mensional, so any bundle filling it has a two-dimensional cross section.) At any event $\mathcal{P}_0$, in any local Lorentz frame there, on a "slice of simultaneity" $x^0 =$ constant, measure the cross-sectional area $\mathscr{A}$ of the bundle. (Note: the area being measured is perpendicular to $\boldsymbol{k}$ in the three-dimensional Euclidean sense; it can be thought of as the region occupied momentarily by a group of photons propagating along, side by side, in the $\boldsymbol{k}$ direction.) Show that the area $\mathscr{A}$ is the same, at a given event $\mathcal{P}_0$, *regardless of what Lorentz frame is used to measure it; but the area changes from point to point along the ray* $\mathcal{C}_0$ *as a result of the rays' divergence away from each other or convergence toward each other:*

$$\partial_{\boldsymbol{k}} \mathscr{A} = (\boldsymbol{\nabla} \cdot \boldsymbol{k}) \mathscr{A}. \tag{22.36}$$

Then show that $\mathscr{A} a^2$ is a constant everywhere along the ray $\mathcal{C}_0$ ("conservation of photon flux"). [Hints: (i) Any vector $\boldsymbol{\xi}$ connecting adjacent rays in the bundle is perpendicular to $\boldsymbol{k}$, because $\boldsymbol{\xi}$ lies in a surface of constant θ and $\boldsymbol{k} \cdot \boldsymbol{\xi} = \langle \tilde{\boldsymbol{k}}, \boldsymbol{\xi} \rangle = \langle \boldsymbol{d}\theta, \boldsymbol{\xi} \rangle =$ (change in θ along $\boldsymbol{\xi}$) $= 0$. (ii) Consider, for simplicity, a bundle with rectangular cross section as seen in a specific local Lorentz frame at a specific event $\mathcal{P}_0$ [edge vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ with $\boldsymbol{v} \cdot \boldsymbol{w} = 0$ (edges perpendicular) and $\boldsymbol{v} \cdot \boldsymbol{e}_0 = \boldsymbol{w} \cdot \boldsymbol{e}_0 = 0$ (edges in surface of constant time) and $\boldsymbol{v} \cdot \boldsymbol{k} = \boldsymbol{w} \cdot \boldsymbol{k} = 0$ (since edge vectors connect adjacent rays of the bundle)]. Show pictorially that in any other Lorentz frame at $\mathcal{P}_0$, the edge vectors are $\boldsymbol{v'} = \boldsymbol{v} + \alpha \boldsymbol{k}$ and $\boldsymbol{w'} = \boldsymbol{w} + \beta \boldsymbol{k}$ for some α and β. Conclude that in all Lorentz frames at $\mathcal{P}_0$ the cross section has identical shape and identical area, and is spatially perpendicular to the direction of propagation ($\boldsymbol{k} \cdot \boldsymbol{v} = \boldsymbol{k} \cdot \boldsymbol{w} = 0$). (iii) By a calculation in a local Lorentz frame show that $\partial_{\boldsymbol{k}} \mathscr{A} = (\boldsymbol{\nabla} \cdot \boldsymbol{k}) \mathscr{A}$. (iv) Conclude from $\partial_{\boldsymbol{k}} a = -\frac{1}{2}(\boldsymbol{\nabla} \cdot \boldsymbol{k}) a$ that $\partial_{\boldsymbol{k}} (\mathscr{A} a^2) = 0$.]

Exercise 22.14. FOCUSING THEOREM

The cross-sectional area $\mathscr{A}$ of a bundle of rays all lying in the same surface of constant phase changes along the central ray of the bundle at the rate (22.36) (see Figure 22.1).

(a) Derive the following equation ("*focusing equation*") for the second derivative of $\mathscr{A}^{1/2}$:

$$\frac{d^2 \mathscr{A}^{1/2}}{d\lambda^2} = -\left(|\sigma|^2 + \frac{1}{2} R_{\alpha\beta} k^\alpha k^\beta \right) \mathscr{A}^{1/2}, \tag{22.37}$$

where λ is affine parameter along the central ray ($\boldsymbol{k} = d/d\lambda$), and the "magnitude of the shear of the rays", $|\sigma|$, is defined by the equation

$$|\sigma|^2 \equiv \frac{1}{2} k_{\alpha;\beta} k^{\alpha;\beta} - \frac{1}{4} (k^\mu{}_{;\mu})^2. \tag{22.38}$$

[Hint: This is a vigorous exercise in index manipulations. The key equations needed in the manipulations are $\mathscr{A}_{,\alpha} k^\alpha = (k^\alpha{}_{;\alpha}) \mathscr{A}$ [equation (22.36)]; $k^\alpha{}_{;\beta} k^\beta = 0$ [geodesic equation (22.32) for rays]; $k_{\alpha;\beta} = k_{\beta;\alpha}$ [which follows from $k_\alpha \equiv \theta_{,\alpha}$]; and the rule (16.6c) for interchanging covariant derivatives of a vector.]

(b) Show that, in a local Lorentz frame where $\boldsymbol{k} = \omega(\boldsymbol{e}_t + \boldsymbol{e}_z)$ at the origin,

$$|\sigma|^2 = \frac{1}{4} (k_{x,x} - k_{y,y})^2 + (k_{x,y})^2. \tag{22.39}$$

Thus, $|\sigma|^2$ is nonnegative, which justifies the use of the absolute value sign.

(c) *Discussion:* The quantity $|\sigma|$ is called the *shear* of the bundle of rays because it measures the extent to which neighboring rays are sliding past each other [see, e.g., Sachs (1964)]. Hence, the focusing equation (22.37) says that shear focuses a bundle of rays (makes $d^2 \mathscr{A}^{1/2}/d\lambda^2 < 0$); and spacetime curvature also focuses it if $R_{\alpha\beta} k^\alpha k^\beta > 0$, but defocuses it if $R_{\alpha\beta} k^\alpha k^\beta < 0$. (When a bundle of toothpicks, originally circular in cross section, is squeezed into an elliptic cross section, it is sheared.)

(d) Assume that the energy density $T_{\hat{0}\hat{0}}$, as measured by any observer anywhere in spacetime, is nonnegative. By combining the focusing equation (22.37) with the Einstein field equation, conclude that

$$\frac{d^2\mathscr{A}^{1/2}}{d\lambda^2} \leq 0 \begin{pmatrix}\text{for any bundle of rays, all in the same}\\ \text{surface of constant phase, anywhere in}\\ \text{spacetime}\end{pmatrix} \tag{22.40}$$

(*focusing theorem*). This theorem plays a crucial role in black-hole physics (§34.5) and in the theory of singularities (§34.6).

§22.6. KINETIC THEORY IN CURVED SPACETIME*

The stars in a galaxy wander through spacetime, each on its own geodesic world line, each helping to produce the spacetime curvature felt by all the others. Photons, left over from the hot phases of the big bang, bathe the Earth, bringing with themselves data on the homogeneity and isotropy of the universe. Theoretical analyses of these and many other problems are unmanageable, if they attempt to keep track of the motion of every single star or photon. But a statistical description gives accurate results and is powerful. Moreover, for most problems in astrophysics and cosmology, the simplest of statistical descriptions—one ignoring collisions—is adequate. Usually collisions are unimportant for the large-scale behavior of a system (e.g., a galaxy), or they are so important that a fluid description is possible (e.g., in a stellar interior).

Consider, then, a swarm of particles (stars, or photons, or black holes, or . . .) that move through spacetime on geodesic world lines, without colliding. Assume, for simplicity, that the particles all have the same rest mass. Then all information of a statistical nature about the particles can be incorporated into a single function, the "*distribution function*" or "*number density in phase space*", $\mathscr{N}$.

Volume in phase space for a group of identical particles

Define $\mathscr{N}$ in terms of measurements made by a specific local Lorentz observer at a specific event $\mathscr{P}_0$ in curved spacetime. Give the observer a box with 3-volume $\mathscr{V}_x$ (and with imaginary walls). Ask the observer to count how many particles, N, are inside the box *and* have local-Lorentz momentum components p^j in the range

$$P^j - \frac{1}{2}\Delta p^j < p^j < P^j + \frac{1}{2}\Delta p^j.$$

(He can ignore the particle energies p^0; since all particles have the same rest mass m, energy

$$p^0 = (m^2 + \mathbf{p}^2)^{1/2}$$

*For more detailed and sophisticated treatments of this topic, see, e.g., Tauber and Weinberg (1961), and Lindquist (1966), Marle (1969), Ehlers (1971), Stewart (1971), Israel (1972), and references cited therein. Ehlers (1971) is a particularly good introductory review article.

is fixed uniquely by momentum.) The volume in momentum space occupied by the N particles is $\mathcal{V}_p = \Delta p^x \, \Delta p^y \, \Delta p^z$; and *the volume in phase space is*

$$\mathcal{V} \equiv \mathcal{V}_x \mathcal{V}_p. \tag{22.41}$$

Lorentz invariance of volume in phase space

Other observers at $\mathcal{P}_0$, moving relative to the first, will disagree on how much spatial volume $\mathcal{V}_x$ and how much momentum volume $\mathcal{V}_p$ these same N particles occupy:

$$\mathcal{V}_x \text{ and } \mathcal{V}_p \text{ depend on the choice of Lorentz frame.} \tag{22.42}$$

However, all observers will agree on the value of the product $\mathcal{V} \equiv \mathcal{V}_x \mathcal{V}_p$ ("volume in phase space"):

> The phase-space volume $\mathcal{V}$ occupied by a given set of N identical particles at a given event in spacetime is independent of the local Lorentz frame in which it is measured. (22.43)

(See Box 22.5 for proof.) Moreover, as the same N particles move through spacetime along their geodesic world lines (and through momentum space), the volume $\mathcal{V}$ they span in phase space remains constant:

Liouville's theorem (conservation of volume in phase space)

> The $\mathcal{V}$ occupied by a given swarm of N particles is independent of location along the world line of the swarm (*"Liouville's theorem in curved spacetime"*). (22.44)

(See Box 22.6 for proof.)

Number density in phase space (distribution function)

More convenient for applications than the volume $\mathcal{V}$ in phase space occupied by a given set of N particles is the *"number density in phase space"* (*"distribution function"*) in the neighborhood of one of these particles:

$$\mathcal{N} \equiv N/\mathcal{V}. \tag{22.45}$$

On what does this number density depend? It depends on the location in spacetime, $\mathcal{P}$, at which the measurements are made. It also depends on the 4-momentum $\boldsymbol{p}$ of the particle in whose neighborhood the measurements are made. But because the particles all have the same rest mass, $\boldsymbol{p}$ cannot take on any and every value in the tangent space at $\mathcal{P}$. Rather, $\boldsymbol{p}$ is confined to the "forward mass hyperboloid" at $\mathcal{P}$:

$$\boldsymbol{p}^2 = m^2; \qquad \boldsymbol{p} \text{ lies inside future light cone.}$$

Thus,

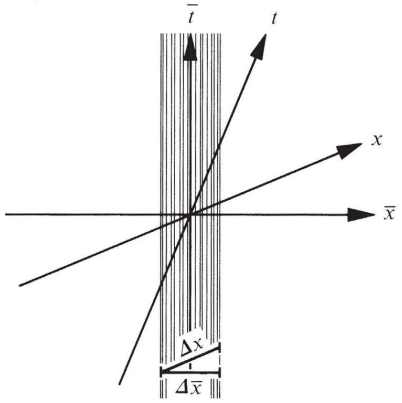
$$\mathcal{N} = \mathcal{N}\left[\begin{pmatrix}\text{location, } \mathcal{P},\\ \text{in spacetime}\end{pmatrix}, \begin{pmatrix}\text{4-momentum } \boldsymbol{p}\text{, which must lie}\\ \text{on the forward mass hyperboloid}\\ \text{of the tangent space at } \mathcal{P}\end{pmatrix}\right]. \tag{22.46}$$

Pick some one particle in the swarm, with geodesic world line $\mathcal{P}(\lambda)$ [λ = (affine parameter) = (proper time, if particle has finite rest mass)], and with 4-momentum

Box 22.5 VOLUME IN PHASE SPACE

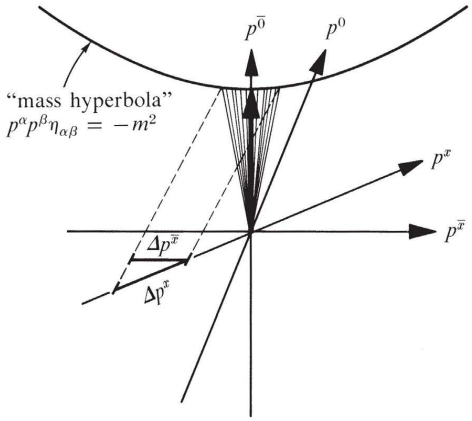
A. For Swarm of Identical Particles with Nonzero Rest Mass

Pick an event $\mathcal{P}_0$, through which passes a particle named "John" with a 4-momentum named "***P***". In John's local Lorentz rest frame at $\mathcal{P}_0$ ("barred frame", $\bar{S}$), select a small 3-volume, $\mathcal{V}_x \equiv \Delta\bar{x}\,\Delta\bar{y}\,\Delta\bar{z}$, containing him. Also select a small "3-volume in momentum space," $\mathcal{V}_{\bar{p}} \equiv \Delta p^{\bar{x}}\,\Delta p^{\bar{y}}\,\Delta p^{\bar{z}}$ centered on John's momentum, which is $P^{\bar{x}} = P^{\bar{y}} = P^{\bar{z}} = 0$. Focus attention on all particles whose world lines pass through $\mathcal{V}_{\bar{x}}$ and which have momenta $p^{\bar{j}}$ in the range $\mathcal{V}_{\bar{p}}$ surrounding $P^{\bar{j}} = 0$.



Examine this bundle in another local Lorentz frame ("unbarred frame", S) at $\mathcal{P}_0$, which moves with speed β relative to the rest frame. Orient axes so the relative motion of the frames is in the x and $\bar{x}$ directions. Then the space volume $\mathcal{V}_x$ occupied in the new frame has $\Delta y = \Delta\bar{y}$, $\Delta z = \Delta\bar{z}$ (no effect of motion on transverse directions), and $\Delta x = (1 - \beta^2)^{1/2}\,\Delta\bar{x}$ (Lorentz contraction in longitudinal direction). Hence $\mathcal{V}_x = (1 - \beta^2)^{1/2}\mathcal{V}_{\bar{x}}$ ("transformation law for space volumes") or, equivalently [since $P^0 = m/(1 - \beta^2)^{1/2}$]:

$$P^0\mathcal{V}_x = m\mathcal{V}_{\bar{x}} = \begin{pmatrix}\text{constant, independent}\\ \text{of Lorentz frame}\end{pmatrix}.$$



A momentum-space diagram, analogous to the spacetime diagram, depicts the momentum spread for particles in the bundle, and shows that $\Delta p^x = \Delta p^{\bar{x}}/(1 - \beta^2)^{1/2}$. The Lorentz transformation from $\bar{S}$ to S leaves transverse components of momenta unaffected; so $\Delta p^y = \Delta p^{\bar{y}}$, $\Delta p^z = \Delta p^{\bar{z}}$. Hence $\mathcal{V}_p = \mathcal{V}_{\bar{p}}/(1 - \beta^2)^{1/2}$ ("transformation law for momentum volumes"); or, equivalently

$$\frac{\mathcal{V}_p}{P^0} = \frac{\mathcal{V}_{\bar{p}}}{m} = \begin{pmatrix}\text{constant, independent}\\ \text{of Lorentz frame}\end{pmatrix}.$$

Although the spatial 3-volumes $\mathcal{V}_x$ and $\mathcal{V}_{\bar{x}}$ differ from one frame to another, and the momentum 3-volumes $\mathcal{V}_p$ and $\mathcal{V}_{\bar{p}}$ differ, the volume in six-dimensional phase space is Lorentz-invariant:

$$\mathcal{V} \equiv \mathcal{V}_{\bar{x}}\mathcal{V}_{\bar{p}} = \mathcal{V}_x\mathcal{V}_p.$$

It is a frame-independent, geometric object!

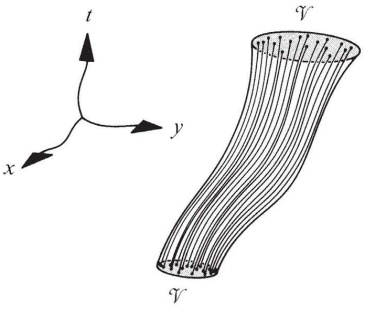
B. For Swarm of Identical Particles with Zero Rest Mass

Examine a sequence of systems, each with particles of smaller rest mass and of higher velocity relative to a laboratory. For every bundle of particles in each system, $P^0\mathcal{V}_x$, $\mathcal{V}_p/P^0$, and $\mathcal{V}_x\mathcal{V}_p$ are Lorentz-invariant. Hence, in the limit as $m \longrightarrow 0$, as $\beta \longrightarrow 1$, and as $P^0 = m/(1 - \beta^2)^{1/2} \longrightarrow$ finite value (particles of zero rest mass moving with speed of light), $P^0\mathcal{V}_x$ and $\mathcal{V}_p/P^0$ and $\mathcal{V}_x\mathcal{V}_p$ are still Lorentz-invariant, geometric quantities.

