.12, Ex. 21.22]
York, J. W., Jr., 1972a, personal communication. [Ex. 21.8, 21.9, 21.24]
York, J. W., Jr., 1972b, "The role of conformal 3-geometry in the dynamics of gravitation," *Phys. Rev. Letters 28,* 1082–1085. [§§21.6, 21.11, Ex. 21.24]
York, J. W., Jr., 1973a, "Conformally invariant orthogonal decomposition of symmetric tensors on Riemannian manifolds and the initial-value problem of general relativity," *J. Math. Phys. 14,* in press. [§21.11]
York, J. W., Jr., 1973b, personal communication. [§§21.10, 21.11]
York, J. W., Jr., *cited by name in* §21.1, Ex. 21.8, 21.9. *See also* O'Murchadha and York 1973.
Yourgrau, W., and S. Mandelstam, 1968, *Variational Principles in Dynamics and Quantum Theory,* W. B. Saunders, Philadelphia, Pa. [Box 21.1]
Yu, J. T. *See* Peebles and Yu 1970.
Zapolsky, H. S., 1968, "Can the redshift of quasistellar objects be gravitational?" *Astrophys. J. Lett. 153,* 163–169. [Ex. 25.29]
Zaret, H., *quoted at end of* Chap. 44.
Zeeman, E. C., 1964, "Causality implies the Lorentz group," *J. Math. Phys. 5,* 190–493. [Box 2.1]
Zel'dovich, Ya. B., 1963. "The theory of the expanding universe as originated by A. A. Friedmann," *Usp. Fiz. Nauk, 80,* 357–390. English translation in *Sov. Phys.—Uspekhi. 6,* 475–494 (1964). [§29.2]
Zel'dovich, Ya. B., 1967, "Cosmological constant and elementary particles," *Zh. Eksp. & Teor. Fiz. Pis'ma 6,* 883–884. English translation in *Sov. Phys.—JETP Lett. 6,* 316–317. [§17.3, Ex. 17.5, Box 17.2 part 6]
Zel'dovich, Ya. B., 1968, "The cosmological constant and the theory of elementary particles," *Usp. Fiz. Nauk 95,* 209–230. English translation in *Sov. Phys.—Uspekhi 11,* 381–393. [Ex. 17.5]
Zel'dovich, Ya. B., 1970, "Particle production in cosmology," *Zh. Eksp. Teor. Fiz.—Pis'ma 12,* 443–447. English translation in *Sov. Phys.—JETP Lett. 12,* 307–311. [§30.5, Ex. 27.12]
Zel'dovich, Ya. B., 1971, "Particle creation by gravitational fields in collapse and singularities," *Comm. Astrophys. & Space Phys. 3,* 179–184. [Ex. 27.12]
Zel'dovich, Ya. B., 1972, "Creation of particles and antiparticles in an electric and gravitational field," in Klauder 1972, pp. 277–288. [§30.8, Ex. 27.12]
Zel'dovich, Ya. B. *See also* Bisnovatyi-Kogan and Zel'dovich 1969, Braginsky, Zel'dovich, and Rudenko 1969, *and* Doroshkevich, Zel'dovich, and Novikov 1965.
Zel'dovich, Ya. B., and O. H. Guseynov, 1965, "Collapsed stars in binaries," *Astrophys. J. 144,* 840–841. [Box 33.3]
Zel'dovich, Ya. B., and I. D. Novikov, 1966, "The hypothesis of cores retarded during expansion and the hot cosmological model," *Astron. Zh. 43,* 758–760. English translation in *Sov. Astron.—A. J. 10,* 602–603. [Box 33.3]
Zel'dovich, Ya. B., and I. D. Novikov, 1971, *Relativistic Astrophysics, Vol. I: Stars and Relativity,* Univ. of Chicago Press, Chicago, Ill. [§§22.2, 23.1, 24.1, 24.3, 24.4, 24.6, 25.6, 32.1, Boxes 24.1, 24.2]
Zel'dovich, Ya. B., and I. D. Novikov, 1974, *Relativistic Astrophysics, Vol. II: The Universe and Relativity,* Univ. of Chicago Press, Chicago, Ill., in press. [§§23.1, 28.1, 28.2, 30.1]
Zel'dovich, Ya. B., and M. A. Podurets, 1965, "The evolution of a system of gravitationally interacting point masses," *Astron. Zh. 42,* 963–973. English translation in *Sov. Astron.—A. J. 9,* 742–749 (1966). [§§24.6, 25.7, Boxes 24.1, 25.9]
Zel'dovich, Ya. B., and Yu. P. Rayzer, 1967, *Physics of Shock Waves and High-temperature Hydrodynamic Phenomena,* Academic Press, New York. [§22.2, Box 22.2]
Zel'dovich, Ya. B., and A. A. Starobinsky, 1971, "Particle production and vacuum polarization in an anisotropic gravitational field," *Zh. Eksp. Teor. Fiz. 61,* 2161–2175. English translation in *Sov. Phys.—JETP 34,* 1159–1166 (1972). [§30.5]
Zeno, *cited by name in* Box 30.1; *see* A. Grünbaum, 1967, *Modern Science and Zeno's Paradoxes,* Wesleyan Univ. Press, Middletown, Conn.
Zerilli, F. J., 1970a, "Effective potential for even-parity Regge-Wheeler gravitational perturbation equations," *Phys. Rev. Lett. 24,* 737–738. [Box 33.3]
Zerilli, F. J., 1970b, "Gravitational field of a particle falling in Schwarzschild geometry analyzed in tensor harmonics," *Phys. Rev. D 2,* 2141–2160. [§36.5, Fig. 36.2, Box 33.3]
Zimmer, H., 1946, *Myths and Symbols in Indian Art and Civilization,* Bollingen Foundation, Washington, D.C.; paperback reprint, Harper and Row, New York, 1962. [Box 27.7]
Zimmerman, B. A. *See* Fletcher *et al.* 1967.
Zwicky, F. *See* Baade and Zwicky 1934a *and* 1934b.