Box 22.6 CONSERVATION OF VOLUME IN PHASE SPACE

Examine a very small bundle of identical particles that move through curved spacetime on neighboring geodesics. Measure the bundle's volume in phase space, $\mathcal{V}$ ($\mathcal{V} = \mathcal{V}_x\mathcal{V}_p$ in any local Lorentz frame), as a function of affine parameter λ along the central geodesic of the bundle. The following calculation shows that

$$d\mathcal{V}/d\lambda = 0 \qquad \left(\begin{array}{l}\text{"Liouville theorem in}\\ \text{curved spacetime"}\end{array}\right).$$



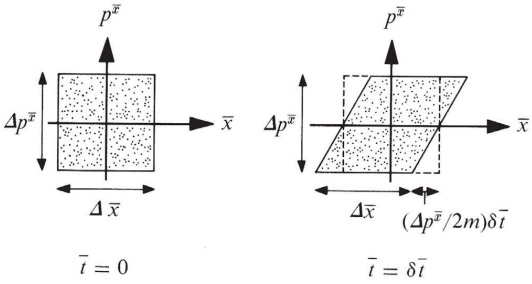
Proof for particles of finite rest mass: Examine particle motion during time interval $\delta\tau$, using local Lorentz rest frame of central particle. All velocities are small in this frame, so

$$p^{\bar{j}} = m dx^{\bar{j}}/d\bar{t}.$$

Hence (see pictures) the spreads in momentum and position conserve $\Delta\bar{x}\,\Delta p^{\bar{x}}$, $\Delta\bar{y}\,\Delta p^{\bar{y}}$, and $\Delta\bar{z}\,\Delta p^{\bar{z}}$; i.e.,

$$\frac{d\mathcal{V}}{d\tau} = \frac{\delta(\Delta\bar{x}\,\Delta\bar{y}\,\Delta\bar{z}\,\Delta p^{\bar{x}}\,\Delta p^{\bar{y}}\,\Delta p^{\bar{z}})}{\delta\bar{t}} = 0.$$

But $\tau = a\lambda + b$ for some arbitrary constants a and b; so $d\mathcal{V}/d\lambda = 0$.



Each particle moves with speed $d\bar{x}/d\bar{t}$ proportional to height in diagram

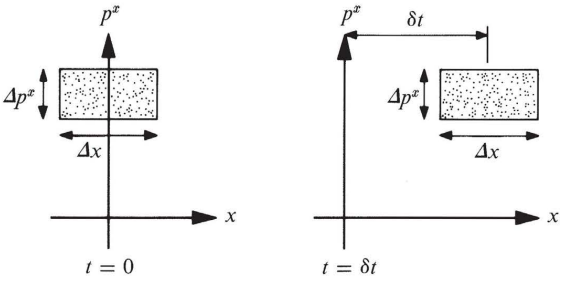
$$d\bar{x}/d\bar{t} = p^{\bar{x}}/m,$$

and conserves its momentum, $dp^{\bar{x}}/d\bar{t} = 0$. Hence the region occupied by particles deforms, but maintains its area. Same is true for $(y - p^{\bar{y}})$ and $(z - p^{\bar{z}})$.

Proof for particles of zero rest mass. Examine particle motion in local Lorentz frame where central particle has $\boldsymbol{P} = P^0(\boldsymbol{e}_0 + \boldsymbol{e}_x)$. In this frame, all particles have $p^y \ll p^0$, $p^z \ll p^0$, $p^x = p^0 + O([p^y]^2/P^0) \approx P^0$. Since $p^\alpha = dx^\alpha/d\lambda$ for appropriate normalization of affine parameters (see Box 22.4), one can write $dx^j/dt = p^j/p^0$; i.e.,

$$\frac{dx}{dt} = 1 + O([p^y/P^0]^2 + [p^z/P^0]^2) \approx 1,$$

$$\frac{dy}{dt} = \frac{p^y}{P^0}, \qquad \frac{dz}{dt} = \frac{p^z}{P^0}.$$



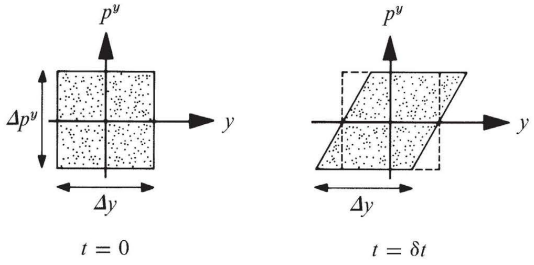
Each particle ("photon") moves with $dx/dt = 1$ and $dp^x/dt = 0$ in the local Lorentz frame. Area and shape of occupied region are preserved.

Hence (see pictures) $\Delta x\,\Delta p^x$, $\Delta y\,\Delta p^y$, and $\Delta z\,\Delta p^z$ are all conserved; and

$$\frac{d\mathcal{V}}{dt} = \frac{\delta(\Delta x\,\Delta y\,\Delta z\,\Delta p^x\,\Delta p^y\,\Delta p^z)}{\delta t} = 0.$$

But t and the affine parameter λ of central particle are related by $t = P^0\lambda$ [cf. equation (16.4)]; thus

$$d\mathcal{V}/d\lambda = 0.$$



Particle ("photon") speeds are proportional to height in diagram

$$dy/dt = p^y/P^0,$$

and $dp^y/dt = 0$. Hence, occupied region deforms but maintains its area. Same is true of $z - p^z$.

$\boldsymbol{p}(\lambda)$. Examine the density in phase space in this particle's neighborhood at each point along its world line:

$$\mathcal{N} = \mathcal{N}[\mathcal{P}(\lambda), \boldsymbol{p}(\lambda)].$$

Calculate $\mathcal{N}(\lambda)$ as follows: (1) Pick an initial event $\mathcal{P}(0)$ on the world line, and a phase-space volume $\mathcal{V}$ containing the particle. (2) Cover with red paint all the particles contained in $\mathcal{V}$ at $\mathcal{P}(0)$. (3) Watch the red particles move through spacetime alongside the initial particle. (4) As they move, the phase-space region they occupy changes shape extensively; but its volume $\mathcal{V}$ remains fixed (Liouville's theorem). Moreover, no particles can enter or leave that phase-space region (once in, always in; once out, always out; boundaries of phase-space region are attached to and move with the particles). (5) Hence, at any λ along the initial particle's world line, the particle is in a phase-space region of unchanged volume $\mathcal{V}$, unchanged number of particles N, and unchanged ratio $\mathcal{N} = N/\mathcal{V}$:

$$\frac{d\mathcal{N}[\mathcal{P}(\lambda), \boldsymbol{p}(\lambda)]}{d\lambda} = 0. \tag{22.47}$$

Collisionless Boltzmann equation (kinetic equation)

This equation for the conservation of $\mathcal{N}$ along a particle's trajectory in phase space is called the "collisionless Boltzmann equation," or the "kinetic equation."

Photons provide an important application of the Boltzmann equation. But when discussing photons one usually does not think in terms of the number density in phase space. Rather, one speaks of the "specific intensity" I_ν of radiation at a given frequency ν, flowing in a given direction, $\boldsymbol{n}$, as measured in a specified local Lorentz frame:

$$I_\nu \equiv \frac{d(\text{energy})}{d(\text{time})\,d(\text{area})\,d(\text{frequency})\,d(\text{solid angle})}. \tag{22.48}$$

Distribution function for photons expressed in terms of specific intensity, I_ν

(See Figure 22.2). A simple calculation in the local Lorentz frame reveals that

$$\mathcal{N} = h^{-4}(I_\nu/\nu^3), \tag{22.49}$$

Invariance and conservation of I_ν/ν^3

where h is Planck's constant (see Figure 22.2). Thus, if two different observers at the same or different events in spacetime look at the same photon (and neighboring photons) as it passes them, they will see different frequencies ν ("doppler shift," "cosmological red shift," "gravitational redshift"), and different specific intensities I_ν; but they will obtain identical values for the ratio I_ν/ν^3. Thus I_ν/ν^3, like $\mathcal{N}$, is invariant from observer to observer and from event to event along a given photon's world line.

EXERCISES

Exercise 22.15. INVERSE SQUARE LAW FOR FLUX

The *specific flux* of radiation entering a telescope from a given source is defined by

$$F_\nu = \int I_\nu \, d\Omega, \tag{22.50}$$

where integration is over the total solid angle (assumed $\ll 4\pi$) subtended by the source on the observer's sky. Use the Boltzmann equation (conservation of I_ν/ν^3) to show that $F_\nu \propto$ (distance from source)$^{-2}$ for observers who are all at rest relative to each other in flat spacetime.

Exercise 22.16. BRIGHTNESS OF THE SUN

Does the surface of the sun look any brighter to an astronaut standing on Mercury than to a student standing on Earth?

Exercise 22.17. BLACK BODY RADIATION

An "optically thick" source of black-body radiation (e.g., the surface of a star, or the hot matter filling the universe shortly after the big bang) emits photons isotropically with a specific intensity, as seen by an observer at rest near the source, given (Planck radiation law) by

$$I_\nu = \frac{2h\nu^3}{e^{h\nu/kT} - 1}. \tag{22.51}$$

Here T is the temperature of the source. Show that any observer, in any local Lorentz frame, anywhere in the universe, who examines this radiation as it flows past him, will also see a black-body spectrum. Show, further, that if he calculates a temperature by measuring the specific intensity I_ν at any one frequency, and if he calculates a temperature from the shape of the spectrum, those temperatures will agree. (Radiation remains black body rather than being "diluted" into "grey-body.") Finally, show that the temperature he measures is redshifted by precisely the same factor as the frequency of any given photon is redshifted,

$$\frac{T_{\text{observed}}}{T_{\text{emitted}}} = \left(\frac{\nu_{\text{observed}}}{\nu_{\text{emitted}}}\right) \text{ for a given photon.} \tag{22.52}$$

[Note that the redshifts can be "Doppler" in origin, "cosmological" in origin, "gravitational" in origin, or some inseparable mixture. All that matters is the fact that the parallel-transport law for a photon's 4-momentum, $\nabla_{\boldsymbol{p}}\boldsymbol{p} = 0$, guarantees that the redshift $\nu_{\text{observed}}/\nu_{\text{emitted}}$ is independent of frequency emitted.]

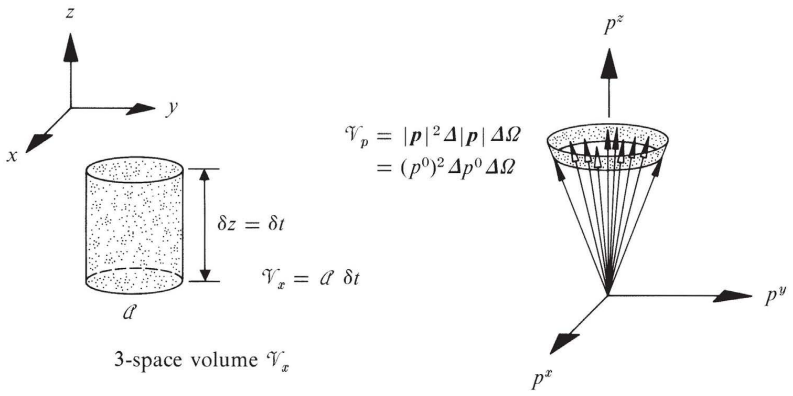


Figure 22.2.

Number density in phase space for photons, interpreted in terms of the specific intensity I_ν. An astronomer has a telescope with filter that admits only photons arriving from within a small solid angle $\Delta\Omega$ about the z-direction, and having energies between p^0 and $p^0 + \Delta p^0$. The collecting area, $\mathcal{A}$, of his telescope lies in the x, y-plane (perpendicular to the incoming photon beam). Let δN be the number of photons that cross the area $\mathcal{A}$ in a time interval δt. [All energies, areas, times, and lengths are measured in the orthonormal frame ("proper reference frame; §13.6) which the astronomer Fermi-Walker transports with himself along his (possibly accelerated) world line—or, equivalently, in a local Lorentz frame momentarily at rest with respect to the astronomer.] The δN photons, just before the time interval δt begins, lie in the cylinder of area $\mathcal{A}$ and height $\delta z = \delta t$ shown above. Their spatial 3-volume is thus $\mathcal{V}_x = \mathcal{A}\ \delta t$. Their momentum 3-volume is $\mathcal{V}_p = (p^0)^2\, \Delta p^0\, \Delta\Omega$ (see drawing). Hence, their number density in phase space is

$$\mathcal{N} = \frac{\delta N}{\mathcal{V}_x \mathcal{V}_p} = \frac{\delta N}{\mathcal{A}\ \delta t (p^0)^2 (\Delta p^0)\ \Delta\Omega} = \frac{\delta N}{h^3\, \mathcal{A}\ \delta t\, \nu^2\, \Delta\nu\ \Delta\Omega}$$

where ν is the photon frequency measured by the telescope ($p^0 = h\nu$).

The specific intensity of the photons, I_ν (a standard concept in astronomy), is the energy per unit area per unit time per unit frequency per unit solid angle crossing a surface perpendicular to the beam: i.e.,

$$I_\nu = \frac{h\nu\, \delta N}{\mathcal{A}\ \delta t\ \Delta\nu\ \Delta\Omega}.$$

Direct comparison reveals $\mathcal{N} = h^{-4}(I_\nu/\nu^3)$.

Thus, conservation of $\mathcal{N}$ along a photon's world line implies conservation of I_ν/ν^3. This conservation law finds important applications in cosmology (e.g., Box 29.2 and Ex. 29.5) and in the gravitational lens effect (Refsdal 1964); see also exercises 22.15–22.17.

Exercise 22.18. STRESS-ENERGY TENSOR

(a) Show that the stress-energy tensor for a swarm of identical particles at an event $\mathcal{P}_0$ can be written as an integral over the mass hyperboloid of the momentum space at $\mathcal{P}_0$:

$$\boldsymbol{T} = \int (\mathcal{N}\boldsymbol{p} \otimes \boldsymbol{p})(d\mathcal{V}_p/p^0), \tag{22.53}$$

$$\frac{d\mathcal{V}_p}{p^0} \equiv \frac{dp^x\, dp^y\, dp^z}{p^0} \text{ in a local Lorentz frame.} \tag{22.54}$$

(Notice from Box 22.5 that $d\mathcal{V}_p/p^0$ is a Lorentz-invariant volume element for any segment of the mass hyperboloid.)

(b) Verify that the Boltzmann equation, $d\mathcal{N}/d\lambda = 0$, implies $\nabla \cdot \boldsymbol{T} = 0$ for any swarm of identical particles. [*Hint:* Calculate $\nabla \cdot \boldsymbol{T}$ in a local Lorentz frame, using the above expression for $\boldsymbol{T}$, and using the geodesic equation in the form $Dp^\mu/d\lambda = 0$.]

Exercise 22.19. KINETIC THEORY FOR NONIDENTICAL PARTICLES

For a swarm of particles with a wide distribution of rest masses, define

$$\mathcal{N} = \frac{\Delta N}{\mathcal{V}_x \mathcal{V}_p \, \Delta m}, \tag{22.55}$$

where $\mathcal{V}_x$ and $\mathcal{V}_p$ are spatial and momentum 3-volumes, and ΔN is the number of particles in the region $\mathcal{V}_x \mathcal{V}_p$ with rest masses between $m - \Delta m/2$ and $m + \Delta m/2$. Show the following.

(a) $\mathcal{V}_x \mathcal{V}_p \, \Delta m$ is independent of Lorentz frame and independent of location on the world tube of a bundle of particles.

(b) $\mathcal{N}$ can be regarded as a function of location $\mathcal{P}$ in spacetime and 4-momentum $\boldsymbol{p}$ inside the future light cone of the tangent space at $\mathcal{P}$:

$$\mathcal{N} = \mathcal{N}(\mathcal{P}, \boldsymbol{p}). \tag{22.56}$$

(c) $\mathcal{N}$ satisfies the collisionless Boltzmann equation (kinetic equation)

$$\frac{d\mathcal{N}[\mathcal{P}(\lambda), \boldsymbol{p}(\lambda)]}{d\lambda} = 0 \qquad \text{along geodesic trajectory of any particle.} \tag{22.57}$$

(d) $\mathcal{N}$ can be rewritten in a local Lorentz frame as

$$\mathcal{N} = \frac{\Delta N}{[(p^0/m)\, \Delta x \, \Delta y \, \Delta z][\Delta p^0 \, \Delta p^x \, \Delta p^y \, \Delta p^z]}. \tag{22.58}$$

(e) The stress-energy tensor at an event $\mathcal{P}$ can be written as an integral over the interior of the future light cone of momentum space

$$T^{\mu\nu} = \int (\mathcal{N} p^\mu p^\nu) m^{-1} \, dp^0 \, dp^1 \, dp^2 \, dp^3 \tag{22.59}$$

in a local Lorentz frame (Track-1 notation for integral; see Box 5.3);

$$\boldsymbol{T} = \int (\mathcal{N} \boldsymbol{p} \otimes \boldsymbol{p}) m^{-1} * 1 \qquad \text{in frame-independent notation}$$

$$= \int (\mathcal{N} \boldsymbol{p} \otimes \boldsymbol{p}) m^{-1} \, \boldsymbol{d}p^0 \wedge \boldsymbol{d}p^1 \wedge \boldsymbol{d}p^2 \wedge \boldsymbol{d}p^3 \tag{22.59$'$}$$

in a local Lorentz frame (Track-2 notation; see Box 5.4).

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

Preview of the rest of this book

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.

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).

Static stars must be spherical

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.

§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)$/2\pi$.

(3) derived more rigorously

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.

Exercise 23.1. ISOTROPIC COORDINATES AND NEWTONIAN LIMIT

EXERCISE

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.

The form of any metric can reveal the nature of the coordinates being used

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

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 $\mathcal{P}$ for which the value $r(\mathcal{P})$ is desired; and one can then calculate

(2) r measures surface area of sphere

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

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.

(3) t has 3 special geometric properties

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

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

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.]

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

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;

*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 ($r, \theta, \phi = \text{const.}$) 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

$$\rho = \rho(r) = \text{density of mass-energy in rest-frame of fluid;}$$

$$p = p(r) = \text{isotropic pressure in rest-frame of fluid;}$$

$$n = n(r) = \text{number density of baryons in rest-frame of fluid;} \quad (23.11)$$

$$u^\mu = u^\mu(r) = \text{4-velocity of fluid;}$$

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

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

(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) $\boldsymbol{u}$

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

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 $T^{\mu\nu}$:

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

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

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

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

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:

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

(1) in neutron stars

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

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

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

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

(2) in white dwarfs

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

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

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

(p_s, ρ_s: 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.

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:

Equation for Λ derived

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

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),

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

Equation of hydrostatic equilibrium rewritten in "OV" form

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

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

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 ρ, p, n, Φ, Λ. 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:

(energy gained from gravity) $\equiv -$(gravitational potential energy)

$$= \int_0^R (\rho r \Phi_{,r}) 4\pi r^2 \, dr = \int_0^R (\rho m/r) 4\pi r^2 \, dr$$

$$= -\frac{1}{2} \int_0^R (\rho \Phi) 4\pi r^2 \, dr = \frac{1}{8\pi} \int_0^\infty (\Phi_{,r})^2 4\pi r^2 \, dr$$

$$= 3 \int_0^R 4\pi r^2 p \, dr.$$

§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

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

Spacetime outside star possesses "Schwarzschild" geometry

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

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

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

Box 23.2 (*continued*)

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

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

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.

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

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

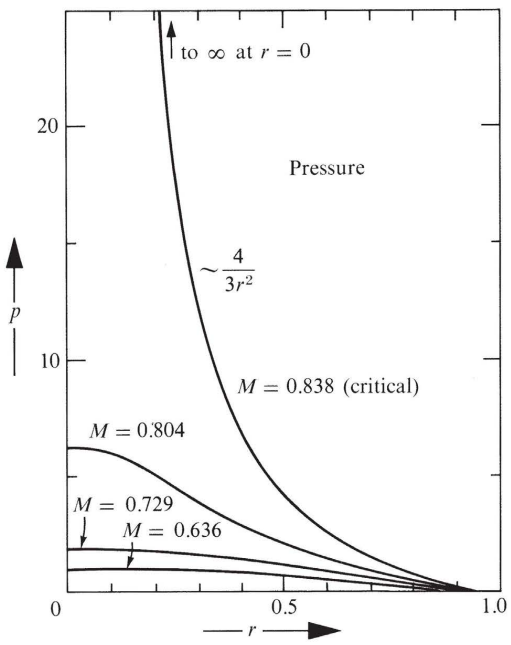
$$R_{\text{lim}} = (9/4)M_{\text{lim}} = (3\pi\rho_0)^{-1/2}, \tag{9}$$

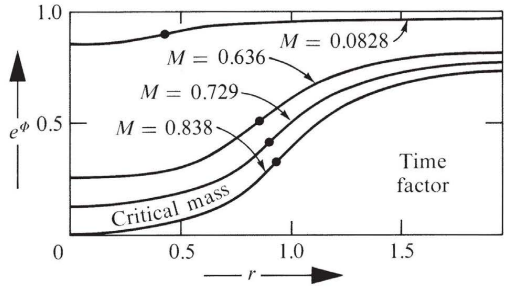
$$(2M/R)_{\text{lim}} = 8/9. \tag{10}$$

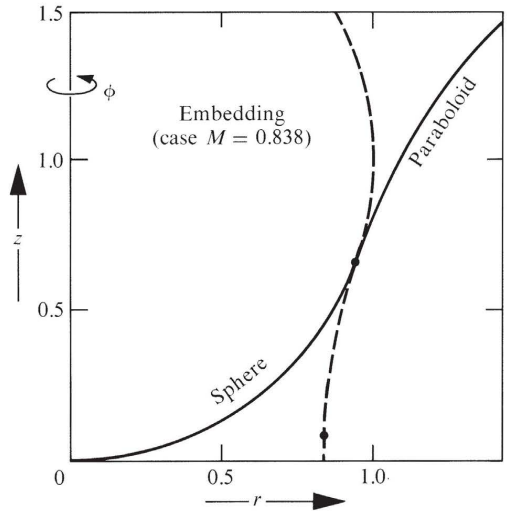
No star of uniform density can have a mass and radius exceeding these limits. These limits are purely relativistic phenomena; no such limits occur in Newtonian theory. (3) Inside the star the space geometry (geometry of a hypersurface $t = constant$) is that of a three-dimensional spherical surface with 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 $dr/d(\text{proper radial distance}) = dr/d\ell = 1$. 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

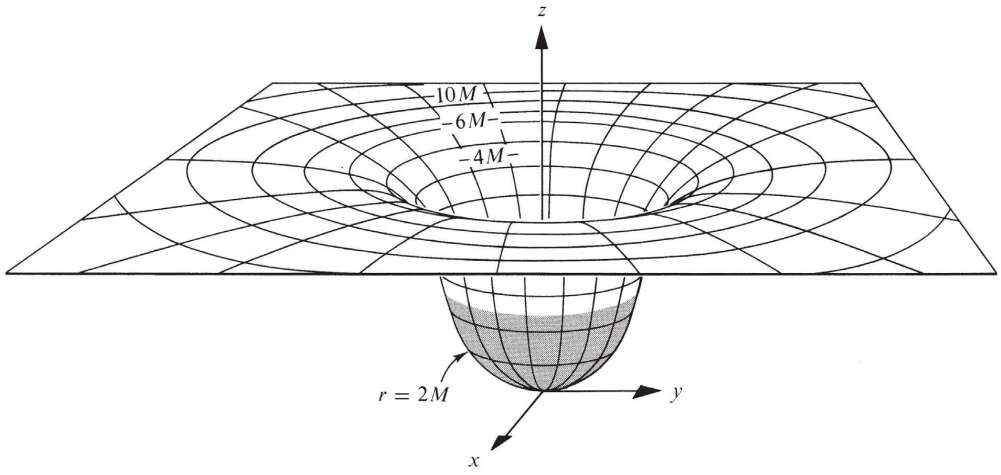


Figure 23.1.
Geometry within (grey) and around (white) a star of radius $R = 2.66M$, schematically displayed. The star is in hydrostatic equilibrium and has zero angular momentum (spherical symmetry). The two-dimensional geometry

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

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,

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

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

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 ($A: M^4 \longrightarrow M^4 : \mathcal{P} \longrightarrow A\mathcal{P}$), 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 $\mathcal{P}$, form the set $s = \mathcal{S}(\mathcal{P}) = \{A\mathcal{P} \in M^4 | A \in SO(3)\}$ of all points equivalent to $\mathcal{P}$ 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 $\mathcal{S}: M^4 \longrightarrow M^2 : \mathcal{P} \longrightarrow s = \mathcal{S}(\mathcal{P})$ allows one to obtain, from $R: M^2 \longrightarrow \mathcal{R} : s \longrightarrow R(s)$ [the "circumference" function on M^2 as defined by equation (1)], a corresponding function $R: M^4 \longrightarrow \mathcal{R} : \mathcal{P} \longrightarrow R(\mathcal{S}(\mathcal{P}))$ on M^4 which in some cases can eventually be used as a coordinate on M^4. (Note: $\mathcal{R}$ denotes here the real numbers.)

Now assume (iii)† there is a spherically symmetric 4-velocity field $\boldsymbol{u}$, defined so that if $\mathcal{P} = \mathcal{C}(\tau)$ is one trajectory of $\boldsymbol{u}$ with $\boldsymbol{u} = d/d\tau$, then each curve $\mathcal{P} = A\mathcal{C}(\tau)$ 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 $r = r(\mathcal{S}(\mathcal{P}))$ 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 $\boldsymbol{df} \cdot \boldsymbol{dg}$ 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, $L_{\nu}=3.78\times10^{41}$ ergs/sec. Calculations indicate that the total radiation, visible and invisible, is of the order $L_{\tau}=10^{7}L_{\nu}=3.78\times10^{48}$ ergs/sec. The supernova therefore emits during its life a total energy $E_{\tau}\geq10^{5}L_{\tau}=3.78\times10^{53}$ 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 $\sigma=2\times10^{-3}$ erg/cm² sec. The observational values are about $\sigma=3\times10^{-3}$ 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

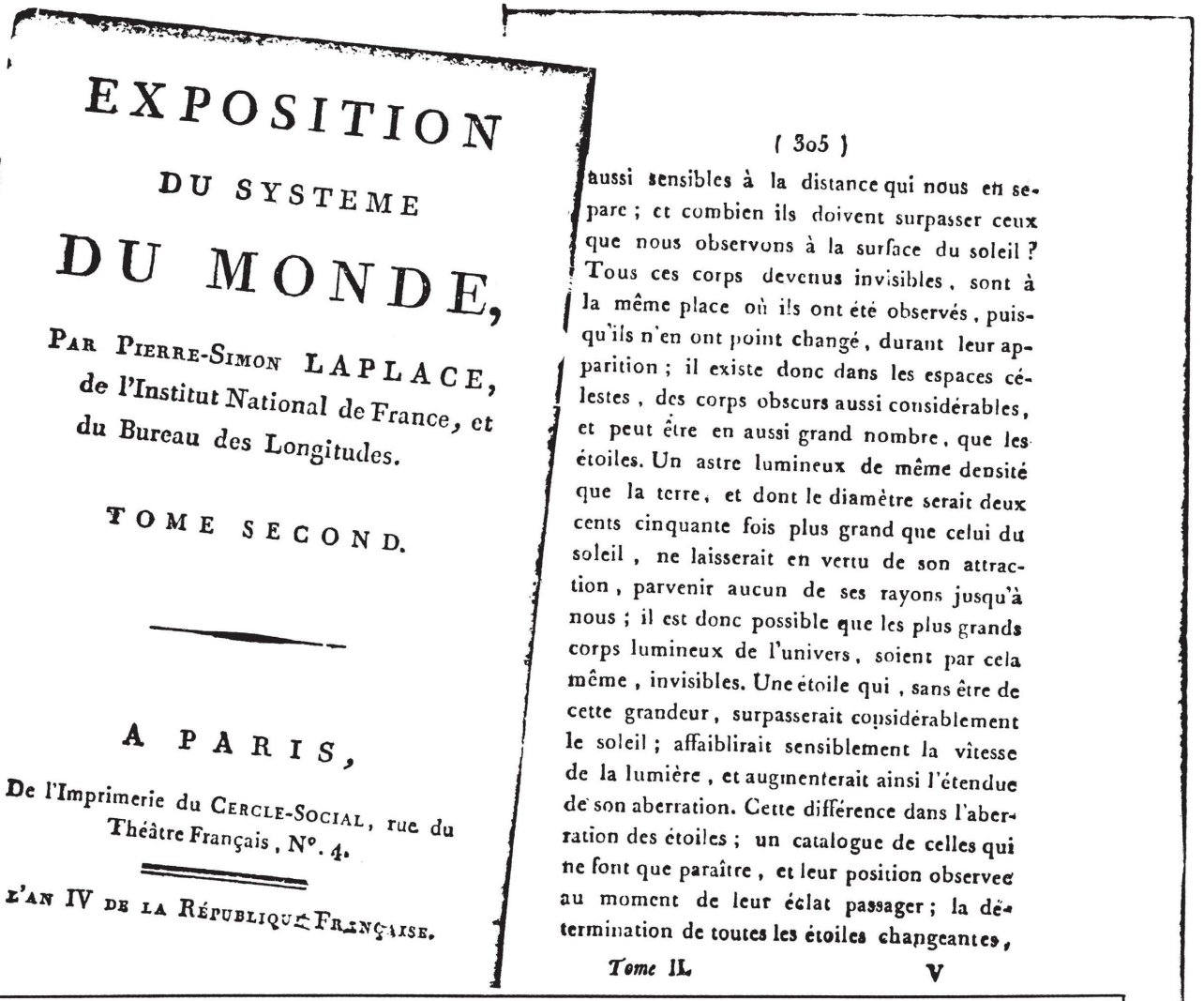
On Massive Neutron Cores

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.

(305)

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 vîtesse 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,

Tome II. V

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

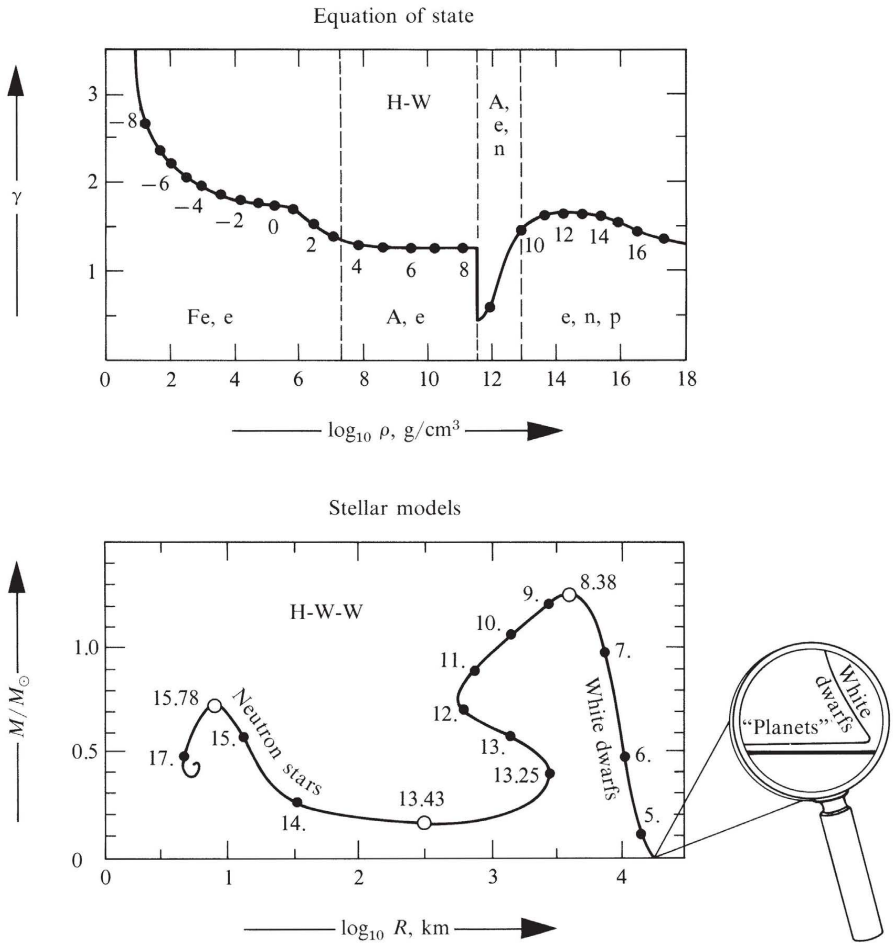


Figure 24.2.

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

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

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

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

the correct upper limits are probably in the range

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

[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

Neutron star as a pulsar

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.]

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

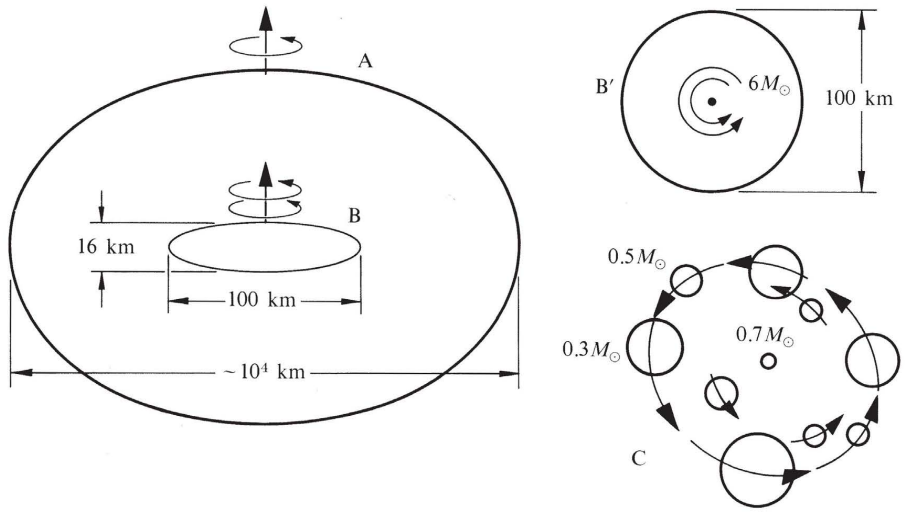


Figure 24.3.
"Collapse, pursuit, and plunge scenario" [schematic from Ruffini and Wheeler (1971b)].

- 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

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

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

$$\delta\bar{F}_{\text{buoy}} = \frac{\delta\bar{p}}{R} - \frac{\bar{p}}{R^2}\,\delta R = -(3\Gamma_1 + 1)\frac{\bar{p}}{R}\frac{\delta R}{R} = -(3\bar{\Gamma}_1 + 1)\bar{F}_{\text{buoy}}\left(\frac{\delta R}{R}\right),$$

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

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 $M > 60M_\odot$

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 $\varepsilon_{\rm HCNO} \propto T_c^{11}$.) 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," $\delta F_{\rm grav}$, 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" $\Gamma_{1\,\rm crit}$ of gravitational forces is increased above the Newtonian value of 4/3; thus,

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

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:

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

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

$$k = \left(\begin{array}{l}\text{effective}\\ \text{spring constant}\end{array}\right) = \left(\begin{array}{l}\text{contribution}\\ \text{of "elastic}\\ \text{forces"}\end{array}\right) - \left(\begin{array}{l}\text{contribution}\\ \text{of gravity}\end{array}\right)$$

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

(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

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

i.e.,

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

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 $\delta\bar{F}_{\text{grav}}$ and $\delta\bar{F}_{\text{buoy}}$ 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 unim-

portant 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).