SUBJECT INDEX

D

E

F

J

K

L

M

N

O

P

Q

R

S

T

U

V

TABLE OF SIGN CONVENTIONS

Reference	g sign	Riemann	Einstein	Spacetime four-dimensional indices
Landau, Lifshitz (1962) "spacelike convention"	+	+	+	latin
Landau, Lifshitz (1971) "timelike convention"	−	+	+	latin
Misner, Thorne, Wheeler (1973; this text)	+	+	+	greek
Adler, Bazin, Schiffer (1965)	−	−	−	greek
Anderson (1967)	−	−	−[b]	greek
Bergmann (1942)	−	−[a]	−	greek
Cartan (1946)		−	−	
Davis (1970)	−	+	−	latin
Eddington (1922)	−	+	−	greek
Ehlers (1971)	+	+	+	latin
Einstein (1950)	−	+	−	greek
Eisenhart (1926)		+	−	
Fock (1959)	−	−[a]	−	greek
Fokker (1965)	−	−	+	latin
Hawking and Ellis (1973)	+	+	+	latin
Hicks (1965)		+	+	
Infeld, Plebanski (1960)	−	+	−	greek
Lichnerowicz (1955)	−	+	+	greek
McVittie (1956)	−	+	−	greek
Misner (1969a)	+	+	+	greek
Møller (1952)	+	−	−	latin
Pauli (1958)	+	−	−	latin
Penrose (1968)	−	−	−	latin
Pirani (1965)	−	−	−	latin
Robertson, Noonan (1968)	+	+	−	latin
Sachs (1964)	±	+	+	latin
Schild (1967)	−	+	−	latin
Schouten (1954)		−	+	
Schroedinger (1950)	−	+	−	latin
Synge (1960b)	+	+	−	latin
Thorne (1967)	−	+	+	greek
Tolman (1934a)	−	+	−	greek
Trautman (1965)	−	−	−	latin
Weber (1961)	+	+	+	greek
Weinberg (1972)	+	−	−	greek
Weyl (1922)	−	+	+	latin
Wheeler (1964a)	+	+	+	greek

[a] Unusual index positioning on ***Riemann*** components gives a different sign for $R_{\mu\nu\alpha\beta}$.
[b] Note: his $\kappa < 0$ is the negative of the gravitational constant.