Overview of this chapter

§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:

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

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*)

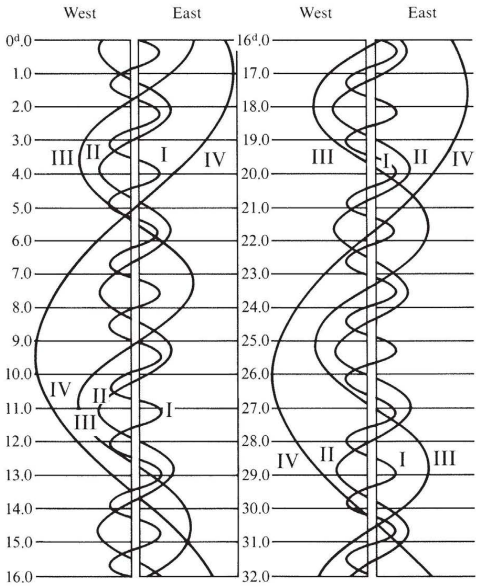


Figure 25.1.
Jupiter's satellites, as followed from night to night with field glasses or telescope, provide an opportunity to check for oneself the central ideas of gravitation physics in the Newtonian approximation (distances large compared to Schwarzschild radius). For the practically circular orbits 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*	
	1.04×10^{-5}	2.39×10^{-6}	1.99×10^{-7}	
	3.46×10^{-16}	7.95×10^{-17}	6.63×10^{-18}	

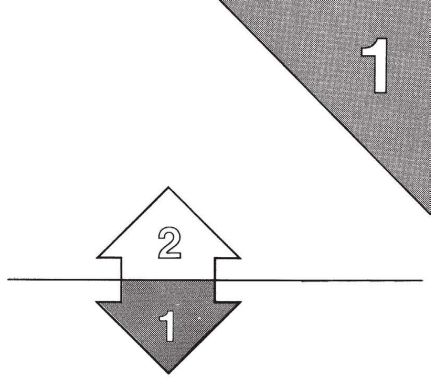


Figure 25.2.
Effective potential for motion of a test particle in the Schwarzschild geometry of a concentrated mass M. Energy, in units of the rest mass μ of the particle, is denoted $\tilde{E} = E/\mu$; angular momentum, $\tilde{L} = L/\mu$. The quantity r denotes the Schwarzschild r coordinate. The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

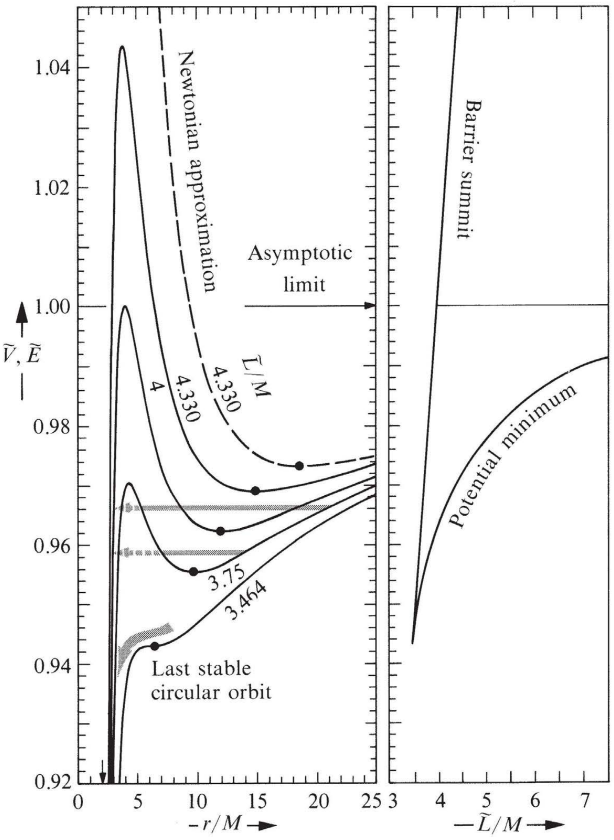
The diagram at the right gives values of the minimum and maximum of the potential as they depend on the angular momentum of the test particle. The roots of $\partial\tilde{V}/\partial r$ are given in terms of the "reduced angular momentum parameter" $L^\dagger = \tilde{L}/M = L/M\mu$ by

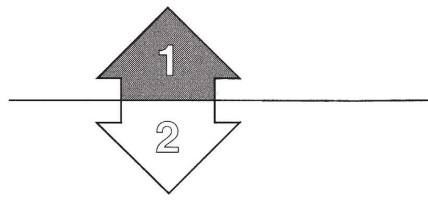
$$r = \frac{6M}{1 + (1 - 12/L^{\dagger 2})^{1/2}},$$

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

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

(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 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)

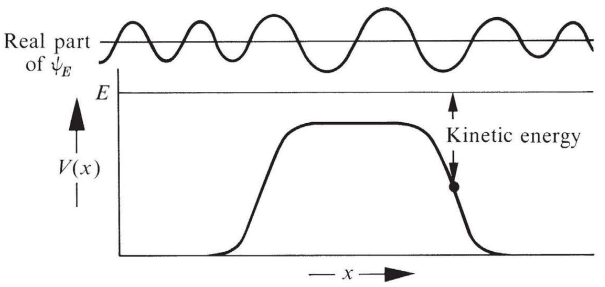
Box 25.3 THE HAMILTON-JACOBI DESCRIPTION OF MOTION: NATURAL BECAUSE RATIFIED BY THE QUANTUM PRINCIPLE

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 $\hbar = (1.6 \times 10^{-33}$ 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,

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

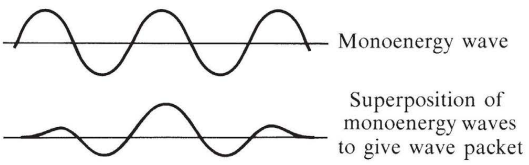
7. It is of no help in localizing the probability distribution that the solution of this equation for a particle of energy E is extraordinarily simple,

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

(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 mono-frequency (monoenergy) terms, according to a prescription qualitatively of the form

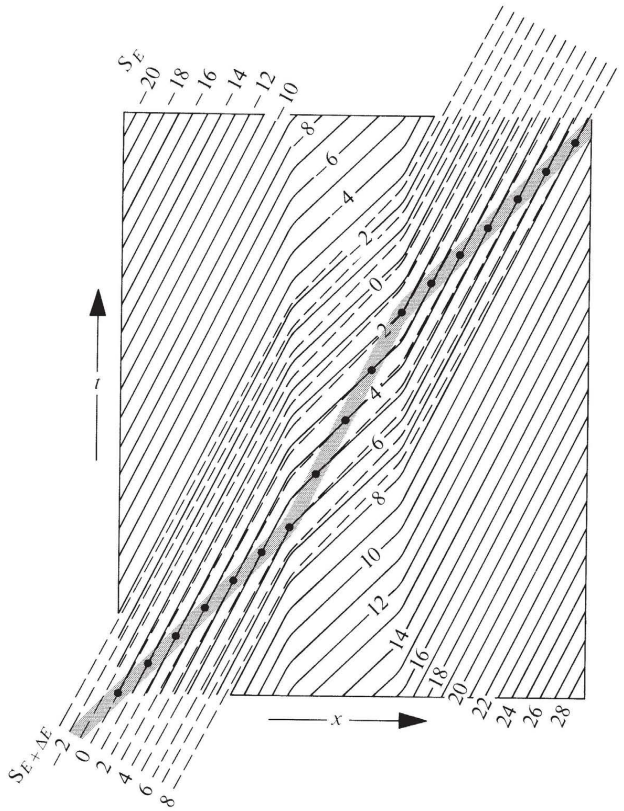
$$\psi(x, t) = \psi_E(x, t) + \psi_{E+\Delta E}(x, t) + \cdots \tag{4}$$



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

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

Equation of Hamilton-Jacobi for propagation of wave crests:

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

Box 25.4 (*continued*)

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

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

(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 $iS/\hbar = i\mu\widetilde{S}/\hbar$ is exponent in approximate expression for the probability amplitude.)

Constructive interference of waves:

(1) with slightly different $\widetilde{E}$ values (impose "condition of constructive interference" $\partial\widetilde{S}_{\widetilde{p}_\phi, \widetilde{L}, \widetilde{E}}(t, r, \theta, \phi)/\partial\widetilde{E} = 0$) 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," $\widetilde{L}$ (impose condition of constructive interference $\partial\widetilde{S}_{\widetilde{p}_\phi, \widetilde{L}, \widetilde{E}}(t, r, \theta, \phi)/\partial\widetilde{L} = 0$) 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," $\widetilde{p}_\phi$ (impose condition $\partial\widetilde{S}/\partial\widetilde{p}_\phi = 0$) gives correlation between θ and ϕ,

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

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 $\widetilde{L}$ stands perpendicular to this plane. The projection of this angular momentum along the z-axis is $\widetilde{p}_\phi = \widetilde{L}\cos\alpha$ (definition of orbital inclination, α). Straight line connecting origin with particle cuts unit sphere in a point $\mathcal{P}$. As time runs on, $\mathcal{P}$ 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 $\mathcal{P}$ 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:

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

whence

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

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

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

(greater than 1 for positive $\tilde{E}$, hyperbolic orbit; equal to 1 for zero $\tilde{E}$, parabolic orbit; less than 1 for negative $\tilde{E}$, 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)
$$a = \frac{\tilde{L}^2/M}{1 - e^2} = \frac{M}{(-2\tilde{E})}; \tag{10}$$

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

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

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

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

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

Box 25.4 (*continued*)

Time as correlated with position:

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

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

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

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:*

$$p_{K} = \boldsymbol{p}\cdot\boldsymbol{\xi} = \text{constant}. \tag{25.5}$$

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,

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

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

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

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 $\lambda = \tau/\mu$.

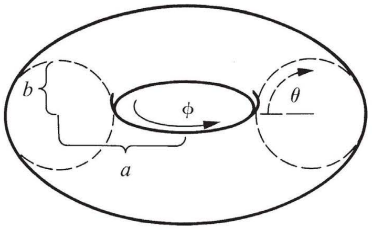
Terminology: "cyclic coordinate," "conjugate momentum"

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)

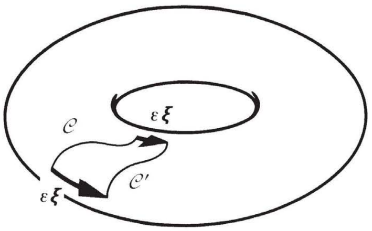
A. On a given manifold (e.g., spacetime, or the donut pictured here), in a given coordinate system, the metric components are independent of a particular coordinate x^K. Example of donut:



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

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

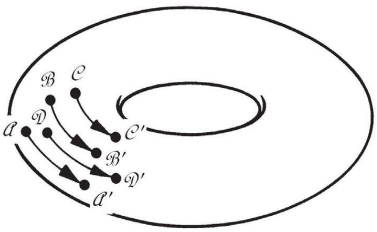
B. Translate an arbitrary curve $\mathcal{C}$ through the infinitesimal displacement



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

to form a new curve $\mathcal{C}'$. In coordinate language $\mathcal{C}$ is $\theta = \theta(\lambda)$, $\phi = \phi(\lambda)$; while $\mathcal{C}'$ 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 $\mathcal{C}$ and $\mathcal{C}'$ have the same length (see text).

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



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

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 $\boldsymbol{\xi}(\mathcal{P})$ generates an isometry if and only if it satisfies Killing's equation $\xi_{(\alpha;\beta)} = 0$.

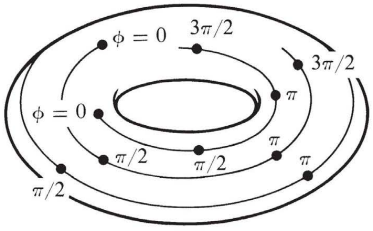
E. If $\boldsymbol{\xi}(\mathcal{P})$ generates an isometry (i.e. if $\boldsymbol{\xi}$ is a "Killing vector"), then the curves

$$\mathcal{P}(x^K, \underbrace{\alpha_1, \ldots, \alpha_n}_{\text{labels to tell "which" curve}})$$

[parameter on curve]

to which $\boldsymbol{\xi}$ is tangent [$\boldsymbol{\xi} = (\partial\mathcal{P}/\partial x^K)_{\alpha_1,\ldots,\alpha_n}$] 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 [$\mathcal{P}(x^K, \alpha_1, \ldots, \alpha_n) \longrightarrow \mathcal{P}(x^K + \Delta x^K, \alpha_1, \ldots, \alpha_n)$; "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 specializa-

tion 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"

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

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

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

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

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

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 $[\mathcal{K}, L_K]$ 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.

§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 25.12. REDSHIFT BY TIMED PULSES **EXERCISE**

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),

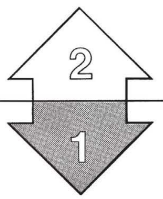
$$\left(\frac{dr}{d\tau}\right)^2 + \tilde{V}^2(r) = \tilde{E}^2, \tag{25.16a}$$

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

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

Qualitative features of orbits diagnosed from effective-potential diagram

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).



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

A. Equations Governing Orbit (see text for derivation)

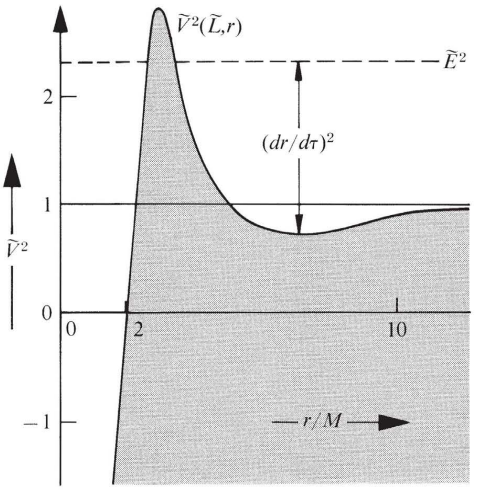
1. Effective-potential equation for radial part of motion:

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

$\tilde{E}$ = (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,"

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

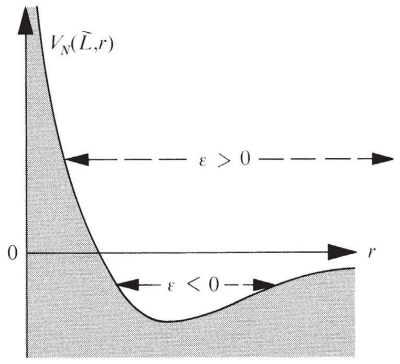
2. Speak not about $\tilde{V}^2(\tilde{L}, r)$ but instead about the Newtonian effective potential,

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

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

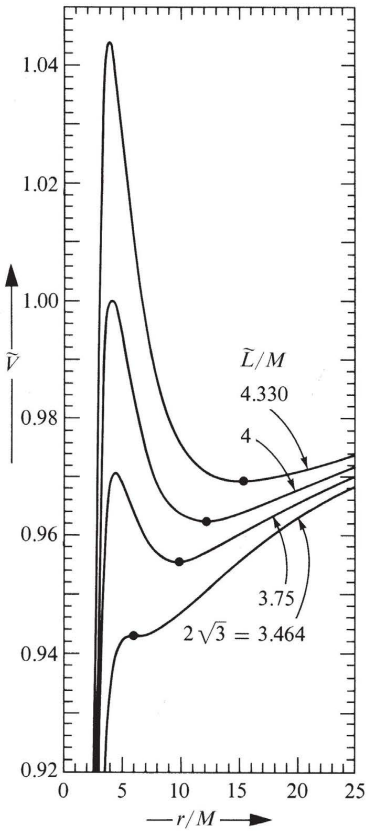
$$\tilde{E}^2 < \tilde{V}_{\max}{}^2 = (1 - 2u_m)(1 + L^{\dagger 2}u_m{}^2),$$

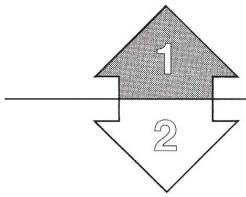
$$u_m \equiv \frac{1 + \sqrt{1 - 12/L^{\dagger 2}}}{6}$$

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, $r^*(t)$

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

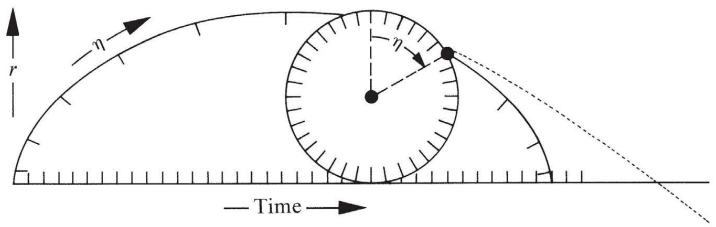


Figure 25.3.
A cycloid gives the relation between proper time and Schwarschild r coordinate for a test particle falling straight in toward center of gravitational attraction of negligible dimensions. The angle of turn of the wheel as it rolls on the base line and generates the cycloid is denoted by η.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

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

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

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.

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

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 $r = 2M$:

$r/2M$	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 $\tilde{V}$ of (25.33) and of Figure 25.2 replotted as a function of the tortoise coordinate. The approach of $\tilde{V}$ to zero at $r = 2M$ shows up as an exponential approach of $\tilde{V}$ 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

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

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 $\tilde{E}$ is enough to

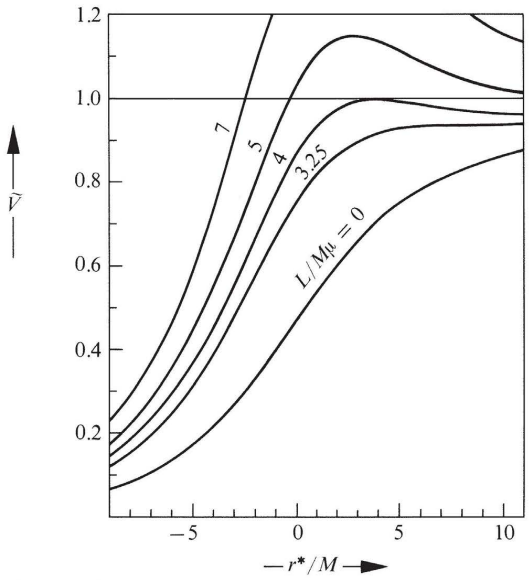


Figure 25.4.
Effective potential for motion in Schwarzschild geometry, expressed as a function of the tortoise coordinate, for selected values of the angular momentum of the test particle. The angular momentum L is expressed in units $M\mu$, where M is the mass of the black hole and μ the mass of the 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-

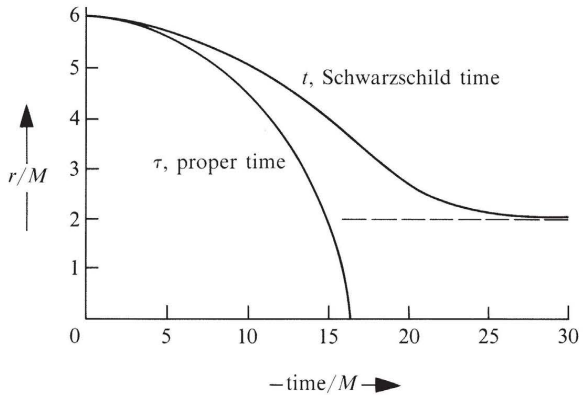


Figure 25.5.
Fall toward a Schwarzschild black hole as described (a) by a comoving observer (proper time τ) and (b) by a faraway observer (Schwarzschild-coordinate time t). In the one description, the point $r = 0$ is attained, and quickly [see equation (25.28)]. In the other description, $r = 0$ is 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" $\tilde{V}$ in its dependence on r (Figure 25.2) when one is interested in proper time; or the same effective potential $\tilde{V}$ 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 $\tilde{L}$ 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

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

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

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

(3) nearly circular orbits: periastron shift

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:

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

(4) qualitative features of angular motion

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

(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 $\tilde{V}_{\min}$,

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

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

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).]

Scattering of incoming particles:

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:

(1) Rutherford (nonrelativistic) cross section

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

(2) new features due to relativistic gravity

(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

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 $\tilde{L}$ 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 $\tilde{S}(t, r, \theta) = 0$ for $t = 0$ and for values $\tilde{L} = 4M$, $\tilde{E} = 1$ (or for $\tilde{L} = 2\sqrt{3}M$, $\tilde{E} = (8/9)^{1/2}$, or for some other equally simple set of values for these two parameters). Show that the whole set of surfaces of constant $\tilde{S}$ 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

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

(photon limit for $\tilde{E} \longrightarrow \infty$) 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

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

and

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

individually go to infinity; but the ratio

(1) impact parameter defined

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

goes to the finite value

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

(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

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

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,

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

(3) effective potential

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 $B(r)$:

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

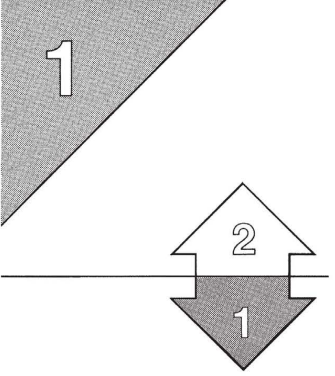
A ray with zero impact parameter (head-on impact), or any impact parameter less than $b_{\text{crit}} = \min[B(r)] = 3\sqrt{3}M$, can get to any and all r values.

(4) critical impact 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 $\mu \longrightarrow 0$.

(5) affine parameter

(*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:

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

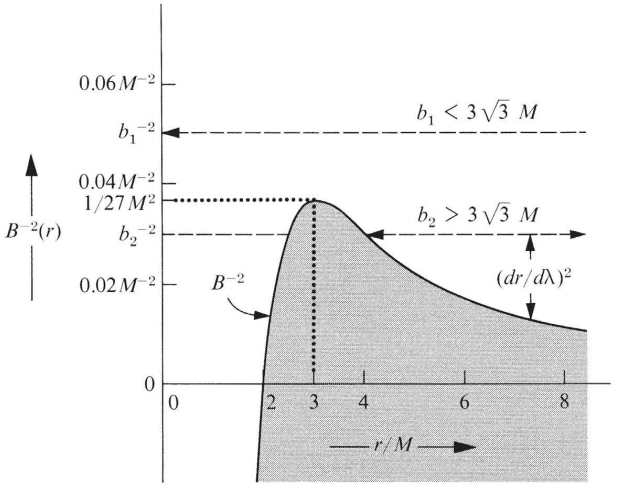
$$B^{-2}(r) = r^{-2}(1 - 2M/r);$$

b = (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)]:

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

$$v_{\hat{\phi}} = \frac{|g_{\phi\phi}|^{1/2}\, d\phi/d\lambda}{|g_{00}|^{1/2}\, dt/d\lambda} = b/B;$$

$$(v_{\hat{r}})^2 + (v_{\hat{\phi}})^2 = 1;$$

$\delta \equiv$ (angle between propagation direction and radial direction)
$= \cos^{-1} v_{\hat{r}} = \sin^{-1} v_{\hat{\phi}}$.

To be able to cross over the potential barrier, the particle must have $b < 3\sqrt{3}M$, or $v_{\hat{\phi}}{}^2 B^2 < 27M^2$, or $\sin^2\delta < 27M^2/B^2$. This result, restated:

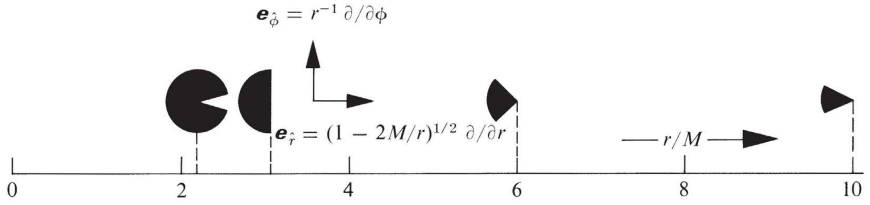
1. *A particle of zero rest mass at $r < 3M$ will eventually escape to infinity, rather than be captured by a black hole at $r = 2M$ if and only if v_r is positive and*

$$\sin\delta < 3\sqrt{3}MB^{-1}(r).$$

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,

$$\lambda_{\text{new}} = L\lambda, \tag{25.62}$$

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

(6) equations for orbit

$$\frac{d\phi}{d\lambda} = \frac{1}{r^2}, \tag{25.63}$$

$$\frac{dt}{d\lambda} = \frac{1}{b(1 - 2M/r)}. \tag{25.64}$$

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.

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

(7) 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

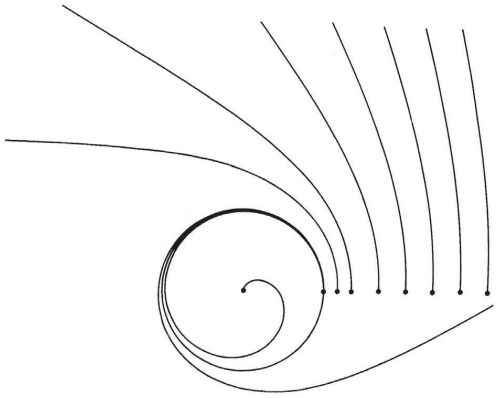


Figure 25.6.
The orbit of a photon in the "equatorial plane" of a black hole, plotted in terms of the Schwarzschild coordinates r and ϕ, for selected values of the turning point of the orbit, $r_{TP}/M = 2.99$, 3.00 (unstable circular orbit), 3.01, 3.5, 4, 5, 6, 7, 8, 9. The impact parameter is given 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.

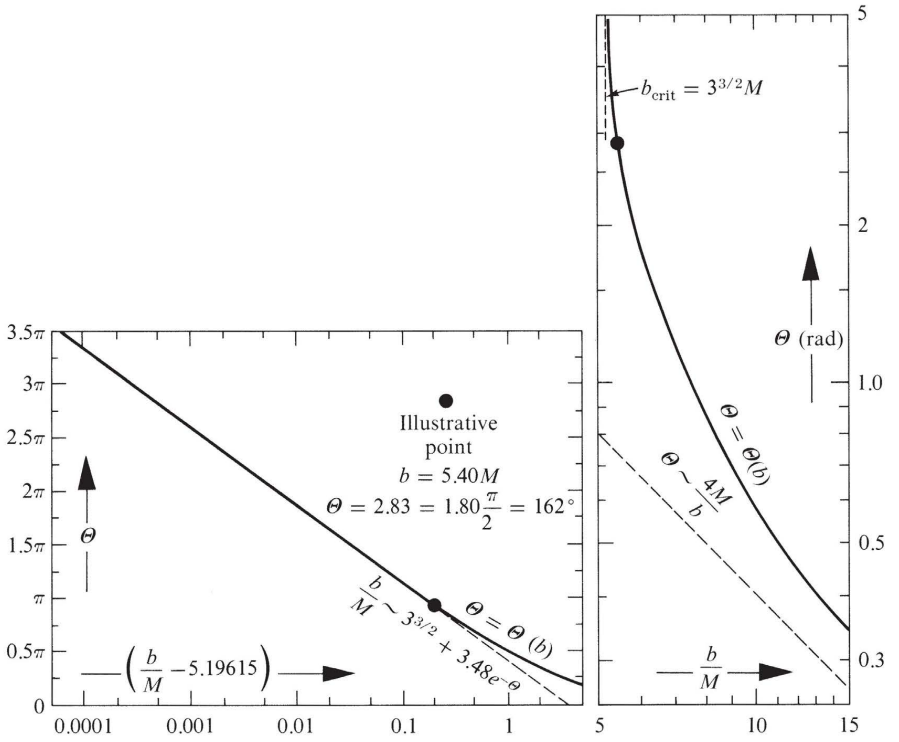


Figure 25.7.

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.

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:

(1) foundations for analysis

$$ds^2 = -e^{2\Phi}\,dt^2 + e^{2\Lambda}\,dr^2 + r^2\,d\Omega^2; \qquad \Phi = \Phi(r), \qquad \Lambda = \Lambda(r). \tag{25.73}$$

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

$$\boldsymbol{\omega}^{\hat{t}} = e^{\Phi}\,\mathbf{d}t, \qquad \boldsymbol{\omega}^{\hat{r}} = e^{\Lambda}\,\mathbf{d}r, \qquad \boldsymbol{\omega}^{\hat{\theta}} = r\,\mathbf{d}\theta, \qquad \boldsymbol{\omega}^{\hat{\phi}} = r\sin\theta\,\mathbf{d}\phi. \tag{25.74}$$

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

$$\mathscr{N} = H(E, L_x, L_y, L_z, \mu).$$

But this general solution is not spherically symmetric. For example, the distribution function

$$\mathscr{N} = H(E, \mu, L_z)\,\delta(L_y)\,\delta(L_x),$$

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

$$\mathscr{N} = F(E, L, \mu). \tag{25.78}$$

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 redshifting 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).

(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

$$\begin{aligned} \rho \equiv T^{\hat{0}\hat{0}} &= (\text{total density of mass-energy}) \\ &= 4\pi \int F(e^{\Phi}p^{\hat{0}}, rp^{\hat{T}}, \mu)(p^{\hat{0}2}p^{\hat{T}}/p^{\hat{r}})\, dp^{\hat{T}}\, dp^{\hat{0}}\, d\mu, \end{aligned} \tag{25.81a}$$

$$\begin{aligned} P_T \equiv \frac{1}{2}(T^{\hat{\theta}\hat{\theta}} + T^{\hat{\phi}\hat{\phi}}) &= T^{\hat{\theta}\hat{\theta}} = T^{\hat{\phi}\hat{\phi}} = (\text{tangential pressure}) \\ &\quad \text{[by spherical symmetry]} \\ &= 2\pi \int F(e^{\Phi}p^{\hat{0}}, rp^{\hat{T}}, \mu)[(p^{\hat{T}})^3/p^{\hat{r}}]\, dp^{\hat{T}}\, dp^{\hat{0}}\, d\mu, \end{aligned} \tag{25.81b}$$

$$\begin{aligned} P_r \equiv T^{\hat{r}\hat{r}} &= (\text{radial pressure}) \\ &= 4\pi \int F(e^{\Phi}p^{\hat{0}}, rp^{\hat{T}}, \mu)(p^{\hat{r}}p^{\hat{T}})\, dp^{\hat{T}}\, dp^{\hat{0}}\, d\mu. \end{aligned} \tag{25.81c}$$

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 $\mathcal{N} = F(E, L, \mu)$, where

E = energy-at-infinity of a star,

L = angular momentum of a star,

μ = 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

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, $\tilde{T} = T/\mu_0$. Replace $E_{\max}$ by the maximum energy per unit rest mass, $\tilde{E}_{\max} = E_{\max}/\mu_0$. Then the clusters are parametrized by (M_0, $\tilde{T}$, μ_0, $\tilde{E}_{\max}$). When one now doubles μ_0, holding M_0, $\tilde{T}$, $\tilde{E}_{\max}$ 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: $\tilde{T}$ and $\tilde{E}_{\max}$.

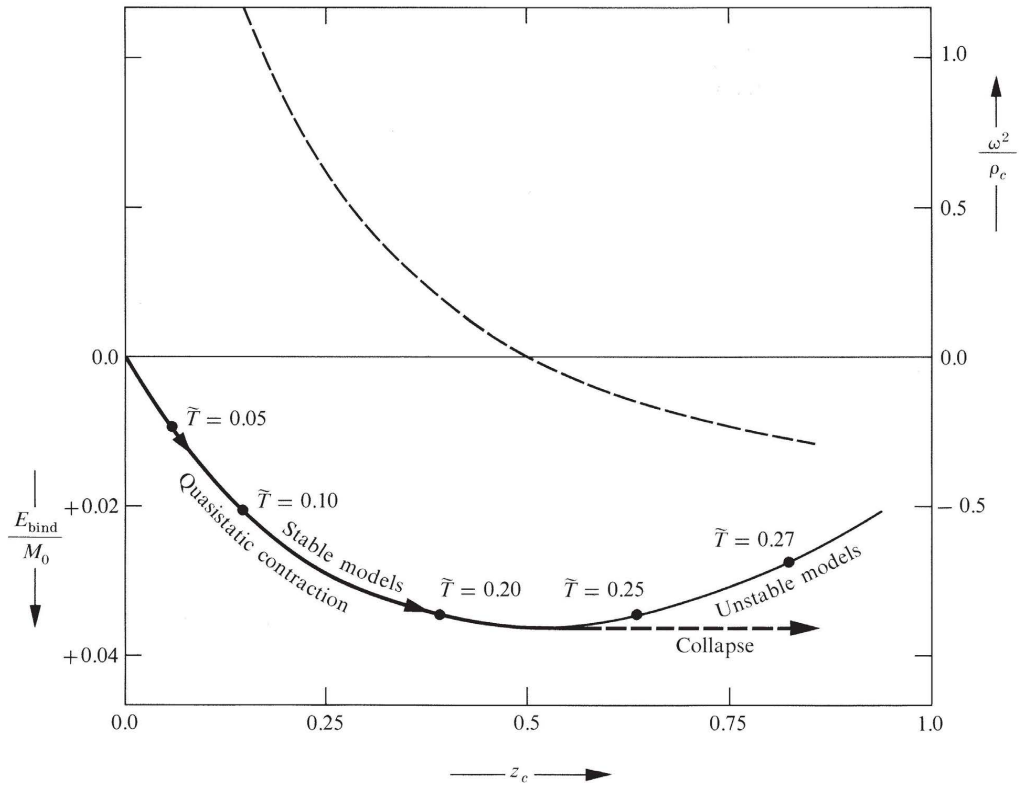
3. Consider as an instructive special case [Zel'dovich and Podurets (1965)] the one-parameter sequence with $\tilde{E}_{\max} = 1 - \frac{1}{2}\tilde{T}$. The following figure, computed by Ipser (1969), plots for this sequence the fractional binding energy,

$$E_{\text{bind}}/M_0 \equiv (M_0 - M)/M_0 \tag{6}$$

(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 $\tilde{T} = T/\mu_0$.

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.

2

CHAPTER 26

STELLAR PULSATIONS

This chapter is entirely Track 2, but it neither depends on nor prepares for any other chapter.