SIGN CONVENTIONS

This book follows the "Landau-Lifshitz Spacelike Convention" (LLSC). Arrows below mark signs that are "+" in it. The facing table shows signs that other authors use.

$$+\boldsymbol{g} = -(\boldsymbol{\omega}^0)^2 + (\boldsymbol{\omega}^1)^2 + (\boldsymbol{\omega}^2)^2 + (\boldsymbol{\omega}^3)^2$$

g sign
(col. 2)

$$+\mathcal{R}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{\nabla}_{\boldsymbol{u}} \boldsymbol{\nabla}_{\boldsymbol{v}} - \boldsymbol{\nabla}_{\boldsymbol{v}} \boldsymbol{\nabla}_{\boldsymbol{u}} - \boldsymbol{\nabla}_{[\boldsymbol{u},\boldsymbol{v}]}$$

$$+R^{\mu}{}_{\nu\alpha\beta} = \partial_\alpha \Gamma^{\mu}{}_{\nu\beta} - \partial_\beta \Gamma^{\mu}{}_{\nu\alpha} + \Gamma^{\mu}{}_{\sigma\alpha}\Gamma^{\sigma}{}_{\nu\beta} - \Gamma^{\mu}{}_{\sigma\beta}\Gamma^{\sigma}{}_{\nu\alpha}$$

Riemann sign
(col. 3)

$$+R_{\mu\nu} = R^{\alpha}{}_{\mu\alpha\nu}$$

quotient of ***Einstein***
and ***Riemann*** signs

$$\textbf{\textit{Einstein}} = +8\pi \boldsymbol{T}$$

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = +8\pi T_{\mu\nu}$$

Einstein sign
(col. 4)

$$T_{\hat{0}\hat{0}} = \boldsymbol{T}(\boldsymbol{e}_{\hat{0}}, \boldsymbol{e}_{\hat{0}}) > 0$$

all authors agree
on this "positive
energy density" sign

The above sign choice for ***Riemann*** is convenient for coordinate-free methods, as in the curvature operator $\mathcal{R}(\boldsymbol{u}, \boldsymbol{v})$ above, in the curvature 2-forms (equation 14.19), and for matrix computations (exercise 14.9). The definitions of ***Ricci*** and ***Einstein*** with the signs adopted above are those that make their eigenvalues (and $R \equiv R^{\mu}{}_{\mu}$) positive for standard spheres with positive definite metrics.

SOME USEFUL NUMBERS IN CONVENTIONAL AND GEOMETRIZED UNITS*

Fundamental Constants

Gravitation constant	$G = (6.673 \pm 0.003) \times 10^{-8}\ \mathrm{cm^3/g\ sec^2} = 1$
Speed of light	$c = (2.997\ 924\ 562 \pm 0.000\ 000\ 011) \times 10^{10}\ \mathrm{cm/sec} = 1$
useful combinations	$G/c^2 = 0.7425 \times 10^{-28}\ \mathrm{cm/g} = 1.476\ 64\ \mathrm{km}/M_\odot = 1$
	$c^5/G = 3.629 \times 10^{59}\ \mathrm{erg/sec} = 2.030 \times 10^5 M_\odot c^2/\mathrm{sec} = 1$ (emission factor)
	$G/c = 2.226 \times 10^{-18}\ \mathrm{cm^2\ Hz/g} = 1$ (receptor factor)
	$c^2/G^{1/2} = 3.479 \times 10^{24}\ \mathrm{gauss\ cm} = 3.479 \times 10^{24}\ \mathrm{statvolt} = 1$
Planck's reduced constant	$\hbar = (1.054\ 592 \pm 0.000\ 008) \times 10^{-27}\ \mathrm{g\ cm^2/sec} = 2.612 \times 10^{-66}\ \mathrm{cm^2}$
Planck distance, L^*	$(\hbar G/c^3)^{1/2} = 1.616 \times 10^{-33}\ \mathrm{cm}$
Planck time, T^*	$(\hbar G/c^5)^{1/2} = 5.391 \times 10^{-44}\ \mathrm{sec}$
Planck mass, M^*	$(\hbar c/G)^{1/2} = 2.177 \times 10^{-5}\ \mathrm{g}$
Planck density, M^*/L^{*3}	$c^5/\hbar G^2 = 5.157 \times 10^{93}\ \mathrm{g/cm^3}$
Quantum of charge	$e = (4.803\ 25 \pm 0.000\ 02) \times 10^{-10}\ (\mathrm{g\ cm^3/sec^2})^{1/2} = 1.381 \times 10^{-34}\ \mathrm{cm}$
Reciprocal fine-structure constant, $1/\alpha$	$\hbar c/e^2 = 137.0360 \pm 0.0002$
Electron rest mass	$m_e = (9.109\ 56 \pm 0.000\ 05) \times 10^{-28}\ \mathrm{g} = 8.1873 \times 10^{-7}\ \mathrm{erg} = 0.511\ 004\ \mathrm{MeV}$
	$= 5.9301 \times 10^9\ \mathrm{K} = 6.764 \times 10^{-56}\ \mathrm{cm}$
Proton rest mass	$M_p = (1.672\ 614 \pm 0.000\ 012) \times 10^{-24}\ \mathrm{g} = 1.503\ 27 \times 10^{-3}\ \mathrm{erg} = 0.938\ 259\ \mathrm{GeV}$
	$= 1.088\ 83 \times 10^{13}\ \mathrm{K} = 1.2419 \times 10^{-52}\ \mathrm{cm}$
Bohr radius	$a_0 = \hbar^2/m_e e^2 = 0.529\ 177 \times 10^{-8}\ \mathrm{cm}$
Reduced Compton wavelength	$\lambda\!\!\!^{-}_e = \hbar/m_e c = 3.861\ 59 \times 10^{-11}\ \mathrm{cm}$
Classical electron radius	$r_0 = e^2/m_e c^2 = 2.817\ 94 \times 10^{-13}\ \mathrm{cm}$
Atomic energy unit	$e^2/a_0 = m_e e^4/\hbar^2 = 4.359\ 83 \times 10^{-11}\ \mathrm{erg} = 27.2116\ \mathrm{eV} = 3.157\ 86 \times 10^5\ \mathrm{K} = 3.602 \times 10^{-60}\ \mathrm{cm}$

Conversion Factors (see also "fundamental constants," above)

Distance	$1\ \mathrm{pc} = 3.0856 \times 10^{18}\ \mathrm{cm}$; $1\ \mathrm{lt\text{-}yr} = 0.94605 \times 10^{18}\ \mathrm{cm}$; $1\ \mathrm{A.U.} = 1.495\ 985 \times 10^{13}\ \mathrm{cm}$
Time	$1\ \mathrm{yr} = 3.155\ 692\ 6 \times 10^7\ \mathrm{sec}$; $1\ \mathrm{day} = 86\ 400.\ \mathrm{sec}$; $1\ \mathrm{sidereal\ day} = 86\ 164.091\ \mathrm{sec}$
Mass, energy, temperature	$1\ \mathrm{eV} = 1.160\ 48 \times 10^4\ \mathrm{K} = 1.602\ 192 \times 10^{-12}\ \mathrm{ergs} = 1.782\ 68 \times 10^{-33}\ \mathrm{g} = 1.324 \times 10^{-61}\ \mathrm{cm}$

Electromagnetic Radiation

Blackbody radiation	
energy density $= aT^4$,	$a = 7.5647 \times 10^{-15}$ erg cm^{-3} K^{-4}
emittance $= \sigma T^4$,	$\sigma = \frac{1}{4}ac = 5.6696 \times 10^{-5}$ erg cm^{-2} sec^{-1} K^{-4}
peak of $dE/d\lambda$ spectrum	$\lambda_{\max} T = 0.289\ 79$ cm K $= (0.289\ 79$ Å$)(10^8$ K$)$
Free-free emittance of ionized hydrogen,	$\epsilon = (6 \times 10^{20}$ erg/g sec$)(\rho_{\mathrm{g\,cm^{-3}}})\, T_{\mathrm{K}}{}^{1/2}$
Free-free opacity of ionized hydrogen,	$\bar{\kappa}_{\mathrm{ff}} = (0.8 \times 10^{23}$ cm^2/g$)(\rho_{\mathrm{g\,cm^{-3}}})\, T_{\mathrm{K}}{}^{-7/2}$
Thomson cross section,	$\sigma_T = (8\pi/3)r_0{}^2 = 0.665\ 24 \times 10^{-24}$ cm^2
Electron scattering opacity of ionized hydrogen,	$\kappa_{\mathrm{es}} = 0.398$ cm^2/g