The *raison d'etre* of this chapter

§26.1. MOTIVATION

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

Setting up the analysis of stellar pulsations

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."

a. Equilibrium Configuration

Equilibrium configuration of star

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):

$$ds^2 = -e^{2\Phi_o}\,dt^2 + e^{2\Lambda_o}\,dr^2 + r^2(d\theta^2 + \sin^2\theta\,d\phi^2), \tag{26.1a}$$

$$\Lambda_o' = \frac{1}{2r}(1 - e^{2\Lambda_o}) + 4\pi r\rho_o e^{2\Lambda_o}, \tag{26.1b}$$

$$p_o' = -(\rho_o + p_o)\Phi_o', \tag{26.1c}$$

$$\Phi_o' = -\frac{1}{2r}(1 - e^{2\Lambda_o}) + 4\pi r p_o e^{2\Lambda_o}. \tag{26.1d}$$

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

Coordinates for perturbed configuration

$$ds^2 = -e^{2\Phi}\,dt^2 + e^{2\Lambda}\,dr^2 + r^2(d\theta^2 + \sin^2\theta\,d\phi^2), \tag{26.2}$$
$$\Phi = \Phi(t, r), \qquad \Lambda = \Lambda(t, r).$$

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

$$\begin{aligned} \Phi(t,r) &= \Phi_0(r) + \delta\Phi(t,r), \qquad \Lambda(t,r) = \Lambda_0(r) + \delta\Lambda(t,r), \\ p(t,r) &= p_0(r) + \delta p(t,r), \qquad \rho(t,r) = \rho_0(r) + \delta\rho(t,r), \\ & n(t,r) = n_0(r) + \delta n(t,r). \end{aligned} \tag{26.3a}$$

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 $\nabla \cdot \boldsymbol{T} = 0$, 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

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

Eulerian perturbations defined

Lagrangian perturbations defined

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

$$\Delta p(t, r) = p[t, r + \xi(t, r)] - p_o(r) \approx \delta p + p_o'\xi; \tag{26.4a}$$

$$\Delta \rho(t, r) = \rho[t, r + \xi(t, r)] - \rho_o(r) \approx \delta \rho + \rho_o'\xi; \tag{26.4b}$$

$$\Delta n(t, r) = n[t, r + \xi(t, r)] - n_o(r) \approx \delta n + n_o'\xi. \tag{26.4c}$$

§26.4. INITIAL-VALUE EQUATIONS

a. Baryon Conservation

Derivation of initial value equations:
(1) for baryon perturbations Δn and δn

The law of baryon conservation, $\nabla \cdot (n\boldsymbol{u}) = 0$ (§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 $\boldsymbol{u} \cdot \nabla n = \nabla_{\boldsymbol{u}} n = dn/d\tau$, one can rewrite the conservation law as

$$dn/d\tau = -n(\nabla \cdot \boldsymbol{u}).$$
$$\uparrow$$
[derivative of n along fluid world line]

In terms of Δn, the perturbation measured by an observer moving with the fluid, this equation can be rewritten as

$$\frac{d\,\Delta n}{d\tau} = -n(\nabla \cdot \boldsymbol{u}). \tag{26.5}$$

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

$$\frac{d}{d\tau} = \boldsymbol{u} = u^\alpha \frac{\partial}{\partial x^\alpha}, \qquad \nabla \cdot \boldsymbol{u} = \frac{1}{\sqrt{-g}}(\sqrt{-g}u^\alpha)_{,\alpha},$$

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

(4) for metric perturbations $\delta\Lambda$ and $\delta\Phi$

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

$$\boldsymbol{\omega}^{\hat{t}} = e^{\Phi}\,\mathbf{d}t, \qquad \boldsymbol{\omega}^{\hat{r}} = e^{\Lambda}\,\mathbf{d}r, \qquad \boldsymbol{\omega}^{\hat{\theta}} = r\,\mathbf{d}\theta, \qquad \boldsymbol{\omega}^{\hat{\phi}} = r\sin\theta\,\mathbf{d}\phi, \quad (26.12)$$

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:

$$G_{\hat{r}\hat{t}} = 2(\dot{\Lambda}/r)e^{-(\Lambda+\Phi)} \overset{\text{[linearized]}}{=} 2r^{-1}e^{-(\Lambda_o+\Phi_o)}\,\delta\dot{\Lambda}; \quad (26.13a)$$

$$G_{\hat{r}\hat{r}} = 2(\Phi'/r)e^{-2\Lambda} + r^{-2}(e^{-2\Lambda} - 1) \quad (26.13b)$$
$$= (G_{\hat{r}\hat{r}})_o + 2r^{-1}e^{-2\Lambda_o}\,\delta\Phi' - 2e^{-2\Lambda_o}(2r^{-1}\Phi_o' + r^{-2})\,\delta\Lambda.$$
$$\text{[linearized]}$$

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

$$T_{\hat{r}\hat{t}} = -(\rho_o + p_o)e^{\Lambda_o - \Phi_o}\dot{\xi}, \qquad T_{\hat{r}\hat{r}} = p_o + \delta p. \quad (26.14)$$

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

$$\delta\Lambda = -4\pi(\rho_o + p_o)re^{2\Lambda_o}\xi = -(\Lambda_o' + \Phi_o')\xi. \quad (26.15)$$

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 Φ_o', reduces to

$$\delta\Phi' = -4\pi\Gamma_1 p_o r^{-1}e^{2\Lambda_o + \Phi_o}(r^2e^{-\Phi_o}\xi)' + [4\pi p_o' r - 4\pi(\rho_o + p_o)]e^{2\Lambda_o}\xi. \quad (26.16)$$

This is the initial-value equation for $\delta\Phi$.

§26.5. DYNAMIC EQUATION AND BOUNDARY CONDITIONS

Derivation of equation of motion for fluid displacement ξ

The dynamic evolution of the fluid displacement $\xi(t, r)$ is governed by the Euler equation (22.13):

$$(\rho + p) \times (\text{4-acceleration}) = -(\text{projection of } \boldsymbol{\nabla}p \text{ orthogonal to } \boldsymbol{u}). \quad (26.17)$$

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

$$(\rho_o + p_o)e^{2(\Lambda_o - \Phi_o)}\ddot{\xi} = -\delta p' - (\delta\rho + \delta p)\Phi_o' - (\rho_o + p_o)\,\delta\Phi'. \tag{26.18}$$

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

$$W\ddot{\zeta} = (P\zeta')' + Q\zeta, \tag{26.19}$$

where ζ is a "renormalized displacement function," and W, P, Q are functions of radius determined by the structure of the equilibrium star:

$$\zeta \equiv r^2 e^{-\Phi_o}\xi; \tag{26.20}$$

$$W \equiv (\rho_o + p_o)r^{-2}e^{3\Lambda_o + \Phi_o}; \tag{26.21a}$$

$$P \equiv \Gamma_1 p_o r^{-2} e^{\Lambda_o + 3\Phi_o}; \tag{26.21b}$$

$$Q \equiv e^{\Lambda_o + 3\Phi_o}\left[\frac{(p_o')^2}{\rho_o + p_o}r^{-2} - 4p_o' r^{-3} - 8\pi(\rho_o + p_o)p_o r^{-2}e^{2\Lambda_o}\right]. \tag{26.21c}$$

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

$$(\xi/r) \text{ finite or zero in limit as } r \longrightarrow 0 \tag{26.22a}$$

[see (26.9) and (26.11)]; also, it must leave the pressure equal to zero at the star's surface, which means

$$\Delta p = -\Gamma_1 p_o r^{-2} e^{\Phi_o}(r^2 e^{-\Phi_o}\xi)' \longrightarrow 0 \text{ as } r \longrightarrow R \tag{26.22b}$$

[surface radius] $= R$

[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 inte-

grating 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:

$$\omega^2 = \text{extremal value of } \left\{ \frac{\int_0^R (P\zeta'^2 - Q\zeta^2)\, dr}{\int_0^R W\zeta^2\, dr} \right\}, \tag{3}$$

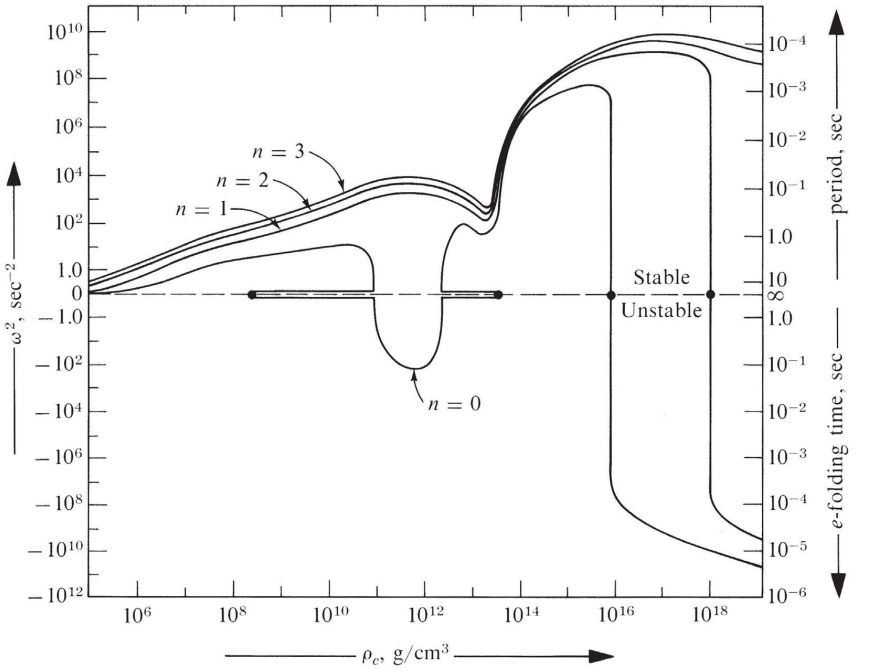
where ζ is varied over all functions satisfying the boundary conditions (2). [See e.g., §12.3 of Mathews and Walker (1965) for discussion of the equivalence between this variational principle and the original eigenvalue problem.]

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,

$$\left[\begin{array}{l}\text{stability against}\\ \text{adiabatic radial}\\ \text{perturbations}\end{array}\right] \Longleftrightarrow \left[\int_0^R (P\zeta'^2 - Q\zeta^2)\, dr > 0 \text{ for all functions } \zeta \text{ satisfying (2)}\right]. \tag{4}$$

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 $I = \int(\rho_0 r^2)4\pi r^2\, dr$ 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

$$\omega^2 = \frac{\int_0^R p_0[\Gamma_1 r^4\eta'^2 + (3\Gamma_1 - 4)(r^3\eta^2)'](1 + \Lambda_0 + 3\Phi_0)\, dr - \int_0^R F_0\eta^2\, dr}{\int_0^R \rho_0 r^4(1 + 3\Lambda_0 + \Phi_0 + p_0/\rho_0)\eta^2\, dr}, \tag{6}$$

where

$$F_0 \equiv 8\pi r^4 p_0\rho_0 + 8rm_0p_0 + \rho_0 m_0^2, \qquad \eta = \zeta/r^3 = (\xi/r)(1 - \Phi_0), \tag{7}$$

and $m_0(r)$ 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 $\xi = \epsilon r e^{\Phi_0} \approx \epsilon r(1 + \Phi_0)$, 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 $\Phi = \Phi_o$ and $\Lambda = \Lambda_o$ (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 $\boldsymbol{\nabla} \cdot \boldsymbol{E} = 4\pi\rho$ 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

$$a_{\max} = 4M/3\pi. \tag{5}$$

Box 27.1 (*continued*)

The decisive component of the Einstein field equation, in the terms of this notation, becomes

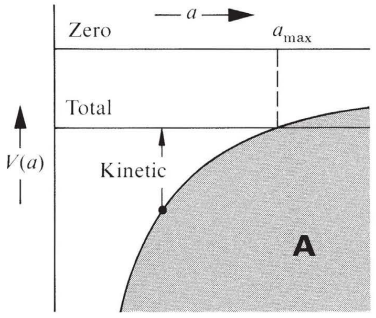
$$\frac{6}{a^2}\left(\frac{da}{dt}\right)^2 + \frac{6}{a^2} = \frac{6a_{\max}}{a^3}$$

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 $a_{\max}$ 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*)

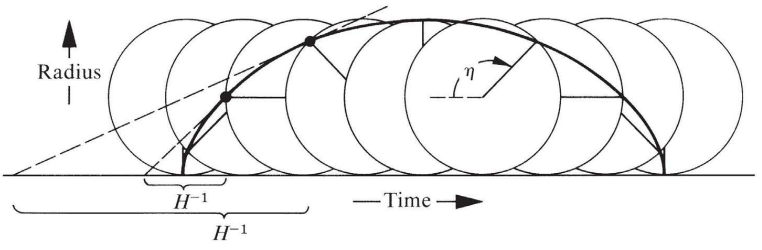
deceleration of the expansion of the universe. Therefore define an expansion parameter a^* as the ratio between the radius of any one of these spheres now and the radius of the same sphere at some fiducial instant; thus, $a^* = r/r_0$ is to be considered as independent of the particular sphere under 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,

$$\begin{aligned} a &= \frac{1}{2}a_{\max}(1 - \cos\eta), \\ t &= \frac{1}{2}a_{\max}(\eta - \sin\eta). \end{aligned} \tag{9}$$

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

$$\rho_H = \frac{3}{8\pi} \frac{1}{(1.7 \times 10^{28}\ \text{cm})^2} \text{ or } \rho_{H,\text{conv}} = 5 \times 10^{-30}\ \text{g/cm}^3 \tag{12}$$

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.

§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

$$\rho_{\rm cr} \lesssim 10^{-33}\ {\rm g/cm^3}, \tag{27.5}$$

(4) density of intergalactic gas

and, perhaps, gas in intergalactic space with

$$\rho_{\rm ig} \lesssim 10^{-28}\ {\rm g/cm^3}. \tag{27.6}$$

[Delineating more sharply the value of $\rho_{\rm ig}$ 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

$$\rho_{\rm mag} \lesssim 10^{-35}\ {\rm g/cm^3} \tag{27.7}$$

(6) radiation density

(corresponding to $B_{\rm avg} \lesssim 10^{-6}\ G$), and radiation (electromagnetic radiation, neutrino radiation, and perhaps gravitational radiation) totaling, one estimates,

$$\rho_{\rm rad} \simeq 10^{-33}\ {\rm g/cm^3}. \tag{27.8}$$

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,

$$\begin{aligned} \rho_{\rm microwave} &= 3p_{\rm microwave} = aT^4 \\ &= 4 \times 10^{-34}\ {\rm g/cm^3}. \end{aligned} \tag{27.9}$$

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 (~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 of fluid idealization of matter in universe

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

$$\boldsymbol{T} = (\rho + p)\boldsymbol{u} \otimes \boldsymbol{u} + p\boldsymbol{g}. \tag{27.10}$$

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.

Spacelike hypersurface as generalization of "moment of time"

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

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

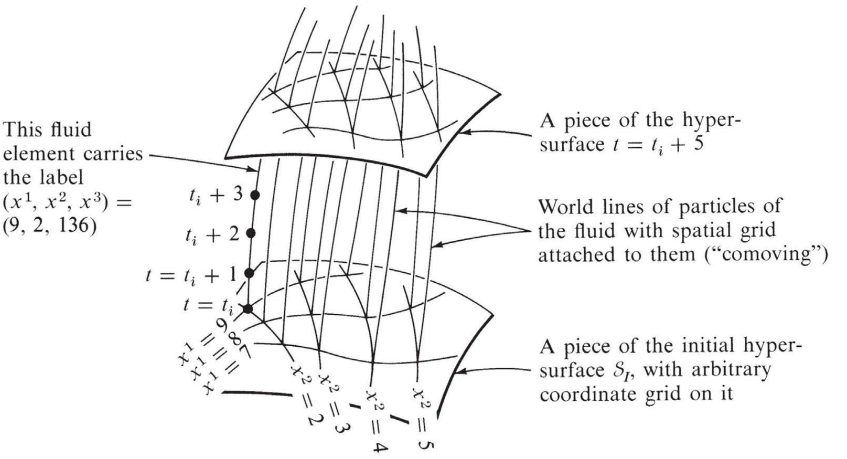


Figure 27.1.
Comoving, synchronous coordinate system for the universe, as constructed in §27.4 of the text. Key features of such a coordinate system are as follows (see §§27.4 and 27.5). (1) The spatial coordinates move with the fluid, and the time coordinate is proper time along the fluid world lines; i.e., the coordinate description of a particular fluid world line is

$$(x^1, x^2, x^3) = \text{constant}, \quad x^0 \equiv t = \tau + \text{constant}.$$

(τ: proper time measured along world line)

(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, $a(t)$ = "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

$$ds^2 = -dt^2 + g_{ij}\,dx^i\,dx^j. \tag{27.14}$$

Form of the line element in this coordinate system

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 27.2. SYNCHRONOUS COORDINATES IN GENERAL **EXERCISE**

In an arbitrary spacetime manifold (not necessarily homogeneous or isotropic), pick an initial spacelike hypersurface 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 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,

$$ds^2 = -dt^2 + a^2(t)\gamma_{ij}(x^k)\, dx^i\, dx^j. \tag{27.20}$$

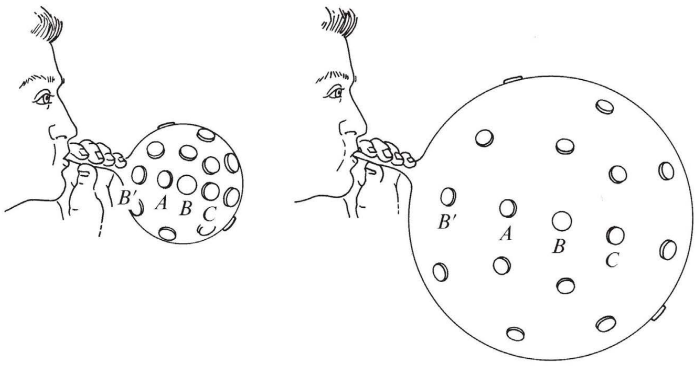


Figure 27.2.
Inflation of a balloon covered with pennies as a model for the expansion of the universe. Each penny A may well consider itself to be the center of the expansion because the distance from A to any neighbor B or C increases the more the more remote that neighbor was to begin with ("the Hubble 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).]