The Universe

Solar mass	$M_\odot = (1.476\ 64 \pm 0.000\ 02) \times 10^5$ cm $= 1.989 \times 10^{33}$ g $= 4.9255\ \mu$sec
Solar radius	$R_\odot = (6.9598 \pm 0.0007) \times 10^{10}$ cm
Solar luminosity	$L_\odot = (3.90 \pm 0.04) \times 10^{33}$ erg/sec $= (1.07 \times 10^{-26})c^5/G$
Earth-Sun distance	1 A.U. $= (1.495\ 985 \pm 0.000\ 005) \times 10^{13}$ cm $= 499.007$ sec
Earth mass	$M_\oplus = (5.977 \pm 0.004) \times 10^{27}$ g $= 0.4438$ cm
Earth mean radius	$R_\oplus = 6.371\ 03 \times 10^8$ cm
Milky Way (Galaxy) mass	$M_{\mathrm{G}} = 1.8 \times 10^{11} M_\odot = 2.7 \times 10^{16}$ cm $= 0.028$ lt-yr
Distance to Galactic center	$r_\odot = 10$ kpc $= 30\ 000$ lt-yr
Galaxy diameter	$D_{\mathrm{G}} \sim (30$ to $50)$ kpc $\sim (100{,}000$ to $150{,}000)$ lt-yr
Galaxy thickness at Sun's location	$h_{\mathrm{G}} \approx 500$ pc ≈ 1500 lt-yr
Nearby galaxies: distance and mass	Large Magellanic Cloud, 52 kpc, $10^{10.1} M_\odot$; Small Magellanic Cloud, 54 kpc, $10^{9.2} M_\odot$; Andromeda Nebula (M31), 570 kpc, $10^{11.5} M_\odot$.
Clusters of galaxies: members, distance, Diameter	Local group, 16 members; $d \sim D \sim 0.4$ Mpc; Virgo cluster, 2500 members, $d \sim 11$ Mpc, $D \sim 12$ Mpc; "typical" cluster, 130 members, $D \sim 3$ Mpc.
Hubble constant	$H = 55$ km sec^{-1} Mpc^{-1} $= (1.7 \times 10^{28}$ cm$)^{-1} = (18 \times 10^9$ yr$)^{-1}$
Universal density of luminous matter	$\langle \rho \rangle = 2 \times 10^{-31}$ g/cm^3

* Numerical data based on, or calculated from Söding *et al.* (1971) and Allen (1963), with supplements from the literature.

SCIENCE—PHYSICS

GRAVITATION

Charles W. MISNER
UNIVERSITY OF MARYLAND

Kip S. THORNE
CALIFORNIA INSTITUTE OF TECHNOLOGY

John Archibald WHEELER
PRINCETON UNIVERSITY

"Einstein's description of gravitation as curvature of spacetime led directly to that greatest of all predictions of his theory, that the universe itself is dynamic. Physics still has far to go to come to terms with this amazing fact and what it means for man and his relation with the universe."

JOHN ARCHIBALD WHEELER

Put as simply as possible, this is a book on Einstein's theory of gravity (general relativity). It is the first textbook on the subject that uses throughout the modern formalism and notation of differential geometry, and it is the first book to document in full the revolutionary techniques developed during the past decade to test the theory of general relativity. (Such tests include the use of atomic clocks, long baseline radio interferometry, interplanetary radar, lunar laser ranging, and interplanetary spacecraft.) Among the topics treated in depth are relativistic stars and star clusters and their possible roles in pulsars and quasars, recent developments in cosmology, gravitational collapse and black holes, and the generation, propagation, and detection of gravitational waves.

All three authors are uniquely qualified to have written this authoritative work: during the past decade, they and their research groups have been responsible for approximately half of the theoretical developments in the areas of concentration mentioned. Here is a book that is without competition and without peer—a book that belongs on the shelves and in the libraries of all who are seriously interested in modern science.

CHARLES W. MISNER is Professor of Physics at the University of Maryland. His Ph.D. (1957) is from Princeton. He has taught at Princeton and Maryland and has been a visiting professor at the California Institute of Technology, Cambridge University, and Oxford University.

KIP S. THORNE is Professor of Theoretical Physics at the California Institute of Technology and Adjunct Professor of Physics at the University of Utah. He has also taught at the University of Chicago. He received his Ph.D. from Princeton in 1965.

JOHN ARCHIBALD WHEELER is Joseph Henry Professor of Physics at Princeton University. His Ph.D. (1933) is from the Johns Hopkins University. More about him will be found in *Magic Without Magic: John Archibald Wheeler* ("A Collection of Essays in Honor of His Sixtieth Birthday"), John R. Klauder, Editor (W. H. Freeman and Company, 1972).

W. H. FREEMAN AND COMPANY
660 Market Street, San Francisco, California 94104
58 Kings Road, Reading, England RG1 3AA

Cover drawing by Kenneth Gwin.

ISBN 0-7167-0344-0