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

$$\frac{\left(\begin{array}{l}\text{proper distance between}\\ \text{two particles of the "cos-}\\ \text{mological fluid" at time } t\end{array}\right)}{\left(\begin{array}{l}\text{proper distance between}\\ \text{same two particles}\\ \text{at time } t_A\end{array}\right)} = a_A(t).$$

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:

$$\frac{\left(\begin{array}{l}\text{proper distance between}\\ \text{two particles of the "cos-}\\ \text{mological fluid" at time } t\end{array}\right)}{\left(\begin{array}{l}\text{proper distance between}\\ \text{the same two particles}\\ \text{at time } t_B\end{array}\right)} = a_B(t).$$

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:

$$\frac{a_A(t_D)}{a_A(t_C)} = \left(\begin{array}{l}\text{relative expansion}\\ \text{from } C \text{ to } D\end{array}\right) = \frac{a_B(t_D)}{a_B(t_C)}.$$

§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 ${}^{(3)}\boldsymbol{Riemann}$ 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

$$ {}^{(3)}R_{ijkl} = K(\gamma_{ik}\gamma_{jl} - \gamma_{il}\gamma_{jk});\ K = \text{"curvature parameter"} = \text{constant}. \quad (27.21) $$

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),

$$ \begin{aligned} d\sigma^2 &= \gamma_{ij}\,dx^i\,dx^j = K^{-1}[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\,d\phi^2)] \text{ if } K > 0, \\ d\sigma^2 &= \gamma_{ij}\,dx^i\,dx^j = d\chi^2 + \chi^2(d\theta^2 + \sin^2\theta\,d\phi^2) \text{ if } K = 0, \\ d\sigma^2 &= \gamma_{ij}\,dx^i\,dx^j = (-K)^{-1}[d\chi^2 + \sinh^2\chi(d\theta^2 + \sin^2\theta\,d\phi^2)] \text{ if } K < 0. \end{aligned} \quad (27.22) $$

Absorb the factor $K^{-1/2}$ or $(-K)^{-1/2}$ into the expansion factor $a(t)$ [see exercise 27.3], and define the function

$$ \begin{aligned} \mathcal{S} &\equiv \sin\chi, && \text{if } k \equiv K/|K| = +1 \text{ ("positive spatial curvature")}, \\ \mathcal{S} &\equiv \chi, && \text{if } k \equiv K = 0 \text{ ("zero spatial curvature")}, \\ \mathcal{S} &\equiv \sinh\chi, && \text{if } k \equiv K/|K| = -1 \text{ ("negative spatial curvature")}. \end{aligned} \quad (27.23) $$

Thus write the full spacetime geometry in the form

$$ \begin{aligned} ds^2 &= -dt^2 + a^2(t)\gamma_{ij}\,dx^i\,dx^j, \\ \gamma_{ij}\,dx^i\,dx^j &= d\chi^2 + \mathcal{S}^2(d\theta^2 + \sin^2\theta\,d\phi^2), \end{aligned} \quad (27.24) $$

and the 3-curvatures of the homogeneous hypersurfaces in the form

$$ {}^{(3)}R_{ijkl} = [k/a^2(t)][\gamma_{ik}\gamma_{jl} - \gamma_{il}\gamma_{jk}]. \quad (27.25a) $$

The curvature parameter K, after this renormalization, is evidently

$$ K = k/a^2(t). \quad (27.25b) $$

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 normaliza-

tion 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

$$d\sigma^2 = a^2[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)]. \tag{1}$$

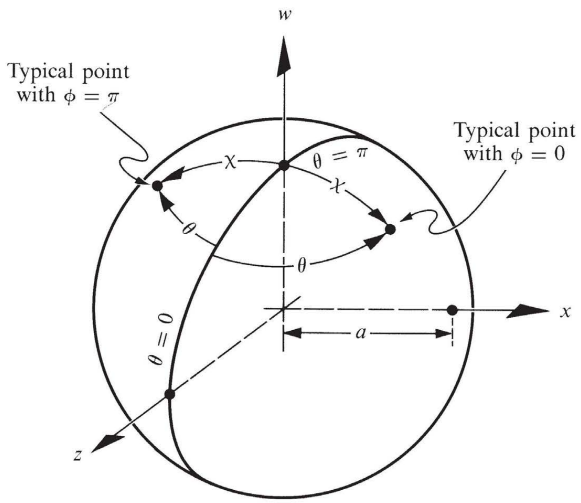
To visualize this 3-geometry, imagine embedding it in a four-dimensional Euclidean space (such embedding possible here; not possible for general three-dimensional manifold; only four freely disposable functions $[w, x, y, z]$ of three variables $[\alpha, \beta, \gamma]$ are at one's disposal to try to reproduce six prescribed functions $[g_{mn}(\alpha, \beta, \gamma)]$ of those same three variables).

The embedding is achieved by

$$\begin{aligned} w &= a\cos\chi, \qquad z = a\sin\chi\cos\theta, \\ x &= a\sin\chi\sin\theta\cos\phi, \\ y &= a\sin\chi\sin\theta\sin\phi, \end{aligned} \tag{2}$$

since it follows that

$$\begin{aligned} d\sigma^2 &\equiv dw^2 + dx^2 + dy^2 + dz^2 \\ &= a^2[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)]. \end{aligned} \tag{3}$$



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).

Box 27.2 (*continued*)

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

$$\begin{aligned} 0 &\le \chi \le \pi, \\ 0 &\le \theta \le \pi, \\ 0 &\le \phi \le 2\pi \end{aligned}$$

(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

$$\begin{aligned} x &= a\chi \sin\theta \cos\phi, \\ y &= a\chi \sin\theta \sin\phi, \\ z &= a\chi \cos\theta, \end{aligned} \tag{7}$$

the metric is

$$d\sigma^2 = dx^2 + dy^2 + dz^2. \tag{8}$$

The entire hypersurface is swept out by

$$\begin{aligned} 0 &\le \chi < \infty, \\ 0 &\le \theta \le \pi, \\ 0 &\le \phi \le 2\pi; \end{aligned} \tag{9}$$

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

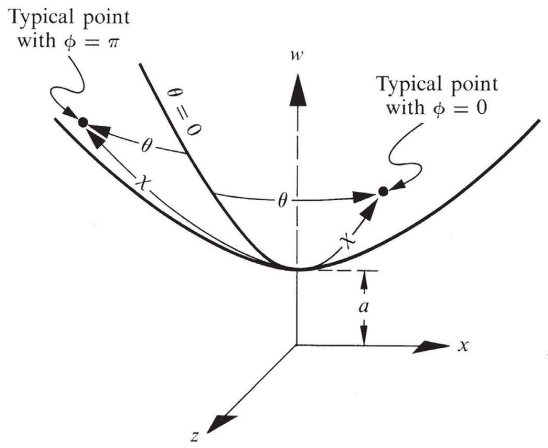
$$\begin{aligned} w &= a\cosh\chi, \qquad z = a\sinh\chi\cos\theta, \\ x &= a\sinh\chi\sin\theta\cos\phi, \\ y &= a\sinh\chi\sin\theta\sin\phi; \end{aligned} \tag{12}$$

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

$$\frac{(\text{proper surface area})}{4\pi\,(\text{proper distance})^2} = \frac{A}{4\pi\,\ell^2} = \frac{4\pi a^2 \sinh^2\chi}{4\pi a^2 \chi^2} \approx \left(\frac{e^{\ell/a}}{2\,\ell/a}\right)^2 \longrightarrow \infty. \quad (14)$$

The entire hypersurface is swept out by

$$0 \le \chi < \infty, \quad 0 \le \theta \le \pi, \quad 0 \le \phi \le 2\pi \quad (15)$$

(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

$$ds^2 = -dt^2 + a^2(t)[dx^2 + dy^2 + dz^2]. \quad (16)$$

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

$$V = a^3 L^3,$$

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 $\nabla \cdot \mathbf{T} = 0$.

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, $\mathbf{u} \cdot (\nabla \cdot \mathbf{T}) = 0$, 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, $(\mathbf{g} + \mathbf{u} \otimes \mathbf{u}) \cdot (\nabla \cdot \mathbf{T}) = 0$, gives the force equation ("Euler equation")

$$(\rho + p) \times (\text{4-acceleration}) = -(\text{component of } \nabla p \text{ orthogonal to } \mathbf{u}). \tag{27.28b}$$

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 $\mathcal{P}$, it would distinguish a direction in the homogeneous hypersurface at $\mathcal{P}$.

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

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

$$\rho_r V^{4/3} = \text{const} = \underbrace{\frac{\rho_r}{V^{-1/3}}}_{\substack{\text{energy } hc/\lambda \text{ of one photon, up to a factor of proportionality}}} V \left(\begin{array}{c}\text{constancy of number} \\ \text{of photons}\end{array}\right) \tag{27.33b}$$

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,

$$\boldsymbol{\omega}^{\hat{t}} \equiv \mathbf{d}t, \qquad \boldsymbol{\omega}^{\hat{\chi}} \equiv a(t)\,\mathbf{d}\chi, \qquad \boldsymbol{\omega}^{\hat{\theta}} \equiv a(t)\Sigma\,\mathbf{d}\theta, \qquad \boldsymbol{\omega}^{\hat{\phi}} \equiv a(t)\Sigma \sin\theta\,\mathbf{d}\phi. \quad (27.36)$$

Evaluation of the Einstein field equation for a homogeneous universe:

The result [see equations (5) of Box 14.5] is

$$G_{\hat{t}\hat{t}} = 3\left(\frac{a_{,t}}{a}\right)^2 + \frac{3k}{a^2}, \quad (27.37a)$$

$$G_{\hat{\chi}\hat{\chi}} = G_{\hat{\theta}\hat{\theta}} = G_{\hat{\phi}\hat{\phi}} = -2\frac{a_{,tt}}{a} - \left(\frac{a_{,t}}{a}\right)^2 - \frac{k}{a^2}, \quad (27.37b)$$

$$G_{\hat{\mu}\hat{\nu}} = 0 \text{ if } \mu \neq \nu. \quad (27.37c)$$

(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:

$$T_{\hat{t}\hat{t}} = \rho, \quad (27.38a)$$

$$T_{\hat{\chi}\hat{\chi}} = T_{\hat{\theta}\hat{\theta}} = T_{\hat{\phi}\hat{\phi}} = p, \quad (27.38b)$$

$$T_{\hat{\mu}\hat{\nu}} = 0 \text{ when } \mu \neq \nu. \quad (27.38c)$$

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

$$dt = \frac{da}{[-k + (8\pi/3)a^2\rho(a) + \underbrace{(\Lambda/3)a^2}_{\text{omit}}]^{1/2}}. \tag{27.42}$$

In terms of a as a new time parameter, it follows from this formula that the metric takes the form [Hughston (1969)]

$$ds^2 = \frac{-(da)^2}{-k + (8\pi/3)a^2\rho(a) + \underbrace{(\Lambda/3)a^2}_{\text{omit}}} + a^2[d\chi^2 + \Sigma^2(d\theta^2 + \sin^2\theta\, d\phi^2)]. \tag{27.43}$$

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

$$d\eta = \frac{dt}{a(t)}. \tag{27.44}$$

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:

$$\eta = \int_0^t \frac{dt}{a(t)}; \tag{27.45}$$

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

$$ds^2 = a^2(\eta)[-d\eta^2 + d\chi^2 + \Sigma^2(d\theta^2 + \sin^2\theta\, d\phi^2)]. \tag{27.46}$$

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:

$$\begin{aligned} H_o &= \frac{55 \text{ km/sec}}{(299{,}793 \text{ km/sec})(3.0856 \times 10^{24} \text{ cm or } 3.2615 \times 10^{6} \text{ yr of time})} \\ &= 0.59 \times 10^{-28} \text{ per cm of light-travel time} \\ &\qquad \text{or } 5.6 \times 10^{-11} \text{ fractional expansion per yr}, \end{aligned} \tag{27.50}$$

$H_o^{-1} = 1.7 \times 10^{28}$ cm of light-travel time or 18×10^{9} 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

$$\left(\frac{da}{dt}\right)^2 - \frac{a_{\max}}{a} = -1, \qquad \begin{pmatrix}27.52;\ \text{matter}\\ \text{dominates}\end{pmatrix}$$

or

$$\left(\frac{da}{dt}\right)^2 - \frac{a^{*2}}{a^2} = -1. \qquad \begin{pmatrix}27.52;\ \text{radiation}\\ \text{dominates}\end{pmatrix}$$

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 re-contraction 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 $\left(\frac{da}{dt}\right)^2 + V(a) = -1$	$V(a) = -\frac{a_{\max}}{a}$	$V(a) = -\frac{a^{*2}}{a^2}$
Value of constant in this "potential" in terms of conditions at some standard epoch	$a_{\max} = \frac{8\pi}{3} a_o^3 \rho_{mo}$	$a^{*2} = \frac{8\pi}{3} a_o^4 \rho_{ro}$
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)	$a = \frac{a_{\max}}{2}(1 - \cos\eta)$	$a = a^* \sin\eta$
	$t = \frac{a_{\max}}{2}(\eta - \sin\eta)$	$t = a^*(1 - \cos\eta)$
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 $H^{-1} = \frac{a}{(da/dt)} = \frac{a^2}{(da/d\eta)}$	$\frac{a_{\max}}{2}\frac{(1 - \cos\eta)^2}{\sin\eta}$	$a^*\frac{\sin^2\eta}{\cos\eta}$

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	$t = \dfrac{a_{\max}}{12}\eta^3$ $a = \dfrac{a_{\max}}{4}\eta^2$ $a = \left(\dfrac{9a_{\max}t^2}{4}\right)^{1/3}$	$t = \dfrac{a^*}{2}\eta^2$ $a = a^*\eta$ $a = (2a^*t)^{1/2}$
Other features of expansion at early times	$H^{-1} = \dfrac{a_{\max}}{8}\eta^3 = 1.5t$ $\rho_m = \dfrac{a_{\max}}{(8\pi a^3/3)}$ $= \dfrac{1}{6\pi t^2} = \dfrac{3H^2}{8\pi}$	$H^{-1} = a^*\eta^2 = 2t$ $\rho_r = 3p_r = \dfrac{a^{*2}}{(8\pi a^4/3)}$ $= \dfrac{3}{32\pi t^2} = \dfrac{3H^2}{8\pi}$

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

$$\rho_r(\text{cm}^{-2}) = \frac{G}{c^4}\rho_{r,\text{conv}}(\text{erg/cm}^3) = \frac{Gf\pi^2}{c^4 120}\frac{(kT)^4}{\hbar^3 c^3}. \tag{27.54}$$

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],	$f_{ev} = 7$;
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

$$T_{\text{em}} > T_{\nu} > T_{g}. \tag{27.56}$$

T_{em} has been measured to be 2.7 K; T_{ν} is calculated to be $(4/11)^{1/3}\, T_{\text{em}} = 1.9$ 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 $\rho_{r,\text{decoupled}}\, a^4$ or the product $T^4_{r,\text{decoupled}}\, a^4$. Compare the temperature of this particular radiation now to the temperature of the same radiation at any chosen fiducial time t_{fid} after its era of decoupling. Find

$$T_{r,\text{fid}} = \frac{a_{\text{now}}}{a_{\text{fid}}}\, T_{r,\text{now}} = (1 + z) T_{r,\text{now}}. \tag{27.57}$$

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

$$\left[\left(\frac{T_{r,\text{now}}}{10^{10}\ \text{K}}\right)(1 + z)\right]^2 \left(\frac{t_{\text{fid}}}{1\ \text{sec}}\right) = 1.19. \tag{27.58b}$$

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,

$$\frac{20 \times 10^9 \text{ lyr}}{10 \times 10^9 \text{ lyr}} = \frac{H^{-1}}{t} = \frac{\frac{a_{\max}}{2}\frac{(1 - \cos\eta)^2}{\sin\eta}}{\frac{a_{\max}}{2}(\eta - \sin\eta)} \tag{27.59}$$

or

$$\eta = 1.975 \text{ (or } 113.2°) \tag{27.60}$$

(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 $(3/8\pi a_o^2) + (3H_o^2/8\pi)$,	$(7.67 + 3.33) \times 10^{-58}$ cm^{-2} $= 11.00 \times 10^{-58}$ cm^{-2} or 14.8×10^{-30} g/cm^3;
Volume today, $2\pi^2 a_o^3$,	38.3×10^{84} cm^3;
Density at maximum $(3/8\pi a^2) + (3H^2/8\pi)$,	$(3.70 + 0.00) \times 10^{-58}$ cm^{-2} $= 5.0 \times 10^{-30}$ g/cm^3;
Volume at maximum,	114×10^{84} cm^3;
Rate of increase of radius today,	13.19×10^9 lyr/20×10^9 yr $= 0.66$ lyr/yr;
Rate of increase of volume today,	1.82×10^{68} cm^3/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.

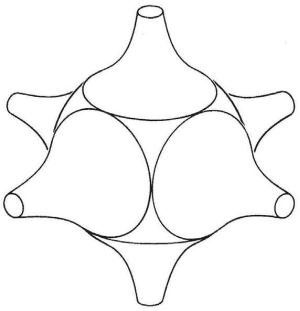


Figure 27.3.
Many Schwarzschild zones are fitted together to make a closed universe. This universe is dynamic because a test particle at the interface between two zones rises up against the gravitational attraction of each and falls back under the gravitational attraction of each. Therefore the two centers 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],

$$r = \frac{R}{2}(1 - \cos\eta),$$
$$\tau = \frac{R}{2}\left(\frac{R}{2M}\right)^{1/2}(\eta - \sin\eta). \tag{27.61}$$

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.

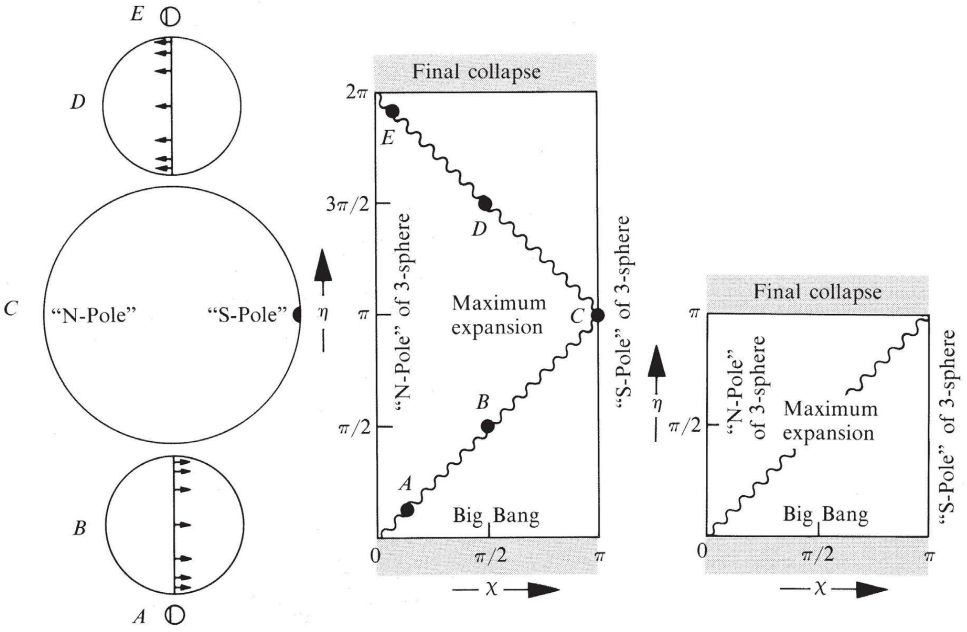


Figure 27.4.
Use of "arc parameter" η as a time coordinate and hyperpolar angle χ as a space coordinate to describe travel of a photon ($\pm 45°$ line) in a Friedmann universe that is matter-dominated (center) or radiation-dominated (right). The burst of photons is emitted from the "*N*-pole" of the 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 $\Lambda = 0$:

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.

(1) expansion factor as function of time

(2) early stage—same as for closed universe

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.

(3) late stage—expansion forever

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

$$\frac{1}{a^2}\left(\frac{da}{dt}\right)^2 - \frac{1}{a^2} = 0. \tag{27.67}$$

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 $a = a_0$ now,

$$a_0/a = 1 + z \tag{27.73}$$

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,"

$$\left[\frac{d}{dt}\left(\frac{a(t)}{a_0}\right)\right]^2 + V(a/a_0) = -\frac{k}{a_0^2} \equiv -K_0. \tag{27.74}$$

Here the quantity

$$V(a/a_0) \equiv -\frac{8\pi}{3}\left[\rho_{mo}\left(\frac{a_0}{a}\right) + \rho_{ro}\left(\frac{a_0}{a}\right)^2\right] - \frac{1}{3}\Lambda\left(\frac{a}{a_0}\right)^2 \tag{27.75}$$

acts as an "effective potential" for the dynamics of the expansion. The constant term K_0 represents one sixth of the intrinsic curvature of the model universe today. Its negative, $-K_0$, 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_0) and from the value of K_0.

(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_0) 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.

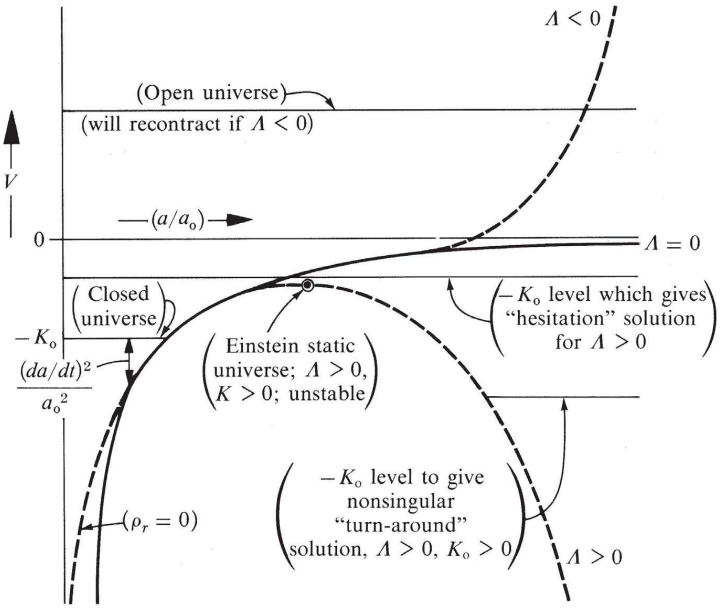
Other non-Einsteinian cosmologies:

(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 $\pm 3 \times 10^{-28}$ g/cm^3; $\rho_{mo,\text{conv}} = 10^{-30}$ g/cm^3 or $\rho_{mo,\text{conv}} = 10^{-28}$ g/cm^3; and $(a_0^{-1}\, da/dt)^2 = H_0^2 = (1/20 \times 10^9 \text{ yr})^2$ or 3.8×10^{-29} 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 $$a = \text{constant} \times \exp[(\Lambda/3)^{1/2}t].$$

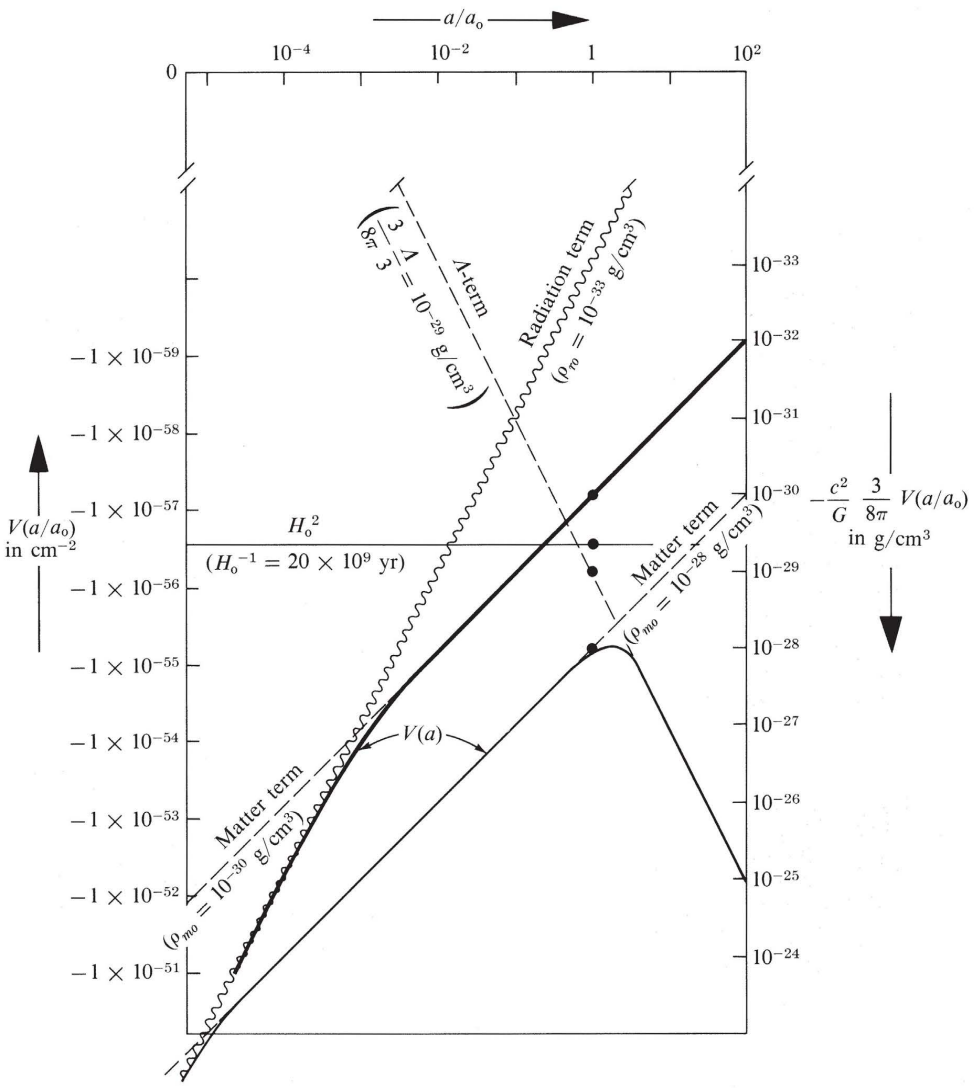
Figure 27.5. (*facing page*)
Log-log plot of the effective potential $V(a)$ of equation (27.75) and Box 27.5 as it enters the generalized Friedmann equation

$$\left[\frac{d}{dt}\left(\frac{a(t)}{a_o}\right)\right]^2 + V\left(\frac{a}{a_o}\right) = -\frac{k}{a_o^2} = -K_o.$$

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 $\dot{a}^2(t)/\dot{a}_o^2 + K_o H_o^{-2}$, where $\dot{a} \equiv da/dt$. As evaluated "today" $(a/a_o = 1)$ this ratio has the value $1 + K_o H_o^{-2}$. 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 $1 + K_o H_o^{-2}$, hence the value of $K_o H_o^{-2}$, 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, $1 + K_o H_o^{-2} = 0.223$. From this follows $K_o = -0.777\ H_o^2$ (open or hyperbolic universe), hence $k = -1$ and $a_o = (k/K_o)^{1/2} = (1/0.777)^{1/2}\ 20 \times 10^9\ \text{yr} = 22.7 \times 10^9\ \text{yr}$. For the high-density model universe, with $\rho_{mo} = 10^{-28}\ \text{g/cm}^3$, one similarly finds $1 + K_o H_o^{-2} = 24.5$, hence $K_o = +23.5 H_o^2$, hence $k = +1$ (closed universe) and $a_o = (k/K_o)^{1/2} = (1/23.5)^{1/2}\ 20 \times 10^9\ \text{yr} = 4.12 \times 10^9\ \text{yr}$. 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 $\dot{a}^2(t)/\dot{a}_o^2 + K_o H_o^{-2}$, falls from its "present value" of $1 + K_o H_o^{-2}$ to $0 + K_o H_o^{-2}$; 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 $q \equiv -(1/H^2 a)(d^2a/dt^2)$ has the value

$$q = -a^2(a^2 - a_{\min}^2)^{-1} < -1.$$

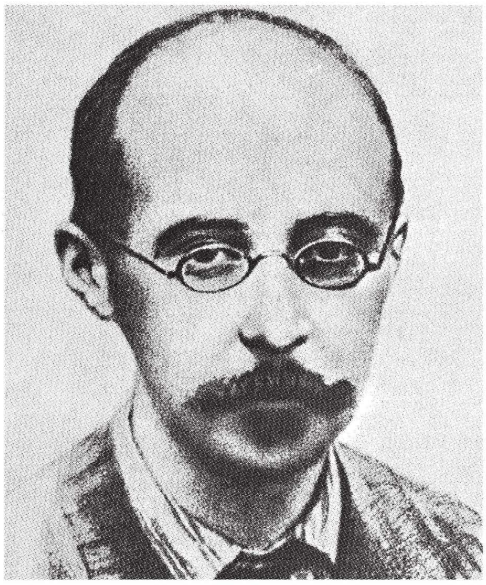
Exercise 27.14. "HESITATION" UNIVERSE

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 $\sim$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_o^{-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_0) [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-

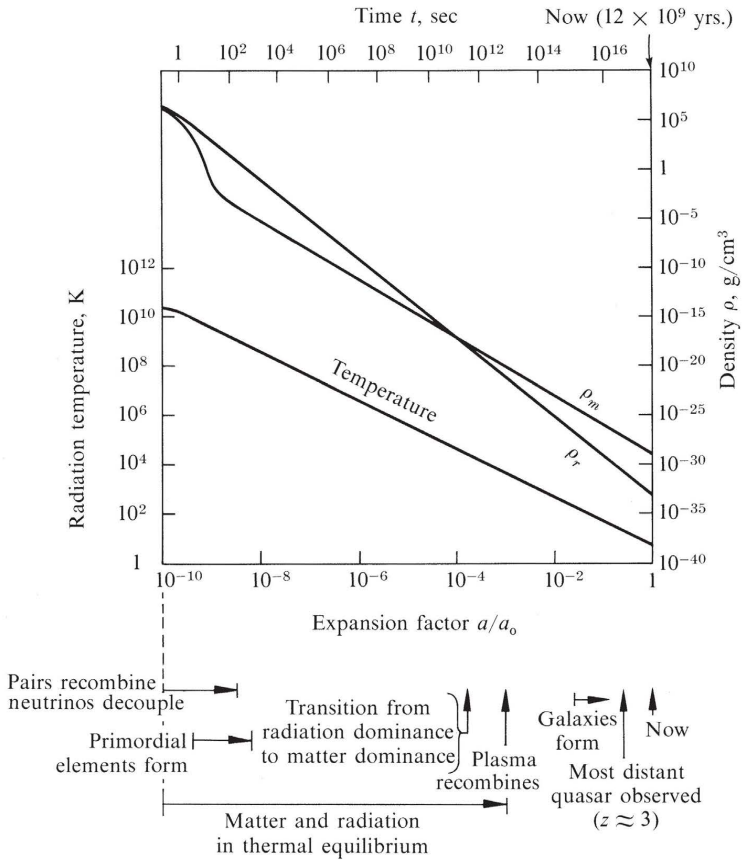


Figure 28.1.
Evolution of the universe into its present state, according to the standard hot big-bang model. The curves are drawn assuming

$$\rho_{mo} = 5 \times 10^{-30}\ \text{g/cm}^3, \qquad \rho_{ro} = 10^{-33}\ \text{g/cm}^3, \qquad k = 0;$$

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.)

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.

(3) decoupling and free propagation of gravitons and neutrinos ($t \lesssim 1$ sec.)

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 ($2 \text{ sec.} \lesssim t \lesssim 1{,}000 \text{ sec.}$)

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 ($1{,}000 \text{ sec.} \lesssim t \lesssim 10^5$ 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 = ρ_{mo}, between 10^{-28} and 2×10^{-31} g/cm^3;
$k = 0$ or $+1$ or -1;
temperature of electromagnetic radiation today = 2.7 ± 0.1 K.
Total radiation density [observed photons, plus neutrinos and gravitons that presumably originated in big bang in thermal equilibrium with photons] = ρ_{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) × (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, $N_{\text{tot}}(z)$, i.e., if the changes in the relative luminosity distribution at 2,500 Å

$$f(z, L_\nu) \equiv [1/N_{\text{tot}}(z)]N(z, L_\nu)$$

were negligible, and if the universe today is characterized by $\sigma_0 = q_0 = 1$ (see Chapter 29 for notation), then the data show

$$N_{\text{tot}}(z = 2) \approx 50 N_{\text{tot}}(z = 0.5).$$

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

$$N_{\text{tot}}(z) \propto (1 + z)^6,$$

$$N_{\text{tot}}[z(t)] \propto 10^{5(t_0 - t)/t_0}.$$

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_{ro} 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.

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.

Expansion forever vs. recontraction of universe

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 $\Lambda \neq 0$:

$$\text{Expansion forever} \Longleftrightarrow \Lambda \geq \Lambda_{\text{crit}} \equiv \begin{cases} 0 & \text{if } k \leq 0, \\ (4\pi\rho_{mo}a_o^3)^{-2} & \text{if } k > 0; \end{cases}$$

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 $\sigma_o - q_o$.

Box 29.1 (*continued*)

3. If $\Lambda = 0$

(a) $q_o > \frac{1}{2} \iff \rho_{mo} > \rho_{\text{crit}} \equiv \frac{3}{8\pi} H_o^2 \iff k > 0$ (positive curvature; "closed" universe)

$\iff$ universe will eventually recontract;

(b) $q_o = \frac{1}{2} \iff \rho_{mo} = \rho_{\text{crit}} \equiv \frac{3}{8\pi} H_o^2 \iff k = 0$ (zero curvature; "flat" universe)

$\implies$ universe will expand forever;

(c) $q_o < \frac{1}{2} \iff \rho_{mo} < \rho_{\text{crit}} \equiv \frac{3}{8\pi} H_o^2 \iff k < 0$ (negative curvature; "open" universe)

$\implies$ universe will expand forever.

4. If $\Lambda \neq 0$

(a) $\sigma_o > \frac{1}{3}(q_o + 1) \iff k > 0$ (positive curvature; "closed" universe),

and in this case,

$$\sigma_o - q_o \geq \frac{1}{{\sigma_o}^2}\left(\sigma_o - \frac{q_o + 1}{3}\right)^3 \iff \text{universe will expand forever,}$$

$$\sigma_o - q_o < \frac{1}{{\sigma_o}^2}\left(\sigma_o - \frac{q_o + 1}{3}\right)^3 \iff \text{universe will eventually recontract;}$$

(b) $\sigma_o = \frac{1}{3}(q_o + 1) \iff k = 0$ (zero curvature; "flat" universe),

and in this case,

$\sigma_o \geq q_o \iff$ universe will expand forever,

$\sigma_o < q_o \iff$ universe will eventually recontract;

(c) $\sigma_o < \frac{1}{3}(q_o + 1) \iff k < 0$ (negative curvature; "open" universe),

and in this case,

$\sigma_o \geq q_o \iff$ universe will expand forever,

$\sigma_o < q_o \iff$ 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"

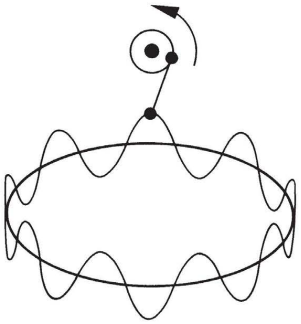
No one has ever put forward a satisfactory explanation for the cosmological redshift other than the expansion of the universe (see below). The idea has been proposed at various times by various authors that some new process is at work ("tired light") in which photons interact with atoms or electrons 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."

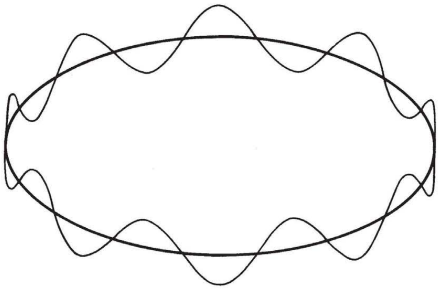
(3) If there does exist any such decay process, then simple arguments of special relativity that Zel'dovich attributes to M. P. Bronshtein, and spells out in detail, demand the relationship

$$\begin{pmatrix}\text{probability per}\\ \text{second of}\\ \text{"photon decay"}\end{pmatrix} = \frac{\begin{pmatrix}\text{a universal constant with}\\ \text{the dimensions sec}^{-2}\end{pmatrix}}{(\text{frequency of photon in sec}^{-1})}.$$

> "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, $a(t_e) = a_{\text{em}}$;
wavelengths small, $\lambda(t_e) = \lambda_{\text{em}}$.



Reception:
universe larger, $a(t_r) = a_{\text{rec}}$;
wavelengths larger, $\lambda(t_r) = \lambda_{\text{rec}}$;
number of nodes in standing wave unchanged;

$$n = \text{constant} = \frac{2\pi a_{\text{rec}}}{2\pi\lambda_{\text{rec}}} = \frac{a_{\text{em}}}{\lambda_{\text{em}}}.$$

Figure 29.1.
Redshift as an effect of standing waves. The ratio of wavelengths, $\lambda_{\text{rec}}/\lambda_{\text{em}}$, is identical with the ratio of dimensions, $a_{\text{rec}}/a_{\text{em}}$ 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

$$\lambda_{\text{rec}} = t_{rB} - t_{rA}, \tag{29.8}$$

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

$$\frac{\lambda_{\text{rec}}}{a(t_r)} = \frac{\lambda_{\text{em}}}{a(t_e)}; \tag{29.10}$$

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